



UNIMORE

UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

University of Modena and Reggio Emilia
Department of Engineering “Enzo Ferrari”

XXXVIII cycle of the Doctorate School in
Information and Communication Technologies (ICT)

Multimodal Models for Understanding and Generating Visually-Rich Documents

Ph.D. Dissertation

FABIO QUATTRINI

Advisor: Prof. Rita Cucchiara

Co-Advisor: Dott. Silvia Cascianelli

PhD Programme Coordinator: Prof. Luigi Rovati

Advisor:

Prof. Rita Cucchiara University of Modena and Reggio Emilia

Co-Advisor:

Dott. Silvia Cascianelli University of Modena and Reggio Emilia

Director of the School:

Prof. Luigi Rovati University of Modena and Reggio Emilia

Review Committee:

Prof. Alicia Fornés Universitat Autònoma de Barcelona
Prof. Giorgos Sfikas University of West Attica

The work described in this thesis has been carried out within the International Doctorate in Information and Communication Technologies (ICT), at the AImageLab research laboratory of the University of Modena and Reggio Emilia.

This dissertation was typeset by the author using $\text{\LaTeX} 2_{\epsilon}$, originally developed by Leslie Lamport and based on Donald Knuth's $\text{\TeX}$. The body text is set in 11 point Egenolff-Berner Garamond, a revival of Claude Garamont's humanist typeface.

Copyright © 2026 by Fabio Quattrini

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission from the author.

To my family, my friends, my mentors.

You believed in me before I did.

You made me who I am.

*As you set out for Ithaka, hope your road is a long one, full of adventure, full of
discovery.*

— C. P. Cavafy, *Ithaka* (trans. E. Keeley & P. Sherrard)

Multimodal Models for Understanding and Generating Visually-Rich Documents

ABSTRACT

A defining human trait is the ability to convey complex information through writing and drawing. This has enabled the creation of documents, the foundational medium for preserving and transmitting knowledge across time and space.

This dissertation presents multimodal methods, datasets, benchmarks, and evaluation tools for understanding, generating, and editing Visually Rich Documents, spanning both cultural heritage sources and modern digital content. The term ‘Visually Rich Documents’ refers to pages where text, figures, graphics, and layout jointly carry meaning. When dealing with historical sources, which are often centuries old, reliable analysis first requires reducing the complexity of the physical substrate to make downstream tasks tractable. Therefore, on the analysis side, we introduce a binarization and restoration model agnostic to language, script, support, and resolution, reducing bleed-through, noise, and material artifacts in 2D document images while preserving informative strokes. We further extend this spectral approach to 3D tomographic volumes of the Herculaneum papyri, demonstrating preliminary applicability to ink detection on carbonized ancient scrolls. As many Visually Rich Documents encode meaning through spatial organization, we develop a multi-page parsing system that extracts regions, relations, and visual semantics without relying on an external optical text recognition model, enabling the parsing of complex, multi-page layouts. To foster research in ultra-low-resource settings, we curate a dataset for the Coptic language and benchmark current methods. Having established methods for document understanding, we turn to generation, targeting the core constituents of Visually Rich Documents: styled text and figures. We analyze existing evaluation approaches for styled text generation and propose an evaluation score specifically sensitive

to calligraphic and typographic style, addressing a systematic flaw in the use of generic image generation metrics for this task. We then introduce methods for controllable handwritten and typewritten styled text generation, conditioning on content, style, and geometric variation. For visual content with extreme aspect ratios, such as panoramas, banners, and wide illustrations, we adapt large-scale pretrained diffusion models in a training-free manner, coordinating multiple generation views to produce globally coherent outputs at arbitrary aspect ratios without introducing new trainable parameters. Additionally, we tackle illustrations, characterized by irregular shapes and semi-transparent elements, and propose a pipeline to create, from a textual prompt, transparency-aware (RGBA) assets ready for compositing in complex layouts. Finally, we address multi-instance infographic editing by introducing an attention-partitioning mechanism for flow matching models that disentangles concurrent edits across multiple localized regions, preventing attribute leakage and enabling single-pass editing. We accompany this contribution with the first large-scale benchmark for multi-instance infographic editing, comprising images with up to 285 simultaneous edit instances, an order of magnitude beyond prior benchmarks.

With this thesis, we aim to help close the loop between analysis, generation, and editing, enabling new applications in archive digitization, dataset augmentation, and content creation for Visually Rich Documents, while showing that models, metrics, and benchmarks developed for one stage of the pipeline can directly support the others.

Architetture Multimodali per la Comprensione e la Generazione di Documenti Visivamente Complessi

SOMMARIO

Un tratto distintivo dell'essere umano è la capacità di trasmettere informazioni complesse tramite scrittura e disegno. Questo ha reso possibile la creazione di documenti, il mezzo fondamentale per preservare e trasmettere la conoscenza attraverso il tempo e lo spazio.

Questa tesi di dottorato presenta architetture multimodali di deep learning per la comprensione e la generazione di Documenti Visivamente Complessi, comprendo sia le fonti del patrimonio culturale che i contenuti digitali moderni. Il termine 'Documenti Visivamente Complessi' si riferisce a pagine in cui testo, figure, grafica e layout contribuiscono congiuntamente al significato. Nel caso delle fonti storiche, che spesso hanno secoli di storia, un'analisi affidabile richiede di ridurre la complessità del substrato fisico per rendere trattabili i processi a valle. Sul fronte dell'analisi, introduciamo un modello di binarizzazione e restauro agnostico rispetto a lingua, sistema di scrittura, supporto e risoluzione, che riduce il passaggio d'inchiostro (bleed-through), il rumore e gli artefatti del materiale nelle immagini 2D, preservando i tratti informativi. Estendiamo inoltre questo approccio spettrale ai volumi tomografici 3D, dimostrando un'applicabilità preliminare al rilevamento dell'inchiostro su antichi rotoli carbonizzati. Poiché molti Documenti Visivamente Complessi codificano il significato tramite la disposizione spaziale, sviluppiamo un sistema di parsing multi-pagina che estrae regioni, relazioni e semantica visiva senza ricorrere a un modello esterno di OCR, consentendo il parsing di layout complessi e multi-pagina. Per promuovere la ricerca in contesti con risorse estremamente limitate (ultra-low-resource), curiamo un dataset per la lingua Copta ed effettuiamo il benchmarking dei metodi attuali.

Una volta definiti i metodi per la comprensione dei documenti, passiamo alla generazione, concentrandoci sui componenti fondamentali dei Documenti Visivamente Complessi: il testo stilizzato e le figure. Analizziamo gli approcci di valutazione esistenti per la generazione di testo stilizzato e proponiamo uno score di valutazione sensibile allo stile calligrafico e tipografico, affrontando un limite sistematico nell'uso di metriche generiche per questo compito. Introduciamo poi metodi per la generazione controllabile di testo stilizzato manoscritto e dattiloscritto, condizionando su contenuto, stile e variazione geometrica. Per i contenuti visivi con proporzioni estreme, come panoramiche, banner e illustrazioni orizzontali, adattiamo modelli diffusivi pre-addestrati su larga scala in modo privo di addestramento aggiuntivo, coordinando molteplici viste di generazione per produrre output globalmente coerenti a proporzioni arbitrarie. Inoltre, affrontiamo le illustrazioni, caratterizzate da forme irregolari ed elementi semitrasparenti, e proponiamo un'architettura per creare, da un prompt testuale, asset con trasparenza (RGBA), pronti per il compositing in layout complessi. Infine, affrontiamo le infografiche, la forma visiva più sofisticata di informazione documentale, introducendo un meccanismo di partizionamento dell'attenzione per i modelli di flow matching che disaccoppia le modifiche concorrenti su più regioni localizzate, prevenendo la contaminazione degli attributi e abilitando l'editing multi-istanza in un unico passaggio. Presentiamo inoltre il primo benchmark su larga scala per questo compito, comprendente immagini con fino a 285 istanze da modificare simultaneamente.

Con questa tesi, puntiamo a contribuire a chiudere il cerchio tra analisi e generazione, abilitando nuove applicazioni nell'arricchimento dei dataset (dataset augmentation), nella digitalizzazione degli archivi e nella creazione di contenuti per i Documenti Visivamente Complessi.

Contents

ABSTRACT	I
SOMMARIO	III
1 INTRODUCTION	1
1.1 Motivations and Challenges	2
1.2 Research Objectives	4
1.3 Contributions and Thesis Organization	5
I Understanding VRD	9
2 LITERATURE SURVEY	11
2.1 Document Image Enhancement and Restoration	11
2.1.1 Document Binarization	12
2.1.2 Digital Document Restoration	13
2.1.3 Frequency-Domain Processing	14
2.2 Visual Rich Document Understanding	15
2.3 Handwritten Text Recognition	17
3 DOCUMENT RESTORATION WITH FAST FOURIER CONVOLUTIONS	21
3.1 Document Image Binarization with FourBi	22
3.1.1 Introduction	22
3.1.2 Proposed Approach	24
FourBi Architecture	24
Training and Inference	27
Post-Processing Strategy	28
3.1.3 Experimental Analysis	28
Ablation Analysis	30
Comparison with the State of the Art	34
3.1.4 Summary	39
3.2 Extending FFCs to Volumetric Data: Ink Detection on the Her- culaneum Papyri	40

3.2.1	Proposed Approach	43
	Volumetric Fast Fourier Convolution	43
	Ink Detection Architecture	45
	Training and Inference	46
3.2.2	Experimental Analysis	46
	Experiment Setup	48
	Results	49
	Visualizations	52
3.3	Conclusion	54
4	MULTI-PAGE DOCUMENT PARSING WITH μ GAT	55
4.1	Introduction	56
4.2	Proposed Approach	60
4.2.1	μ gat Model	60
4.2.2	Training and Fine-tuning Data	62
4.3	Experiments	65
4.3.1	Implementation and Training Details	65
4.3.2	Experiment Setup	65
4.3.3	Results	66
4.4	Conclusion	71
5	SCAM: A SAHIDIC COPTIC DATASET FOR LOW-RESOURCE HTR	73
5.1	Introduction	74
5.2	Proposed Dataset	76
5.2.1	Data Preparation	77
5.2.2	Dataset Characteristics	78
5.3	Experimental Analysis	81
5.3.1	Results	83
5.4	Conclusion	88
II Generating VRD		89
6	LITERATURE SURVEY	91
6.1	Generative Models for Image Synthesis	91
6.1.1	Generative Adversarial Networks	91
6.1.2	Diffusion Models	92
6.1.3	Autoregressive Models	93
6.2	Handwritten Text Generation	94

6.3	Controllable Diffusion for Specific Applications	100
7	HWD	103
7.1	Introduction	104
7.2	Handwriting Distance	106
7.2.1	Text Images Representation	106
7.2.2	Perception-Aware Feature Distance	108
7.3	Experimental Analysis	109
7.3.1	Considered Datasets	110
7.3.2	Sensitivity to the Handwriting	111
7.3.3	Sensitivity to the Number of Samples	112
7.3.4	Styled HTG Models Evaluation	114
7.3.5	Ablation Analysis	114
7.3.6	Sensitivity to the Visual Appearance	115
7.3.7	Computation Time Comparison	115
7.4	Conclusion	117
8	AUTOREGRESSIVE STYLED TEXT IMAGE GENERATION	119
8.1	Zero-Shot Styled Text Image Generation with Emuru	120
8.1.1	Method	122
	Training Data	123
	VAE	124
	Autoregressive Transformer	125
	Inference-time Behavior	126
8.1.2	Experiments	127
	Results	130
8.1.3	Summary	134
8.2	Improving Autoregressive HTG with Eruku	135
8.2.1	Eruku Architecture	136
	VAE Image Tokenizer	137
	Autoregressive Text Encoder	138
	Autoregressive Image Decoder	138
8.2.2	Eruku Training	140
	Training Strategy	140
	Training Data	142
8.2.3	Eruku Experiments	142
	Eruku Results	144
8.3	Conclusion	150

9	MERGING AND SPLITTING DIFFUSION PATHS	151
9.1	Introduction	152
9.2	Preliminaries	154
9.3	Proposed Approach	156
9.4	Experiments	159
9.4.1	Results	160
9.5	Conclusions and Discussion	169
10	ALFIE: RGBA IMAGE GENERATION WITH LIMITED RESOURCES	171
10.1	Introduction	172
10.2	Preliminaries	174
10.3	Inference-Time Illustration Generation	176
10.4	Experiments and Results	181
10.4.1	Results	183
10.5	Conclusions	188
III	Editing VRD	189
11	LITERATURE SURVEY	191
11.1	Generative Editing Models	191
12	INSTANCE-DISENTANGLED ATTENTION FOR MULTI-INSTANCE EDIT- ING	195
12.1	Introduction	196
12.2	Proposed Approach	198
12.2.1	Preliminaries	199
12.2.2	Instance-Disentangled Attention	200
12.2.3	Efficient Multi-Prompt Independent Encoding	203
12.2.4	Domain-Specific Fine-Tuning	204
12.3	Infographics Editing Benchmark	205
12.4	Experimental Analysis	207
12.4.1	Results	210
12.5	Conclusion	216
IV	Takeaways	219
13	CONCLUSIONS	221

13.1	Summary of Contributions	221
13.2	Open Challenges and Future Directions	224
13.3	Final Remarks	226
13.3.1	Publications and Achievements	227
13.4	Ph.D. Activities	228
13.4.1	Exchange Periods	228
13.4.2	Research Collaborations	228
13.4.3	Participations to National and International Projects	228
13.4.4	Conference Attendances	228
13.4.5	Seminars and Workshops	229
13.4.6	Schools	229
13.4.7	Conference Service	229
13.4.8	Master Thesis Co-advising	230
APPENDICES		231
APPENDIX A HWD		233
A.1	Additional Results on the Sensitivity to Handwriting	233
A.2	Further Comparison Between HTG Approaches	233
A.3	Alternative Metric Distances for HWD	234
A.4	Additional Results on the Sensitivity to the Number of Samples	235
APPENDIX B ZERO-SHOT STYLED TEXT IMAGE GENERATION WITH EMURU		237
B.1	Further Details about the Training Dataset	237
B.2	Details about the Considered Datasets	240
B.3	Details about the Considered Scores	242
B.4	Analysis on the Style Input Length	244
B.5	Analysis on the Styled Output Length	244
B.6	Detailed Comparison Between VAEs	249
B.7	Multi-stage Training Ablation	249
B.8	Additional Quantitative Results	250
B.9	Additional Qualitative Results	250
B.10	Editing Application	258
APPENDIX C IMPROVING AUTOREGRESSIVE HTG WITH ERUKU		261
C.1	Style Text Reliance Analysis	261
C.2	Ablation on Second Stage Training	262
C.3	Results Including Color Correction	262

APPENDIX D	MERGING AND SPLITTING DIFFUSION PATHS	265
D.1	Effect of the Application Stage	265
D.2	Effect of the Number of Application Steps	266
D.3	Details of the User Study	266
D.4	Further comparison with Direct Inference	267
D.5	Combination with Attention Scaling	268
D.6	Further Results at Different Aspect Ratios	273
D.7	Further Qualitative Results	273
D.8	Further Discussion of the Limitations	274
APPENDIX E	INSTANCE-DISENTANGLED ATTENTION FOR MULTI-INSTANCE EDITING	295
E.1	Masking Application Strategy	295
E.2	Implementation Details	295
E.3	Attention Masking Implementation Analysis	297
E.4	Extended LLM-as-a-judge Evaluation	299
E.5	Prompts for Qualitative Results in the Main Chapter	300
E.6	Additional Qualitative Results	302
APPENDIX F	COLOR PALETTES	311
F.1	μ gat — <i>Alexander Mosaic</i> , c. 100 BCE	311
F.2	SCAM — <i>Rest on the Flight into Egypt</i> , Luc-Olivier Merson, 1879	313
F.3	IDAttn — <i>Le Sacre de Napoléon</i> , Jacques-Louis David, 1807	314
LIST OF PUBLICATIONS		315
BIBLIOGRAPHY		317
GLOSSARY		361
ACKNOWLEDGEMENTS		367

1

Introduction

HUMAN beings have recorded and transmitted knowledge, culture, and history through visual forms for tens of thousands of years, from cave paintings to written artifacts such as ancient manuscripts inscribed on papyrus scrolls, medieval codices, archival registers documenting civic life, and modern infographics. These documents, which we refer to as Visually Rich Documents (VRD), combine textual and visual elements to communicate complex information. With the accelerating digitization of cultural heritage and the exponential growth of digital content creation, the ability to automatically analyze, understand, generate, and edit such documents has become essential to preserve our past and shape our digital future.

VRD present distinct challenges that differentiate them from other Computer Vision domains: they exhibit structured layouts with semantic relationships between elements; they contain text rendered in diverse scripts, languages, and calligraphic styles; they suffer from degradation due to age, conservation conditions, or digitization artifacts. Furthermore, they often exist in specialized domains where annotated training examples are scarce or entirely unavailable. Historical documents add further complexity through non-standard layouts, decorative ele-

ments, preservation-related constraints, and linguistic variations that reflect their cultural and temporal contexts.

While significant advances have been made in Handwritten Text Recognition (HTR) and Document Parsing for modern, well-preserved documents, state-of-the-art techniques often falter when confronted with historical manuscripts, low-resource languages, or documents with complex visual-textual interactions. Moreover, the generation of realistic VRD, whether synthesizing handwritten or type-written text in arbitrary styles, creating visual content with irregular shapes and aspect ratios, or editing multi-element compositions while maintaining global coherence, remains an open challenge. The difficulty lies not only in producing visually plausible outputs but in doing so with fine-grained control, style consistency, and semantic awareness.

This dissertation sets out to address these complex issues by developing methods, datasets, and evaluation protocols that span the complete pipeline of VRD processing: from analysis and understanding, through controllable generation, to precise editing. Importantly, these contributions are not isolated. For example, the Handwriting Distance score, which we have introduced for styled text generation is later reused to evaluate multi-instance infographic editing, illustrating how tools developed for one stage of the pipeline can directly inform another. By making these connections explicit, we aim to enable new applications in cultural heritage digitization, synthetic data augmentation for model training, content creation for digital design, and low-resource document processing.

1.1 MOTIVATIONS AND CHALLENGES

(C1) Analyzing degraded and low-resource documents. A primary motivation for this research emerges from the needs of cultural heritage institutions and the growing demand for automated document processing. Manual annotation by humanities scholars does not scale to large volumes of data, demanding considerable time and expertise. When dealing with documents in low-resource languages or scripts, such as Coptic manuscripts from 11th-century Egypt or ancient Greek scrolls carbonized by the eruption of Mount Vesuvius, annotated training

datasets are either scarce or entirely non-existent. Standard supervised learning approaches fall short in these scenarios. Even when digitized, the physical degradation of such resources, along with complex layouts and visual artifacts, poses substantial obstacles to automatic transcription and layout analysis.

(C2) Generating styled text and evaluating style fidelity. Moving beyond analysis, the synthesis and generation of VRD are motivated by the need to create realistic styled text for data augmentation, document restoration, and controlled content creation when real examples of a script, writer, or style are limited. Handwritten Text Generation (HTG) therefore demands computational models that can capture not only the semantic content of the text but also the stylistic variations of individual handwriting, including the pressure, slant, spacing, and fluidity that characterize a specific writer’s style. Critically, models must learn to disentangle genuine calligraphic features from background artifacts or noise. Traditional metrics struggle to capture this distinction adequately, highlighting a significant gap in how we evaluate generative fidelity in this domain. This evaluation gap is not isolated: a recurring observation across our contributions is that standard off-the-shelf scores often fail to capture task-specific quality, whether for handwriting style fidelity, text recognition in scripts written without word delimiters, or edit correctness in dense infographic layouts. Devising principled, domain-aware evaluation protocols is therefore as important as advancing the models themselves.

(C3) Controlling graphical generation beyond fixed natural-image assumptions. Considering visual content generation, documents such as posters, illustrations, and infographics demand coherent layouts with precise spatial arrangements and variable aspect ratios. Furthermore, modern design workflows require native support for transparency channels to enable the seamless integration of generated assets. Traditional image generative models, which are primarily designed for fixed-resolution, opaque natural images, face fundamental limitations when confronted with these specialized requirements.

(C4) Editing multiple document elements without leakage or loss of coherence. Finally, the editing of VRD introduces additional complexity. Unlike natural image editing, which typically involves manipulating a single salient

object or applying uniform style transfer, document editing often requires coordinated changes across many elements, for instance translating all textual components of an infographic or updating several labels, icons, and visual attributes in a structured layout. One must change multiple text elements or visual components simultaneously while preserving overall coherence, proper alignment, and stylistic consistency. Existing image editing methodologies based on attention control or diffusion guidance often fail to disentangle these multiple instances properly. This failure leads to attribute leakage, where modifications intended for one specific element inadvertently corrupt the appearance of others. For infographics and structured documents, fine-grained localized edits must rigorously respect layout constraints and maintain the overarching semantic relationships between elements. Moreover, naive approaches to multi-instance editing, such as applying edits sequentially or running one diffusion branch per instance, scale linearly in time and memory with the number of instances, quickly becoming computationally infeasible when dozens of elements must be edited simultaneously.

1.2 RESEARCH OBJECTIVES

Within the broad objective of advancing the analysis, generation, and editing of Visually Rich Documents, this dissertation formally explores a set of interconnected research avenues aimed at overcoming the aforementioned challenges.

(Q1) First, motivated by **C1**, we seek to understand how to effectively analyze and parse complex documents, particularly in scenarios characterized by severe degradation or limited annotated data. This involves investigating new architectural designs, such as volumetric analysis for 3D document scans and long-range attention mechanisms for multi-page parsing, to overcome the limitations of standard 2D, single-page methodologies. A central focus is on defining robust strategies for processing low-resource languages, asking how we can best leverage limited data architectures to transcribe ancient and underrepresented scripts without relying on massive labeled corpora.

(Q2, Q3) Second, motivated by **C2** and **C3**, we aim to advance controllable generation for visually rich content under realistic computational and data con-

straints. A first objective is to determine how generative models can adapt to unseen handwriting styles in a zero-shot setting, producing realistic text without expensive per-style fine-tuning. Concurrently, we investigate how to systematically evaluate the performance of these models, aiming to formulate metrics that accurately reflect genuine style fidelity detached from background noise. In the broader context of graphical generation, we seek to adapt existing paradigms, such as diffusion models, to produce content with variable aspect ratios and transparency channels while preserving coherence and without incurring prohibitive retraining costs.

(Q4) Lastly, motivated by **C4**, we explore the challenges of multi-instance manipulation in document editing. Our goal is to investigate how to achieve disentangled control over multiple elements within a single document or image. We aim to design methodologies that grant fine-grained editing capabilities over textual and visual elements, ensuring that localized interventions do not disrupt the global layout, structural integrity, or semantic coherence of the input image.

1.3 CONTRIBUTIONS AND THESIS ORGANIZATION

This dissertation makes contributions across three fundamental aspects of VRD processing: analysis, generation, and editing. The overarching goal is to bridge the gap between understanding and creation. Our work not only enables novel applications in archive digitization and content creation but also advances the theoretical understanding of how to design models tailored to the unique characteristics of complex structured documents.

The thesis is organized into three main parts, mirroring the contribution trajectory of the research.

PART I: UNDERSTANDING VISUALLY RICH DOCUMENTS.

In Part I, we tackle the analysis and understanding of VRD, focusing on their restoration, parsing, and recognition across diverse conditions and domains. This part moves from recovering degraded visual evidence, to making document structure explicit across pages, to transcribing historically constrained material under low-resource conditions.

Chapter 2 provides a comprehensive survey of the current literature in document analysis, covering document binarization, restoration, VRD understanding, and HTR, establishing the foundation for our subsequent contributions.

In Chapter 3, we present **FourBi**, a document binarization model based on Fast Fourier Convolution (FFC). Designed to be resolution-agnostic and robust to heavy degradation, FourBi captures both local texture and global structural patterns. We further extend this approach to **vFFC**, a volumetric variant devised to perform ink detection on 3D tomographic scans of carbonized ancient Greek scrolls, demonstrating the efficacy of spectral convolutions beyond 2D mediums.

Chapter 4 addresses the challenge of multi-page document parsing. We introduce **μgat**, a network that leverages global attention mechanisms across pages to capture long-range dependencies and contextual relationships that single-page models inherently miss.

In Chapter 5, we introduce **SCAM**, a novel low-resource dataset of Sahidic Coptic manuscripts from 11th-century Egypt. We use this benchmark to evaluate current HTR models, highlighting the complexities of limited annotated data and providing insights into transfer learning strategies for historical scripts.

Alongside the proposed models, this thesis also contributes reusable research infrastructure in the form of the SCAM benchmark for low-resource handwritten text recognition, the HWD score for style-sensitive evaluation, and the Infographic Editing Benchmark for stress-testing dense multi-instance editing at scales beyond prior benchmarks.

PART II: GENERATING VISUALLY RICH DOCUMENTS.

In Part II, the focus shifts to the generation of VRD, encompassing both synthetic handwritten text and broader visual content synthesis. The progression here runs from evaluation to control: we first establish how style fidelity should be measured, then study how controllable generation can be extended from styled text to wider visual assets under minimal retraining assumptions.

Chapter 6 surveys the literature on generative models, highlighting the evolution of approaches targeting HTG and visual content creation, while critically analyzing the challenges related to style control and background disentanglement.

In Chapter 7, we introduce **HWD** (Handwriting Distance), an evaluation

score specifically designed to assess style fidelity in HTG. Unlike generic scores, HWD explicitly penalizes models that memorize background artifacts, providing a more reliable and task-specific measure for generative handwriting models.

Chapter 8 presents **Emuru** and **Eruku**, autoregressive models designed for zero-shot styled text image generation. Trained on vast synthetic datasets, these models are capable of generating handwritten text in entirely unseen calligraphic styles at inference time. Their autoregressive formulation allows them to generate variable-length text images and to determine when generation should stop, adapting the output to the requested content rather than relying on a fixed canvas length.

Shifting to graphical content, Chapter 9 introduces **MAD**, a training-free approach for generating coherent panoramas with variable aspect ratios using diffusion models. By strategically merging and splitting denoising paths across overlapping regions, MAD averts seam artifacts and preserves global coherence in ultra-wide or ultra-tall outputs.

In Chapter 10, we present **Alfie**, a pipeline that extends text-to-image capabilities to generate RGBA illustrations with native transparency. Operating without expensive architectural retraining, Alfie produces assets ready for direct compositing into graphic design and infographic layouts.

Across these chapters, the common thread is controllable generation with minimal retraining: HWD establishes how to measure style fidelity, Emuru and Eruku address zero-shot styled text generation, MAD adapts diffusion models to arbitrary aspect ratios at inference time, and Alfie extends the same low-cost adaptation logic to transparency-aware asset creation.

PART III: EDITING VISUALLY RICH DOCUMENTS.

In Part III, we address the editing of VRD, synthesizing the analytical and generative capabilities developed in the preceding sections. This final part closes the loop by turning localization and controllable rendering into scalable multi-instance editing for dense, document-centric layouts.

Chapter 11 reviews the literature on generative image editing and recent advances in attention-based control, multi-instance editing, and flow matching models.

In Chapter 12, we present **IDAttn** (Instance-Disentangled Attention), a multi-instance image editing approach. IDAttn introduces an inference-time attention-partitioning strategy for disentangling joint attention across multiple instances, preventing attribute leakage. For particularly challenging out-of-domain scenarios such as dense infographic text editing, lightweight LoRA domain adaptation, training fewer than 1% of parameters, provides substantial further gains. Furthermore, to stress-test multi-instance editing at scale, we introduce the **Infographic Editing Benchmark**, the first benchmark for multi-instance infographic editing, featuring up to 285 simultaneous edit instances per image, far exceeding any prior natural image editing benchmark in density and complexity.

FINAL REMARKS.

Finally, Part IV provides a summary of the main takeaways of this dissertation. We reflect on the lessons learned from our holistic approach to VRD processing, arguing that the true potential of the field lies in systems where understanding, generation, and editing continuously inform and enhance one another. By addressing the multifaceted challenges of text and image manipulation within structured mediums, this dissertation lays the groundwork for more adaptable, resilient, and unified document processing pipelines.

Part I

Understanding VRD

2

Literature Survey

THIS part of the thesis focuses on analysis and understanding of Visual Rich Documents, encompassing document image enhancement, restoration, and interpretation. Visual Rich Document Understanding (VrDU) requires addressing multiple interconnected challenges, from low-level image processing tasks such as binarization and restoration of degraded historical documents, to high-level semantic tasks such as parsing and handwritten text recognition. In this literature survey, we review the key research directions that inform our contributions: document image enhancement and restoration techniques, VrDU datasets and models, and handwritten text recognition approaches.

2.1 DOCUMENT IMAGE ENHANCEMENT AND RESTORATION

Document image enhancement and restoration encompass techniques for improving the visual quality of degraded documents, from simple binarization of modern scanned documents to complex restoration of severely damaged historical artifacts. These tasks share common challenges, namely dealing with degrada-

tion, noise, and information loss, and increasingly leverage similar computational strategies, particularly frequency-domain processing and deep learning architectures.

2

2.1.1 DOCUMENT BINARIZATION

HEURISTICS-BASED APPROACHES. Among the strategies that exploit a hand-crafted model of the image appearance, the most commonly applied is thresholding, which consists in determining a global threshold [266, 35] or a local adaptive threshold to deal with the non-uniform background typical of document images [320]. Note that hybrid thresholding algorithms exist, such as Wolf [392], which use both local and global statistics. Other binarization paradigms are based on energy minimization [401, 402] or on edge detection for modeling strokes [92, 414].

LEARNING-BASED APPROACHES. Most learning-based approaches to Document Binarization exploit convolutional neural networks (CNNs). These are employed as denoising autoencoders [429, 141] or as encoder-decoders to perform Image Processing [180, 101] or Segmentation [362, 381]. In this work, we treat Document Binarization as a Segmentation task and devise a U-Net-like model [310] as convolutional architecture, but with the novelty of featuring FFC layers. Another successful research line entails treating the Document Binarization task as a generative one by applying Conditional Generative Adversarial Networks (GANs) to obtain a binarized image starting from a degraded one in a style-transfer fashion [430, 341, 356, 342]. In the context of Document Image Binarization, GANs are also employed to alleviate the training data scarcity issue connoting the task. In particular, these models are used for realistic training data augmentation [39] or to achieve self-supervision via unpaired Image-to-Image Translation with Cycle-GANs [439]. Recent preliminary approaches employ Diffusion Models [136, 413, 132] in a similar way. To exploit long-range dependencies in combination with local information, which is difficult to achieve with standard convolutions, recent approaches [339, 340, 410] resort to Vision Transformer-based architectures [104]. The encouraging performance of such methods sug-

gests that combining global and local information is a promising strategy for Document Binarization. In this work, we explore this strategy by proposing to apply Fast Fourier Convolutions (FFCs), which, contrary to standard convolutions, can model both global and local dependencies and are less computationally demanding than Vision Transformers (ViTs).

Regardless of the technique applied, the performance can be enhanced with some post-processing steps. Typically, these consist in applying morphological operators (*e.g.* Connected Component Analysis, Erosion, Dilation, Shrink, and Swell filtering) [228, 246] or ensembling the output of multiple binarization algorithms [154, 249, 13]. In this work, we combine the output of our model run on overlapping patches of the document of interest to reduce the error on the borders, where artifacts due to discontinuity are most evident.

2.1.2 DIGITAL DOCUMENT RESTORATION

VIRTUAL UNWRAPPING. DDR encompasses Digital Imaging, Image Processing, and Computer Vision techniques for non-invasive content recovery from severely damaged ancient documents [103, 54, 276, 56, 428, 123]. One of these techniques is Virtual Unwrapping [323], which is mostly applied to documents that are too fragile to handle and analyze, as in the case of scrolls. First, the entire scroll is scanned, typically with X-ray micro-computed tomography (micro-CT) [185]. Then, each scroll sheet in the volumetric representation is segmented and projected into a 2D image. Depending on the specific use case, an additional texturing step (such as ink detection) can be performed. Early works operated on small contrived samples with semi-manual pipelines [324, 213]. The first fully-automated solution was proposed by Samko *et al.* [316], who developed a novel graph cut method for parchment scrolls sheets segmentation. In [153, 36], the authors read the contents of a damaged silver scroll by exploiting the specific characteristics of the material, such as the engraved and ruled text. Finally, the carbonized En-Gedi scroll was read in 2015 by Seales *et al.* [325].

The Herculaneum papyri represent a particularly challenging case for DDR [322]. Firstly, the papyrus layers have been compressed, crumpled, and deformed by

the carbonization process, making sheet segmentation not straightforward. Secondly, the text is mostly written using carbon-based ink, which is almost invisible on the carbonized papyrus support in the X-ray scans. Thus, traditional virtual unwrapping, which heavily relies upon the visibility of ink or simple layers structure, is not applicable. In a recent work, Parker *et al.* [269] proved the detectability of carbon-based ink in micro-CT scans of the Herculaneum papyri using a 3D convolutional neural network [167] trained on subvolumes to detect ink in the central voxel. A further step towards DDR of Herculaneum papyri is due to Parsons *et al.* [270], who proposed an open dataset containing volumetric scans of such collection. In this work, we tackle the problem of ink detection on the surface volumes of Herculaneum papyri fragments.

2.1.3 FREQUENCY-DOMAIN PROCESSING

Previous approaches [257, 12, 214] have successfully leveraged the spectral information of document images to address the difficulty of recognizing and recovering structures such as edges, strokes, and corners. Both classical [257] and learning-based methods [12, 214] benefit from frequency-domain representations for document enhancement and restoration tasks. In recent works, the Discrete Wavelet Transform [238], successfully applied to Super-Resolution [434], has been used for Document Binarization to represent the document image in different frequency sub-bands either to perform Segmentation [12] or to obtain a ground truth for a generator in an adversarial training setting [214].

In this work, the main component of our architecture is the FFC, which exploits the Fast Fourier Transform (FFT). The FFC operator combines 2D convolutions and 2D Discrete Fourier Transform (DFT) [83] and has been applied to spatial data for Computer Vision tasks such as inpainting [138], semantic segmentation [38], and super-resolution [331, 420]. Unlike previous works that use spectral information to represent the images, we employ it as an operator inside our end-to-end architecture for both binarization and restoration tasks. For the volumetric restoration of Herculaneum papyri, we extend the 2D FFC operator proposed by Chi *et al.* [75] to process volumetric data. Some attempts have

been made to apply FFC to spatio-temporal data [75]. Nonetheless, volumetric data are different from spatio-temporal ones, and thus, require ad hoc solutions [371, 294, 126, 267]. Therefore, we argue that a solution based on spatio-temporal FFCs is not optimal for modeling the correlations between dimensions in volumetric data. In sight of this, we propose the vFFC operator, which features 3D convolutions and 3D Discrete Fourier Transform and better handles this kind of data.

2.2 VISUAL RICH DOCUMENT UNDERSTANDING

DATASETS FOR DOCUMENT UNDERSTANDING. The field of VrDU encompasses both general and task-specific datasets [200, 374, 200, 137, 433, 206, 165, 240, 378, 377]. The large scale of these datasets and the emergence of the Transformer [380] has revolutionized Document AI [406, 407, 158, 188, 94, 204, 48, 359, 20, 21, 197, 234]. However, these benchmarks have been shown to suffer from single-domain scope, limited document types, and single-page samples [378, 377]. With MP-DOCVQA [366], DUDE [378], and RVL-CDIP_{MP} [377], the research community has started to consider the multi-page case for sub-tasks like (Document) Visual Question Answering and Classification. In this work, we focus on parsing, which refers to translating a document’s image into a structured markup representation [48]. In [48], most of the training documents are parsed from arXiv. In particular, the authors process the source files by using $\text{\LaTeX}$ XML* to standardize and convert them into HTML5. Then, they convert the HTML files into Markdown and split the pages by exploiting heuristics and fuzzy matching. However, this procedure results in independent single-page samples, and the parsing model trained on such data is only capable of handling single pages without considering that some elements can span multiple pages. To overcome this limitation, in this work, we exploit a specifically-collected dataset of academic documents containing both contiguous and sparse pages to allow the training and benchmarking of flexible models capable of managing the context when provided.

* dlmf.nist.gov/LaTeXML/

DOCUMENT UNDERSTANDING METHODS. VrDU has attracted increasing interest in recent years and is dominated by Transformer-based models [380]. Methods can be categorized whether they include an external OCR input or not. In the former category, the LayoutLM family [406, 407, 158] is based on the Transformer Encoder BERT [100] and performs masked layout prediction. In DocFormer [20, 21] and UDOP [359], Transformers are trained to perform several pre-training tasks and align the textual and visual information. On the other hand, OCR-free methods have recently shown that competitive results can be achieved without using an external OCR [94, 188], which also increases generalization to unseen languages and domains. In parsing, the recent Nougat approach [48] has achieved state-of-the-art performance on academic document parsing. This architecture is based on a Swin Encoder [224] that takes as input single-page images and outputs embedding vectors. These are then processed by a mBART [201] decoder that autoregressively outputs the Markdown code by using the same tokenizer as [360]. Note that all the approaches mentioned above, including Nougat, entail considering documents as single pages. This is often not the case in real-world applications, where documents exhibit parts (*e.g.* paragraphs, references, or footnotes) that can sit between multiple pages. Therefore, in our proposed approach, we adopt the Nougat architecture and extend it to process multi-page inputs to increase the consistency of its output.

MULTI-PAGE DOCUMENT UNDERSTANDING. Some previous works have proposed approaches to extend pretrained VrDU models working on single pages for handling multi-page inputs, especially for the DQA task (with the multi-page DocVQA variant of the task [179]). For example, in Hi-VT5 [366], the authors fine-tuned the Document Image Transformer model [204] to encode the input the single-page images into multiple global tokens, which are then fed to a T5 [299] model, trained to align layout and text representations [46]. However, such global tokens contain too coarse information, which can be sufficient for VQA but would be unsatisfactory for document parsing since this latter task requires detailed representations. Another example is GRAM [47], in which a single-page DocFormerv2 [21] encoder is extended to use multiple pages as input by inserting global tokens in the page representation and alternating attention operations

over the single pages and the global tokens. In this work, we take inspiration from the Perceiver [163] model and implement an adapter that extracts relevant information from the previous and next pages and inserts them into learnable tokens, which are then concatenated to the representation of the current page so that the decoder can parse it while keeping into consideration some information from the context.

2.3 HANDWRITTEN TEXT RECOGNITION

Handwritten Text Recognition (HTR) can be approached at various levels of granularity, including characters [162, 79, 80] (useful for ideogrammatic languages or rare scripts, such as ciphers), words [129, 40, 349, 329], lines [274, 293, 382, 51, 77, 251, 50], paragraphs [50, 49, 391], or full pages [80, 251]. Among these, line-level HTR is the most widely adopted, typically assuming pre-segmented text lines [274, 382, 51, 293, 77]. These systems can also be integrated with layout analysis and segmentation to achieve paragraph- or page-level recognition [50, 251, 417]. In this work, we focus on pre-segmented line-level HTR, given its popularity and easy integration into page-level pipelines.

EVOLUTION OF HTR METHODS. Early HTR systems were based on Hidden Markov Models (HMMs) to model sequences of explicitly-isolated character images and n -gram language models to predict textual outputs [369, 370]. Such an explicit character isolation approach was then surpassed by end-to-end approaches, starting from the Deep Learning-based solution proposed by Graves *et al.* [129], which consists of a multidimensional LSTM network that generates 2D feature maps directly from handwritten images, which are then reduced to 1D sequences and decoded using Connectionist Temporal Classification (CTC) algorithm. This approach dominated the field [49, 252, 96] until computationally simpler and faster alternatives were proposed [329, 293], using CNNs for feature extraction and 1D-LSTMs for sequence modeling. These backbones combining CNNs and recurrent models remain popular due to their strong performance and efficiency [274, 382, 51, 77, 81]. Nonetheless, a more recent direction models HTR as a sequence-to-sequence problem, where the image is rep-

resented as a sequence of patches and decoded into text using encoder-decoder architectures [350, 247, 427, 11]. Models implementing this paradigm can be trained by using cross-entropy loss, CTC loss, or both. Transformers [380] have also been adopted for HTR to replace recurrent neural networks entirely, often leveraging large-scale pretraining on synthetic or real text data [176, 390, 205]. A representative example of this trend is TrOCR, a Transformer-based architecture that uses a Vision Transformer (ViT) encoder and an autoregressive text decoder pre-trained on synthetic data and fine-tuned for both printed and handwritten text. Fully convolutional networks have also been explored as a recurrence-free alternative, often incorporating GateBlocks to capture contextual dependencies similar to recurrent networks but with improved parallelism and speed. Finally, it is worth mentioning that the emergence of Large Multimodal Models, such as GPT-4V [412], has opened new avenues for HTR by enabling models to process and understand both visual and textual information. Preliminary explorations suggest that these models can perform HTR tasks with minimal task-specific supervision, although they currently lag behind specialized systems in accuracy and robustness, especially in low-resource settings.

BENCHMARKS FOR HTR. Designing effective HTR systems requires large and diverse datasets that reflect the visual complexity of handwriting and the linguistic richness of the task. Widely used modern-language benchmarks include the IAM [239] and RIMES [22] datasets, which contain English and French lines, respectively. These datasets are written by multiple authors under controlled conditions. IAM writers transcribe sentences from the Lancaster-Oslo/Bergen corpus [171], while RIMES contributors follow a script to simulate customer service communication.

For historical HTR, the ICFHR14 [317], ICFHR16 [318], and ICFHR18 [344] datasets are among the most notable. ICFHR14 includes English lines from the Bentham Papers [60], mostly authored by Jeremy Bentham and collaborators. ICFHR16 (also called READ) focuses on ancient German manuscripts from the Ratsprotokolle collection, written over three centuries. ICFHR18 includes documents in German and Italian, from diverse time periods and collections, with test sets defined at the document level. Other historical datasets include Ro-

drigo [326] and Germana [273], which contain Spanish text written by single authors, and the LAM dataset [59], which has been introduced to support research on ancient Italian single-author manuscripts with temporal variability in writing.

Other small historical datasets exist, which are commonly adopted for exploring low-resource HTR scenarios, which is also the focus of this work. Examples of these datasets are George Washington [114], Parzival [113], Saint Gall [115], Esposalles [309], and Leopardi [58], which contain historical documents written by a few or single authors, often from narrow time spans, in a relatively well-represented language (English, Latin, and Italian). Particularly challenging examples of low-resource scenarios are explored via datasets of ciphered manuscripts, such as Borg and Copiale [31]. These contain text in Latin and German, respectively, rendered in uncommon characters (either existing or abstract symbols). To further explore the low-resource HTR task, we propose SCAM, a dataset of manuscripts written in an underrepresented, extinct language using uncommon symbols (Sahidic Coptic), whose main characteristics compared to existing datasets are synthesized in Section 2.3.

Taken together, the literature reviewed in this chapter motivates the structure of Part I. Chapter 3 addresses image-level context for restoration and ink detection through frequency-domain operators, Chapter 4 addresses inter-page context for document parsing, and Chapter 5 addresses data scarcity and script-specific difficulty in low-resource handwritten text recognition. In this sense, the part progresses from recovering document content, to structuring it, to transcribing it under realistic historical constraints.

	Lines	Lexicon	Period	Language	Authors
IAM [239]	10,373	9,749	Modern	English	Many
RIMES [22]	12,111	8,760	Modern	French	Many
Washington [114]	656	1,471	1755	English	Two
Saint Gall [113]	1,410	5,436	ca 890-900	Latin	One
Esp. Index [309]	1,563	1,725	1491-1495	Catalan	One
Leopardi [58]	2,459	5,067	1818-1832	Italian	One
Parzival [114]	4,477	4,934	ca 1265-1300	German	Two
Esp. Licenses [309]	5,447	3,465	1616-1619	Catalan	One
ICFHR14 [317]	11,473	9,716	ca 1760-1832	English	Many
ICFHR16 [318]	10,550	8,120	1470-1805	German	Many
ICFHR18 [344]	14,803	23,198	Mixed*	German, Italian	Many
Rodrigo [326]	20,357	17,300	1545	Spanish	One
Germana [273]	20,529	27,100	1891	Spanish [†]	One
LAM [59]	25,823	23,428	1691-1750	Italian	One
SCAM (Ours)	3240	-	1002-1003	Sahidic Coptic	One

1 - Both Medieval and Modern.

2 - A small number of lines are in different languages.

Table 2.1: Characteristics of existing line-level benchmark datasets compared to ours.

3

Restoring Documents by Leveraging Fast Fourier Convolutions

MODELING both local and global information is crucial in several Document Analysis and Computer Vision tasks, especially when dealing with degraded documents. Standard convolutions, while effective at capturing local patterns, struggle to model global dependencies due to their fixed receptive field. In this chapter, we explore the use of Fast Fourier Convolution (FFC) as a powerful alternative that combines spatial and spectral information to expand the receptive field in the frequency domain. We first present FourBi, an efficient fully-convolutional model for Document Image Binarization that exploits FFCs to handle both local and global information on 2D document images. Then, we extend the FFC operator to volumetric data, proposing the volumetric Fast Fourier Convolution (vFFC), and apply it to the challenging task

This Chapter is related to the publications “F. Quattrini *et al.*, Binarizing Documents by Leveraging both Space and Frequency, ICDAR 2024” [4] and “F. Quattrini *et al.*, Volumetric Fast Fourier Convolution for Detecting Ink on the Carbonized Herculaneum Papyri, ICCVW 2023” [2]. See the list of Publications on page 315 for more details.

of ink detection on the carbonized Herculaneum papyri. Together, these works demonstrate the versatility and effectiveness of frequency-domain reasoning for document analysis, from 2D binarization to 3D ink detection.

3

3.1 DOCUMENT IMAGE BINARIZATION WITH FOURBI

Document Image Binarization is a well-known problem in Document Analysis and Computer Vision, although it is far from being solved. One of the main challenges of this task is that documents generally exhibit degradations and acquisition artifacts that can greatly vary throughout the page. Nonetheless, even when dealing with a local patch of the document, taking into account the overall appearance of a wide portion of the page can ease the prediction by enriching it with semantic information on the ink and background conditions. In this respect, approaches able to model both local and global information have been proven suitable for this task. In particular, recent applications of Vision Transformer (ViT)-based models, able to model short and long-range dependencies via the attention mechanism, have demonstrated their superiority over standard Convolution-based models, which instead struggle to model global dependencies. In this work, we propose an alternative solution based on the recently introduced FFC, which overcomes the limitation of standard convolutions in modeling global information while requiring fewer parameters than ViTs. We validate the effectiveness of our approach via extensive experimental analysis considering different types of degradations. We release code and checkpoints at ^{*}.

3.1.1 INTRODUCTION

Documents with good visual quality are easier to handle when performing Document Analysis tasks, which is why the problems and approaches to Document Enhancement have long been explored. However, this remains a challenging task, especially for old and highly damaged documents, for which automatically achieving good visual quality can benefit efficient content understanding. Due

^{*}<https://github.com/aimagelab/FourBi>

to preservation conditions and time, these documents exhibit various types of degradations, such as bleed-through ink, ink oxidation, faded ink, stains, smears, scratches, and creases in the paper. Moreover, the digitization process can introduce artifacts due to image compression and non-ideal acquisition conditions such as uneven illumination, contrast variation, and blur [363]. Document Binarization is a typical Document Enhancement task which involves determining whether each pixel in the document image pertains to ink or to the background. This is a critical step both for facilitating downstream Document Analysis tasks [343, 303, 82, 98, 166, 339, 414, 191, 279, 277] and for enhancing the readability of the document by the end-users of digital archives and libraries. One of the main challenges in Document Binarization is the variability with which the aforementioned degradations appear within the page. For this reason, early-proposed image processing approaches based on estimating and applying a global threshold [266, 35] have been surpassed by methods working on local document patches, on which an adaptive threshold is calculated and applied. Learning-based solutions, generally based on CNNs, enhanced the performance by refining the prediction on the single pixel based on its neighbors. Nonetheless, a potential drawback of working on document patches is that, in case they contain too different degradations, some artifacts can appear at their borders when restitching them together to form the whole document image.

For this reason, considering both global and local information has been proven beneficial by the promising performance of hybrid thresholding methods [392], Image-to-Image translation based on GANs [430, 341, 356], and multi-scale Convolutional approaches for Segmentation [362, 381]. Note that global dependencies are difficult to handle with standard Convolution-based models due to the fixed size and locality of their receptive field. A strategy to integrate global information is to exploit attention mechanisms over the pixels. This strategy has been successfully applied also for Document Binarization [339, 340, 410]. However, this strategy requires a large amount of training data, which is difficult to collect for document images, and usually leads to large models, thus being inefficient.

To overcome these limitations, we devise an efficient and fully-convolutional model to integrate global information for Document Binarization by exploiting

the Fast Fourier Convolution [75]. This operator combines spatial and spectral information to expand the receptive field in the frequency domain. The benefits of FFCs have been exploited for Inpainting [355] and Super-Resolution [331] but, to the best of our knowledge, FFCs have not been employed for Document Binarization. However, we argue that their capability to encode both local and global information and model pseudo-periodic patterns and textures is particularly suitable for documents. In fact, text lines have a specific pseudo-periodic pattern, and paper has a very peculiar texture, different from that of artifacts and degradations. Moreover, due to their image-wide receptive field in the spectral domain, FFCs are more robust to differences in resolution and scale between training and test images compared to standard convolutions [355]. Thus, we propose to train our model with 256×256 patches, as customary in Document Binarization, but use larger patches at inference time. This allows obtaining more training samples from the same annotated document and seamlessly exploiting more global information at inference time for better binarizing the patch. We demonstrate the effectiveness of our proposed approach and its superiority with respect to both standard Convolution- and Transformer-based approaches via extensive experimental analysis on multiple benchmark datasets, whose contained documents exhibit a number of different degradations.

3.1.2 PROPOSED APPROACH

In this work, we treat the Document Binarization task as a two-class segmentation problem, with a high imbalance towards the background class and high intra-class variability. To tackle the task, we devise a novel approach consisting of a U-Net-like model featuring FFC layers, which we call *FourBi* (see Fig. 3.1).

FOURBI ARCHITECTURE

The proposed model takes as input square image patches, which are down-scaled via three Down-scale Blocks. Each Down-scale Block contains three sub-blocks, consisting of a Convolutional layer (Conv), a Batch Normalization (BN) layer, and a Rectified Linear Unit (ReLU) activation function. The output of the last

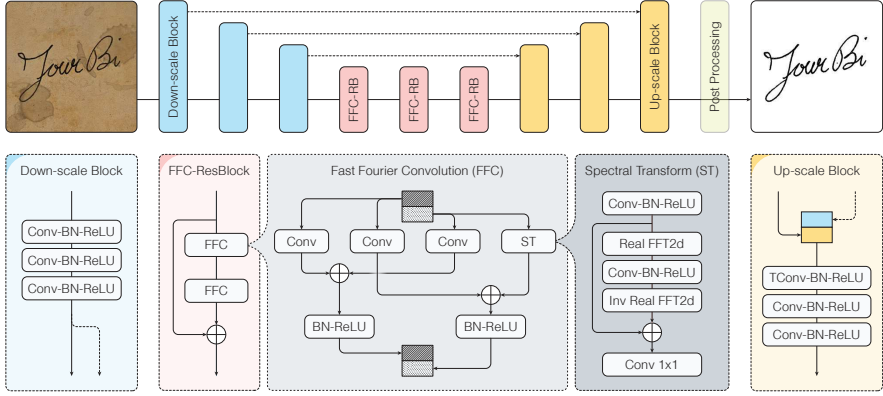


Figure 3.1: Overview of our Document Binarization approach exploiting FFCs (FourBi). The architecture consists in a U-Net-like model in which the outputs of the down-scaling and the up-scaling blocks are combined via concatenation. The central part of the model features FFC Residual Blocks (FFC-RB). The FFC operator in these blocks combines local spatial information with global frequency information, which results in semantic-preserving precise binarization.

Down-scale Block is fed to a stack of FFC Residual Blocks (FFC-RB), which apply the FFC operator described below. The feature map from the last FFC-RB is then up-scaled via three Up-scale Blocks. Each of these blocks takes as input the output of the preceding Up-scale Block concatenated with the output of the corresponding Down-scale Block. The so-obtained tensor is fed to a Transpose Convolution (TConv) layer, followed by a BN layer, a ReLU, and two other Conv-BN-ReLU blocks. The predicted tensor contains continuous values ranging from 0 to 1, representing the probability of each pixel being background or ink. From this, we obtain a binarized image by applying a threshold set equal to 0.5.

FAST FOURIER CONVOLUTION FOR DOCUMENT BINARIZATION. To obtain a high-quality binarized version of a document, we need the global context of the whole image in order to better model the patterns connoting it. In fact, the paper texture, the ink intensity, the color artifacts of the image scan, and the text orientation are aspects that change from document to document but are roughly similar in different parts of the same document page. This must be taken into account to distinguish the ink pixels from the background ones. However, work-

ing with document images also requires precision when binarizing fine parts like strokes and characters in order not to change the semantic content of the text in the document. In light of this, the global context has to be mixed with local information to classify the single pixel correctly. Therefore, we propose to resort to the FFC operator proposed in [75], which divides the network stream into a global and a local branch. The global branch exploits the Fast Fourier Transform to enlarge the receptive field, while the local branch handles finer local patterns. In particular, the input tensor to the FFC layer is split into two chunks along the channel dimension, one for the local branch and one for the global branch. This way, the two branches can focus on different aspects, and different information can be stored in different parts of the tensor. The local branch consists of two Convolutional layers, while the global branch contains a Convolutional layer and a Spectral Transform (ST) component. The ST block exploits the two-dimensional Real Fast Fourier Transform (RFFT) operator (Real FFT2d) and applies a convolution in the frequency domain. Note that the Real FFT is generally faster than the classical FFT for real-valued signals because it takes advantage of the Hermitian symmetry property. Specifically, the block works as follows. Given an input tensor $\mathbf{X}$ with shape $[H \times W \times C]$, it is applied the Real FFT2d:

$$\text{Real FFT2d} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{C}^{H \times \frac{W}{2} \times C},$$

to obtain a complex-valued tensor $\mathbf{Z}$, *i.e.* a tensor containing tuples that represent the real part $\mathbf{R}$ and the imaginary part $\mathbf{I}$ of the input:

$$\text{Real FFT2d}(\mathbf{X}) = \mathbf{Z} = \mathbf{R} + i\mathbf{I}.$$

Then, the values in these tuples are stacked to form a real-valued tensor, *i.e.*

$\mathbb{C}^{H \times \frac{W}{2} \times C} \rightarrow \mathbb{R}^{H \times \frac{W}{2} \times 2C}$. This tensor is fed to a Convolutional block featuring a Convolutional layer with kernel size 1×1 :

$$\text{Conv}_{1 \times 1} \circ \text{BN} \circ \text{ReLU} : \mathbb{R}^{H \times \frac{W}{2} \times 2C} \rightarrow \mathbb{R}^{H \times \frac{W}{2} \times 2C}.$$

Finally, the output of this block is reshaped again into a complex-valued ten-

sor, *i.e.* $\mathbb{R}^{H \times \frac{W}{2} \times 2C} \rightarrow \mathbb{C}^{H \times \frac{W}{2} \times C}$ and fed to a two-dimensional Inverse Real Fast Fourier Transform (IRFFT) operator (Inv Real FFT2d):

$$\text{Inv Real FFT2d} : \mathbb{C}^{H \times \frac{W}{2} \times C} \rightarrow \mathbb{R}^{H \times W \times C}.$$

The output of the Convolutional layer in the global branch is summed with the output of one of the convolutions in the local branch, and the output of the ST is summed to the output of the other. The resulting tensors are fed to separate BN-ReLU blocks and then stacked to obtain the output of the FFC block. Thanks to the combination of local spatial information and global frequency information, the FFC allows handling and distinguishing global pseudo-periodic patterns, such as those typical of the paper support and of the text lines, without disregarding local patterns that need to be taken into account when binarizing detailed parts of the image such as the strokes.

TRAINING AND INFERENCE

Since obtaining ground truth information for Document Binarization is costly and time-consuming, publicly available datasets for the task usually contain only a few annotated pages. Therefore, as customary for this task, we enlarge the training set by training on 256×256 image patches. We train our model with the Charbonnier loss [64], which approximates the Mean Square Error (MSE) loss when the prediction error is smaller than the hyperparameter ε and the Mean Absolute Error (MAE) loss when the error is bigger [34]. Note that both the MSE and MAE loss are usually employed in Document Binarization [429, 430, 39, 101, 339, 414]. For this reason, we adopt this loss that has been found more beneficial than these two in tasks such as image Super-Resolution [195]. The loss is:

$$\mathcal{L} = \sqrt{(I - I')^2 + \varepsilon^2},$$

where I and I' are the ground truth and the predicted image, respectively, and the hyperparameter ε is set equal to 10^{-6} .

At inference time, we exploit the robustness of the multi-scale representations



Figure 3.2: We combine the predictions of our model on overlapping patches at each point based on the distance between the point of interest and the center of each patch so that the artifacts at the patches borders are reduced.

learned by FFCs when applied to test images larger than the training ones and feed the model with 512×512 patches from the test images. The increased global context helps the model distinguish between different degradation types and ink.

POST-PROCESSING STRATEGY

In the commonly adopted setting, also applied in this work, the visual quality enhancement of an entire document image is performed by working on separate local patches and then recomposing them. This leads to most of the errors occurring on the borders, causing discontinuities and artifacts in the reconstructed document image. Therefore, to enhance the performance of our approach, we introduce overlap between the patches used for prediction and keep only the innermost predictions of the model. Specifically, at inference time, we run our model over overlapping patches and recombine their prediction of each pixel by assigning to that pixel the value that it assumes in the patch to whose center it is closer. The process is shown in Fig. 3.2. In our experiments, we set the overlap to 256 (half the patch size), which has been determined empirically.

3.1.3 EXPERIMENTAL ANALYSIS

IMPLEMENTATION DETAILS. Before training, we extract 384×384 image patches from the entire document image, with stride 192 pixels. Then, during training, to further enlarge the training set, we apply a random rotation of $\pm 10^\circ$ to these patches, and random resize and crop to obtain the 256×256 patches for training. Finally, randomly alter the image brightness, contrast, saturation, and hue with

factor 0.5. We train the model by using the Adam Weight Decay (AdamW) optimizer [225] with $\beta_1=0.9, \beta_2=0.95$, batch size 8, and weight decay 0.05. Moreover, we set the learning rate to $1.5 \cdot 10^{-4}$ and use a cosine scheduler with 10 linear warm-up steps.

CONSIDERED DATASETS. Various datasets have been proposed in the literature for the Document Binarization task. However, they all contain only a small number of documents because of the complexity of producing the ground-truth annotation masks. This poses a challenge when training deep learning networks, as these require large amounts of data to model the several possible degradations affecting the documents. Moreover, it is important that the set of documents used in training features a number of diverse degradations, text style, ink, and background. For this reason, in this work, we consider a number of publicly available datasets containing both handwritten and typewritten documents, from different periods and with different paper supports and degradations. In particular, we use the ten public benchmark datasets released over the years for the Document Image Binarization Contest (DIBCO), namely DIBCO09 [122], H-DIBCO10 [285], DIBCO11 [286], H-DIBCO12 [287], DIBCO13 [288], H-DIBCO14 [265], H-DIBCO16 [290], DIBCO17 [291], H-DIBCO18 [289], and DIBCO19 [292], each containing from 10 to 20 document images, for a total of 139. The documents feature several degradations representing real-world cases, such as variable background intensity, shadows, smear, smudge, low-contrast, bleed-through, or show-through. The documents in these datasets contain typed, machine-printed, and handwritten characters in the Latin alphabet, except for DIBCO19, which also contains non-Latin text. Note that these datasets are from disjoint sets of documents and feature different types and levels of degradation. Moreover, our training set also includes images from the Bickley Diary [99], Synchromedia Multispectral Ancient Document Images Dataset (SMADI) [143], and Nabuco [217] datasets. These datasets contain documents in the Latin alphabet, both handwritten and typewritten, thus enforcing the variability of our training set.

For evaluation, we consider three modern datasets (DIBCO11, DIBCO13, and DIBCO17) and five historical ones (H-DIBCO10, H-DIBCO12, H-DIBCO14,

H-DIBCO16, and H-DIBCO18). We adopt a leave-one-out strategy, using all other DIBCO datasets for training. Note that we use DIBCO19 only for testing because the patterns of the characters are too different from Latin scripts. Thus, we can evaluate the generalization capabilities of our approach to the non-Latin alphabet domain. In this respect, we also evaluate the performance of our model and some competitors on the non-Latin scripts Persian Heritage Image Binarization Dataset (PHIBD) [256, 26] and Irish Script on Screen Bleed-Through Database (ISOS-BTD) [314] datasets.

EVALUATION SCORES. We evaluate our models by using the F-Measure (FM), the pseudo-F-Measure (pFM) [264], the Peak Signal-to-Noise Ratio (PSNR), which measures the similarity of two images, and the Distance Reciprocal Distortion (DRD) [227], which measures the visual distortion of binary images.

ABLATION ANALYSIS

ADVANTAGES OF THE FFC OVER THE STANDARD CONVOLUTION. To assess the effectiveness of FFCs over standard convolutions (Convs), we define a baseline with the same architecture as FourBi that contains solely Convs. We swap the ST module with a 1×1 convolution to maintain the global-local branch division. The 1×1 convolution, in contrast to the ST module, focuses on local aspects of the input. Therefore, the Conv-based baseline has the global and local branches logically swapped, where the global branch has the 3×3 convolutions, and the local one has the 1×1 convolution. Fig. 3.3 shows the performance comparison of the two models. As we can see, the global receptive field of the FFC allows the model to outperform the Conv-based baseline.

ROLE OF THE TEST PATCH SIZE. We ran an ablation analysis to assess the capability of our model to learn robust representations and exploit the global context provided by larger patches at inference time. In particular, we evaluate our model and the corresponding Convolutional baseline on patches whose size ranges from 128×128 to 704×704 (see Fig. 3.3). We observe that the scores get better consistently, reaching the best value around the patch size of 512.

The same behavior emerges also in Fig. 3.4, which shows how the Conv-based

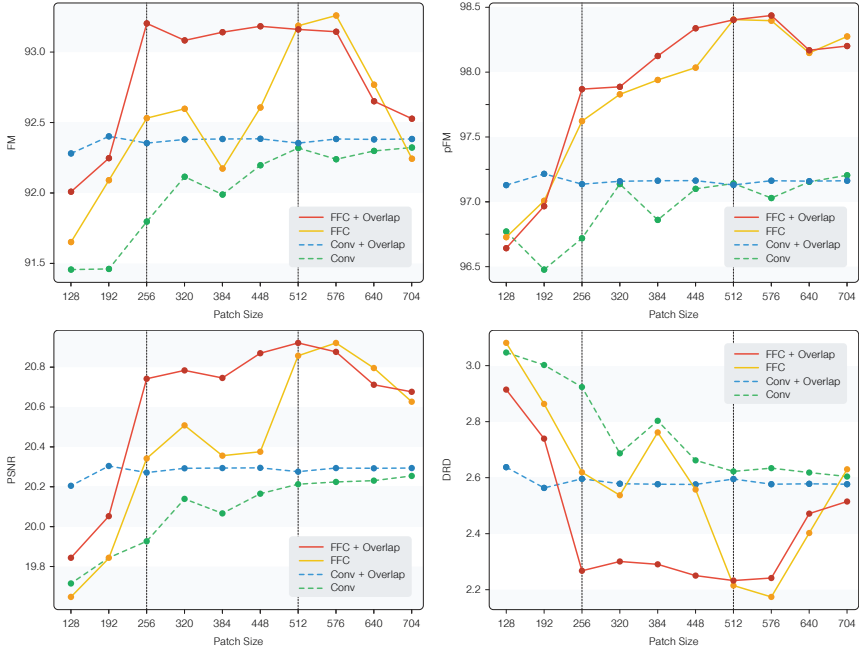


Figure 3.3: FM (top-left), pFM (top-right), PSNR (bottom-left), and DRD (bottom-right) values obtained by our FFC-based model and a baseline variant exploiting only standard convolutions when tested on patches of different sizes from H-DIBCO18, with or without overlap.

approach does not exploit the global information of the whole input patch and processes the images locally due to a limited receptive field. On the other hand, FourBi, thanks to the FFC and its infinite receptive field, increases the performance on increasing the input patch size. From Fig. 3.4 it can be also observed that the local patch contains blurred text, which both the FFC- and Convolution-based models incorrectly classify as foreground. In the lower part of the figure, we can see that with a larger patch, the FFC-based network leverages the surrounding context to identify that the innermost patch contains bleed-through text and correctly classifies it as background. Conversely, the Convolutional baseline does not benefit from the increased patch size and still misclassifies the blurred text as foreground.

Finally, in Tab. 3.1, we compare the capabilities of FourBi and the Transformer-based SotA DocEnTr [339] to model long-range dependencies. When the lat-



Figure 3.4: Results of the Conv-based baseline and our FFC-based model on 256×256 and 512×512 patches. Correct predictions are in green, errors in red.

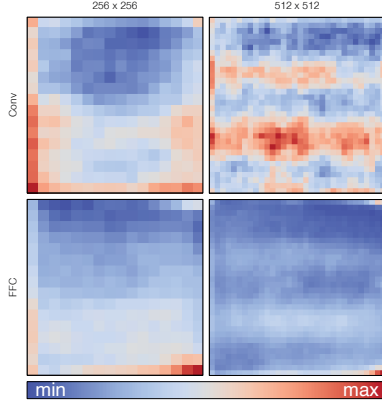


Figure 3.5: Average error distribution of our FFC-based model and the Conv-based baseline on multiple 256×256 and 512×512 patches from different datasets.

	Training patch	Inference patch	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
DocEnTr [339]	256×256	256×256	92.53	95.15	19.11	2.37
FourBi			92.94	97.29	19.26	2.32
DocEnTr [339]	512×512	512×512	92.20	94.93	18.91	2.45
FourBi		512×512	93.38	97.29	19.46	2.16
FourBi		512×512	<i>93.27</i>	<i>97.01</i>	<i>19.36</i>	<i>2.21</i>

Table 3.1: Performance comparison of a Transformer-based model against our FFC-based model on DIBCO17. Best results are in bold, second-best in italic.

ter is trained with larger patch sizes, the performance worsens, showing that the model has difficulties capturing fine-grained details of the documents. On the other hand, FourBi, already outperforming DocEnTr on the patch size 256×256 , benefits from working on larger patches, even at test time only.

BENEFITS OF THE POST-PROCESSING STRATEGY. We perform a study of the most common errors induced by working on single patches (see Fig. 3.5). In particular, for a given patch size, we evaluate the FFC-based model and the Convolutional baseline on several datasets and compute the average absolute error for each pixel location. Then, we discretize the resulting image in 16×16 bins and

compute the average error distribution. The values are independently normalized for visualization purposes. From the figure, we observe that the FFC-based model is more accurate on the borders with a larger patch size. The global receptive field allows it to consider the whole information of the patch. On the other hand, the Convolutional baseline is less precise in the predictions when fed with a larger patch, and the increased error compromises the potential benefit gained from a bigger context. Fig. 3.3 also highlights that the applied overlapping procedure leads to better performance in almost all patch sizes. In particular, we notice that on the Conv-based baseline, the post-processing procedure alone saturates the performance. This behavior suggests that the main weakness of the Conv-based baseline is the predictions close to the edges. Therefore, with the overlap procedure, the number of border pixels is minimized.

ALTERNATIVE LOSS FUNCTIONS. In Tab. 3.2, we report a comparison of the performance obtained by training our model with other loss functions commonly adopted for Document Binarization. In particular, we consider the Binary Cross-Entropy (BCE), the MAE, and the MSE as alternatives to the adopted Charbonnier loss. We recall that the Charbonnier loss behaves as the MAE for high errors and as the MSE for low errors. From the table, it emerges that the error-dependent behavior of the Charbonnier loss benefits training and leads to the best performance in terms of all metrics apart from the FM, where it leads to second-best results (comparable to those obtained when training with the BCE loss). Note that, for Document Binarization, the pFM is more relevant than the FM since it models the readability of the binarized document by penalizing more the errors made near the ground truth ink strokes. In light of this, we can conclude that the Charbonnier loss is a suitable objective function also for this task.

	FM↑	pFM↑	PSNR↑	DRD↓
BCE	93.38	97.18	<i>20.63</i>	<i>2.37</i>
MSE	92.68	97.40	20.56	2.56
MAE	91.83	<i>98.00</i>	20.27	2.74
Charbonnier	<i>93.19</i>	98.41	20.86	2.21

Table 3.2: Ablation analysis when training FourBi with different loss functions on the H-DIBCO18 dataset. Best results in bold, second-best in italic.

COMPUTATIONAL COMPLEXITY. Finally, we analyze the computational complexity of our approach. Note that, given a square image of size n and C_i channels, the Real FFT2d and the Inv Real FFT2d operations have time complexity $O(C_i \cdot n^2 \cdot \log n)$. The convolution has time complexity $O(C_i \cdot C_o \cdot n^2 \cdot k^2)$, where C_o is the output channels and k is the kernel size. The self-attention has time complexity $O(n^2 \cdot d)$, where d is the representation dimension. Nevertheless, the time complexity is not a complete indicator when comparing completely different architectures. In fact, the inductive biases of each operator, its characteristics, and how much its computation can be parallelized influence how deep and big the network needs to be in order to obtain a certain score. Therefore, to analyze the efficiency of our model, in Tab. 3.3, we report runtimes obtained by processing a single image on a single NVIDIA 2080Ti GPU and the number of parameters of the considered models fed with 256×256 patches. We report this information for our approach, the baseline exploiting standard convolution, and two SotA solutions, one based on a fully-convolutional GAN (DE-GAN [341]) and the other on a ViT (DocEnTr [339]).

	params	runtime
DE-GAN [342]	31.0M	11.2 ms
DocEnTr [339]	68.8M	3.8 ms
Conv. Baseline	15.8M	4.7 ms
FourBi	17.6M	9.1 ms

Table 3.3: Computational complexity and performance of SotA approaches and the Conv-based baseline on the H-DIBCO18 dataset, all fed with 256×256 patches.

COMPARISON WITH THE STATE OF THE ART

QUANTITATIVE COMPARISON ON LATIN-CHARACTERS DOCUMENTS. We evaluate our method on eight DIBCO datasets. The results are reported in Tabs. 3.4 to 3.11. For space reasons, we sorted the competitors by PSNR, and we kept for each table the first top ten methods. Moreover, to better highlight the consistency of our method, we include for each table the *rank* of FourBi compared to the competitors. It can be observed that our method achieves top performance on most

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [285]	91.50	93.58	19.78	-
Guo <i>et al.</i> [132]	86.75	89.74	17.93	3.62
Zhang <i>et al.</i> [426]	85.70	90.61	17.96	3.60
Feng [112]	88.16	90.96	18.25	3.41
Du & He [105]	88.47	90.87	18.41	3.22
Guo <i>et al.</i> [131]	88.56	93.50	18.56	3.06
Jia <i>et al.</i> [168]	91.75	95.36	19.84	2.17
Suh <i>et al.</i> [351]	93.92	96.53	21.18	1.50
cGANs [430]	93.92	96.53	21.18	1.50
GDB [414]	95.19	96.62	21.98	1.32
D ² BFormer [410]	96.23	97.49	23.10	0.95
FourBi	96.43	98.60	23.37	0.92
Rank	1 st	1 st	1 st	1 st

Table 3.4: Quantitative comparison on the H-DIBCO10 dataset.

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [286]	88.74	-	17.97	5.36
Jia <i>et al.</i> [168]	91.65	95.56	18.88	2.66
HDSN [381]	92.58	94.67	19.16	2.38
Suh <i>et al.</i> [351]	93.44	96.18	19.97	1.93
GDB [414]	93.44	95.82	20.01	2.25
cGANs [430]	93.81	95.70	20.26	1.81
Lin <i>et al.</i> [214]	94.08	97.08	20.51	1.75
Dang <i>et al.</i> [92]	95.61	97.34	22.09	1.48
DocEnTr [339]	94.37	96.15	20.81	1.63
D ² BFormer [410]	94.82	96.62	21.27	1.56
Text-DIAE [340]	95.01	96.86	21.29	1.48
FourBi	96.03	98.70	22.26	1.19
Rank	1 st	1 st	1 st	1 st

Table 3.5: Quantitative comparison on the H-DIBCO11 dataset.

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [287]	92.85	-	21.80	2.66
Du & He [105]	89.40	91.16	18.80	3.66
Jia <i>et al.</i> [168]	92.96	95.76	20.43	2.30
Suh <i>et al.</i> [351]	94.50	97.36	21.78	1.73
cGANs [430]	94.96	96.15	21.91	1.55
Jemni <i>et al.</i> [166]	95.18	94.63	22.00	1.62
GDB [414]	95.80	97.03	22.62	1.32
Rezaeezad <i>et al.</i> [306]	96.25	97.58	23.27	1.11
DocEnTr [339]	95.31	96.29	22.29	1.60
D ² BFormer [410]	96.28	96.90	23.35	1.26
Text-DIAE [340]	96.52	97.04	23.66	1.10
FourBi	97.07	98.78	24.29	0.94
Rank	1 st	1 st	1 st	1 st

Table 3.6: Quantitative comparison on the H-DIBCO12 dataset.

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [288]	92.12	94.19	20.68	3.10
Du & He [105]	89.94	93.97	19.26	3.24
Jia <i>et al.</i> [168]	93.28	96.58	20.76	2.01
HDSN [381]	93.43	95.34	20.82	2.26
cGANs [430]	93.86	96.47	21.53	2.32
Suh <i>et al.</i> [351]	94.75	97.20	21.79	1.66
Lin <i>et al.</i> [214]	95.24	97.51	22.27	1.59
GDB [414]	95.19	96.37	22.58	1.78
Dang <i>et al.</i> [92]	95.96	98.13	23.14	1.43
DE-GAN [341]	99.50	99.70	24.90	1.10
D ² BFormer [410]	95.96	96.96	22.80	1.30
FourBi	96.71	98.27	24.17	1.13
Rank	2 nd	2 nd	2 nd	2 nd

Table 3.7: Quantitative comparison on the H-DIBCO13 dataset.

metrics. In the tables, we isolate Transformer-based methods, which share with our approach the capability to model both long-range and short-range dependencies. Further, our model is adaptable to different image resolutions and has no strict constraints on the input size.

QUANTITATIVE COMPARISON ON NON-LATIN-CHARACTERS DOCUMENTS. We report a quantitative performance analysis on the non-Latin DIBCO19, PHIBD, and ISOS-BTD datasets, which were not included in any training set, in Tab. 3.12. We compare our approach, both alone and combined with the overlap-based post-processing strategy, with two SotA approaches: the fully-convolutional, generative DE-GAN [341] and the ViT-based DocEnTr [339]. It can be observed the same trend as for the considered Latin-characters training datasets. Specifi-

	FM↑	pFM↑	PSNR↑	DRD↓
Winning Entry [265]	96.88	97.65	22.66	0.90
Feng [112]	92.99	95.54	19.42	2.26
Du & He [105]	93.48	95.49	19.80	2.02
Jia <i>et al.</i> [168]	94.89	97.68	20.53	1.5
Suh <i>et al.</i> [351]	96.19	97.13	21.77	1.14
cGANs [430]	96.41	97.55	22.12	1.07
Lin <i>et al.</i> [214]	86.65	98.19	22.27	0.96
GDB [414]	97.66	97.26	23.13	1.21
HDSN [381]	96.66	97.59	23.23	0.79
DocEnTr [339]	97.16	98.28	22.99	0.16
D ² BFormer [410]	97.70	98.44	23.91	0.64
FourBi	98.19	99.16	25.18	0.55
Rank	1 st	1 st	1 st	2 nd

Table 3.8: Quantitative comparison on the H-DIBCO14 dataset.

	FM↑	pFM↑	PSNR↑	DRD↓
Winning Entry [290]	88.72	91.84	18.45	3.86
Feng [112]	88.88	90.30	18.45	4.99
Zhang <i>et al.</i> [426]	88.04	90.89	18.58	4.00
HDSN [381]	90.01	93.44	18.74	3.91
cGANs [430]	89.77	94.85	18.80	3.85
Jia <i>et al.</i> [168]	90.01	93.72	19.00	4.03
GDB [414]	90.41	94.70	19.00	3.34
Suh <i>et al.</i> [351]	91.11	95.22	19.34	3.25
Du & He [105]	91.51	91.81	19.62	3.16
Jemni <i>et al.</i> [166]	94.95	94.55	21.85	1.56
D ² BFormer [410]	90.96	93.30	19.24	3.22
FourBi	91.30	96.62	19.74	3.02
Rank	2 nd	1 st	2 nd	2 nd

Table 3.9: Quantitative comparison on the H-DIBCO16 dataset.

cally, FourBi achieves competitive or better performance compared to DE-GAN and DocEnTr. The difference in the charset prevents the networks from exploiting a priori knowledge possibly learned on the other datasets, which might allow them to recognize *e.g.* whether the text is flipped (bleed-through) from the context. Therefore, the networks can only exploit the image appearance to classify each pixel correctly. Nonetheless, thanks to the FFC, FourBi can exploit global information, which helps it better model the overall visual appearance.

QUALITATIVE COMPARISON WITH THE STATE OF THE ART. We qualitatively compare our approach, DE-GAN [341], and DocEnTr [339], on Latin and non-Latin documents. We show the results with one of the challenging images of H-DIBCO18 [289] in Fig. 3.6-Right. As we can see, both competitors produce images with speckles on the background, while our model, thanks to its ability to manage the global information, produces less noisy predictions and recognizes the page limit on top. Nonetheless, this sample is heavily stained on the top, and all three methods fail at accurately binarizing part of the head loop on the capital letter “G” and the page number. Finally, we qualitatively compare our approach on images from challenging ancient datasets. Different from those used in training, these documents contain non-Latin text. Thus, our model would not be able to exploit any potentially learned characters appearance information when binarizing those images. In Fig. 3.6-Left, it can be observed that FourBi accurately removes severe degradations such as faded ink and holes. Moreover, it shows su-

perior capabilities in handling bleed-through ink, exploiting the global context.

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [291]	91.04	92.86	18.28	3.40
cGANs [430]	90.73	92.58	17.83	3.58
Suh <i>et al.</i> [351]	90.95	94.65	18.40	2.93
Lin <i>et al.</i> [214]	90.95	93.79	18.57	2.94
Dang <i>et al.</i> [92]	92.08	94.99	18.72	2.84
DE-GAN [341]	97.71	98.23	18.74	3.01
Rezabehzad <i>et al.</i> [306]	93.01	95.42	19.24	2.29
GDB [414]	94.32	96.58	20.04	1.79
DocEnTr [339]	92.53	95.15	19.11	2.37
D ² BFormer [410]	93.52	95.09	19.35	2.12
Text-DIAE [340]	93.84	95.71	19.64	1.93
FourBi	93.81	96.57	19.66	2.03
Rank	4 th	3 rd	2 nd	3 rd

Table 3.10: Quantitative comparison on the DIBCO17 dataset.

	FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
Winning Entry [289]	88.34	90.24	19.11	4.92
DE-GAN [341]	77.59	85.74	16.16	7.93
Dang <i>et al.</i> [92]	91.26	93.97	19.81	3.42
GDB [414]	91.09	94.57	19.92	3.07
Lin <i>et al.</i> [214]	91.66	95.53	20.02	2.81
Suh <i>et al.</i> [351]	91.86	96.25	20.03	2.60
Jemni <i>et al.</i> [166]	92.41	94.35	20.18	2.60
Rezabehzad <i>et al.</i> [306]	92.47	95.99	20.29	2.50
D ² BFormer [410]	88.84	93.42	18.91	3.99
DocEnTr [339]	90.59	93.97	19.46	3.35
Text-DIAE [340]	91.32	94.44	19.95	3.21
FourBi	93.16	98.40	20.92	2.23
Rank	1 st	1 st	1 st	1 st

Table 3.11: Quantitative comparison on the H-DIBCO18 dataset.

		FM $\uparrow$	pFM $\uparrow$	PSNR $\uparrow$	DRD $\downarrow$
PHIBD	DE-GAN [341]	69.83	71.58	12.33	28.30
	DocEnTr [339]	59.28	83.81	13.91	15.89
	FourBi	<i>80.07</i>	<i>84.05</i>	<i>17.31</i>	<i>14.07</i>
	FourBi + Overlap	80.36	90.66	17.41	13.22
ISOS-BTD	DE-GAN [341]	85.19	91.55	<i>12.59</i>	13.25
	DocEnTr [339]	<i>79.34</i>	<i>97.48</i>	12.44	14.12
	FourBi	78.23	98.75	12.46	14.12
	FourBi + Overlap	78.88	98.75	12.62	<i>13.64</i>
DIBCO19	DE-GAN [341]	57.96	57.30	14.43	12.21
	DocEnTr [339]	67.20	77.13	14.12	10.20
	FourBi	63.15	<i>84.99</i>	<i>14.94</i>	<i>9.89</i>
	FourBi + Overlap	<i>63.56</i>	85.82	15.02	9.66

Table 3.12: Quantitative comparison on the considered non-Latin character datasets. Best results in bold, second-best in italic for each of them.



Figure 3.6: Qualitative comparison between FourBi and other SotA approaches on datasets containing ancient documents. Starting from the left column, we test FourBi on the non-Latin datasets PHIBD [256, 26], ISOS-BTD [314], and DIBCO19 [292]. In the rightmost column, we test FourBi on the Latin dataset DIBCO18 [289]. All models were trained on Latin-script datasets.

3.1.4 SUMMARY

In this chapter, we have proposed a fully-convolutional Document Binarization approach that can handle both global and local information to predict more precisely whether a pixel is background or ink. We achieve this capability by exploiting the FFC operator in a U-Net-like architecture. The FFC layers allow handling both local and global information, thus leading to results that are comparable or superior to those achieved with Transformer-based models. Moreover, we have observed that the common strategy to divide the document image into patches and treat them separately leads to artifacts at the borders of the patches. In light of this, we have applied a simple yet effective post-processing strategy consisting in combining the predictions of our model on overlapping patches. The results of deep experimental analysis on multiple benchmark datasets demonstrate the suitability of our approach, together with the possibility of applying the FFC operator to patches that are much wider than those used in training without performance drop, further improving the performance by reducing the artifacts at the borders.

3.2 EXTENDING FFCs TO VOLUMETRIC DATA: INK DETECTION ON THE HERCULANEUM PAPYRI

3

Recent advancements in Digital Document Restoration (DDR) have led to significant breakthroughs in analyzing highly damaged written artifacts. Among those, there has been an increasing interest in applying Artificial Intelligence techniques for virtually unwrapping and automatically detecting ink on the Herculaneum papyri collection. This collection consists of carbonized scrolls and fragments of documents, which have been digitized via X-ray tomography to allow the development of ad-hoc deep learning-based DDR solutions. In this work, we propose a modification of the Fast Fourier Convolution operator for volumetric data and apply it in a segmentation architecture for ink detection on the challenging Herculaneum papyri, demonstrating its suitability via deep experimental analysis. To encourage the research on this task and the application of the proposed operator to other tasks involving volumetric data, we release our weights and code[†].

Some of the most valuable sources of information we have about ancient cultures and populations are the manuscripts and, in general, the artifacts with writings and pictures that survived history [296, 59, 236, 279]. For this reason, even the smallest of such objects is precious for its potentially unique and impactful content. Due to the fragility of the medium and their long history, most of the ancient manuscripts found by archaeologists can be extremely degraded and irreversibly damaged, and thus, challenging to handle and analyze. These challenges are even more critical for those ancient documents that were written on scrolls, which also necessitate being unrolled for reading.

Digital Document Restoration (DDR) aims to provide access to these kinds of documents and has therefore gained great interest from researchers and practitioners [103, 54, 276, 56, 428, 123]. Among the DDR techniques, Virtual Unwrapping can be applied when the textual content is physically unreachable, as in the case of fragile scrolls. Starting from a digital 3D volumetric representa-

[†]<https://github.com/aimagelab/vffc>

tion obtained with X-ray micro-computed tomography [185], this method entails reconstructing the 2D text image, allowing researchers, scholars, and the general public to visually inspect and study these historical artifacts. Early works on the application of Virtual Unwrapping were limited to specific use cases and entailed semi-manual pipelines [324, 213]. Later, fully automated digital unwrapping was applied to undamaged or partially damaged scrolls made of parchments [316, 348, 347, 312], bamboo [346, 345], papyri [18, 17], silver [153, 36], and the famous En-Gedi scroll [325], charred and charcoaled by a fire in early middle ages.

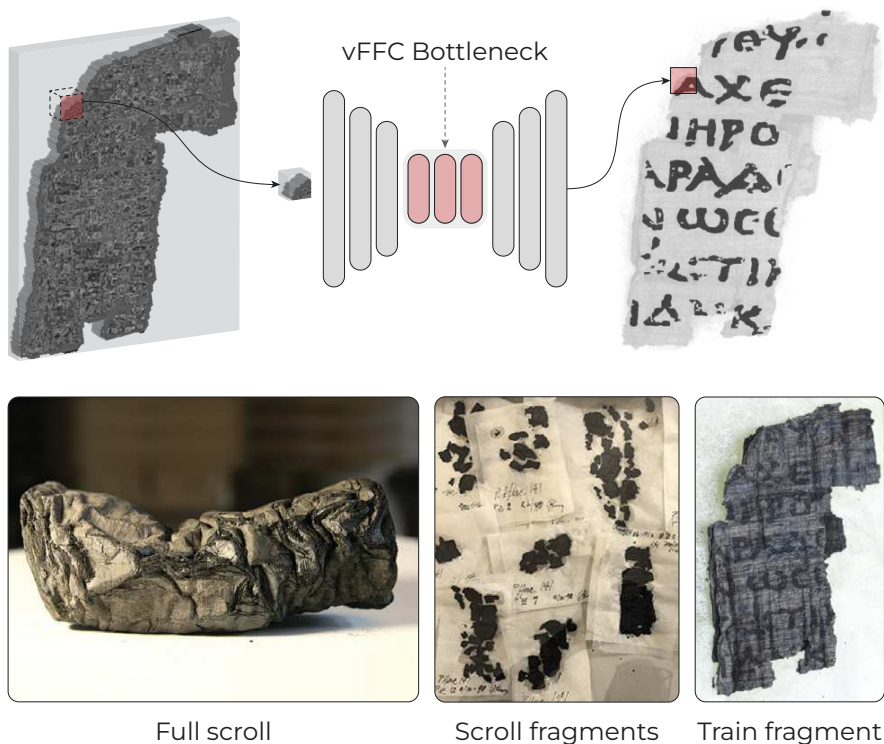


Figure 3.7: We propose an architecture featuring our devised volumetric Fast Fourier Convolution (vFFC) operator for the ink detection task on the fragments of carbonized Herculanum papyri (parts of this image are from the official project website <https://scrollprize.org/>).

When discussing ancient irreparably damaged scrolls, few examples carry as much significance and present as many challenges as the Herculanum papyri.

This collection of more than 1800 manuscripts, recovered from the Herculaneum Villa of the Papyri, presents a particularly important case due to its unique preservation state and the special interest of scholars [172]. Carbonized and buried in ashes during the Vesuvius eruption of 79 C.E., they represent the only intact surviving library from antiquity discovered in its original location [172]. Among the posed challenges, the ink is carbon-based, different from the metal-based ink in the En-Gedi scroll. Thus, the response of the ink to X-rays is not seemingly distinguishable from that of the carbonized papyrus support, making traditional Virtual Unwrapping ineffective.

In the sight of this, in 2023, Parsons *et al.* [270] proposed to push forward DDR on this challenging collection by resorting to modern Computer Vision, Document Analysis, and Artificial Intelligence techniques. To this end, they developed and released EduceLab-Scrolls, an open dataset containing volumetric scans of rolled scrolls and detached fragments from the Herculaneum papyri, which is the object of a dedicated ongoing competition[‡]. The goal is to develop advanced Virtual Unwrapping strategies to unroll these challenging scrolls and then perform ink detection on the unrolled sheets. Meanwhile, the fragments can be used to develop ink detection algorithms. In fact, for the fragments, it has been possible to obtain infrared images and, thus, ink maps that can be used for training supervised deep learning models. Thanks to their effort, this novel task on such challenging data is receiving increasing interest from the Artificial Intelligence community.

In this respect, we tackle the task via an efficient, fully-Convolutional model featuring blocks inspired by FFCs [75] [4], which we modify to handle volumetric data (see Fig. 3.7). Note that the FFC operator exploits the spectral information of the input to expand its receptive field in the frequency domain and to handle pseudo-periodic patterns. Spatial FFC has been employed for tasks on 2D images but, to the best of our knowledge, it has never been applied to volumetric data. For this reason, we propose a modification to the original FFC operation that makes it able to handle volumetric data and thus be applied to the Herculaneum papyrus fragments scans for ink detection. Through a deep evaluation

[‡]<https://scrollprize.org/>

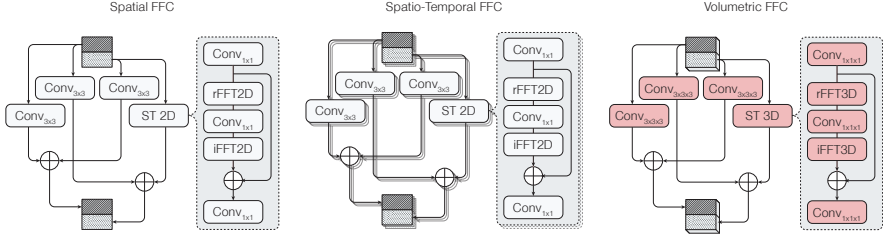


Figure 3.8: The standard spatial FFC (left) can be adapted to work on spatio-temporal data by replicating it along the channel dimension (center). Our proposed vFFC (right) is designed to handle 3D volumetric data directly by combining 3D convolutions and 3D FFTs. For simplicity, we omit the batch normalization and ReLU operations in the schemes.

analysis, we provide useful insights on the challenging task of ink detection on Herculaneum papyri and on the proposed approach, which we demonstrate to be suitable for the task.

3.2.1 PROPOSED APPROACH

Our goal is to detect the presence of ink in the volumetric representation of a carbonized papyrus sheet. Our model takes as input such volumes and is expected to output an ink map with the same width and height of the volume, whose elements contain the probability of ink being in the corresponding papyrus surface. To tackle this task, we devise a U-net-like architecture, whose details are given in Sec. 3.2.1, featuring a variant of the FFC operator that we devise to handle volumetric data as described below.

VOLUMETRIC FAST FOURIER CONVOLUTION

The FFC proposed by Chi *et al.* [75], which here we refer to as spatial FFC, is a neural operator that combines convolution and Fourier Transform to perform local reasoning in the space domain and non-local reasoning in the frequency domain. This information processing expands its receptive field and makes it suitable for handling pseudo-periodic patterns in the data. The operator has also been applied to spatio-temporal data [75]. Here, we refer to this variant as Spatio-Temporal FFC (stFFC). The vFFC variant devised in this work presents the same

properties of the spatial FFC operator but can directly handle volumetric data. A visual comparison between these mentioned operators is reported in Fig. 3.8.

The vFFC operator takes as input a tensor $\mathbf{X} \in \mathbb{R}^{D \times H \times W \times 2C}$, where $2C$ is the number of channels, D is the depth dimension, and H and W are the spatial dimensions. This tensor is split into two parts along the channel dimension, which are fed to two interconnected branches: a local branch and a global branch. Splitting the input tensor into two chunks allows the encoding of different information in separate regions of the tensor. In addition, this approach permits the global and local branches to specialize in different aspects because they do not share the same inputs. The *local branch* contains two 3D convolutional layers with kernel size 3 and is in charge of modeling local volumetric information. In the *global branch*, the input is mapped into the spectral domain to model global information. This branch consists of a 3D convolution with kernel size 3 and a 3D Spectral Transform (ST 3D) block.

The ST 3D block exploits a three-dimensional real Fast Fourier Transform (FFT3D) [83]. Specifically, the FFT3D is performed across the depth, height, and width dimensions of the input tensor:

$$\text{FFT3D}(\mathbf{X}) = \mathbf{Z} = \mathbf{R} + i\mathbf{I},$$

where $\mathbf{Z}$ is a complex tensor, with real and imaginary parts $\mathbf{R}, \mathbf{I} \in \mathbb{R}^{D \times H \times \frac{W}{2} \times C}$. Then, $\mathbf{R}$ and $\mathbf{I}$ are stacked together along the channel axis, thus obtaining a tensor

$$\mathbf{R} || \mathbf{I} \in \mathbb{R}^{D \times H \times \frac{W}{2} \times 2C}$$

that is then fed to a $1 \times 1 \times 1$ convolutional layer. Batch normalization and ReLU activation functions are applied to the output of the convolution. The resulting tensor is reshaped into a complex-valued tensor $\mathbf{Z}' \in \mathbb{C}^{D \times H \times \frac{W}{2} \times C}$, and the inverse real Fast Fourier Transform (iFFT3D) is computed to obtain the global branch output:

$$\text{iFFT3D}(\mathbf{Z}') = \mathbf{X}' \in \mathbb{R}^{D \times H \times W \times C}.$$

The output of each branch is summed to the output of the other, and the

resulting tensors are fed to separate batch normalization and ReLU operation before being stacked together to obtain the final vFFC output tensor.

INK DETECTION ARCHITECTURE

3

Rather than having a different response to X-rays than the papyrus support, the carbon-based ink in the Herculaneum scrolls ink modifies the substrate structural patterns in subtle ways, as shown in [269]. In particular, parts containing ink have different types of cracks and densities and are thicker and smoother than empty ones. In this challenging case, it is important to consider both local and global contexts. The subtle patterns are, in fact, very localized, but with a global context, the model is able to consider holistically information from the whole subvolume. Moreover, we argue that the vFFC operator, which is able to handle pseudo-periodic patterns, is suitable for modeling this kind of data. We treat ink detection as a pixel-wise classification task, and we use an encoder-bottleneck-decoder architecture composed of a 3D convolutional encoder, a vFFC-based bottleneck, and a 2D convolutional decoder. To accurately localize the signal and fuse high-level and low-level features, we employ skip connections between the encoder and the decoder, similar to U-Net [310], and vFFC layers in the latent space of the network.

Specifically, we employ a 3D ResNet-34 [138] to encode the subvolumes. The output from the last block is fed to 3 vFFC Residual Blocks, each combining two vFFC layers with a residual connection. The $\mathbb{R}^{D \times H \times W \times C}$ feature maps from the bottleneck are collapsed to $\mathbb{R}^{H \times W \times C}$ by averaging the depth dimension. The same applies to the encoder activations in the skip connections. Then, a 2D decoder reconstructs the ink map. This decoder is obtained by stacking four blocks made of 2-fold bilinear interpolations and convolutions, operating on the concatenation of the output of the previous layer and the features from the corresponding encoder layer. Since the fragments have very high resolution but represent small objects, the decoder does not reconstruct the full-scale ink map but rather a 4-fold downsampled one. This approach also limits the computational weight of the model.

TRAINING AND INFERENCE

We train our model on subvolumes (3D patches) of the fragments scans both to limit the computational complexity of our model and to obtain more samples from the same fragment. Note that this strategy is customary for other data-constrained tasks such as document binarization [362] or medical images segmentation [173, 97]. Moreover, we apply a number of data augmentation operations (described in Sec. 3.2.2) in order to reduce the risk of overfitting and help the model generalize to unseen data. Our adopted loss function is a combination of the Dice loss and weighted binary cross entropy (WBCE). The Dice score coefficient has been proposed for unbalanced segmentation [248], while the weighted binary cross-entropy is customary for classification. Let N be the number of voxels, $p_i \in P$ the predicted binary segmentation map, and $g_i \in G$ the ground-truth ink image. The loss can be written as:

$$\mathcal{L} = \text{Dice} + \text{WBCE},$$

with

$$\text{Dice} = \frac{2 \sum_n^N p_n g_n + \varepsilon}{\sum_n^N p_n^2 + \sum_n^N g_n^2 + \varepsilon},$$

$$\text{WBCE} = -\frac{1}{N} \sum_{n=1}^N w [g_n \log p_n + (1 - g_n) \log(1 - p_n)]$$

where w is the weight attributed to the ink class. The predicted ink map contains values ranging from 0 to 1, representing the ink presence probability. From this, we obtained a binarized image by applying a threshold of 0.5.

3.2.2 EXPERIMENTAL ANALYSIS

In this section, we present the experiment setup concerning the ink detection task on the fragments of the recently proposed EduceLab-Scrolls dataset and detail our training strategy. Moreover, we present an extensive analysis of the proposed components, both to explore their contribution and give some intuitions on what is more suitable for this task.



Figure 3.9: Fragments contained in the considered dataset, in different modalities.

EXPERIMENT SETUP

DATASET. As mentioned above, in this work, we focus on the surface volumes of the fragments in the EduceLab-Scrolls [270] (see Fig. 3.9). These scans have been performed with X-ray micro-CT and thus are very detailed. With a $3.2\mu\text{m}$ voxel size, they have resolutions in the order of thousands of pixels and weigh several Gigabytes. In particular, we use the three publicly available fragments released for the Vesuvius Challenge on Ink Detection related to the EduceLab-Scrolls dataset project. These documents, broken and detached from the scrolls during destructive physical unrolling attempts, have visible ink and, thus, it has been possible to obtain a ground truth ink map, also thanks to infrared analysis. The scans are all composed of 65 slices but have different spatial sizes: 6330×8181 for fragment 1, 9506×14830 for fragment 2, and 5249×7606 pixels for fragment 3. In our experiments, we use fragment 1 for test and the other two for training.

EVALUATION METRICS. Considering the novelty of the task and its similarities to other tasks, we use a combination of metrics to evaluate our models. First of all, we adopt the score proposed in [270] to evaluate the ink detection performance, *i.e.* the F_β score, which weighs more the precision than the recall and is defined as follows:

$$F_\beta = \frac{(1 + \beta^2)pr}{\beta^2 p + r},$$

where p and r are precision and recall, respectively, and the parameter $\beta=0.5$. Moreover, we employ some scores commonly used in Document Binarization, namely the pseudo-F-Measure (pFM) [264] and the Peak Signal-to-Noise Ratio (PSNR), which measures the similarity of two images.

IMPLEMENTATION DETAILS. Considering that most of the information is in the central slices and to reduce computational impact, we use 16 slices for training. Note that using more slices would not straightforwardly improve performance: in fact, fragments have hidden text, in papyrus layers fused in the back, that could be present in the surface slice if we go too deep. Taking also into account that the ink is a localized signal, we train on $16 \times 256 \times 256$ subvolumes.

For optimization, we use the AdamW [225] optimizer with $\beta_1 = 0.9, \beta_2 =$

0.95, and the OneCycle scheduler [332] with learning rate 0.003. We set the batch size to 4 and apply floating point 16 mixed precision to speed up training. For regularization, we use Drop Path [196] with rate 0.1, Channel Dropout [367], with rate and maximum dropped channels both set to 0.5, and additional data augmentation strategies as described in Sec. 3.2.2. We set the ink weight in the WBCE to 1 and train the models for 20 epochs, which takes 24 hours with an NVIDIA RTX A5000. At test time, for the entire test fragment, we perform the prediction on $16 \times 256 \times 256$ subvolumes and keep only the prediction on the inner $16 \times 128 \times 128$ part to reduce border artifacts.

DATA AUGMENTATION. We perform data augmentation in order to enhance model generalization and robustness. In particular, we obtain the training subvolumes starting from a 3D lattice of the whole fragment scans, with cells of size $32 \times 512 \times 512$ and stride 64. During each training iteration, we extract subvolumes of size $16 \times 256 \times 256$ from a random 3D position of the lattice cells. The rationale behind this strategy is twofold. In the spatial dimension, the random crop increases the variability in the training data, preventing the network from being exposed to the same patches repetitively. In the depth dimension, exposing the network to different slices ensures that the model learns to discern patterns across all depth coordinates. Indeed in a real-world application, the surface volumes would be extracted from segmented sheets in a scroll. This process is prone to errors and misalignment. Thus, there is rarely correspondence between depth coordinates in different samples. By enforcing depth invariance, the network is able to model patterns independently of specific depth locations. Moreover, we randomly apply horizontal and vertical flip, 90° random rotation, and transposition to obtain the corresponding Dihedral group D_4 transformations, further increasing diversity in spatial patterns.

RESULTS

In this section, we report an analysis aimed at identifying the most relevant elements of the training procedure and assessing the effectiveness of our proposed vFFC-based model for the ink detection task.

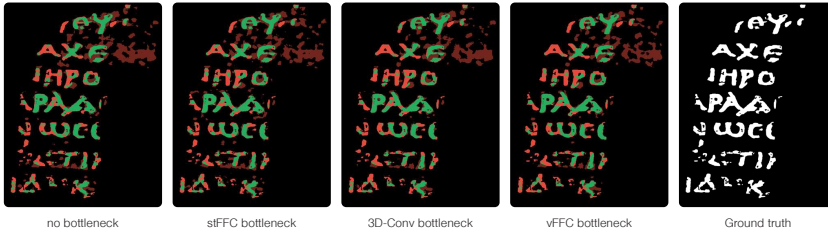


Figure 3.10: Qualitative results of the considered variants on the test fragment. Red indicates the missed ink predictions, green indicates the pixels correctly identified as ink, and maroon indicates the background pixels incorrectly classified as ink.

TRAINING STRATEGY. First, we analyze the contribution of the applied data augmentation strategies. The results of this analysis are reported in Tab. 3.13. It can be observed that the Dihedral transformations bring the most benefits to the training by inducing the larger variability. From these experiments on fragment 1, the benefit of the random crop is not evident. Nonetheless, when testing on the public and the private test set of the Kaggle Ink Detection Challenge associated with the EduceLab-Scrolls dataset, this regularization strategy has been proven beneficial for increasing the generalization capability of our model in handling fragments in which the relevant information is localized in different slices. The effect of the random crop, especially on the depth axis, is visible from the activation maps discussed in Sec. 3.2.2. Then, we perform an ablation on the training loss terms, whose results are reported in Tab. 3.14. It emerges that the WBCE leads to good performance, and the combination with the Dice helps refine the results. Moreover, giving more weight to the ink class leads to worse performance, despite the ink class being much less represented than the background.

ARCHITECTURE. We compare the proposed architecture with a number of baselines in Tab. 3.15. To assess the effectiveness of vFFCs, we compare our proposed model against variants featuring different kinds of bottlenecks. In particular, we consider a variant without bottleneck layers, one with stFFCs, and one with 3D-Convs. Overall, using the proposed vFFC in the bottleneck leads to the best performance. Arguably, this is due to the more precise prediction of the background pixels, as can also be observed from the qualitative comparison in Fig. 3.10 (see,

Dihedral Transform	Random Crop	Channel Dropout	F_β	pFM	PSNR
-	-	-	0.37	0.53	9.37
✓	-	-	0.46	0.57	10.38
✓	✓	-	0.46	0.57	9.62
✓	✓	✓	0.47	0.58	10.09

Table 3.13: Ablation analysis on the augmentation strategies adopted.

Loss	w	F_β	pFM	PSNR
Dice	-	0.40	0.56	9.47
WBCE	1	0.44	0.56	10.34
WBCE + Dice	5	0.41	0.53	8.20
WBCE + Dice	2	0.45	0.56	9.86
WBCE + Dice	1	0.47	0.58	10.09

Table 3.14: Ablation analysis on the training loss function.

e.g. the top-right part of the prediction map).

KAGGLE INK DETECTION CHALLENGE RESULTS. We participate in the Kaggle Vesuvius Challenge - Ink Detection[§] with a boosted version of our model. In particular, we maintain the architecture but employ a different training and inference strategy to fully exploit the training fragments available. In particular, we split each available fragment into two parts in order to obtain a total of six subsets. Then, we train six versions of our model in a k-fold strategy and set the prediction threshold to 0.8. We then submit an ensemble of these models, combined via majority voting on each pixel with four votes out of six. The submission scored silver medal (*i.e.* within the top-5% submissions) in the competition, scoring around $F_\beta=0.70$ on the public test set and $F_\beta=0.60$ on the private one.

[§]www.kaggle.com/competitions/vesuvius-challenge-ink-detection

Bottleneck	F_β	pFM	PSNR
-	0.46	0.58	9.87
stFFC	0.45	0.58	9.76
3D-Conv	0.46	0.58	10.04
vFFC	0.47	0.58	10.09

Table 3.15: Quantitative comparison between our model featuring vFFCs in the bottleneck and baselines with different bottlenecks.

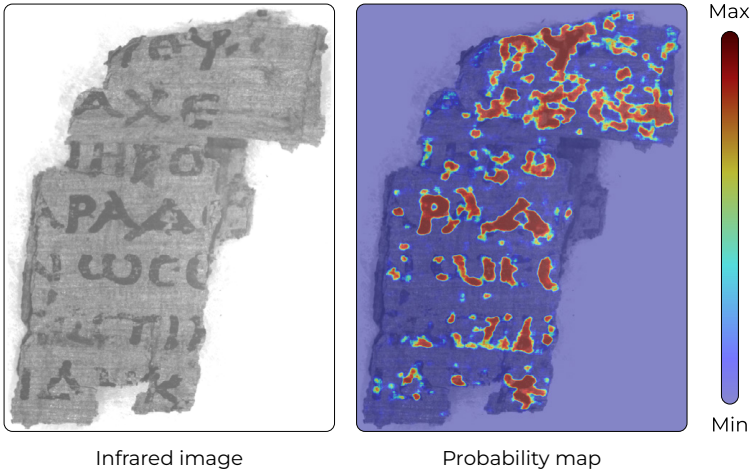


Figure 3.11: Prediction probability map of our proposed model on the test fragment.

VISUALIZATIONS

For reference, Fig. 3.11 shows the ink prediction probability map. Even though the threshold is set to 0.5, we argue that this non-thresholded visualization can be helpful for manual visual inspection of the fragments.

Finally, we study the model class activation map by using LayerCam [169]. In particular, we consider the whole depth size for a given 2D patch on the papyrus surface, obtaining a volume $\mathbf{V}$ with shape $65 \times 256 \times 256$. Then, we extract subvolumes of size $\mathbf{v} \subset \mathbf{V}$ with shape $16 \times 256 \times 256$ and stride set to 1, covering the whole Z axis, and feed them to the network. We compute the mean activation value on the spatial dimensions for each depth slice and show the re-

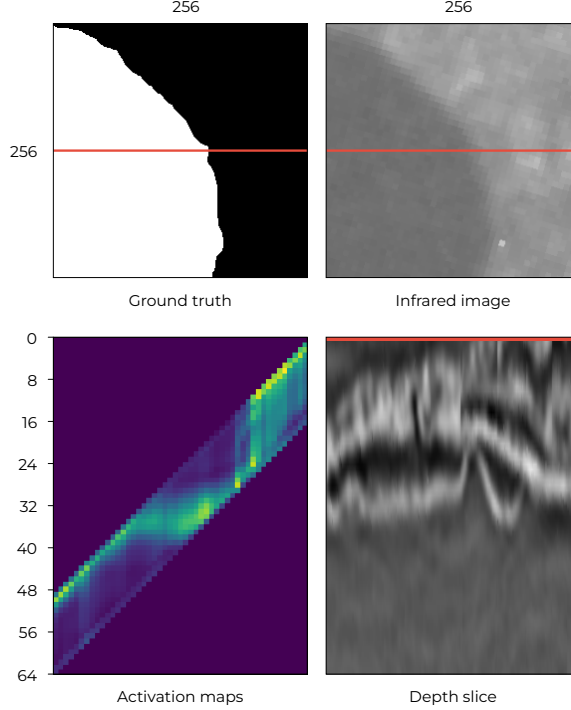


Figure 3.12: Activation maps by varying the depth and keeping the spatial coordinates of the subvolume fed to our model (blue represents the minimum value, yellow the maximum). For reference, we report the corresponding ground-truth ink map, the infrared image of the surface, and the coronal slice of the subvolume.

sult in Fig. 3.12. On the x-axis of the activation maps plot, we report the value of the starting $z \in Z$ coordinate of the subvolume. On the y-axis, we plot the absolute $z \in Z$ depth coordinate. As we can see, the plot is divided into three regions, corresponding to the upper ($0 \leq z < 24$), middle ($24 \leq z < 40$), and lower ($40 \leq z \leq 65$) parts of the papyrus volume $\mathbf{V}$. When we feed to the model slices corresponding to small values of z , they are all informative; thus, the network does not look specifically at any of them but rather combines the patterns. In the center-most region of the plot, the network consistently focuses on the slice with $z \approx 30$ for several passes of the subvolume. We argue that this is the main manifestation of the depth invariance that we instill with the subvolume random crop. This way, the network can account for segmentation misalignment in real-world

scenarios, learning to recognize important slices independently from their relative position in the subvolume. Then, with growing values of z , the depth slices contain less and less information, so the network focuses on the slices with relatively lower values of z . We provide visualizations of the volume $\mathbf{V}$ in Fig. 3.12: from the depth slice, we can see that the information is for the most part in the low z coordinates, while the infrared image and the ground truth show where ink is present in the surface and changes the inner texture of the papyrus support.

3.3 CONCLUSION

In this chapter, we have explored the use of Fast Fourier Convolution (FFC) for two document analysis tasks. First, with FourBi, we have proposed a fully-convolutional Document Binarization approach that exploits FFCs in a U-Net-like architecture to handle both global and local information, achieving results comparable or superior to those of Transformer-based models while being more efficient. We have also demonstrated that a simple overlapping post-processing strategy effectively reduces artifacts at patch borders, and that FFCs can seamlessly operate on patches larger than those used during training. Then, building upon the FourBi experience, we have devised the vFFC operator, a modification of the standard FFC designed to handle volumetric data directly. We have incorporated the vFFC in an architecture for the DDR sub-task of ink detection on volumetric scans of carbonized papyri fragments from the EduceLab-Scrolls dataset. Through experimental analysis, we have assessed the effectiveness of this approach and contributed to the emerging research interest in DDR on the challenging Herculaneum papyri data. Overall, these works demonstrate the versatility and effectiveness of frequency-domain reasoning for document analysis, from 2D binarization to 3D ink detection. The vFFC operator could also be applied to other tasks and scenarios involving volumetric data (*e.g.* medical imaging), which we leave for future work.

4

Multi-Page Document Parsing with μ gat

REGESTA are catalogs of summaries of other documents and, in some cases, are the only source of information about the content of such full-length documents. For this reason, they are of great interest to scholars in many social and humanities fields. In this work, we focus on Regesta Pontificum Romanorum, a large collection of papal registers. Regesta are Visually-Rich Documents (VRDs), where the layout is as important as the text content to convey the contained information through the structure, and are inherently multi-page documents. Among Digital Humanities techniques that can help scholars efficiently exploit regesta and other documental sources in the form of scanned documents, Document Parsing has emerged as a task to process document images and convert them into machine-readable structured representations, usually markup language. However, current models focus on scientific and business

This Chapter is related to the publication “F. Quattrini*, C. Zaccagnino* *et al.*, μ gat: Improving Single-Page Document Parsing by Providing Multi-Page Context, ECCVW 2024” [8]. *Equal contribution. See the list of Publications on page 315 for more details.

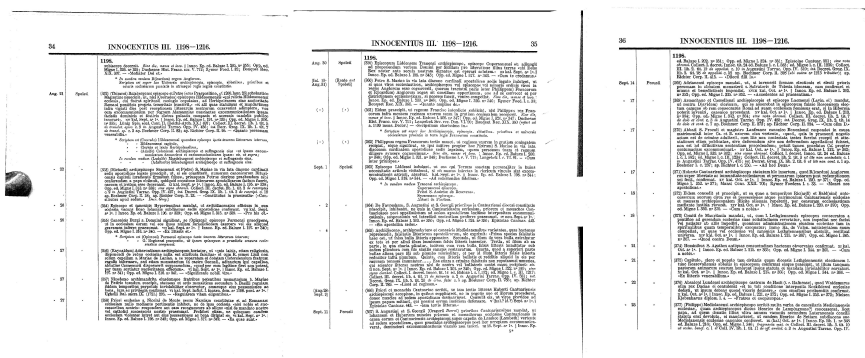


Figure 4.1: Three consecutive pages from the Regesta Pontificum Romanorum: summaries are both simple-layout paragraphs and text in a tabular layout, and the corresponding information about date and place is given via layout. Such entries (summaries and metadata) span multiple pages, with interleaved information and layout elements.

documents, and most of them consider only single-paged documents. To overcome this limitation, in this work, we propose μ gat, an extension of the recently proposed Document Parsing Nougat architecture, which can handle elements spanning over the single page limits. Specifically, we adapt Nougat to process a larger, multi-page context, consisting of the previous and the following page, while parsing the current page. Experimental results, both qualitative and quantitative, demonstrate the effectiveness of our proposed approach also in the case of the challenging Regesta Pontificum Romanorum. To foster research in multi-page document parsing, we release our code for both the dataset preparation and the proposed approach at ^{*}.

4.1 INTRODUCTION

A vast majority of Cultural Heritage and historical data comes in the form of digitized document images. Among those, *regesta* collections are a valuable source of information for scholars of many different fields of humanities and social sciences (e.g. history, religious studies, linguistic and literary studies, etc.). In fact, regesta serve both as organized collections of sources and data related to a vast range of

^{*}<https://github.com/aimagelab/mugat>

topics, places, and individuals and as indexing tools for scholars to find the historical documents preserved by archives all around the world. Specifically, regesta are chronologically ordered catalogs of summaries of historical documents, which, in some cases, are the only existing source of information about the full-length documents they summarize. A regestum is a highly structured summary of a document, both visually and textually. Such documents have been created by following a precise and consolidated methodology, which involves the extraction of information from specific sections of the full-length document and its translation and codification into a language shared by the community of scholars at the time the regestum has been edited. This methodology to summarize historical documents has been specifically formalized and used to collect, index, and summarize medieval documents. Regesta collections have been compiled by intellectuals and scholars starting from the 19th century until the present, creating a rich corpus of historical data. However, only scanned versions of the regesta collections are available, not in machine-readable format. This is because this wealth of historical data is preserved in printed texts from the 19th and 20th centuries, with fonts and layouts that are difficult to process with traditional OCR systems. Specifically, regesta are characterized by difficult-to-read ancient fonts, extensive use of abbreviations and acronyms, and, peculiarly, a multi-column layout (different from edition to edition). In fact, regesta are documents whose layout is as important as the content since it helps to correctly correlate the contained information through the structure. In this sense, *regesta are visually rich documents*.

Visually-rich Document Understanding (VrDU) is an area of research that aims to analyze and extract information from document images. Unlike other vision-based data, the text is an integral part of the image. Therefore, to effectively perform this task, it is necessary to consider and combine textual, visual, and spatial information. Given the huge range of different layouts, conservation status, scan quality, and content of documents, existing approaches usually entail pertaining models to perform tasks like parsing and masked language/visual modeling [406, 407, 158, 188, 94, 204, 48] on large-scale general datasets like IIT-CDIP [200] or CC-PDF [374]. Then, these models are fine-tuned on specific datasets for different subtasks, such as Document Classification [200, 137],

Document Layout Analysis [433, 206], Key Information Extraction [165], and Document Question Answering (DQA) [240]. Despite being an active research area, there are still many challenges and open questions concerning scalability, generalization, and sub-task definitions [378, 377]. One aspect that is receiving increasing research interest is that in most settings (and relative datasets), the considered documents are single-paged or seen as an independent set of single pages. This limitation severely hinders the capabilities of the resulting developed models in real-world scenarios, both in terms of performance (as they cannot leverage multi-page information and wrongly consider each page as independent) and in terms of possible applications [378, 377]. In this respect, some recent works have tackled this problem by also proposing new multi-page document datasets for VrDU [378, 219], Classification [377], and DQA [378]. In particular, multi-page DQA is attracting interest, and some approaches have been proposed that extend single-page pretrained architectures to handle multi-paged inputs [366, 47].

Particularly relevant in the context of VrDU is the Document Parsing task, which consists in extracting a structured representation of a document in markup language. This task is considered both for pre-training [94, 188] and as the final goal [48, 354, 197, 234] in VrDU. Academic documents are particularly interesting for the VrDU community [48, 354] because of the challenges they pose, which make them a valuable tool for developing general and effective VrDU models. Most of these documents are images in PDF format, obtained from scanned documents produced before the age of the Internet, which usually present complicated and highly variable layouts with tables, formulas, and graphs intertwined with text. Moreover, these documents are almost always multi-page, with text split between pages and interleaved with images or tables and cross-page references. However, state-of-the-art works are trained to parse single pages. We argue that, due to the multi-page nature of most visually-rich documents, when parsing a page, the models can benefit from taking into account the *context* around the page, which we define as the previous and next page.

In this work, we treat academic document parsing as a proxy task for obtaining a model able to parse regesta collections without fine-tuning on such documents. In fact, due to their information-carrying layout and the underrepresented lan-

guage in which they are written (medieval Latin), obtaining a large amount of high-quality annotated regesta pages requires multidisciplinary skills and, therefore, is costly and time-consuming. Thus, a language-agnostic, multi-page document parsing model that can be used off-the-shelf on regesta would be highly valuable for the Digital Humanities community. For this reason, in this work, we tackle multi-page document parsing by proposing an adaptation strategy to make pretrained OCR-free single-page document parsing models able to handle multiple pages as context, increasing the consistency of the parsing output. We consider the recently proposed Nougat [48] architecture as the base single-page document parsing model and call our approach *μgat* (*multi-page Nougat*).

As a use-case, we consider the *Regesta Pontificum Romanorum* [284], edited by August Potthast between 1874 and 1895 (see some pages in Fig. 4.1). This is a well-known collection of summaries of documents produced and issued by popes and pontifical chanceries (papal letters, bulls, decretals, *etc.*) from the origins of the Latin Church to the beginning of the 14th Century. By providing detailed records of papal decisions, letters, and administrative actions, they serve as a scholarly resource for the study of the Latin Church and, more broadly, for the history of ancient and medieval Europe. The *Regesta Pontificum Romanorum* can be seen as a multi-page, three-column table where each page has a header bearing the pontiff who authored the summarized documents and the year. The table entries are relative to different full-length documents and report the following information: 1) the edit date (month and day) of the document ; 2) the place where it was edited; and 3) the text of the regestum, *i.e.* the summary. The regestum itself has a specific structure and contains the following information in a fixed order: addressee; summary of the contents; date according to the ancient calendar; references to find the document in the original version (on parchment) or in a printed edition, and the incipit of the text. The loss of any of these elements or a reversal of their order would make the summary incomprehensible and the original medieval document impossible for scholars to find. Therefore, for its relevance for humanities studies and for the challenges posed for VrDU, we choose this collection as our use-case to validate the proposed approach for multi-page context-enhanced document parsing.

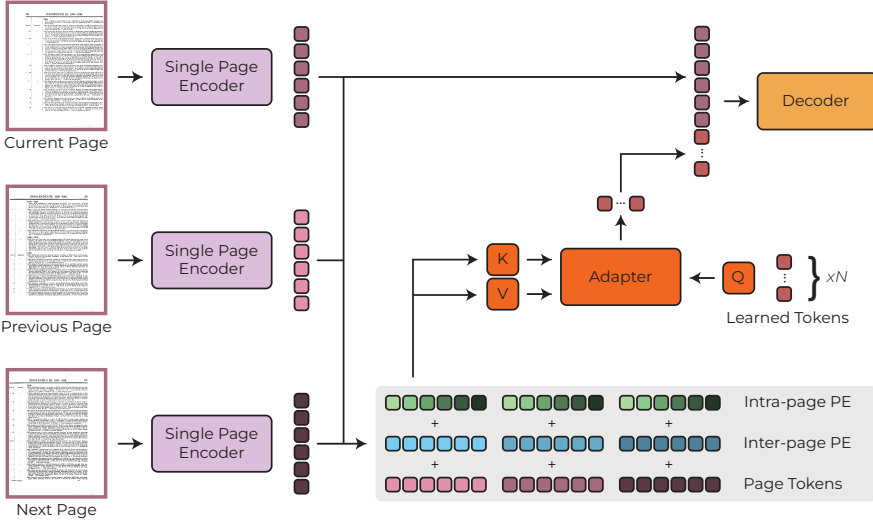


Figure 4.2: Our proposed μ gat architecture. The Visual Encoder embeds the previous, current, and next page. These embeddings are then concatenated and, after the addition of inter-page and intra-page positional embeddings, fed into the key and value of the adapter’s cross-attention layers, which extract information into the learned N tokens representing compressed information about the pages. These tokens are then appended to the current page tokens and fed to the auto-regressive Decoder.

4.2 PROPOSED APPROACH

In this section, we first provide the details about the Document parsing architecture that we use as the base model and the Adapter model we devise to enhance its performance by exploiting multiple pages as context when parsing the page of interest. Then, we describe the semi-automatic procedure we adopt to collect increasingly challenging data used for training and validating our model. These include complex-layout academic papers, long tables, and pages from the Regesta Pontificum Romanorum.

4.2.1 μ GAT MODEL

BASE ARCHITECTURE. Our approach is based on the idea of introducing multi-page context as input to encoder-decoder document parsing models to increase the consistency of their output by exploiting information from multiple pages.

In this work, we build upon the recently proposed single-page parsing Nougat [48] architecture. This is an Encoder-Decoder Transformer [380] specialized in document parsing, inspired by Donut [188]. In particular, the encoder is a Visual Encoder implemented as a Swin Transformer [224], which takes as input a document image $\mathbf{x} \in \mathbb{R}^{H_0 \times W_0}$ and embeds it into a sequence of P vectors $\mathbf{z} \in \mathbb{R}^d$. The decoder is implemented as a mBART[201], which auto-regressively generates a sequence of tokens by attending the embedded image via cross-attention and its generated tokens via self-attention. The output tokens are then projected to the same vocabulary as the one used in [360]. The model is trained to output parsed documents formatted in Markdown, with additional $\text{\LaTeX}$ syntax to represent mathematical equations and have greater flexibility when representing tables.

In this work, we introduce an Adapter module between the Encoder and the Decoder. In particular, we propose to keep the Nougat Encoder frozen and train the Adapter while also fine-tuning the Nougat Decoder. This reduces the resources needed to extend the parsing model, both in terms of training time and annotated data required. In particular, the proposed Adapter takes as input the Encoder embeddings of three consecutive pages and extracts relevant information into N learned tokens using the cross-attention mechanism. These tokens are then passed to the Decoder alongside the P embedding vectors extracted by the Encoder for the current page that is being parsed. We provide an overview of our pipeline in Fig. 4.2, providing details of the Adapter module in the following.

ADAPTER MODULE. In order to use multiple pages as input for parsing, a simple approach would be to concatenate multiple page embeddings and feed them directly to the Decoder [378]. However, our Decoder, as well as most of those used in document parsing solutions, feature attention operators, which scale quadratically with the input size. Therefore, this naïve solution would lead to a significant computation overhead. In light of this, we propose to use a dedicated Adapter module. We implement the adapter as a Perceiver [163] architecture, which takes as input the Encoder outputs for the previous, current, and next pages and consists of a series of L layers, in which cross-attention and feed-forward operations are performed. We train this adapter to generate N d -dimensional tokens that

summarize information on the content of neighboring pages, which is useful for parsing the current page.

Specifically, we run the Visual Encoder for each of the three input pages and concatenate the resulting embeddings. Then, we sum a learnable inter-page positional encoding, whose values are the same for all the tokens referring to the same page, and a learnable intra-page positional encoding, whose values are the same for the tokens in the same position of their respective page. The result is linearly projected to the dimension of the keys and values for cross-attention, which is performed using as queries a linear projection of the N learnable tokens. After the cross-attention, the N output tokens go through a position-wise feed-forward neural network, as in the standard Transformer. We propose to maintain the number of learned tokens much smaller than the number of Encoder embedding vectors in order to limit the computation overhead. In fact, the cross-attention operations inside the Adapter scale linearly with the number of context pages considered and have only N additional elements. Moreover, the overhead is limited also in the Decoder since its input is increased by only $N \ll P$ tokens. Finally, note that when the previous or the next pages are not available (which can be the case for incomplete documents or for the first and the last pages), we feed the adapter with the embeddings of an empty page.

4.2.2 TRAINING AND FINE-TUNING DATA

ACADEMIC PAPERS. Previous works tackling document parsing [48] usually collect paired single-page images and corresponding markup transcription training data by crawling academic papers PDFs and the associated $\text{\LaTeX}$ source files, from which the ground truth markup parsing can be obtained. However, the authors of these works do not release the data they used. Moreover, the multi-page focus of this work requires preserving the information about the order of the pages to allow feeding our model with the context of the current page to be parsed. In light of this, we collect a small dataset with this characteristic by crawling data from arXiv, which we will release to foster future research on VrDU models able to exploit the context. Note that Nougat, which we use as the base

model, was trained on arXiv papers published on the website before January 2024. Thus, for our dataset, we select only papers released after January 2024 to avoid feeding the Encoder model with data already seen during its training. We download 6933 multi-page documents, and we perform the same preprocessing steps as in [48][†], resulting in 7724 annotated pages. In fact, due to the necessity of linking the visual parts of the images to the textual ground truth, after these preprocessing steps, only a portion of the image-markup pairs are deemed sufficiently high-quality and maintained in the dataset, with an acceptance rate of about 47%. In our dataset, there can be four scenarios, depending on the availability of the previous and the next page as context for parsing the current page. These are summarized in Section 4.2.2 and are as follows:

- Only the current page is available (the input to μgat will contain empty-page embeddings to replace the previous and the next page);
- The previous and the current pages are available, but not the next one (this latter is replaced with empty-page embeddings when preparing the input to μgat);
- The current and next pages are available, but not the previous one (this latter is replaced with empty-page embeddings when preparing the input to μgat);
- Both the previous and next page are available for the current page.

When composing our dataset, we make sure to maintain a balanced combination of these scenarios in order to facilitate the training of a more general adapter. Our dataset splits follow the 80-10-10 rule, obtaining 7690 samples for training, 962 for validation, and 913 for test. We provide detailed statistics in Sec. 4.2.2.

TABLES. In order to simulate the regesta use-case more closely, we build a synthetic dataset of three-column, full-page tables. Specifically, we select ten Pandoc[‡] templates[§] and fill them with text taken from parts of the papers in our arXiv

[†]<https://github.com/facebookresearch/nougat>

[‡]<https://pandoc.org/index.html>

[§]Selected among those in <https://github.com/cab-1729/Pandoc-Themes>

prev.	curr.	next	Training	Validation	Test
	✓		1276	156	165
✓	✓		1758	207	210
	✓	✓	1758	208	210
✓	✓	✓	2898	391	328
total			7690	962	913

Table 4.1: arXiv dataset Statistics. For each split, we report the different types of context that can be provided to the model. Given the current page (curr), there can be the previous page (prev.) only, the next page (next) only, and both or neither available.

4

dataset. In this way, the major challenge of the tables dataset will consist in the layout rather than in the text. The table documents are generated by paying attention to having a good variety of styles (in terms of fonts and table decorations) and table format. In this respect, we include both tables with columns of equal width, and more *regesta-like* tables, in which the first two columns are smaller and contain fewer words than the third column. In fact, recall that in the regesta, the first two columns contain information about the date and place, while the third contains the summarized text of the referenced full-length document. In total, we obtain 399 samples that can be used for training or fine-tuning, 36 samples for validation, and 38 for test.

REGESTA PONTIFICUM ROMANORUM. In order to evaluate the performance of our models on the selected regesta use-case, we semi-automatically annotate 12 pages from the *Regesta Pontificum Romanorum inde a 1198 ad 1304*. To this end, we feed GPT-4V with each page image and prompt it to parse the image into a $\text{\LaTeX}$ table that follows the three-column Regesta structure. Afterward, we manually correct both the transcription and formatting errors. Note that annotating the regesta is a time-consuming task, which also requires multidisciplinary expertise (*e.g.*, certain proficiency both in markup and Latin languages). For this reason, we collect a small amount of pages. This set constitutes our evaluation benchmark for the trained models on the challenging regesta use-case.

4.3 EXPERIMENTS

4.3.1 IMPLEMENTATION AND TRAINING DETAILS

We consider Nougat-Small [48] as the baseline model, with Swin-Base as the Visual Encoder [224] and a 4-layer mBART [201] Decoder. Following [48], we render the documents as images at a resolution of 96 DPI and resize and pad them to reach the resolution (896, 672) for compatibility with Swin-Base. This results in $P=588$ embedding vectors with $d = 1024$. The proposed Adapter is composed of 2 layers with 4 learnable tokens. We train the overall μ gat model for 50 epochs on the proposed academic papers dataset. Note that we keep the Encoder weights frozen, train the Adapter by using AdamW [225] as optimizer, with learning rate $5e-4$, and fine-tune the Decoder with the same optimizer but learning rate $5e-5$. Every epoch, the learning rates are reduced by a factor $\gamma=0.9996$ until they reach $2e-6$, with an exponential scheduler. Note that we run an experiment in which we fine-tune on the synthetic tables. To this end, we use a learning rate of $5e-5$ and fine-tune the models for 30 epochs.

4.3.2 EXPERIMENT SETUP

For validating our proposed approach, we consider the three datasets of documents described in Sec. 4.2.2. Recall that these are academic papers (which we refer to as **arXiv**), rendered tables (which we refer to as **tables**), and pages from the Regesta Pontificum Romanorum (which we refer to as **RPR**).

As for the evaluation scores, in line with Document Parsing literature, we report the performance in terms of the following metrics: Edit Distance (**ED**) [199], which measures the number of character substitutions, insertions, and deletions to transform one string into another; BLEU (**B**) [268], which measures the number of matching n-grams between a source and target text; METEOR (**M**) [29], which considers also the recall of the target n-grams; Precision (**P**), and Recall (**R**). Note that these two latter scores are obtained by collecting the words in the ground truth and in the model output prediction into two sets and calculating the number of words in the prediction set that also appear in the ground truth

	arXiv					RPR				
	ED	B	M	P	R	ED	B	M	P	R
Nougat-S	15.6	80.1	86.3	90.4	88.7	75.4	3.2	11.0	36.4	15.0
Nougat-S*	10.4	84.7	89.1	91.9	89.5	75.4	3.2	11.0	36.4	16.0
Nougat-B	15.6	80.1	86.3	90.5	88.7	75.4	3.2	11.0	36.4	16.0
μgat	8.8	87.1	91.4	92.2	91.2	72.1	4.8	11.8	18.9	14.3

Table 4.2: Comparison between μ gat, Nougat-S (our base model) applied in direct transfer and fine-tuned on the same data used to train our model (Nougat-S*), and Nougat-B applied in direct transfer.

(for the precision) and the number of words in the ground truth set that also appear in the prediction (for the recall).

4.3.3 RESULTS

QUANTITATIVE COMPARISON. First, we perform a quantitative comparison against the baseline Nougat-Small (which we refer to as Nougat-S) model, both directly applied to the arXiv and RPR samples and fine-tuned on the training split of arXiv (we refer to this version as Nougat-S*), and the larger Nougat-Base model (which we refer to as Nougat-B). We show the results on arXiv and RPR in Tab. 4.2. As we can see, our model performs better than the competitors. Another aspect that emerges from the results is that the RPR dataset, to which we directly apply all the compared models without any fine-tuning, is still very challenging, even for our model.

ROLE OF THE CONTEXT. Moreover, in Tab. 4.3, we show the validation results for our μ gat model and the Nougat-S* baseline on the arXiv test set further divided depending on the context available for each page to be parsed. As we can see, while the performance of the compared models is very close when the isolated current page is provided, the performance of μ gat consistently increases as more context is provided, indicating that our strategy allows exploiting inter-page structure for parsing.

ROLE OF THE TRAINING OR FINE-TUNING DATASET. To evaluate the role of the training and fine-tuning, we perform an analysis in which we fine-tune the baseline Nougat-S on the arXiv training set used for training μ gat and further

		prev.	curr.	next	ED	B	M	P	R
Nougat-S*			✓		13.4	80.2	85.6	88.0	86.3
	✓		✓		11.2	83.9	88.7	91.7	89.0
		✓	✓	✓	10.9	84.6	88.8	92.5	89.0
	✓	✓	✓	✓	7.6	88.1	91.6	93.7	91.7
μgat			✓		12.3	81.8	87.4	88.0	87.5
	✓		✓		9.3	86.7	91.4	92.5	91.2
		✓	✓	✓	9.1	87.5	91.3	92.7	91.2
	✓	✓	✓	✓	6.5	89.8	93.4	93.8	93.2

Table 4.3: Performance of Nougat-S* and μ gat on pages of the arXiv test set for which it is available just the current page, also the previous or the next, and both.

		arXiv					tables					RPR				
	tr/ft	ED	B	M	P	R	ED	B	M	P	R	ED	B	M	P	R
Nougat-S	-	15.6	80.1	86.3	90.4	88.7	63.3	31.1	59.5	79.7	65.9	75.4	3.2	11.0	36.4	15.0
	arXiv	10.4	84.7	89.1	91.9	89.5	63.7	27.6	41.3	69.6	44.6	75.4	3.2	11.0	36.4	15.0
	arXiv, tables	19.7	75.2	82.7	84.6	83.8	58.7	35.7	57.7	73.8	62.4	75.4	3.2	11.0	36.4	15.0
μgat	arXiv	8.8	87.1	91.4	92.2	91.2	61.4	32.7	51.1	68.1	55.3	72.1	4.8	11.8	18.9	14.3
	arXiv, tables	18.5	76.4	82.2	84.2	82.6	8.6	85.5	94.2	96.4	94.5	84.9	0.7	04.8	23.3	7.1

Table 4.4: Ablation on the dataset used for fine-tuning (ft) or training (tr) the Adapter of μ gat and decoder of Nougat-S (our base model) or μ gat.

fine-tune both models on the tables dataset. From the results reported in Tab. 4.4, we observe that fine-tuning improves the performance of both models but does not help much when directly applying them to the RPR pages. It is worth noting that Nougat-S benefits less than μ gat from fine-tuning, especially on tables.

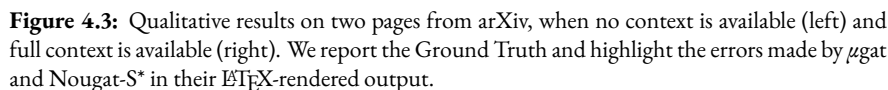
ABLATION ON THE NUMBER OF ADAPTER LAYERS AND TOKENS. In Tab. 4.5, we show an ablation analysis by varying the number of Adapter layers L and of the learnable tokens N . We notice that increasing L beyond 2 does not improve the performance, possibly also due to the small size of the training dataset. As for the number of tokens, we empirically find that 4 is a good trade-off between representation capabilities and generalization.

QUALITATIVE RESULTS. Finally, we report a qualitative analysis of the performance of Nougat-S* and μ gat on the arXiv dataset in Fig. 4.3 and of μ gat on the RPR pages in Fig. 4.4. The images for the ground truth and the models parsing output are obtained by rendering the markdown/L^AT_EXin PDF by using Pandoc.

#layers	#tokens	arXiv						tables						RPR					
		ED	B	M	P	R		ED	B	M	P	R		ED	B	M	P	R	
2	2	10.2	84.9	89.7	91.4	89.8		67.3	24.4	39.2	69.4	43.1		84.8	1.2	4.1	22.7	6.2	
	4	8.8	87.1	91.4	92.2	91.2		61.4	32.7	51.1	68.1	55.3		72.1	4.8	11.8	18.9	14.3	
	8	9.6	86.1	90.7	91.7	90.7		68.5	20.4	35.0	58.6	38.6		76.7	2.4	6.4	16.6	8.4	
4	2	9.4	86.4	90.7	91.9	90.7		64.1	27.5	46.2	71.8	50.4		77.3	2.6	7.1	17.1	9.6	
	4	9.9	85.7	90.4	91.6	90.5		65.4	27.3	40.9	75.6	43.6		78.0	2.6	8.3	24.1	11.3	
	8	9.6	86.0	90.6	91.5	90.6		63.4	31.9	47.4	73.1	51.7		88.3	0.8	2.5	14.4	3.5	

Table 4.5: Ablation on the number of Adapter layers and tokens in our μ gat model.

From Fig. 4.3, we can observe that the output of μ gat for the arXiv data is very close to the ground truth, especially when full context is provided. On the other hand, Nougat-S* misses some more details (like superscript numbers signaling footnotes) and even collapses into repeating the phrase *Milano, Italy* a few times. As for RPR pages (Fig. 4.4), we observe that our model struggles to capture the complex table-like structure but can extract most of the Latin text from the page, with some mistakes. We argue that this is a promising result since it shows that, even though the model was trained almost completely on English text, it exhibits some capability to also extract text in other languages and from low visual quality pages. This ability and the capability to deal with the complex layout of regesta could be further improved by fine-tuning the model on a larger amount of specific data, which we leave for future research.



Original	INNOCENTIIUS III. 1198-1216		INNOCENTIIUS III. 1198-1216	
	1198	1199	1200	1201
	1202	1203	1204	1205
Ground Truth	INNOCENTIIUS III. 1198-1216		INNOCENTIIUS III. 1198-1216	
	1198	1199	1200	1201
	1202	1203	1204	1205
μ gat	INNOCENTIIUS III. 1198-1216		INNOCENTIIUS III. 1198-1216	
	1198	1199	1200	1201
	1202	1203	1204	1205

Figure 4.4: Qualitative results on two pages from the challenging RPR. We report the Ground Truth and highlight the errors made by μ gat in its $\text{\LaTeX}$ -rendered output.

4.4 CONCLUSION

In this work, we have investigated the limitations of current single-page parsing models in handling multi-page documents. This is critical in many real-world scenarios, particularly for documents where inter-page dependencies are crucial, such as regesta collections, which are a valuable tool for historians and humanities scholars. To tackle this limitation, we have proposed a novel approach for context-enhanced document parsing, dubbed μ gat. Our key contribution is the introduction of a novel adaptation strategy for pre-trained, OCR-free single-page document parsing models (specifically, the recently proposed Nougat), which extends these models for effectively utilizing the context provided by surrounding pages, resulting in significantly improved parsing consistency and accuracy.

We have validated the efficacy of our approach on a specifically collected dataset of scientific papers consisting of pages with their context, rendered long tables, and pages from the Regesta Pontificum Romanorum. The results obtained prove that leveraging information from surrounding pages can significantly improve the parsing accuracy of single-page models. These encouraging results can contribute to the development of VrDU solutions able to deal with multi-page and intricately-structured documents, such as ancient documents and regesta, which are particularly relevant to Digital Humanities.

5

A Text Recognition Dataset from Sahidic Coptic Ancient Manuscripts

IN this chapter, we target HTR in low-resource scenarios, which arise from underrepresented languages, rare scripts, and degraded visual conditions typical of historical documents. We introduce SCAM (Sahidic Coptic Ancient Manuscripts), a new line-level dataset of digitized ancient manuscripts written in the extinct Sahidic Coptic dialect. The dataset reflects a realistic and challenging setting, as it combines heterogeneous acquisition conditions across libraries with typical manuscript degradations such as ink fading, bleed-through, and material deterioration. In addition to visual complexity, SCAM poses significant linguistic challenges due to the scarcity of resources for Sahidic Coptic, its uncommon alphabet, and dialect-specific diacritics. To support research in low-resource HTR, we benchmark several state-of-the-art approaches based on different paradigms, highlighting their limitations and strengths in this setting. Our

This Chapter is related to the publication “F. Quattrini*, C. Zaccagnino* *et al.*, A Text Recognition Dataset from Sahidic Coptic Ancient Manuscripts, ICDAR 2026” [7]. *Equal contribution. See the list of Publications on page 315 for more details.

results underline the gap between current HTR performance on well-resourced modern scripts and historically grounded, low-resource scenarios, thus providing a reference point for future developments.

5.1 INTRODUCTION

5 HTR has made substantial progress in recent years, particularly due to the success of Deep Learning and end-to-end trainable architectures [16]. These systems, typically based on CNN-BLSTM-CTC pipelines or Transformer encoders, achieve good recognition performance when trained on large and well-annotated datasets in modern languages and scripts. However, their performance significantly degrades when applied to low-resource scenarios, where data is limited, noisy, or involves rare scripts and underrepresented languages. Such low-resource settings are common when dealing with historical manuscripts, poorly preserved documents, or extinct language scripts. In these contexts, the performance achievable with standard HTR pipelines often drop. To address this limitation, recent work has explored alternative strategies, such as synthetic data generation [178, 40], and more recently, script-agnostic frameworks [27]. Nonetheless, some training is still required to achieve satisfactory performance, which realistically enables further usage of the transcribed documents by Humanities scholars.

In light of these considerations, we introduce a new line-level dataset featuring manuscripts in Sahidic Coptic, an underrepresented language of considerable interest to Coptologists and scholars of biblical studies, since it is among the languages into which the Bible was translated at an early stage. Our dataset is obtained from manuscript images preserved in different libraries and digitized with highly differing settings. This characteristic reflects a common fate of Coptic manuscripts from the White Monastery, many of which were dismantled in the 18th and 19th centuries and sold as individual leaves rather than as complete codices. From the point of view of the resulting images, the different acquisition settings and quality adds on the nuisances that are typical of ancient manuscripts (*e.g.* ink fading, bleed through, scratches, yellowed paper). Moreover, the language of the manuscripts considered is another major challenge of the proposed



Figure 5.1: Sample images from the manuscript used to obtain the SCAM dataset. The top-four rows are from the SCAM-A subset, the bottom ones are from the SCAM-B.

dataset. In fact, Sahidic Coptic is no longer a spoken language, and the texts in the dataset are written in the Coptic script and employ a supralinear system characteristic of this language.

To promote research in low-resource HTR, especially in the challenging scenario of manuscripts from different libraries, written in extinct languages, we devise our dataset, dubbed Sahidic Coptic Ancient Manuscript (SCAM)*. The dataset features text lines extracted from pages preserved in two different archives and digitized under different conditions. Therefore, we organize the dataset into two subsets, dubbed SCAM-A and SCAM-B. Some example images used to build our dataset are reported in Fig. 5.1. In this chapter, we analyze the visual and lin-

*A link will be added upon publication

guistic characteristics of our dataset, and benchmark different-paradigm State-of-the-Art HTR approaches, both to highlight the challenges of these data and the most promising solutions to face them, and to provide a leaderboard for future works.

5.2 PROPOSED DATASET

Some recent works have targeted HTR in low-resource scenarios, also proposing datasets to better explore the task. By low-resource, one can refer to the visual appearance of the manuscript (given *e.g.* by its preservation state, or writer-specific handwriting, or both) or to the language of the manuscript, which is usually the case for historical documents in ancient languages. In this work, we devise a dataset for low-resource HTR that features all the challenges of this scenario. Specifically, our dataset, which we call SCAM, contains lines from 27 leaves, corresponding to 53 pages of an ancient manuscript written by a single author in Sahidic Coptic, an extinct language. Note that by “leaf” we mean a physical parchment sheet, which can accommodate one or two pages, written on both faces. The text is written in *scriptio continua*, meaning that there are no spaces delimiting words.

The parchment manuscript, now known as Coptic Literary Manuscript (CLM) 359, is dated to 1002-1003 C.E., as indicated by the colophon, and comes from the White Monastery in Upper Egypt. To the best of our knowledge, only 49 leaves of the original manuscript have survived, which are now spread across countries, preserved in various libraries and cultural institutions[†]. This manuscript is of considerable importance, as it preserves substantial Coptic translations of Greek canonical and conciliar texts of the fourth and fifth centuries. It is not only a key witness to the development of Coptic canon law, but also an important source for reconstructing earlier stages of the Greek textual tradition.

As a result, each small set of leaves has been digitized under different conditions. This can significantly hinder the performance of an HTR model trained on samples from a certain collection when applied to samples in a different collec-

[†]<https://atlas.paths-erc.eu/manuscripts/359>

tion. To study this behavior, we collect samples from leaves preserved in two different libraries and organize the SCAM dataset into two subsets, dubbed SCAM-A and SCAM-B. The 15 leaves in the SCAM-A subset are digitized with professional scanners, under controlled lighting conditions, and with careful placement of the page in the scanner. On the other hand, the 12 leaves in the SCAM-B collection have been photographed with a relatively good-quality camera, without explicit control over the lighting conditions or the placement of the page. In the following, we describe the process we followed for obtaining the samples from these pages, and analyze the visual and linguistic characteristics of the resulting dataset.

5.2.1 DATA PREPARATION

First, we isolate the line-level samples from all the page images contained in the collected leaves. Specifically, we resort to the popular open-source computer vision annotation tool CVAT[‡] to trace polygons around each line. Note that we prefer using polygons instead of bounding boxes for two main reasons. The first is that some lines are not straight, either because the writer has laid them down without following a perfectly horizontal line or because of digitization artifacts. The second reason is more critical and specific of the manuscript at hand: since the manuscript is in scriptio continua, some of the lines start with bigger, ornate letters to indicate the beginning of a new paragraph or phrase. As a result, straight bounding boxes would lead to glyphs from other line samples be part of the line image of the sample with such slant or beginning letters (see Fig. 5.1 for some examples).

Once the lines have been isolated, an expert Coptologist provided their diplomatic transcriptions for each line, which are associated with the image by using CVAT. Note that, in this process, the Coptologist was able to exploit a special keyboard for Sahidic Coptic. The Sahidic Coptic characters are encoded in Unicode, but some of the graphemes are encoded using sequences of multiple code points, called **grapheme clusters**. This makes it possible to introduce variations

[‡]<https://www.cvat.ai/>

of a character, such as accents. For model training and prediction, we ask the model to predict grapheme clusters and not individual code points, so that each glyph always corresponds to a single character.

5.2.2 DATASET CHARACTERISTICS

STATISTICS. Our dataset comprises 3240 lines, corresponding to 53 pages. It includes a total alphabet of 99 graphemes, with an average of 11.39 ± 2.04 characters per line. The data are partitioned into training, validation, and test splits as follows: the training set contains 2000 lines from 33 pages (89 unique graphemes), the validation set includes 610 lines from 10 pages (59 unique graphemes), and the test set consists of 630 lines from 10 pages (68 unique graphemes).

As stated above, the images in our dataset are split into two subsets, SCAM-A and SCAM-B, depending on the archive in which the leaves are preserved. The characteristics of these subsets are the following:

- **SCAM-A** contains 1445 lines in total, corresponding to 24 pages from 12 leaves. The subset features 87 graphemes, with an average line length of 11.52 ± 2.30 characters. The data are partitioned into training, validation, and test sets as follows: the training split includes 943 lines from 16 pages (76 unique graphemes), the validation split comprises 242 lines from 4 pages (48 unique graphemes), and the test split consists of 260 lines from 4 pages (61 unique graphemes).
- **SCAM-B** comprises a total of 1795 lines, corresponding to 29 pages extracted from 15 leaves. The subset includes 67 distinct graphemes, with an average of 11.29 ± 1.81 characters per line. It is divided into training, validation, and test splits as follows: the training set contains 1057 lines from 17 pages (63 unique graphemes), the validation set includes 368 lines from 6 pages (51 unique graphemes), and the test set consists of 370 lines from 6 pages (50 unique graphemes).

In Section 5.2.2 we quantify the visual and linguistic differences between the subsets.

	FID	KID	HWD	JSD
SCAM-A $\leftrightarrow$ SCAM-B (RGB)	113.62	0.13	1.06	0.49
SCAM-A $\leftrightarrow$ SCAM-B (grayscale)	65.06	0.07	1.04	

Table 5.1: Visual and linguistic differences between the two subsets, expressed in terms of quantitative scores. We consider both the original RGB version of the images and their grayscale version.

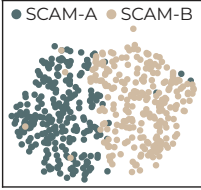


Figure 5.2: t-SNE projection of VGG features extracted from RGB images in our dataset. 250 random points per subset plotted.

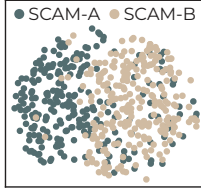


Figure 5.3: t-SNE projection of VGG features extracted from grayscale images in our dataset. 250 random points per subset plotted.

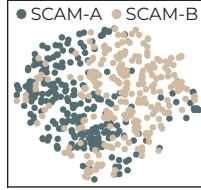


Figure 5.4: t-SNE projection of HWD features extracted from RGB images in our dataset. 250 random points per subset plotted.

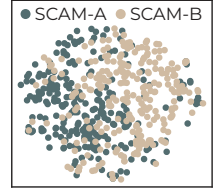


Figure 5.5: t-SNE projection of HWD features extracted from grayscale images in our dataset. 250 random points per subset plotted.

VISUAL CHARACTERISTICS. First, we consider the visual characteristics, and quantify them in terms of Fréchet Inception Distance (**FID**) [147] and the Kernel Inception Distance (**KID**) [44]. Note that these scores, commonly adopted in Image Generation evaluation, measure the dissimilarity between the distributions of visual features extracted from two sets of images, and therefore, we deem them suitable to compare the samples in the SCAM-A and SCAM-B subsets in terms of overall visual appearance. Moreover, we include in the analysis the Handwriting Distance (**HWD**) [1], a recently-proposed score which isolates just the differences in writing style between pairs of images. The first row in the table contains the values of such scores computed on the original, RGB images (such as those in Fig. 5.1). It emerges that the perceptual features are quite distinct and that the writing style in the two subsets is related but not identical. In particular, as shown by [1], typical values for HWD in same-author sets is between 0 and 1.5. This reflects the fact that the manuscript from which SCAM was derived was written by the same person. We deepen our analysis by applying a grayscale con-

version to the images of both subsets, as this is a simple transformation that practitioners can easily adopt with the aim of reducing the visual gap between the subsets. Alternatively, document restoration approaches such as binarization [266], [4] could be applied, but once sufficiently explored for our data. In fact, blindly applying such approaches can result in text content alteration in the lines. As it can be observed from the values in the second row of Section 5.2.2, the difference in visual features is reduced but not entirely eliminated, underlining that the SCAM dataset poses visual challenges that are not trivial to overcome.

These aspects are even clearer when looking at t-SNE projections of visual features extracted from both subsets (Figs. 5.2 to 5.5). In particular, we project features extracted by a VGG trained on natural images and by the VGG backbone of the HWD score, trained to distinguish handwriting styles. We argue that the first feature set captures the overall appearance (*e.g.* color and texture) of the images, while the second set captures the handwriting style characteristics (*e.g.* thickness, slant, and kerning). Note that both feature sets are extracted from the original RGB images in the subsets and from their grayscale version. From Fig. 5.2, we can see that the VGG features on RGB images are clearly divided into two separate clusters, which get closer when computed on grayscale images, as shown in Fig. 5.3. The HWD features, as shown for RGB and grayscale images in Fig. 5.4 and in Fig. 5.5 respectively, are less separated. This suggests that, although the handwriting in the two subsets is the same (the images come from a single-author manuscript), the differences between the images in the two subsets in terms of preservation state and digitization settings have also impacted the characteristics of the handwriting (*e.g.* by altering the thickness).

LINGUISTIC CHARACTERISTICS. In Section 5.2.2 we also report the Jensen-Shannon divergence (**JSD**) computed on n -grams (with $n = 1, \dots, 5$) extracted from the textual transcriptions of the SCAM-A and SCAM-B subsets. This score is considered to obtain a measure of the similarity of the text content of the two subsets. The value of the score suggests that, even though the dataset contains text taken from the same manuscript treating the same topic, there are non-negligible differences in the text content.

We deepen this analysis by plotting the distribution of the graphemes on the

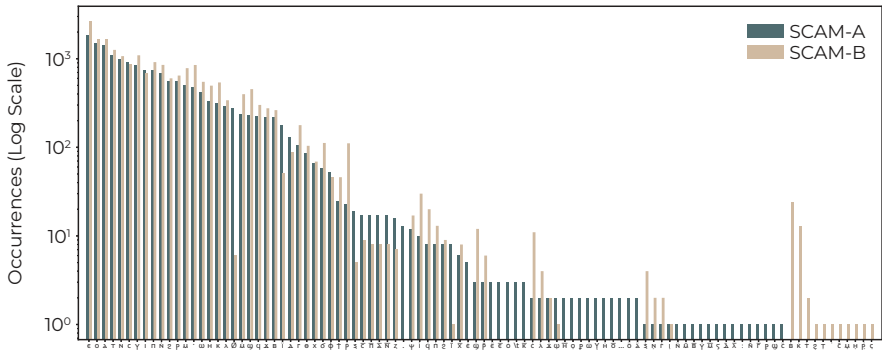


Figure 5.6: Distribution of the graphemes in the two subsets of our dataset, in logarithmic scale. For comparison, we order the graphemes in descending order of frequency as computed in SCAM-A.

two subsets (in Fig. 5.6). It can be observed that, as happens in languages, the most common characters are equally frequent in both subsets. However, both subsets contain unique graphemes that are not present in the other, with SCAM-A having more of such cases (31 graphemes that are not present in SCAM-B, versus 11 graphemes in SCAM-B that are not in SCAM-A).

5.3 EXPERIMENTAL ANALYSIS

This section presents an experimental analysis of the performance of several state-of-the-art models. Evaluation is performed by training each of them on **SCAM-A**, **SCAM-B** and the full **SCAM** dataset. Then, each model variant is tested on both the SCAM-A and SCAM-B subsets separately and jointly.

CONSIDERED SCORES. To assess performance, we report the Character Error Rate (**CER**) and the Sequence Error Rate (**SER**). Following standard conventions in HTR, these scores are derived from the Edit Distance between the predicted text and the corresponding ground truth. Specifically, we compute the minimum number of substitutions, deletions, and insertions required to transform the predicted sequence into the ground truth. To determine the CER and SER for the entire test set, we sum the distances of all samples and normalize by the total number of units in the ground truth. For the CER, the distance is com-

puted at the character level, while for the SER, it is computed on the content of the entire line sample, since our dataset is in scriptio continua.

CONSIDERED APPROACHES. We consider a range of model architectures and sizes in order to showcase the challenging nature of this dataset, and highlight which models perform best on it. In particular, we test the following architectures:

- We include the model proposed in [293], dubbed **CRNN**, which combines a convolutional backbone with BLSTM layers and is trained end-to-end to optimize the CTC loss. This model, which represents one of the most popular paradigms for learning-based HTR, serves as our primary recurrent baseline for sequence modeling without explicit alignment.
- We also consider the Sequence-to-Sequence recurrent attention paradigm by including the approach proposed by **Michael et al.** [247]. This approach combines a CNN-BLSTM encoder with an attention-based LSTM decoder and auxiliary CTC loss, serving as a strong RNN-based attention baseline.
- Additionally, we consider models that use attention operations instead of recurrent components, namely the Vertical Attention Network (**VAN**) proposed in [84] and the Convolutional Self-Attention Network (**C-SAN**) from [93]. These models replace RNNs with gated or self-attention mechanisms to enable parallelization and more efficient training, and assessing whether attention alone can model handwriting dependencies.
- Another strategy to replace the recurrent component of the HTR pipeline entails exploiting a Transformer encoder-decoder, as **LT** [32], **VLT** [33], and the one proposed by **Kang et al.** [176], or a Transformer encoder, as done in **HTR-VT** [210]. In both cases, the visual features are extracted by a convolutional backbone, and the Transformer-based decoder improves parallelization and captures long-range dependencies in handwriting.
- Finally, we consider **TrOCR** [205], which represents the fully Transformer-based, end-to-end paradigm. Specifically, the image features are extracted

with a Transformer encoder, and the transcription is obtained with a Transformer decoder. Both components have been pre-trained on synthetic data.

IMPLEMENTATION DETAILS. For the experimental evaluation, we first finetune TrOCR using the official HuggingFace implementation [393]. For each model size, we initialize the weights from the respective *handwritten* released checkpoint[§]. For the other architectures, we base our code on the HTR models implementations available online[¶] and train them from scratch. To this end, we modify the tokenizer, removing the ASCII transliteration step^{||}, due to its incompatibility with non-Latin scripts and, therefore, our graphemes (Section 5.2.1). Each model is trained (or fine-tuned, in the case of the TrOCR variants) with an early stopping strategy based on the CER computed on the validation set of the dataset used for training, with patience set to 100 epochs.

5.3.1 RESULTS

In the following, we report an experimental analysis, both quantitative (Tab. 5.3 and Sec. 5.3.1) and qualitative (Fig. 5.7) of the considered approaches, and discuss the insights drawn from such analysis.

ARCHITECTURE TYPE MATTERS MORE THAN SCALE FOR RARE SCRIPTS. Across all tables, we can observe that VAN achieves the best CER and SER on both subsets (some examples of its best and worst transcriptions over the different setups are reported in Fig. 5.7). We argue that the gated vertical-attention mechanism is particularly well-suited to the regular layout of the Coptic script glyphs: by attending to vertical strips of the image, VAN naturally aligns with the character-level structure of the lines without relying on explicit language priors. In contrast, TrOCR variants, despite their large-scale synthetic pre-training on Latin-script handwriting, do not achieve competitive results. These models rely on an autoregressive language prior at decoding time, while CTC-based models condition

[§]hf.co/microsoft/trocr-size-handwritten, where $\text{size} \in \{\text{small}, \text{base}, \text{large}\}$

[¶]<https://github.com/carlosl0garrido/HTR-00D/>

^{||}<https://pypi.org/project/Unidecode/>



Figure 5.7: Qualitative results on the considered setups (Train dataset → Test dataset) of the overall best-performing benchmarked model (VAN). We also report the CER on each considered sample and highlight the errors in red.

each predicted frame based on visual features independently. This pattern suggests that Transformer decoders pre-trained on high-resource scripts cannot efficiently adapt their learned language priors to an unseen script under data-scarce fine-tuning, and that model scale is not a substitute for architectural alignment with the task. Additionally, the scriptio continua writing provides a particularly challenging setting, which prevents the model from easily learning word priors. We also note that SER values are consistently high across all models and settings, even when the CER is relatively low. This is a direct consequence of the scriptio continua: with an average line length of 11.39 ± 2.04 characters, even a model achieving ~ 7.5 CER is expected to produce at least one error in roughly half of all lines, counted as a full-line mismatch in the SER.

DATA BREADTH MATTERS MORE THAN MODEL SCALE. In Section 5.3.1, we report results when training on the full SCAM dataset and testing on each subset.

	SCAM		SCAM-A		SCAM-B	
	CER	SER	CER	SER	CER	SER
CRNN [293]	9.47	49.79	9.00	51.54	9.85	48.38
C-SAN [93]	88.06	100.00	87.52	100	88.50	100.0
VAN [84]	5.25	31.99	5.96	36.15	4.68	28.65
HTR-VIT [210]	10.27	59.62	10.98	66.54	9.70	54.05
Kang et al. [176]	11.43	54.18	11.44	57.69	11.42	51.35
Michael et al. [247]	33.35	91.16	36.19	95.39	31.07	87.76
LT [32]	22.59	78.74	21.86	78.85	23.18	78.65
VLT [33]	15.03	62.81	14.89	67.31	15.15	59.19
TrOCR-S [205]	13.12	65.99	17.04	86.87	9.96	49.20
TrOCR-B [205]	12.42	74.39	14.42	77.78	10.81	71.66
TrOCR-L [205]	11.33	72.63	12.82	74.50	10.14	71.12

Table 5.2: Performance obtained by models trained on the entire SCAM training set and tested on the whole test set (SCAM) and on each test subset (SCAM-A and SCAM-B).

This represents the most favorable scenario, with models trained on annotated samples from multiple archives. Training on SCAM brings substantial improvements for most models. We argue that these results confirm that the in-domain failures observed in Tab. 5.3 are primarily due to data scarcity rather than to a fundamental incompatibility with the Sahidic Coptic script. Overall, these results show that using in-domain training data is more important than model capacity, as lightweight VAN and CRNN outperform the pre-trained TrOCR-L.

DOMAIN GENERALIZATION ACROSS SUBSETS. In Section 5.3.1, we evaluate the performance of models trained on one subset and tested on the other. This setting directly mirrors a real-world scenario, where a scholar might annotate leaves from one library, train a model, and then apply it to newly-digitized leaves from a different archive. Note that a scholar does not always have physical access to the archives and, therefore, control over digitization conditions. New leaves might have different acquisition conditions and contain graphemes absent from the training data (*e.g.* SCAM-A and SCAM-B have only partially overlapping grapheme vocabularies of 87 and 67 symbols, respectively, as shown in Fig. 5.6). This is an interesting setting that can be explored with the SCAM dataset. We can observe that training on SCAM-B and testing on SCAM-A is consistently harder than doing the opposite for almost every model. We attribute this to

	A→A		B→B	
	CER	SER	CER	SER
CRNN [293]	13.91	65.39	14.63	64.32
C-SAN [93]	87.52	100.0	9.87	44.05
VAN [84]	7.50	40.77	7.58	41.08
HTR-VIT [210]	15.24	74.23	14.48	71.08
Kang et al. [176]	14.16	65.39	18.38	67.84
Micheal et al. [247]	79.15	98.46	82.00	100
LT [32]	44.28	98.46	46.99	97.30
VL [33]	21.48	80.00	32.94	88.92
TrOCR-S [205]	30.70	98.23	12.11	56.15
TrOCR-B [205]	18.57	81.57	12.36	75.40
TrOCR-L [205]	14.99	78.54	10.59	68.45

Table 5.3: Performance obtained by models trained and tested on the same subset. “A” indicates the SCAM-A subset, “B” the SCAM-B.

the lower grapheme count of SCAM-B. Training on the SCAM-A, which has broader glyph coverage and cleaner images, provides a richer and more generalizable foundation. This asymmetry shows that high-quality library scans with broad grapheme coverage lead to data that can be used to train models that generalize better to field photographs than vice versa. This result offers a guiding principle for practitioners planning transcription pipelines, and defines a challenge for scholars who want to tackle domain generalization in scenarios similar to the one in our dataset.

IMAGE PREPROCESSING HELPS REDUCE DOMAIN GAP. In Section 5.3.1 we evaluate the effect of a simple image preprocessing step, namely conversion to grayscale, on the inter-subset performance. From the results, we can see that this step consistently improves performance across most models. As a notable example, we observe that, in this setting, TrOCR-L reaches comparable performance w.r.t. VAN in CER values, with slightly worse SER values, revealing that bridging the color gap is not sufficient to achieve reliable line-level transcription. This is consistent with the dataset statistics (Section 5.2.2): converting to grayscale nearly halves the FID, but the HWD score barely changes, indicating a residual visual gap. Furthermore, in a scriptio continua text, a single character error in a long boundary-free line leads to a full-line SER mismatch.

	A→B		B→A	
	CER	SER	CER	SER
CRNN [293]	26.63	76.76	31.28	93.85
C-SAN [93]	88.50	100	23.54	88.46
VAN [84]	14.44	55.41	19.94	79.61
HTR-VIT [210]	29.76	89.46	25.31	94.23
Kang et al. [176]	11.44	57.69	32.92	93.08
Micheal et al. [247]	81.48	100	80.51	99.62
LT [32]	64.03	99.73	62.03	99.62
VLT [33]	37.50	88.92	47.28	98.85
TrOCR-S [205]	45.46	98.13	30.37	98.49
TrOCR-B [205]	24.18	92.25	18.36	90.15
TrOCR-L [205]	14.15	79.95	18.36	90.15

Table 5.4: Performance obtained by models trained on a subset and tested on a different subset. “A” indicates the SCAM-A subset, “B” the SCAM-B.

	A→B		B→A	
	CER	SER	CER	SER
CRNN [293]	26.22	78.11	28.63	89.23
C-SAN [93]	87.50	100.00	85.36	100.00
VAN [84]	12.69	52.93	17.96	77.31
HTR-VIT [210]	27.65	85.68	25.52	90.77
Kang et al. [176]	26.91	81.08	35.91	96.54
Micheal et al. [247]	80.81	100.00	81.35	100.00
LT [32]	63.29	98.92	60.01	99.62
VLT [33]	34.13	88.65	51.74	99.62
TrOCR-S [205]	22.73	92.42	25.38	84.23
TrOCR-B [205]	17.73	82.58	18.98	89.31
TrOCR-L [205]	14.84	76.52	15.77	83.70

Table 5.5: Performance obtained by models trained on a subset and tested on a different subset, both in grayscale. “A” indicates the SCAM-A subset, “B” the SCAM-B.

HYBRID DECODERS UNDERPERFORM CTC-ONLY APPROACHES. A consistent finding across all three experimental settings (intra-subset - Tab. 5.3, inter-subset - Section 5.3.1, full-dataset - Section 5.3.1) is the unsatisfactory performance of the three *hybrid* architectures (Michael et al., LT, and VLT), which combine a CTC branch with an additional attention or cross-entropy decoder, compared to approaches relying solely on CTC decoding. From the results, we can infer that the joint CTC+decoder training objective does not converge well on the proposed SCAM dataset. Pure CTC-based models (CRNN, VAN) do not rely on an explicit autoregressive prior at decoding time and are agnostic to word delimiters, making them the most reliable paradigm in this low-resource setting.

5.4 CONCLUSION

In this chapter, we have introduced SCAM, a new line-level dataset in the under-represented Sahidic Coptic language, derived from the 11th-century CLM 359 manuscript and manually annotated by domain experts. By gathering sources from different libraries and digitized under heterogeneous acquisition conditions, we have modeled a realistic and challenging low-resource scenario featuring visual variability, degradation phenomena, and linguistic complexity.

Our experimental evaluation of several state-of-the-art HTR architectures across different sub-splits reveals several findings of broader relevance for Document AI. First, pure CTC-based architectures proved to be the most robust choice for scripto continua manuscripts in resource-scarce languages: by not relying on an explicit autoregressive decoder, they are less prone to error cascades and scale better with limited data than hybrid models, which combine CTC with attention or cross-entropy decoders. Second, the domain gap between SCAM’s two subsets results in models trained on SCAM-B consistently generalizing worse to SCAM-A than vice versa. Third, grayscale preprocessing helps narrow the cross-archive gap when multi-domain data is unavailable but is not a substitute for training on diverse, multi-archive data. When such data is available, models like VAN provide the best accuracy across all settings. When only a single archive is available, simpler architectures like CRNN with grayscale preprocessing can combine robustness and adaptability to cross-domain deployment.

More broadly, these findings underscore the value of SCAM’s dual-archive structure, exposing the domain asymmetry, the grapheme-coverage gap, or the architectural sensitivity of certain models to specific visual distributions. We release SCAM as a benchmark to support future research on low-resource, rare-script, and cross-domain handwritten text recognition, and we hope it will encourage the development of HTR methods better suited to the challenges posed by real historical manuscript collections.

Part II

Generating VRD

6

Literature Survey

THIS part of the thesis focuses on generative approaches for visually-rich documents, with particular emphasis on handwritten text generation and controllable image synthesis. We begin by reviewing the evolution of handwritten text generation methods, from early handcrafted approaches to modern learning-based solutions. We then explore the broader landscape of diffusion models and their applications to controllable generation tasks, including panorama synthesis and transparent image generation.

6.1 GENERATIVE MODELS FOR IMAGE SYNTHESIS

This section reviews the generative modeling paradigms that underpin our contributions, progressing from general-purpose image generation to specialized applications in handwritten text synthesis and controllable generation.

6.1.1 GENERATIVE ADVERSARIAL NETWORKS

GANs [127] have been widely adopted for image generation tasks, achieving impressive results in natural image synthesis [52, 183]. In the context of document

generation, GANs have been successfully applied to handwritten text generation, either unconditioned (for Non-Styled generation) or conditioned on handwriting style examples (Styled generation).

6.1.2 DIFFUSION MODELS

Diffusion models [333, 336, 338, 337, 149, 260, 102] are powerful generative probabilistic models that have achieved drastic performance improvements compared to GANs, especially in text-to-image generation conditioned on natural language prompts [300, 308, 315, 281, 67].

The introduction of Latent Diffusion Models (LDMs) [308] has increased the popularity of diffusion models. LDMs perform the diffusion process in the latent space and thus require fewer computational resources for both training and inference. Such popularity has been further boosted in the text-to-image setting, where the generation is guided by a prompt in natural language (as for, *e.g.*, DALL·E 2 [300], SD-XL [281], and the PixArt family [67, 66]).

Moreover, recently proposed Diffusion Transformers [271, 67, 66, 258, 87] have further increased the scalability and visual generation quality of diffusion models by exploiting a Transformer model instead of the convolutional-attentive U-Net-based noise estimators adopted in previous works.

CONTROLLABILITY AND ADAPTATION. Lots of research efforts have been dedicated to devising efficient fine-tuning strategies to obtain better controllability over the generation process of diffusion models based on additional guiding signals [118, 53, 24, 423] while maintaining the visual quality of the output. Another line of research focuses on achieving zero-shot capabilities by proposing strategies to alter the inference-time behavior of pre-trained diffusion models [244, 230, 23, 250, 373][5].

Some of these approaches entail modifying the noisy latent vectors at each inference step. Some approaches perform mask-guided inpainting by combining the noised image with generated latent vectors [23, 230], or combine multiple independently generated latent vectors for region-based image generation [30]. Other works exploit the gradient of a downstream task-related loss or score [184,

198], possibly computed on the foreseen image obtained from the latent vector.

Some of the proposed strategies for inference-time adaptation exploit the attention layers of pretrained models to perform tasks such as image editing [144, 85], varied-size image generation [170], open-vocabulary segmentation [211, 259], and zero-shot video generation [186]. The role of attention layers in U-Net-based diffusion models has been extensively studied also from an explainability perspective by DAAM [358].

ATTENTION MAP ANALYSIS. Attention maps in diffusion models are not only useful for interpretability, but also for control. DAAM [358] shows that cross-attention maps can be aggregated to localize the image regions associated with prompt tokens, while Prompt-to-Prompt [144] demonstrates that manipulating cross-attention can preserve layout while altering textual attributes. This line of work motivates treating cross- and self-attention as spatial signals that can be repurposed at inference time, which is directly relevant to Chapter 10, where attention maps are used to infer subject localization and transparency structure.

CLASSIFIER-FREE GUIDANCE. Since its introduction [150], CFG has been applied to diffusion- and autoregressive-based image generation [308, 352, 372][5, 3] to improve alignment with the target conditioning. CFG is widely used across various generation tasks, including handwritten text generation [441, 262, 90].

6.1.3 AUTOREGRESSIVE MODELS

Following the success of Transformer-based natural text generation solutions [380, 298], autoregressive image generation has been explored extensively. In most cases, image tokenizers are used to convert continuous images into discrete tokens, and generation is performed as next-token prediction [375, 304, 109, 301, 352, 418]. Recently, LlamaGen [352] and GIVT [372] have shown the potential of continuous autoregressive generative models. In particular, GIVT [372] leverages a continuous-latent β -VAE [148] and a Transformer Decoder to obtain superior performance compared to quantized VAEs. Removing the vector quantization and using continuous latents allows avoiding excessive information compression and the optimization problems in training [375, 63, 159, 245].

In line with this progress, recent approaches are evolving towards the prediction of continuous latent vectors [372, 207, 134, 110, 361, 111, 435].

6.2 HANDWRITTEN TEXT GENERATION

EVOLUTION FROM HANDCRAFTED TO LEARNING-BASED METHODS. The early-proposed approaches to Handwritten Text Generation (HTG) [384, 215, 364, 133] apply handcrafted geometric statistical-based feature extraction on human-made glyphs segmentations, then combine them with appropriate ligatures and render the results with texture and background blending. The two major limitations of these approaches are their inability to render glyphs and ligatures not observed for each style and their reliance on costly human intervention.

HTG approaches can be broadly categorized into two paradigms: *online* and *offline*. Online HTG conceptualizes handwriting as a temporal sequence of strokes. Models in this setting take as input a style representation encoded in the form of stroke trajectories and then predict subsequent trajectories, which are finally rendered as an image [128, 15, 14, 91, 231, 305]. However, the need for specialized hardware such as digitizing tablets to capture stroke-level data makes this approach costly and impractical in scenarios where the reference style is derived from existing manuscripts, such as historical collections. This limitation has motivated the development of offline HTG, where both the style reference and the generated samples are static images. Offline HTG has therefore attracted more research attention and constitutes the focus of this work.

STYLED AND NON-STYLED GENERATION. In recent years, the HTG task has re-gained popularity thanks to learning-based solutions. The first learning-based Non-Styled HTG approach, proposed in [19], entails generating fixed-sized images conditioned on the embedding of the desired textual content but does not control the calligraphic style of the output. Since, different from natural images, handwritten text images are highly variable-sized, subsequent approaches [116] entail concatenating character images.

Styled HTG approaches condition the generation on both the text content and a representation of the style. This capability was introduced in [178], which

proposed to condition the generation on a vector representation of the style. In this respect, the style reference can consist either of a writer identifier [232, 441, 261] (which completely prevents the models from imitating unseen styles) or some reference images in the form of: a paragraph [243], a line [95], fifteen [178, 42, 278, 379] to five [262] words, or a single word [120, 121, 193, 90].

ARCHITECTURAL APPROACHES. Most HTG models leverage one of the generative paradigms discussed in Section 6.1: GAN-based [178, 95, 120, 121, 232, 193], diffusion-based [441, 261, 262, 90], or autoregressive [6] approaches. Early models relied on convolutional backbones, where content and style are encoded separately and fused in later stages [178, 95, 120, 175, 242, 121], preventing those approaches from effectively capturing local writing style and patterns (*e.g.* ligatures or character-specific rendering variations such as repeated letters). To address this issue, recent Transformer-based [380] approaches [42, 278, 379][6] introduce cross-attentions to capture style–content interactions, exploiting the cross-attention mechanism between the style vector representation and the content text representation to entangle the content and the style, thus better rendering local style patterns.

HANDLING LONG SEQUENCES. The majority of existing HTG methods target short sequences, typically single words. As a consequence, when longer outputs are attempted, models often fail to preserve character quality, proportion, and consistency across the output image. Moreover, the words rendered are generated inside fixed-height canvases. This can result in different scales and misaligned baselines of words depending on the fact that they contain or do not contain characters with ascenders and descenders. For this reason, using existing HTG models to obtain long pieces of text is cumbersome and cannot be satisfactorily done by simply stitching images of words. To overcome this limitation, some works have been specifically trained to handle text lines [95, 175][6] and paragraphs [243]. As an alternative applicable to output length-constrained HTG models (such as Diffusion Model-based approaches), the authors in [262] have proposed a specific algorithm to stitch and blend short images to obtain longer text.

AUTOREGRESSIVE HTG. Generative methods for natural images often assume a fixed number of output tokens, either because the image size is fixed or because it is one of the user inputs. On the other hand, for HTG, the length of the generated styled text image depends on the input style and content. This makes the application of autoregressive paradigms, which have shown success in natural image generation, non-trivial for styled HTG, since the model must include a mechanism to determine when generation should stop, in a way that is coherent with both the input style and the desired textual content. For these reasons, autoregressive approaches for styled HTG are still largely unexplored, especially compared to their adoption in the natural image domain.

HTG EVALUATION. As for the performance evaluation, models for HTG are evaluated by exploiting the Fréchet Inception Distance (FID) [147]. The FID is a commonly used score for evaluating the quality of generative models. It exploits image representations extracted from an Inception-v3, which are fit to two multivariate Gaussian distributions, one from real images and the other from generated images. The Kernel Inception Distance (KID) [44] is often reported alongside FID because it avoids the Gaussian approximation and provides an unbiased estimator that is typically more stable on smaller evaluation sets. For this reason, the FID tends to focus more on general image characteristics rather than the shape of handwriting. Furthermore, the backbone network used to compute the FID is trained on ImageNet, which contains natural images whose domain and aspect ratio are completely different from those of handwritten text images, which can result in misleading values. Other adopted metrics are the Geometric Score [187] and the Character Error Rate. The latter measures the readability of the generated text images, which serves as a proxy to express their realism, *i.e.* how similar to a well-formed text they look. However, as discussed in [178, 175, 278], these measures fail to capture all the desired characteristics of a well-generated styled text image.

DATASETS.

The evaluation of HTG models relies on a variety of datasets that differ in language, script, number of writers, image format, and annotation granularity.

In the following, we provide an overview of the datasets employed throughout this thesis for training, evaluating, and benchmarking HTG approaches. These datasets can be broadly grouped into multi-writer datasets, single-author historical datasets, synthetic pretraining datasets, and novel benchmarks specifically devised to assess the generalization capabilities of HTG models.

MULTI-WRITER DATASETS. Multi-writer datasets feature handwriting samples from a large pool of writers and are primarily used to evaluate the ability of Styled HTG models to imitate different calligraphic styles. The following multi-writer datasets are considered in this thesis.

The **IAM** Handwriting Database [239] is the most widely adopted benchmark for HTG. It consists of greyscale scans of English documents handwritten by 657 writers with ink on white paper, subsequently digitized and cleaned. The dataset is available at multiple annotation levels: forms, lines, and words. The line-level version contains 13,353 text line images, while the word-level version provides individual segmented word images. IAM is commonly used for both training and evaluation in HTG, with an average of approximately 20 samples per writer. Due to its widespread adoption, it serves as the primary benchmark for comparing HTG models in the literature.

The **CVL** Database [190] consists of word images obtained from high-resolution RGB scans of English and German manuscripts, written with ink on white paper by 310 writers. The dataset contains 13,473 samples with an average of approximately 43 word samples per writer. The availability of RGB images, as opposed to greyscale scans, makes CVL useful for evaluating the sensitivity of HTG scores and models to color information. In the context of this thesis, the line-level version of CVL is used for zero-shot cross-dataset evaluation.

The **RIMES** Database [22] consists of binarized French postal mail documents handwritten by 1,500 authors. The dataset comprises 12,111 samples at the line level, with an average of approximately 8 samples per writer. RIMES is particularly interesting for evaluating cross-lingual generalization of HTG models, since French text contains accented characters and diacritics that are not present in English handwriting datasets. The binary format of the images further contributes to assessing the robustness of HTG models and evaluation scores to dif-

ferent image preprocessing conditions.

The **KHATT** Database [237] contains binarized images of handwritten Arabic words by 838 writers, for a total of 11,427 samples with an average of approximately 14 samples per author. KHATT poses unique challenges for HTG evaluation, as Arabic is a right-to-left script with connected characters and context-dependent letter shapes. This dataset is primarily used to assess the sensitivity of evaluation scores, such as Handwriting Distance (HWD) and FID, to scripts and languages significantly different from the Latin alphabet on which most backbones are trained.

6

The **BanglaWriting** Dataset [253] is composed of greyscale images of Bengali words handwritten by 212 authors, totaling 17,265 samples with an average of approximately 81 samples per writer. Bengali uses a complex Abugida script with a large set of conjunct characters and diacritical marks, which makes this dataset valuable for evaluating the generalization of handwriting representation models to non-Latin scripts with different stroke patterns and spatial configurations.

The **NorHand** Dataset [236] features text lines extracted from greyscale scans of historical Norwegian documents, written with ink on yellowed paper by 12 authors. Despite the small number of writers, the dataset is relatively large, containing 21,939 text line images (approximately 1,828 per author). The historical nature of the documents, with aged paper backgrounds and variable ink quality, makes NorHand particularly useful for evaluating the robustness of evaluation scores to document-level visual variations that are unrelated to the handwriting style.

SINGLE-AUTHOR HISTORICAL DATASETS. Single-author datasets consist of handwritten documents from a single writer, typically drawn from historical correspondence or manuscripts. These datasets are used to evaluate the numerical stability and bias of HTG evaluation scores, as they provide large pools of images in a single, consistent handwriting style.

The **Saint Gall** Dataset [113] features binarized images of 1,410 lines from a medieval manuscript written in Latin with gothic calligraphy. The distinctive gothic script, with its angular and highly structured letterforms, makes this dataset an important test case for evaluating whether HTG scores can capture

the unique characteristics of non-cursive historical handwriting.

The **George Washington** Dataset [114] contains binarized images of 656 lines from English letters written by American President George Washington. The cursive style is relatively regular but features period-specific calligraphic conventions. This dataset is commonly used in document analysis research and serves as a reference for evaluating score stability on a moderate-sized single-author collection.

The **Rodrigo** Database [326] contains 20,357 lines extracted from greyscale scans of a historical manuscript written in Spanish with ink on ancient paper. It is one of the largest single-author datasets available for handwriting analysis. The substantial size and the specific calligraphic conventions of the featured handwriting make Rodrigo useful for studying the behavior of evaluation scores when a large number of same-style samples is available.

The **ICFHR14** Dataset [317] is a collection of 11,473 lines extracted from greyscale scans of pages written by the English philosopher Jeremy Bentham. The dataset was initially proposed for a handwriting recognition competition and features a cursive, somewhat irregular style with variable spacing and character sizes. Its large size and single-author nature make it suitable for stability analysis of evaluation metrics.

The **Leopardi** Dataset [58] is a collection of 2,459 lines from high-resolution RGB scans of letters by the Italian poet Giacomo Leopardi, written with ink on ancient paper. The RGB format preserves color information about the ink and paper aging, which is useful for assessing whether evaluation scores are influenced by visual appearance aspects unrelated to the calligraphic style.

The **LAM** Dataset [59] includes 25,823 text line images obtained from RGB scans of ancient Italian letters written by Lodovico Antonio Muratori. LAM is one of the largest single-author datasets of handwritten text images available. Thanks to its size, it is used as a primary reference for evaluating the numerical stability of HTG scores with respect to varying numbers of imaging samples, as it allows subsampling experiments that reveal score bias as a function of the dataset size used for computation.

SYNTHETIC TRAINING DATASETS. Training HTG models on purely synthetic

data enables zero-shot generalization to unseen real handwriting styles and overcomes the limitations of small, dataset-specific training. The following synthetic datasets are used for pretraining the models presented in this thesis.

Font² is a large-scale synthetic dataset of text images introduced in [277]. It is obtained by rendering 10,400 English words in 10,400 calligraphic fonts and superimposing them onto paper-like backgrounds, with random geometric distortions and color transformations applied to increase realism. The resulting dataset contains over 100 million samples and is used to pretrain the VGG16 backbone adopted for computing the HWD score. The high visual variability across fonts forces the network to learn features that represent the handwriting style independently of the overall visual appearance, including the textual content, the background texture, and the ink color.

6.3 CONTROLLABLE DIFFUSION FOR SPECIFIC APPLICATIONS

Building upon the diffusion models discussed in Section 6.1.2, this section examines specialized applications requiring additional control mechanisms, including panorama synthesis, document design generation, and transparent image generation.

PANORAMA AND LONG-FORM GENERATION. With most of the available diffusion models trained to generate images with limited aspect ratio ranges, being able to exploit such models to generate larger images, possibly at different, odd aspect ratios, without retraining has received increasing interest [23, 424, 30, 198, 142, 170]. One of the most commonly considered pretrained models in this context is Stable Diffusion [308], which is adapted at inference-time to generate long images by following either an iterative diffusion paradigm [23, 424] or a joint diffusion one [30, 198].

It is worth noting that large-scale pretrained models can generate long images to some extent, but these images lack variability, which hinders their realism. To overcome this limitation, the authors of [30] introduced the joint diffusion

paradigm with MultiDiffusion, an efficient inference-time adaptation strategy to generate long images with variability. This entails generating overlapped squared views and then recombining them in the final panorama image by averaging the noise at the overlap. However, MultiDiffusion can introduce artifacts and lead to perceptually incoherent outputs. To tackle this issue, the recently-proposed SyncDiffusion [198] entails providing guidance in the form of the gradient from a perceptual similarity loss [425] between the central view of the panorama and the other views. However, SyncDiffusion does not provide guidance for semantic coherence. Moreover, its perceptual coherence guidance is computationally expensive to obtain and requires schedulers that compute the foreseen denoised observation, like [337].

DOCUMENT AND DESIGN GENERATION. In the context of generating designs and visually-rich documents, training, fine-tuning, and inference-time adaptation approaches for diffusion models have been proposed. These focus on generating aesthetically-pleasing document layouts [203, 139, 68, 61, 161], layout-aware designs backgrounds [389], user interfaces [160], SVGs [164, 400], and collages [30, 424, 319].

TRANSPARENT IMAGE GENERATION. Several approaches exist to estimate the α channel of a given RGB image [440], which can be binary and indicate foreground and background only (as in segmentation strategies), or continuous between 0 and 1 to indicate the opacity of each pixel (as in from matting algorithms). Proposed approaches range from classical ones exploiting perceptual features of the image (such as color [37] and edges [313]) to more recent learning-based solutions [353, 405, 155, 295, 189, 415]. Note that those approaches require some additional input to guide the prediction of the α channel. For example, image matting approaches [416, 353, 405, 155] need a map of foreground, background, and unknown regions (a trimap), while even recent class-free segmentation approaches [229, 189, 415] need anchor points or visual prompts indicating the subject of interest.

In LayerDiffuse [422], the authors propose a fine-tuning strategy to enable transparent image generation using large-scale pre-trained diffusion models. In

particular, they adjust the latent space of Stable Diffusion XL [281] and finetune both the U-Net and the VAE, by using a loss that preserves the original latent distribution. This method, while effective, requires about 350 A100 hours for fine-tuning (as specified by the authors) and is not flexible to changes in the backbone model.

Taken together, these research threads motivate the structure of Part II. Chapter 7 addresses evaluation by proposing a style-sensitive score for handwritten text generation, Chapter 8 develops zero-shot autoregressive generation of styled text images, Chapter 9 studies inference-time diffusion control for aspect-ratio-flexible image synthesis, and Chapter 10 extends the same controllability perspective to transparency-aware RGBA asset generation. The common theme across these contributions is controllable visual generation under practical constraints, often without expensive retraining.

7

HWD: A Novel Evaluation Score for Styled Handwritten Text Generation

STYLED Handwritten Text Generation (HTG) is an important task in document analysis, aiming to generate text images with the handwriting of given reference images. In recent years, there has been significant progress in the development of deep learning models for tackling this task. Being able to measure the performance of HTG models via a meaningful and representative criterion is key for fostering the development of this research topic. However, despite the current adoption of scores for natural image generation evaluation, assessing the quality of generated handwriting remains challenging. In light of this, we devise the Handwriting Distance (HWD), tailored for HTG evaluation. In particular, it works in the feature space of a network specifically trained to extract handwriting style features from the variable-length input images and exploits a perceptual distance to compare the subtle geometric features of hand-

This Chapter is related to the publication “V. Pippi, F. Quattrini *et al.*, HWD: A Novel Evaluation Score for Styled Handwritten Text Generation, BMVC 2024” [1]. See the list of Publications on page 315 for more details.

writing. Through extensive experimental evaluation on different word-level and line-level datasets of handwritten text images, we demonstrate the suitability of the proposed HWD as a score for Styled HTG. The pretrained model used as backbone will be released to ease the adoption of the score, aiming to provide a valuable tool for evaluating HTG models and thus contributing to advancing this important research area. We release the code and the weights of the convolutional backbone used to compute the HWD score ^{*}.

7.1 INTRODUCTION

Styled HTG entails producing realistic images of arbitrary handwritten text in a desired style given in the form of one or more exemplar style images. Those images can be used to: train models for document analysis tasks (*e.g.* Handwritten Text Recognition [427, 40, 174, 82, 41, 43, 175, 57, 280]) in low-resource scenarios such as ancient languages or documents by specific authors; enhance the user experience in augmented reality scenarios and the public engagement at GLAM institutions (galleries, libraries, archives, and museums); assist physically impaired people in taking notes on electronic devices. Although HTG is receiving increased interest in recent years [116, 178, 42, 278], a precise standard for evaluation has not yet been defined. It is important to note that evaluating the similarity between two writers' calligraphy involves more than considering the overall appearance of the text images. This concerns color and texture of background and ink, stroke thickness, slant, and roundness. Nonetheless, handwriting is characterized by the shape of individual characters, ligatures, and the spacing between characters in the text. Hence, a thorough evaluation procedure should consider all these factors to ensure an accurate and meaningful assessment of the ability of Styled HTG models to imitate a desired handwriting.

The commonly adopted approach is the one proposed by Kang *et al.* [178], which entails exploiting the Fréchet Inception Distance (FID) [147]. The FID is computed on the features extracted by an Inception-v3 ConvNet trained on natural images from ImageNet. Thus, the FID is somehow unsatisfactory for mea-

^{*}<https://github.com/aimagelab/HWD>

asuring the faithfulness in the handwriting style of the generated images but rather captures the overall appearance [178, 175, 278]. Another critical point of applying the FID in the text images domain is that it leverages a backbone trained on images whose aspect ratio is very different from text images. These latter are usually wider than high, while natural images in ImageNet are roughly squared. For this reason, in HTG evaluation, the FID is commonly computed on the beginning part of the text image, discarding the rest. This approach offers invariance with respect to the textual content but is prone to miss artifacts and dissimilarities that appear in the center or at the rightmost part of the image. Therefore, while the FID can help evaluate certain aspects of HTG, it does not provide a complete picture of the quality of the generated handwriting images. Moreover, different studies [44, 76] have shown that the FID is biased on the number of samples used for its calculation, resulting in lower values with more instances and higher and more unstable values when reducing the number of examples.

To address these challenges, we propose a new evaluation score called HWD. The main characteristics of HWD consist of: (1) the use of robust style features extracted by a backbone trained on a large dataset of synthetic text images; (2) the application of a perceptual distance for style comparison; (3) the ability to handle variable-length text images; (4) the numerical stability even when computed on a limited number of samples. This is particularly useful in Styled HTG, where only a few real images per author are usually available. To assess the suitability of the proposed score for the Styled HTG task, we examine the values obtained when comparing sets of text images in the same style with respect to those obtained when comparing images in different styles. We demonstrate that the HWD effectively captures differences in the handwriting and is numerically stable. This makes it more suitable than the FID in expressing the performance of the Styled HTG models. Overall, the HWD could contribute to the field of HTG by providing a tailored and practical evaluation score to measure the realism and faithfulness of generated text images. We release the code and the weights of the convolutional backbone used to compute the HWD score [†].

[†]<https://github.com/aimagelab/HWD>

7.2 HANDWRITING DISTANCE

Inspired by the strategy adopted for natural image generation evaluation, Styled HTG works employ the FID score, adapted as described in [178] and depicted in Figure 7.1 (bottom). In this work, we devise an alternative score for evaluating the performance of Styled HTG models, called HWD. The main characteristics of HWD are the domain-aware image representation strategy and the use of a perceptual distance instead of a distribution distance. The pipeline for computing our proposed score is described below and depicted in Figure 7.1 (top).

7

7.2.1 TEXT IMAGES REPRESENTATION

When evaluating Styled HTG, we consider the set of real images $\mathbf{X}_m = \{\mathbf{x}_{m,i}\}_{i=0}^N$, where N is the number of samples for writer m , and the set $\mathbf{X}'_m = \{\mathbf{x}'_{m,i}\}_{i=0}^{N'}$ of generated images in the style of writer m . In general, the number of generated images, N' , can differ from N .

DOMAIN-SPECIFIC FEATURE EXTRACTION. Given the constrained domain of the images for the HTG task, we propose to use a domain-specific backbone as feature extractor. In particular, we adopt a VGG16 pretrained on Font², a large synthetic dataset of text images built according to the procedure presented in [277]. We choose VGG16 as a backbone for its superiority in extracting meaningful style representations compared to deeper networks featuring skip connections [385]. The pretraining dataset contains more than 100M samples obtained by rendering 10400 English words in 10400 calligraphic fonts and superimposing them to paper-like backgrounds. Random geometric distortions and color transformations are also performed to increase the realism of the images. We train the VGG16 backbone to classify the images according to their calligraphic font. The high visual variability of the datasets forces the network to learn features that represent the handwriting given by the font and disregard the overall visual appearance. In this way, the features extracted by our adopted backbone are strong representations of the handwriting style. For computing the HWD, we feed the pretrained VGG16 with the whole image resized to a height of 32 pixels while

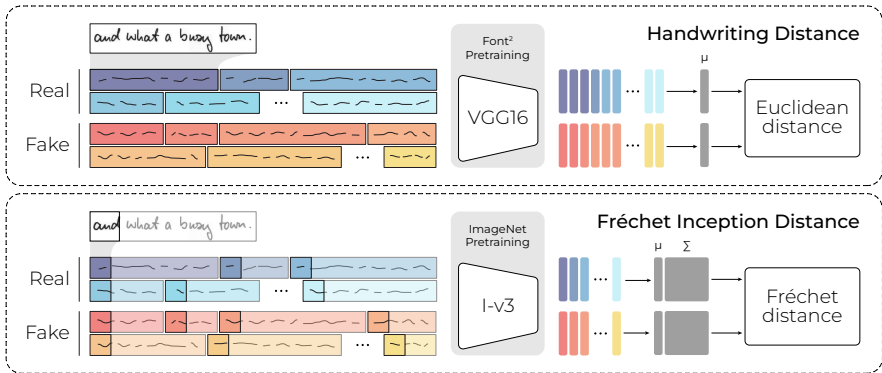


Figure 7.1: Pipelines for the computation of the proposed HWD score (top) and the commonly used FID score (bottom) from a set of real and generated text images with varying widths. Top: the samples are processed by a VGG16 pretrained on a synthetic dataset of text images. The different widths of the inputs result in a different number of feature vectors. The value of HWD is the Euclidean distance between the feature vectors’ means μ . Bottom: the text images are cropped to a square and fed to an Inception-v3 pretrained on ImageNet, obtaining one feature vector per image. The FID is the Fréchet distance between the distributions of the features, represented as multivariate Gaussians with mean μ and covariance matrix Σ .

keeping the aspect ratio. Note that since the images have different widths, we feed them to the network one by one to avoid using padding.

VARIABLE-LENGTH IMAGES REPRESENTATION. To handle the difference in the aspect ratio of the natural images in ImageNet and text images, the strategy adopted in [178] and the following HTG approaches entails feeding to the Inception-v3 backbone only a squared, truncated sample. The text image is reshaped to a height of 32 pixels, keeping the aspect ratio, and then cropped at the beginning to have a fixed width (32 pixels). This strategy has two drawbacks. First, it disregards a large part of the text image. Second, the characters appearing more often at the beginning of words due to language regularity assume more importance than others in computing the score. Consequently, this strategy prevents evaluating the ability of the HTG models to generate images of variable-length texts with the same quality throughout the entire image. To overcome these limitations, we propose processing the whole images with the VGG16 backbone for the HWD. The representations are obtained from the feature maps of the last convolutional block. Such feature maps have shape $1 \times W \times 512$, where

W depends on the input text image width. From those maps, we obtain a set of 512-sized feature vectors that represent the images. As a result, wide images are represented by a larger set of vectors than short ones. When computing the HWD for writer m , all the vectors from all the real images sets are gathered and then averaged (the same applies to the generated images). In the gathering set, there will be more vectors coming from the longer words, and thus, the longer words will have a bigger impact on the mean of the gathering set. In summary, each image for writer m , both real and generated, is represented by a set of vectors. The real images are represented by sets as $\mathbf{y}_{m,i} = \{f(\mathbf{x}_{m,i})_j\}_{j=0}^{W_i}$, and the generated ones by $\mathbf{y}'_{m,i} = \{f(\mathbf{x}'_{m,i})_j\}_{j=0}^{W'_i}$, where $f(\cdot)$ denotes the feature extraction, and W_i and W'_i are the number of vectors extracted from the i -th real and generated images, respectively.

7.2.2 PERCEPTION-AWARE FEATURE DISTANCE

The main idea behind the distribution distance-based evaluation scores, such as the FID, is to evaluate the performance of a generative model by its ability to generate images that match the distribution of the real ones. Our domain of interest is more constrained than the generation of natural images. In fact, Styled HTG entails considering the handwriting, expressed by subtle geometric features other than macroscopic texture. In light of this, we argue that a score capturing the perceptual aspects is more suitable than one based on the distance between feature distributions. Therefore, we employ the Euclidean distance between the averaged feature vectors of the real and generated images in the style of the same writer:

$$Y_m = \frac{\sum_{i=1}^N \sum_{j=1}^{W_i} f(\mathbf{x}_{m,i})_j}{\sum_{i=1}^N W_i} \quad \text{and} \quad Y'_m = \frac{\sum_{i=1}^{N'} \sum_{j=1}^{W'_i} f(\mathbf{x}'_{m,i})_j}{\sum_{i=1}^{N'} W'_i}.$$

For the images in the style of writer m , the HWD is given by:

$$\text{HWD}_m = \|Y_m - Y'_m\|_2.$$

Note that when computed on robust image representations, *e.g.* obtained from a backbone trained on a semantic prediction task, the Euclidean distance is highly predictive of the perceptual similarity between images [425]. Finally, the HWD on datasets containing images in the style of M different authors is obtained as

$$\text{HWD} = \frac{1}{M} \sum_{m=1}^M \text{HWD}_m.$$

The HWD score has the non-negativity, symmetry, triangular inequality properties, but it is not guaranteed that it exhibits the identity of indiscernible elements property since two non-identical images might have $\text{HWD} = 0$ due to their representation via the pre-trained backbone. This is a desirable characteristic for the HTG task, in which text images containing different texts and different backgrounds should be at low (or even zero) HWD if they are written in the same handwriting style.

7.3 EXPERIMENTAL ANALYSIS

For our experimental analysis of the proposed HWD score, we consider images from a number of multi-author and single-author datasets. We compare the HWD score against the FID score in the variant proposed in [178], which is the common approach adopted in Styled HTG. The comparison is performed along two dimensions. First, we assess its capability to recognize corresponding handwriting styles, quantifying the style verification capability with the Overlap coefficient and the Equal Error Rate (EER). Second, we compare the numerical stability of the proposed HWD score and the FID. Additionally, we evaluate current Styled HTG State-of-the-Art models with our score and other common metrics for image generation evaluation. Finally, we conduct extensive ablation analyses on the main components of our approach to investigate their individual contributions. Further results can be found in Chapter A.

7.3.1 CONSIDERED DATASETS

The considered multi-author datasets are described below and in Table 7.1.

IAM. The IAM Database [239] is a collection of greyscale document scans by 657 writers. These are written in English with ink on white paper and cleaned digitally. Here, we use the line-level version of the dataset.

RIMES. The RIMES Database [22] consists of binary French documents by 1500 authors. For this dataset, we use the line-level version.

CVL. The CVL Database [190] features word images obtained from RGB scans of English and German manuscripts, written with ink on white paper by 310 writers.

KHATT. The KHATT Database [237] contains binarized images of handwritten Arabic words handwritten by 838 people.

BANGLAWRITING. The BanglaWriting Dataset [253] is composed of greyscale images of Bengalese words handwritten by 212 authors.

NORHAND. The NorHand Dataset [236] features text lines extracted from greyscale scans of ancient documents written with ink on yellowed paper by 12 Norwegian authors.

The considered single-author datasets are presented below.

SAINT GALL. The Saint Gall Dataset [113] features binary images of 1410 lines from a medieval manuscript written in Latin with gothic calligraphy.

WASHINGTON. The George Washington Dataset [114] contains binary images of 656 lines from English letters written by American President George Washington.

RODRIGO. The Rodrigo Database [326] contains 20357 lines extracted from greyscale scans of an historical manuscript, written in Spanish with ink on ancient paper.

ICFHR14. The ICFHR14 Dataset [317] is a collection of 11473 lines extracted from greyscale scans of ancient pages, written by the English philosopher Jeremy Bentham.

LEOPARDI. The Leopardi Dataset [58] is a collection of 2459 lines from RGB scans of letters by the Italian poet Giacomo Leopardi, written with ink on ancient paper.

LAM. The LAM Dataset [59] includes 25823 text lines images obtained from RGB scans of ancient letters in Italian, written by Lodovico Antonio Muratori.

7.3.2 SENSITIVITY TO THE HANDWRITING

We evaluate the sensitivity of HWD to the handwriting style by splitting the multi-author datasets. We consider half of the samples for each featured writer as references and the other half as if they were the output of a perfect Styled HTG model. Then, we compare the distributions of the HWD and the FID computed on text images of multiple matching and non-matching authors pairs. Note that, in such an ideal case for Styled HTG, both the HWD and the FID should be as close as possible to their best value. Therefore, the more these two distributions are separated, the better the corresponding score captures the handwriting similarity between the considered images.

The obtained distributions are reported in Figure 7.2. We observe that the histograms for the FID show significant overlap and that there is no clear separation between the distributions of matching and non-matching authors pairs. Moreover, except for CVL, which has a higher number of samples per author compared to the other three datasets, the FID values of the corresponding authors' distribution are roughly above 100. Such high FID values highlight the bias of the score when computed on a few samples. On the other hand, the histograms for the HWD are more separated.

In addition, we quantitatively evaluate the style recognition capability by comparing the distributions in Figure 7.2 in terms of the Overlap coefficient, a statistical index that quantifies the overlap between distributions, and the EER, which represents the point where the False Acceptance Rate equals the False Rejection Rate. The results are reported in Table 7.1. As we can see, for both the FID and HWD, the Overlap and ERR are lower on the IAM and CVL datasets than in KHATT and RIMES. We argue the causes to be, respectively, the languages of

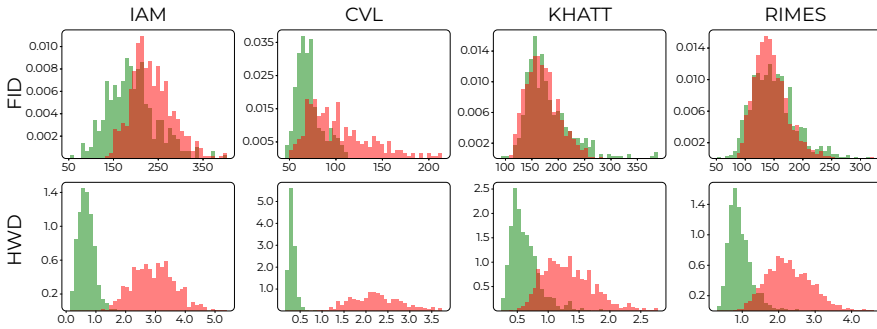


Figure 7.2: Distributions of different scores used to evaluate HTG models when applied on same-author (green) or different-author (red) subsets. The overlap area is in dark red.

	Language	Samples	Authors	Avg. Samples per author	FID		HWD	
					Overlap	EER	Overlap	EER
Norhand	Norwegian	21939	12	1828.25	4.2	4.2	0.0	0.0
BanglaWriting	Bengali	17265	212	81.44	11.6	5.6	6.1	2.9
CVL	English/German	13473	310	43.46	24.7	12.5	0.0	0.0
IAM	English	13353	657	20.32	27.1	13.6	0.7	0.3
KHATT	Arabic	11427	838	13.64	40.3	21.6	12.0	5.9
RIMES	French	12111	1500	8.07	39.1	20.8	7.0	3.3

Table 7.1: Overlap and EER of the FID and HWD values calculated on the images in the considered multi-author datasets.

the datasets and the similar overall appearance of the samples from the same author. In fact, the FID mainly focuses on the latter, while HWD performs feature extraction with a VGG16 pretrained on Font², which contains words in the same language as IAM and CVL. Nonetheless, the HWD score achieves lower Overlap and EER values than the FID in all datasets, including those in languages different from English. We report additional results in Appendix A.1.

7.3.3 SENSITIVITY TO THE NUMBER OF SAMPLES

Numerical stability is an important factor to consider when assessing a score. As argued by [76, 44], the FID exhibits a strong bias towards the number of samples. To assess the stability of HWD, we consider the large single-author LAM dataset and compute the values of the HWD and FID on images from LAM against variably sized subsets of images from ICFHR14, Saint Gall, Leopardi, Rodrigo,

Washington, and LAM itself. We determine the mean of the scores over multiple runs and also consider the range between the 25th and 75th percentiles of values. The results are reported in Figure 7.3.

It can be observed that the FID (left-most plot) shows a significant bias towards the number of images: all the curves start with very high values and decrease slowly until at least around 2000 samples are used to obtain the score. On the other hand, the HWD (right-most plot) is more stable with respect to the subset sizes, reaching a plateau even when computed on around 100 samples. To further investigate the cause of this behavior, we compute the Euclidean distance on the Inception-v3 features (FID w/ Euclidean) and the Fréchet distance on the VGG16 features (HWD w/ Fréchet). In both cases (center-left plot and center-right plot of Figure 7.3, respectively), we can observe a slight bias with respect to the number of samples used for the computation. Moreover, the computation of these scores necessitates roughly one more order of magnitude of images to reach a plateau, compared to the case of the FID and the HWD, respectively. For the FID w/ Euclidean, we argue that this is because its Inception-v3 backbone is applied to the beginning part of the text images. As a result, more samples are needed to effectively represent the author’s handwriting. In the case of the HWD w/ Fréchet, we argue that this score suffers from the numerical instability of the Fréchet Distance, which fits the image representations to multivariate Gaussians.

Further, by comparing the values in the plots in the sight of the exemplar images reported in Figure 7.3, we can make some considerations on the convolutional backbones used as feature extractors. Inception-v3-based scores assign dataset distances according to their overall appearance. For instance, ICFHR14 and Rodrigo (both containing greyscale images) are close to each other, and the same happens for Saint Gall and Washington (both containing binarized images). On the other hand, the VGG16-based scores focus more on the handwriting. Thus, Saint Gall is isolated from the others because of its peculiar gothic calligraphy, while ICFHR14 is closer to Washington and Leopardi than to Rodrigo, reflecting the similarity in the cursive calligraphies. This characteristic is better suited for the Styled HTG task, as we are interested in evaluating a model’s ability to mimic the handwriting style. We report additional results in Appendix A.4.

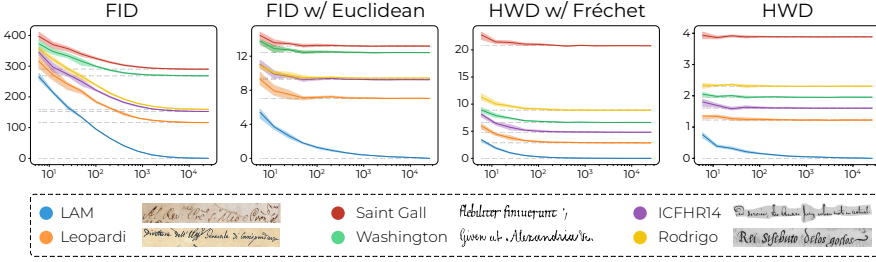


Figure 7.3: Left to right: comparison between FID, a version of the FID exploiting the Euclidean distance (FID w/ Euclidean), a version of HWD exploiting the Fréchet Distance (HWD w/ Fréchet), and HWD with varying number of samples. Lines denote the mean; transparent bands represent the range between the 25th and 75th percentiles, obtained with 10 calculation runs.

7.3.4 STYLED HTG MODELS EVALUATION

For reference, we report the performance of State-of-the-Art Styled HTG models trained on the IAM dataset. We consider the conditional-GAN-based HIGAN+ approach [121], which uses a single image to condition the generation and disentangles text and style, and the few-shot Convolutional-Transformer based models HWT [42] and VATr [278], which capture both global and character-level style-text dependencies by exploiting cross-attention. The latter approach represents the textual input with visual archetypes (*i.e.* Unifont-rendered characters) instead of one-hot vectors, as done in the other two. In Table 7.2, we evaluate the aforementioned models in terms of the proposed HWD and other scores for image generation evaluation. In particular, we consider the FID score; the Geometric Score (GS) [187], which compares the data and model manifold estimates; and the Kernel Inception Distance (KID) [44], which relaxes the Gaussian assumption of the FID and uses the Maximum Mean Discrepancy to compare the distributions of the features.

7.3.5 ABLATION ANALYSIS

Finally, we analyze the effects of the four main components of our proposed score: the backbone, the pretraining dataset, the input image portion, and the distance measure. We perform the ablation analysis on the IAM dataset and report the results in Table 7.3. Looking at the results, we notice that the backbone used

	FID	KID	GS	HWD
HWT	23.36	1.37×10^{-2}	$\mathbf{1.05} \times 10^{-2}$	1.928
HIGAN+	18.21	9.38×10^{-3}	2.15×10^{-2}	1.237
VATr	18.80	$\mathbf{7.06} \times 10^{-3}$	2.19×10^{-2}	0.828

Table 7.2: Scores of the considered HTG models when generating the same or different words as those in the style reference images of the IAM test set. Best performance in bold.

to extract the features plays a crucial role in the separability scores, as the use of VGG16 leads to very good results also when pretrained on ImageNet. A second important aspect is the feature distance score used. The Fréchet distance on the VGG16-extracted features achieves good results on all settings. Nevertheless, it is not influenced by the backbone pretraining and the image portion used to extract the features. On the other hand, the Euclidean distance fully exploits the input information (*i.e.* the quantity and the type of feature vectors) and thus is the best in the HWD setting.

7.3.6 SENSITIVITY TO THE VISUAL APPEARANCE

To assess the sensitivity to handwriting-related visual aspects, we compare the FID and HWD between reference images and increasingly altered ones, taken from the LAM dataset. In particular, the considered alterations entail shear, erosion, and dilation to simulate handwriting slant and strokes thickness. The results are reported in Figure 7.4 and show that HWD is more sensitive than the FID to such visual aspects and that it increases linearly with the alteration intensity, thus enforcing its suitability for evaluating HTG.

7.3.7 COMPUTATION TIME COMPARISON

We consider the computation time, consisting of image representation and distance computation, of FID and HWD on the same hardware and data. Computing the FID score on 25823/25823 real/fake images from the LAM dataset takes 426.12s + 9.03s (image representation + distance computation), while the computation time of HWD is 135.50s + 0.01s.

Backbone	Pretraining Dataset	Image Portion	Distance	IAM	
				Overlap	EER
Inception-v3	ImageNet	Beginning	Fréchet	27.1	13.6
Inception-v3	ImageNet	Beginning	Euclidean	29.6	14.5
Inception-v3	ImageNet	Whole	Fréchet	24.0	11.6
Inception-v3	ImageNet	Whole	Euclidean	8.5	3.9
Inception-v3	Font ²	Beginning	Fréchet	18.8	9.3
Inception-v3	Font ²	Beginning	Euclidean	11.3	4.8
Inception-v3	Font ²	Whole	Fréchet	19.0	9.1
Inception-v3	Font ²	Whole	Euclidean	7.2	3.3
VGG16	ImageNet	Beginning	Fréchet	3.2	1.6
VGG16	ImageNet	Beginning	Euclidean	26.2	13.0
VGG16	ImageNet	Whole	Fréchet	2.8	1.2
VGG16	ImageNet	Whole	Euclidean	6.2	2.9
VGG16	Font ²	Beginning	Fréchet	3.4	1.7
VGG16	Font ²	Beginning	Euclidean	16.5	8.2
VGG16	Font ²	Whole	Fréchet	3.5	1.6
VGG16	Font ²	Whole	Euclidean	0.7	0.3

Table 7.3: Ablation analysis of the main components of the HWD score. Note that the first row is the FID while the last is the complete HWD. Best performance in bold.

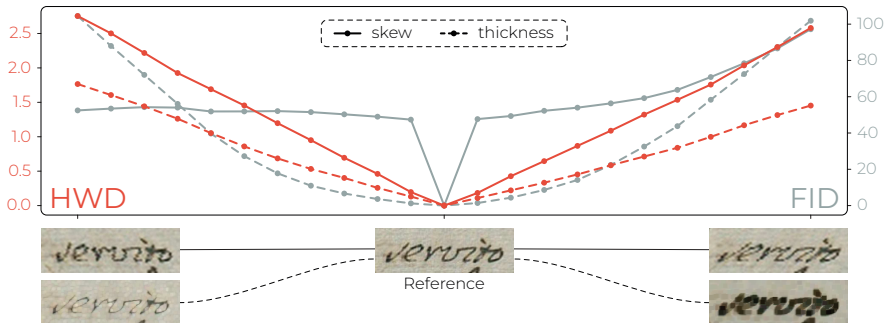


Figure 7.4: FID and HWD by varying handwriting thickness and slant.

7.4 CONCLUSION

In this work, we have proposed HWD, a score specifically designed to evaluate Styled HTG. HWD exploits the features extracted by a convolutional backbone trained on a large synthetic dataset of handwritten text to compare the perceptual differences between handwritings. Moreover, it is designed to work with images of variable lengths, such as those containing text. The results obtained from extensive experimental analysis demonstrate its suitability for evaluating text image generation approaches, its sensitivity to different styles, and its numerical stability. Hopefully, the use of the proposed score, whose implementation and backbone model weights will be made publicly available, will contribute to pushing forward the research on the Styled HTG task. Additional experiments on Nor-Hand and BanglaWriting reported in Chapter A further indicate that HWD retains its discriminative behavior on scripts and document conditions that differ substantially from the English-centric core benchmarks, suggesting broader applicability beyond Latin handwriting alone.

Within this thesis, the score is later reused in Chapter 12 to assess text-rendering consistency in multi-instance infographic editing.



Autoregressive Styled Text Image Generation

STYLED HTG has recently received attention from the Computer Vision and Document Analysis communities, which have developed several solutions, either GAN- or diffusion-based, that achieved promising results. Nonetheless, these strategies fail to generalize to novel styles and have technical constraints, particularly in terms of maximum output length and training efficiency. In this chapter, we present an autoregressive paradigm for text image generation, addressing these limitations in two stages. First, we introduce Emuru, a novel framework that leverages a powerful text image representation model (a Variational Autoencoder) combined with an autoregressive Transformer, trained solely on a diverse, synthetic dataset of English text rendered in over 100,000 type-written and calligraphy fonts, enabling zero-shot generalization to unseen styles.

This Chapter is related to the publications “F. Quattrini*, V. Pippi* *et al.*, Zero-Shot Styled Text Image Generation, but Make It Autoregressive, CVPR 2025” [6] and “C. Zaccagnino, F. Quattrini *et al.*, Autoregressive Styled Text Image Generation, but Make it Reliable, WACV 2026” [10]. * Equal contribution. See the list of Publications on page 315 for more details.

Then, we present Eruku, which improves upon the autoregressive formulation by freeing the model from requiring the transcription of the style image, introducing an explicit stopping mechanism via a dedicated end-of-generation token, and enforcing text adherence through a Classifier-Free Guidance (CFG)-inspired strategy, resulting in more reliable and faithful generation.

8.1 ZERO-SHOT STYLED TEXT IMAGE GENERATION WITH EMURU

The Styled HTG task entails generating images containing user-specified text rendered in a handwriting style resembling one or more reference style samples. In recent years, this task has regained significant attention from the Document Analysis and Computer Vision communities, especially due to its potential applications, ranging from generating synthetic data for other Document Analysis tasks to reproducing written styles to be used in graphic design and assistive technologies [133, 42, 279]. From a technical point of view, handwriting can be treated as a sequence of strokes, as in the Online HTG paradigm, or as a static image, as in the Offline HTG paradigm. Note that Online HTG entails giving the model a style input in the form of strokes trajectories, and letting the model generate strokes trajectories to be rendered on an image. Therefore, Online HTG models are implemented as sequential models [128, 15, 14, 91, 231, 305]. Note that collecting data to train and use Online HTG models involves using digital pens and pads, which is cumbersome and costly and cannot be applied in scenarios where the style to reproduce is from an existing manuscript, *e.g.* historical documents. For these reasons, Offline HTG, where the style input and the generated output are both static images, has gained more popularity and is also the focus of this work.

Despite the increasing research efforts in Offline HTG, generating synthetic images of handwritten text in a writer-specific style is still challenging. Specifically, current State-of-the-Art models: A. Struggle to generalize to handwriting styles that are significantly different from those encountered in training; B. Are implemented by making some dataset-specific assumptions due to technical lim-



Figure 8.1: Our proposed Emuru model can generate images of any-length text lines mimicking out-of-distribution handwriting styles.

iterations of the architecture adopted, *e.g.* normalizing character width or enforcing a maximum text line length; C. Cannot disentangle properly the handwriting style (*i.e.* the shape of the strokes) from the reference image background, and thus, generate undesired artifacts in the output images. To overcome these limitations, in this work, we take a step towards generalizable few-shot styled text image generation and propose Emuru (Fig. 8.1): a continuous-token autoregressive image generation model that can generate text images of any length (B.) with high style fidelity and readability, while increasing generalization to novel styles (A.) and reducing background artifacts (C.). To the best of our knowledge, Emuru is the first autoregressive Offline styled HTG model proposed in the literature.

Emuru consists of a VAE that maps the style sample images in an embedding space where they become a varying-length sequence of continuous vectors, and an autoregressive Transformer Encoder-Decoder that takes as input the desired style embeddings sequence, the text contained in the style image, and the desired text, and iteratively generates an image containing the desired text in the desired style. The generation is implemented as an autoregressive loop that out-

puts visual embeddings, which are then fed to the VAE Decoder to obtain the final output image. Given the autoregressive nature of the model, we train it to automatically establish when to stop the generation, thus removing any technical constraint on the maximum output length obtainable. Both the VAE and the autoregressive Transformer in Emuru are trained (separately, in two stages) on a specifically-designed, purely synthetic dataset. The images in the synthetic dataset are obtained by rendering a large amount of different English text lines (whose characters frequency is designed to be almost uniform) on different background images and with over 100k different fonts, both handwritten (calligraphy) and typewritten. Note that by working with a synthetic dataset, we have all the necessary information for training the VAE to reconstruct the text line image without the background. This way, the learned embeddings encode only the writing style, which gives Emuru the capacity to generate images without background artifacts. Moreover, training style variety and textual character frequency uniformity make Emuru able to generalize better to unseen styles, both typewritten and handwritten, and languages (*i.e.* characters arrangements).

To validate our proposed approach, we perform an extensive experimental analysis that considers multiple real datasets of handwritten text in multiple Latin alphabet languages. Moreover, we introduce a novel dataset of typewritten and handwritten text images to foster the development of HTG models able to generalize also to typewritten styles. The results of our analysis demonstrate the efficacy of the proposed architectural and training solutions for generating single words and entire lines with increased generalization capability to new styles, both handwritten and typewritten with respect to the State-of-the-Art. The code and weights for Emuru are publicly available*.

8.1.1 METHOD

Our framework for text-image generation consists of a powerful text-image representation model (a VAE) combined with an autoregressive Transformer Encoder-Decoder to generate text images conditioned on the desired textual content and

*<https://huggingface.co/blowing-up-groundhogs/emuru>

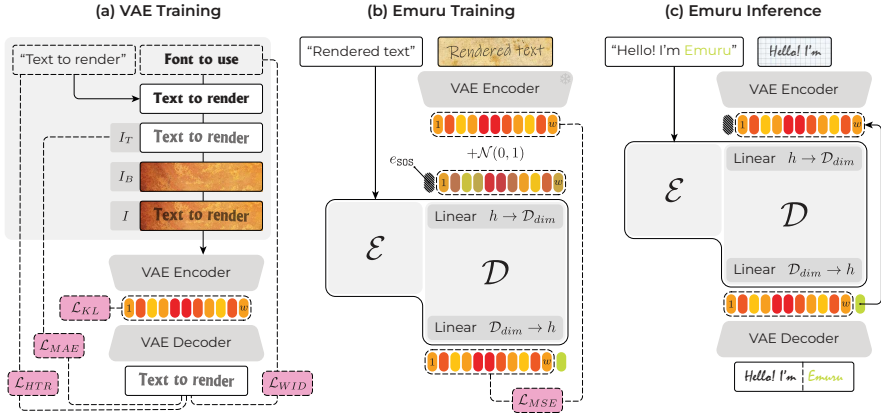


Figure 8.2: Our Emuru approach consists of a VAE (a) and an autoregressive Transformer Encoder-Decoder (b), both trained on a massive synthetic dataset of texts rendered in different fonts. At inference time (c), Emuru is given a reference style image, the text in the reference style image, and the desired text and is tasked to iteratively generate the output styled image, autonomously deciding when to stop.

a style example, both of which can be of any length. Specifically, the VAE maps the styled text images into a dense latent space and reconstructs them without background. The Transformer autoregressively generates embeddings compatible with those in the VAE latent space, conditioned on the embeddings of the reference style image and the desired text. Then, the VAE Decoder can be used to obtain the final image. A schematic representation of our approach is depicted in Fig. 8.2.

TRAINING DATA

State-of-the-Art HTG models are usually trained on a single dataset at a time, which is quite limiting due to the small variety of featured styles and text in the training images. This causes many HTG models to have a performance drop when generating images that are out-of-distribution for the style and the textual content, *e.g.* from a different dataset. In light of these considerations and inspired by the effectiveness of the synthetic pretraining of the style Encoder in [278, 277], we develop a massive synthetic dataset, which we use to train our HTG model. The dataset features RGB images containing text lines written over different back-

grounds. As for the text content, we collect words from multiple English corpora available on the NLTK library[†] and combine 1 to 32 of them into text lines T , making sure that both rare and common words are both well represented. Moreover, we collect over 100k fonts available online and render the text lines using these fonts on greyscale images, where the background is always white. Finally, we apply several random geometric transformations to the images (see Chapter B for details). This way, we obtain the text image I_T for each sample in the synthetic dataset. Then, we superimpose I_T to an RGB image I_B depicting a plausible background for text (*e.g.* paper), randomly selected among a pre-defined set. This way, we obtain the synthetic sample images I with shape $W \times H$ where $H = 64$ and W scales proportionally to the length of the text. In total, we obtain 2.2 million text images. Further details about the dataset are given in Chapter B. Note that we solely use this synthetic dataset for training both the components of our proposed approach: first, the VAE and, in a later stage, the autoregressive Transformer, keeping the VAE frozen.

VAE

First, we train a β -VAE to encode the text images into a compressed latent space. We choose a convolutional architecture similar to those commonly used in Stable Diffusion Models [308]. The VAE Encoder takes as input the images $I^{3 \times W \times H}$ and maps them in a $c \times b \times w$ embedding tensor. Due to the nature of lines of text, *i.e.* arbitrary length but mostly fixed height, we model the images as a w -long sequence of $b \times c$ vectors. Intuitively, each of the w vectors encodes a vertical slice of the text line. The Decoder is fed with these vectors and is tasked to reconstruct the grayscale text image I_T .

We train the VAE by using a Reconstruction loss expressed in terms of L_1 distance between the ground truth I_T image (*i.e.* only the text without background) and the reconstructed output ($\mathcal{L}_{MAE}$) and the Kullback–Leibler divergence loss ($\mathcal{L}_{KL}$). In line with [308, 372], we weigh this latter loss term with $\beta \ll 1$ in order to prioritize reconstruction quality rather than constraining a strongly

[†]<https://www.nltk.org/>

regularized latent space [148]. Additionally, we incorporate supervision signals (namely, a Cross-Entropy loss $\mathcal{L}_{WID}$ and a CTC loss $\mathcal{L}_{HTR}$) from two auxiliary networks. The first is a writing style classification network, pretrained on our synthetic dataset to identify the 100k styles from the reconstructed image, which enforces style fidelity. The second is a Handwritten Text Recognition (HTR) model, also pretrained on the synthetic dataset, tasked to recognize the text T in the reconstructed image to enforce readability.

AUTOREGRESSIVE TRANSFORMER

Once we obtained a strong latent image representation with the VAE, we train a generative model by using an autoregressive Transformer featuring a Transformer Encoder $\mathcal{E}$ and a Transformer Decoder $\mathcal{D}$, architecturally similar to T5 [299]. Both are trained on samples from the devised synthetic dataset. Specifically, at training time, $\mathcal{E}$ takes as input a sequence of characters in the desired content string (T_{out}) tokenized by using the single-character tokenizer. Then, self-attention is computed among these tokens to obtain a sequence of text embeddings $\mathbf{s} = [s_1, \dots, s_k]$. On the other hand, the input to the Transformer Decoder $\mathcal{D}$ is obtained as follows. First, we feed the pretrained, frozen VAE Encoder with the style image I to obtain a w -long sequence of b -dimensional embeddings, $\mathbf{v} = [v_1, \dots, v_w]$. To reduce the risk of exposure bias at inference time, these embeddings are summed to some noise sampled from a Normal distribution. Then, the resulting noisy embeddings are linearly projected into $\mathcal{D}_{dim}$ -dimensional vectors, to which we prepend a special start-of-sequence (SOS) learned embedding. The resulting $\mathbf{e} = [e_{SOS}, e_1, \dots, e_w]$ vectors are then fed to $\mathcal{D}$, which autoregressively predicts a sequence of vectors $\mathbf{e}' = [e'_1, \dots, e'_w]$. To this end, causal masked self-attention is performed on the $\mathbf{e}$'s, and unmasked cross-attention is performed between the $\mathbf{e}$'s and the $\mathbf{s}$'s from $\mathcal{E}$. The so-obtained $\mathbf{e}'$ sequence is projected back into the $\mathbf{v}'$ sequence of b -dimensional vectors that we use to compute the training loss.

We train the Transformer following a teacher-forcing strategy and using the MSE loss between the predicted $\mathbf{v}'$ and ground truth $\mathbf{v}$ embeddings ($\mathcal{L}_{MSE}$). Moreover, to make Emuru handle sequences of different lengths, we take inspi-

ration from the curriculum learning strategy adopted in [175]. The authors proposed to train their HTG model in three stages, with non-overlapping sets of increasingly longer images. Here, we train Emuru in two stages and increase the variability of the training images maximum length. Specifically, we first train on text lines containing from 4 to 7 words, then we fine-tune on lines containing from 1 to 32 words. Note that using only a single loss term in the form of the MSE makes the training process more controllable, thanks to the limited number of hyperparameters involved, and less resource-demanding compared to adversarial and diffusion process-based approaches that also rely on auxiliary networks to obtain regularization signals.

8

INFERENCE-TIME BEHAVIOR

At inference time, we feed Emuru with a reference style image containing a text line (I_{style}) and the concatenation of two strings: one with the text in the reference style image (T_{style}), and one with the desired output text (T_{out}). Emuru is tasked to continue the generation from I_{style} by following the desired text T_{out} , in the style of the reference image. Specifically, at each generation step t , $\mathcal{D}$ predicts the $(w + 1)$ -th embedding (e_{w+1}) based on the embeddings of I_{style} and the t previously-generated embeddings. Note that all training images used for our model have some padding in the form of white spaces since it is required for batched execution. Therefore, a training sample for Emuru is consistently modeled as a sequence of embeddings relative to characters followed by embeddings relative to white padding. Looking at the VAE latent space, it can be observed that the latter are quite distinguishable from the text characters embeddings (see Fig. 8.3). Thus, we stop the iterative generation of embeddings when Emuru emits P consecutive padding embeddings, which act as an implicit end-of-sequence signal. P is set to 10 after an empirical analysis to minimize the risk of early stopping the generation. At the end of the iterative generation, the new sequence of the generated b -dimensional vectors (excluding the last P padding vectors) is fed to the pretrained, frozen VAE Decoder to obtain the final image.

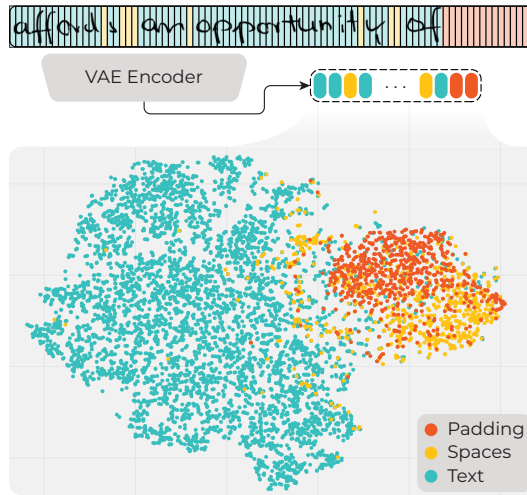


Figure 8.3: Characters and padding VAE embeddings distribution obtained via t-SNE.

8.1.2 EXPERIMENTS

Despite the recently increased interest in HTG, the proposed works on the task do not follow a common, standard evaluation protocol. To fairly assess our contribution against the State-of-the-Art, we re-run the previous approaches for standardizing the evaluation setup. In the following, we give the details of the evaluation protocol we follow and of the competitors considered, which are models for which code and weights were made publicly available. We hope this setting to be embraced by the community as a new standardized benchmark.

CONSIDERED SCORES. To evaluate the performance of our approach by considering different aspects of the HTG task, we adopt multiple scores in our analysis. These are the FID [147] and the Kernel Inception Distance (**KID**) [44], commonly adopted for image generation, and the task-specific HWD [1]. Moreover, we introduce two other scores. The first is the Binarized FID (**BFID**), obtained by computing the FID on binarized reference and generated images to capture mostly the perceptual handwriting style and disregard color or textures. The second is the Absolute Character Error Rate Difference (Δ **CER**). To compute it, we consider the State-of-the-Art text recognition TrOCR-Base [205] model and

compute its CER on the reference and generated images to capture their readability while taking into consideration the style. Intuitively, we want the generated images to be as readable as the reference ones but not significantly more, which could mean that the HTG model has collapsed to a very readable style, different from the reference one. Note that we also considered additional scores for our evaluation, which are reported in Chapter B.

CONSIDERED DATASETS. Our proposed Emuru is trained only on a large synthetic dataset of images containing English text rendered in calligraphy and typewritten fonts. To assess its generalization capabilities, we consider multiple multi-writer datasets and apply Emuru without any training or fine-tuning. These datasets are the commonly adopted **IAM** [239] dataset, both in its word-level and line-level format, and the line-level annotated **CVL** [190] and **RIMES** [22] datasets. Moreover, we devise a dataset of text lines images obtained by rendering song lyrics in multiple Latin-alphabet languages (English, French, German, and Italian) with 100 publicly available fonts[‡], both calligraphy and typewritten, on a white background. To foster the research on styled text image generation strategies able to generalize also to typewritten styles, we will release this dataset, dubbed **Karaoke**. Further details about these datasets are reported in Chapter B. Note that to ensure a consistent and fair evaluation, we define a fixed set of reference style images and desired text to guide the generation for Emuru and the considered competitors, described below.

CONSIDERED STATE-OF-THE-ART APPROACHES. In our comparison, we consider State-of-the-Art approaches for styled HTG following both the adversarial training paradigm and the denoising diffusion paradigm. Note that for consistent evaluation, we select models for which code and weights have been made available by the respective authors. Specifically, we include in our analysis the Convolutional GAN-based **HiGAN+** [121] and **TS-GAN** [95], and the Transformer GAN-based **HWT** [42], **VATr** [278], and **VATr++** [379]. Moreover, we consider the recently-proposed **DiffPen** [262] and **One-DM** [90] HTG Diffusion Models.

[‡]<https://fonts.google.com/>

IMPLEMENTATION DETAILS. We train all the components of our model on a single NVIDIA RTX 4090.

Emuru convolutional VAE has 4 layers for the Encoder and the Decoder, with the output channels for each block set to 32, 64, 128, and 256. The Encoder spatially downsamples the input $I^{3 \times W \times H}$ to a 1-channel latent vector $b \times w$, where $b=H/8$ and $w=W/8$. We use one channel for the latent space to compress the information on the styled images and allow a lightweight Transformer Decoder to effectively handle the resulting latents. We train the VAE for 60k iterations with the Adam Weight Decay (AdamW) [226] optimizer and learning rate $1e-4$. The loss is composed of four terms: $\mathcal{L}_{MAE}$, $\mathcal{L}_{WID}$, $\mathcal{L}_{HTR}$, and $\mathcal{L}_{KL}$, combined with weight 1, 0.005, 0.3, and $\beta = 1e-6$, respectively. Note that for the HTR supervision, we employ noisy teacher-forcing with probability 0.3. As for the auxiliary networks, the writer identification model is a ResNet with 6 blocks trained with the AdamW optimizer, learning rate $1e-4$, and batch size 256 until it reached 60% accuracy on the synthetic dataset. The HTR network is a Transformer Encoder-Decoder [177] trained with AdamW, learning rate $1e-4$, and batch size 128 until it reached 0.25 CER on the synthetic dataset.

The base architecture of the autoregressive Transformer in Emuru is a T5-Large [299] Encoder-Decoder with two linear adapters, one before the input of the Decoder and another that replaces the original T5 Decoder linear classifier. The adapter layers convert the embedding sequence from the VAE feature space $b = 8$ to the T5-Decoder inner dimension $\mathcal{D}_{dim} = 1024$ and vice versa. The text input of the Encoder is tokenized with the single-character tokenizer Google-ByT. For the first stage of training, we adopt the AdamW optimizer with learning rate $1e-4$ and weight decay $1e-2$ and train for 70k iterations with batch size 256, employing noisy teacher-forcing with probability 0.1. All images used in this phase contain from 4 to 7 words and are padded or cropped to a fixed width of 768px. For the second stage of training, we use images containing from 1 to 32 words, padded or cropped to a fixed width of 2048. For this phase, we train for 250 iterations with a virtual batch size of 256 with gradient accumulation, without employing noisy teacher-forcing.

	FID↓	BFID↓	KID↓	Δ CER↓	HWD↓
SD1.5 VAE	29.39	7.36	32.14	0.00	0.77
SD3 VAE	21.90	3.61	23.01	0.00	0.74
Emuru VAE	19.22	1.62	16.35	0.03	0.85

Table 8.1: Comparison between the VAE featured in Emuru and other VAEs used in Diffusion Model-based HTG solutions when reproducing the images in the test set of different HTG datasets. We report the mean of each score obtained over all the considered datasets. The KID is multiplied by 10^3 and the best performance is in bold for readability.

RESULTS

8

VAE RECONSTRUCTION PERFORMANCE. Since our autoregressive model generates only latent vectors, it is clear that the performance of Emuru is tied to the VAE reconstruction quality. Therefore, we start our experiments by evaluating the quality of the VAE, which is tasked to embed the styled text images into the latent space and perform background removal. We perform extensive analysis on the reconstruction task over multiple datasets and compare Emuru VAE to other VAEs pretrained for natural images from State-of-the-Art Diffusion Models, namely Stable Diffusion 1.5 (SD1.5) [308] and Stable Diffusion 3 (SD3) [108]. We report the average scores over our considered datasets in Tab. 8.1 and the dataset-specific results in Chapter B. As it emerges from this analysis, our text images-tailored VAE achieves good generalization and style consistency compared to generalist VAEs, also enabling downstream lightweight token modeling using only $\sim 16\%$ of their parameters and having a single latent channel (note that the VAE has 4 channels in SD1.5 and 16 channels in SD3).

ZERO-SHOT INFERENCE ON IAM. Current HTG models achieve State-of-the-Art performance by training on a single dataset at a time (generally, the word-level IAM). In Tab. 8.2, we compare IAM-trained competitors with Emuru, trained only on the proposed synthetic dataset. We refer to this setting as *zero-shot inference* because the model has never seen IAM samples during training. The performance of our approach is mostly in line with that of existing methods, even achieving top scores for line generation in terms of BFID and HWD. As these

	IAM Words					IAM Lines				
	FID↓	BFID↓	KID↓	ΔCER↓	HWD↓	FID↓	BFID↓	KID↓	ΔCER↓	HWD↓
TS-GAN	129.57	86.45	141.08	0.28	4.22	44.17	19.45	45.42	0.02	3.21
HiGAN+	50.19	21.92	43.39	0.20	3.12	74.41	34.18	77.27	0.00	3.25
HWT	27.83	15.09	19.64	0.15	2.01	44.72	30.26	43.49	0.33	2.97
VATr	30.26	15.81	22.31	0.00	2.19	35.32	27.97	33.61	0.02	2.37
VATr++	31.91	17.15	23.05	0.07	2.54	34.00	21.67	29.68	0.03	2.38
One-DM	27.54	10.73	21.39	0.10	2.28	43.89	21.54	44.48	0.13	2.83
DiffPen	15.54	6.06	11.55	0.06	1.78	12.89	6.87	9.73	0.03	2.13
Emuru	63.61	37.73	62.34	0.19	3.03	13.89	6.19	11.30	0.14	1.87

Table 8.2: Comparison on the word-level and line-level IAM datasets between Emuru and State-of-the-Art approaches trained on IAM. The KID is multiplied by 10^3 and the best performance is in bold for readability.

	CVL Lines					RIMES Lines				
	FID↓	BFID↓	KID↓	ΔCER↓	HWD↓	FID↓	BFID↓	KID↓	ΔCER↓	HWD↓
TS-GAN	42.12	31.97	43.15	0.13	3.07	109.04	36.39	132.90	0.12	3.26
HiGAN+	78.44	39.47	80.39	0.12	3.07	160.57	47.38	183.82	0.14	3.39
HWT	31.22	16.73	26.14	0.38	2.59	118.21	35.26	128.66	0.45	3.36
VATr	34.40	24.64	32.21	0.06	2.36	113.76	30.21	114.21	0.07	3.09
VATr++	35.53	19.87	34.15	0.12	2.18	110.04	35.61	104.05	0.10	2.83
One-DM	60.45	26.58	64.13	0.06	2.66	121.18	36.07	121.67	0.20	3.36
DiffPen	40.40	17.50	38.21	0.01	2.99	89.79	18.25	94.78	0.04	2.58
Emuru	14.39	10.77	12.34	0.13	1.82	26.93	13.26	21.19	0.25	2.18

Table 8.3: Comparison on the line-level CVL and RIMES datasets between Emuru and state-of-the-art approaches trained on IAM. The KID is multiplied by 10^3 and the best performance is in bold for readability.

scores focus on style rather than background, this indicates that the generated lines are more representative of the input style. Moreover, as shown in Fig. 8.4, scores such as FID and KID are not representative of aspects like word alignment [1], which is a weakness of models that generate single words and stitch them together to create a sentence, like *e.g.* DiffPen. Despite not having been trained on IAM samples, Emuru demonstrates strong generalization, performing on par with competitors trained on IAM.

ZERO-SHOT INFERENCE ON CVL AND RIMES. The dataset-specific tailoring of existing methods severely limits their generalization performance, as demonstrated in Tab. 8.3. In this zero-shot evaluation on CVL and RIMES lines, all competing models (which have been trained on IAM) show a marked performance drop. In contrast, Emuru achieves top performance across nearly all met-



Figure 8.4: Qualitative comparison between Emuru, the GAN-based VATr++, and the Diffusion Model-based DiffPen when generating images from the considered datasets. We report the input style image used for guiding the generation and another reference image in the same style. We let the models generate the same text as in the reference to better observe the style imitation capabilities of the models.

rics. Interestingly, despite not being able to capture the input style, some approaches maintain readability in the generated text, as reflected by the relatively stable ΔCER score. These results, in line with those achieved on IAM, demonstrate the generalization capabilities of Emuru, also observed from Fig. 8.4.

ZERO-SHOT INFERENCE ON THE KARAOKE. Finally, we evaluate the generalization capability of all models on calligraphy and typewritten text images by considering our devised Karaoke dataset. Different from IAM, CVL, and RIMES, Karaoke has a diverse set of 100 fonts and no background artifacts. In Tab. 8.4, we evaluate Emuru against the considered competitors on Karaoke, distinguishing between images of text in calligraphy fonts and typewritten fonts. In both cases, the results are in line with the zero-shot inference comparisons in Tab. 8.3, with Emuru achieving the best values across all scores except ΔCER . Arguably, this is because Emuru, like the diffusion-based models, does not have any mechanism to ensure the correctness of the generated text. In Chapter B, we report further experiments, including fine-tuning Emuru with an HTR-based loss. Across

	Karaoke Calligraphy					Karaoke Typewritten				
	FID↓	BFID↓	KID↓	Δ CER↓	HWD↓	FID↓	BFID↓	KID↓	Δ CER↓	HWD↓
TS-GAN	60.30	12.68	64.80	0.23	4.59	141.41	75.78	157.33	0.32	4.70
HiGAN+	125.75	69.41	136.75	0.08	4.90	135.34	63.39	146.34	0.07	5.19
HWT	62.69	43.03	59.35	0.32	4.50	72.78	37.40	62.77	0.37	4.57
VATr	72.22	47.66	67.70	0.05	3.89	80.38	41.02	70.46	0.05	4.14
VATr++	67.16	46.53	58.57	0.01	3.96	76.03	41.69	63.17	0.01	4.15
One-DM	59.73	38.30	56.55	0.04	4.31	70.75	44.06	60.90	0.05	4.80
DiffPen	34.19	25.78	28.91	0.16	4.18	78.07	61.16	67.17	0.14	4.71
Emuru	13.87	7.99	9.24	0.13	2.24	9.85	4.33	5.60	0.11	1.28

Table 8.4: Comparison on the line-level Karaoke dataset, separating calligraphy and typewritten styles, between Emuru and State-of-the-Art approaches trained on IAM. The KID is multiplied by 10^3 and the best performance is in bold for readability.

all quantitative experiments in Tabs. 8.2 to 8.4 and the qualitative comparison in Fig. 8.4, Emuru stands out as the only model that exhibits a consistent value range for the scores and comparable zero-shot performance on all the considered datasets. On the other hand, other methods show significant variability, sometimes completely failing to generate images in the style of the target dataset. This highlights Emuru’s robustness and ability to generalize to diverse scenarios.

EDITING APPLICATION. As mentioned in previous sections, we train Emuru to generalize on unseen styles and to be background-agnostic. Combined, these two characteristics enable Emuru to generate images that are both faithful to the input style and free from background artifacts, which makes them usable in real-world editing scenarios. To illustrate this, in Fig. 8.1, we extract a style sample from a handwritten letter image and task Emuru to render the sentence “*Hello World! I am Emuru: Born to make mistakes, not to fake perfection.*”. Then, we overlay the resulting image on the same letter. Note that the same results cannot be replicated with other models due to the background artifacts they maintain, as shown in Fig. 8.5. DiffPen effectively copies the style of the IAM sample, but it also reproduces the gray background typical of IAM samples. This issue of learning to imitate the background in addition to the text strokes is common in many HTG approaches because they are trained to mimic the training text images as a whole, including their background. In contrast, by leveraging the VAE’s background removal capabilities, Emuru generates images that reproduce only the text style, and thus, that can be easily blended over any custom background

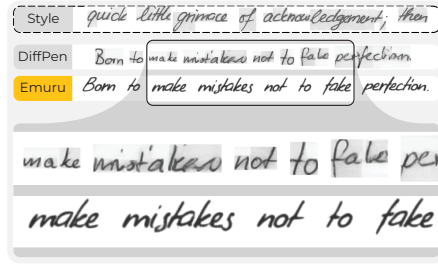


Figure 8.5: Results of image generation from a style image with background artifacts obtained with Emuru and DiffPen. Thanks to the VAE training, Emuru images do not reproduce these artifacts.

image.

8

8.1.3 SUMMARY

Emuru is a novel autoregressive model for Offline styled text image generation. Our approach relies on a powerful text images representation model (a VAE) combined with an autoregressive Transformer, and both are trained on a diverse, synthetic dataset of English text rendered in over 100,000 typewritten and calligraphy fonts. Emuru effectively addresses the limitations of previous approaches, such as: generalization to novel styles, limited or fixed output length, and entanglement between text style and background style. Extensive evaluation considering multiple datasets, both handwritten and typewritten, demonstrates the effectiveness of Emuru in generating both short and long text sequences with high fidelity to unseen styles and its superiority with respect to the State-of-the-Art in terms of generalization capabilities.

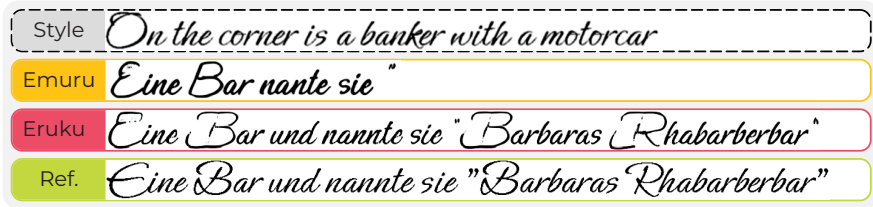


Figure 8.6: Our proposed Eruku model can generate text images with arbitrary length and with great text adherence, respecting both the generation text and with the conditioning writing style.

8.2 IMPROVING AUTOREGRESSIVE HTG WITH ERUKU

While the autoregressive formulation introduced by Emuru provides undeniable advantages like training efficiency and the ability to generate arbitrary-length outputs, it also presents important drawbacks. First, it requires as input the transcription of the style image. This helps the model associate style features with textual content, but creates a strong dependency on accurate transcriptions, which may not be available in real-world scenarios, or may be unreliable when obtained with text recognition networks, which can be imperfect. Moreover, to stop the generation, it relies on a heuristic that entails emitting 10 consecutive padding tokens, which are then discarded. This strategy is somewhat inefficient. Finally, the model often struggles to precisely render the desired text, suffering from issues typical of autoregressive generation models such as repetitions, incomplete sequences, and failure to stop at the correct length (see Fig. 8.6).

Therefore, in this work, we follow the same autoregressive formulation, improving its key aspects. To this end, we introduce modeling novelties to

- free the model from requiring the transcription of the style image as input, making it instead optional;
- provide the model with an explicit stopping mechanism via a single, dedicated end of generation token;
- enforce adherence to the desired text sequence without relying on auxiliary networks for supervision.

We achieve these goals by introducing special visual and textual tokens to guide the generation, and by a Classifier-Free Guidance (CFG)-inspired approach that works only on the textual inputs.

Our method, dubbed *Eruku*, is built upon the Emuru VAE and an autoregressive Transformer trained on a large-scale synthetic dataset of text images. Specifically, the Transformer is trained to iteratively predict VAE-compatible embeddings, including our introduced special tokens, and to generate by exploiting our CFG-inspired mechanism.

We conduct extensive experiments on multiple handwritten and typewritten datasets, all different from the synthetic one used in training. The obtained results show that Eruku achieves robustness to missing or noisy inputs and improves text fidelity while maintaining strong generalization to unseen styles. We release code and weights of our approach [§].

8

8.2.1 ERUKU ARCHITECTURE

Our HTG approach takes as input a reference *style image* (I_s) and the text content to be rendered, dubbed *generation text* (T_g) from here on. Optionally, the model also takes as input the text contained in the style sample, which we call *style text* (T_s). Then, it is tasked to generate a text image I_g containing T_g in the same style as I_s (Fig. 8.7).

Specifically, we generate I_g by using an autoregressive Transformer Encoder-Decoder model operating in the latent space of a Variational Autoencoder (VAE), which acts as an image tokenizer. The Transformer Encoder, $\mathcal{E}$, takes as input T_g , optionally preceded T_s . The Transformer Decoder, $\mathcal{D}$, is fed with the I_s , tokenized by the VAE Encoder. Then, the model autoregressively outputs embeddings of the VAE latent space and stops generating by emitting a special token. Finally, these latent space tokens are decoded by the VAE Decoder, which outputs the desired text image I_g . The details of our pipeline are given below.

[§]<https://aimagelab.github.io/Eruku>

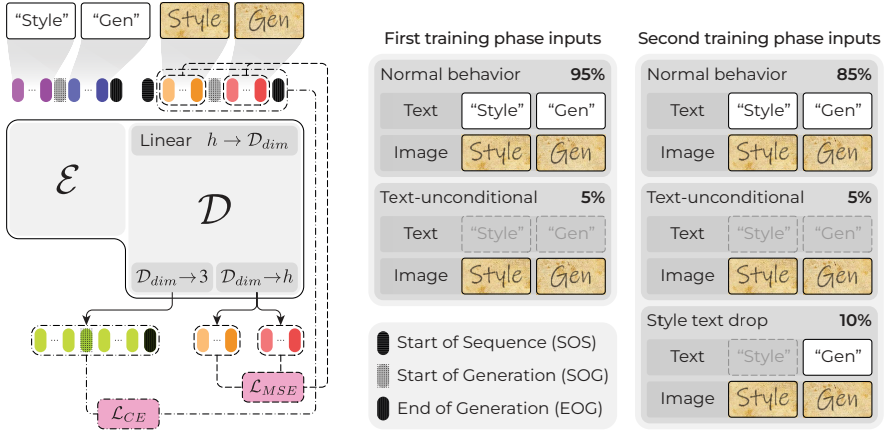


Figure 8.7: Training framework of Eruku, our autoregressive text image generation model. We condition generation on: the textual content of a style image T_s (“Style”), the generation text T_g (“Gen”), and the style image I_s . Eruku is trained on next-token prediction, learning to generate an image containing the generation text T_g with the same writing style as the style image I_s . Providing the style text T_s enables the model to link each character with its representation style, but we also enable generation without the style text T_s by dropping it during inference and using synchronization tokens to separate the sequence components. We represent images with VAE [6] continuous latents. Our model automatically learns to stop generation emitting the *Visual End of Generation token* $\langle \text{EOG} \rangle$.

VAE IMAGE TOKENIZER

To project the style image I_s into a compressed latent space and then convert the embeddings generated by the autoregressive Transformer into the output image I_g , we reuse the continuous β -VAE provided in [6], frozen. Specifically, given an RGB ($3 \times H \times W$) text image I , the VAE Encoder projects it into a latent representation. Given the VAE’s number of channels, c , and its downscaling factor, f , the resulting latent vector will have shape $c \times h \times w$, where $w = W/f$ and $h = H/f$. This is then reshaped into a w -long sequence of $h \cdot c$ vectors, $\mathbf{v} = [v_0, \dots, v_w]$, so that each one encodes a vertical slice of the text image. The VAE Decoder is tasked to reconstruct a grayscale version of the text image without the background, therefore enforcing the latent space to represent text style rather than background content. The model was trained with a combination of reconstruction loss, KL divergence, Cross-Entropy from a pretrained writing style classification network [6], and a CTC loss from a pretrained HTR model [6].

AUTOREGRESSIVE TEXT ENCODER

Eruku autoregressive Transformer Encoder takes as inputs text tokens obtained by tokenizing T_s (if given) and T_g at character-level with the byte-by-byte tokenizer from ByT5 [408], separated and followed by two special tokens. Specifically, the input to the Encoder is:

$$\mathbf{t} = [t_{s,1}, \dots, t_{s,l_s}, \langle \text{SOG} \rangle, t_{g,1}, \dots, t_{g,l_g}, \langle \text{EOG} \rangle].$$

The *textual End of Generation* token $\langle \text{EOG} \rangle$ is a utility token automatically inserted by the tokenizer. The *textual Start of Generation* token $\langle \text{SOG} \rangle$ is an additional special token that we introduce to signal the model that the following tokens are part of the generated image it is tasked to render. The $\langle \text{SOG} \rangle$, coupled with appropriate training, enables the model to learn to ignore the style text if it is unable to match it to the style image, or if the style text is unavailable. The Encoder performs multi-layer self-attention on $\mathbf{t}$ and passes its output to the Autoregressive Decoder.

AUTOREGRESSIVE IMAGE DECODER

Eruku autoregressive Transformer Decoder takes as inputs the visual tokens $\mathbf{v}_s$ obtained by tokenizing the style image I_s with the VAE Encoder, linearly projected into embedding vectors with size compatible with the Transformer dimension ($\mathcal{D}_{dim}$), thus obtaining the sequence $\mathbf{e}_s$. Moreover, we prepend and append two additional learnable embeddings: one for the Start of sequence token, e_{SOS} , and one for the visual Start of generation token, e_{SOG} . Therefore, the input to the Decoder is $\mathbf{e} = [e_{\text{SOS}}, e_{s,1}, \dots, e_{s,w_s}, e_{\text{SOG}}]$.

The Decoder performs multi-layer self-attention on $\mathbf{e}$ and cross-attention between $\mathbf{e}$ and the output of the Transformer Encoder, and iteratively generates a sequence of embeddings $\hat{e}_{g,i} \in \hat{\mathbf{e}}$. At each generation step, the current embedding $\hat{e}_{g,i}$ is linearly projected into two separate vectors: $\hat{s}_i \in \mathbb{R}^3$ and $\hat{v}_{g,i} \in \mathbb{R}^{c \cdot b}$. The value of $\hat{s}_i$ belongs to a dictionary of three elements, $\{\langle \text{SOG} \rangle, \langle \text{IMG} \rangle, \langle \text{EOG} \rangle\}$, within a next-token prediction scheme. In particular:

- Visual Start of Generation token <SOG>, meaning that the model suggests to insert the start of generation token at that point in the sequence. This behavior is mainly useful in training, but it also enables the model to recover cases in which the T_s is incorrect. In this case, we append e_{SOG} to the generated sequence of embeddings and run another generation step.
- Visual token , meaning that the model is continuing to generate image tokens in $\hat{v}_i$. In this case, we append its linear projection $\hat{e}_{g,i}$ to the generated sequence and run another generation step.
- Visual End of Generation token <EOG> meaning that the model suggests that the entire generation text T_g has been rendered in the output image. In this case, we stop the autoregressive generation process.

Note that the visual <SOG> and <EOG> tokens also act as a synchronization signal alongside the corresponding textual <SOG> and <EOG> tokens in the Autoregressive Encoder (see Section 8.2.1). In fact, as detailed in Section 8.2.2, during training Eruku consistently observes paired textual/visual <SOG> and textual/visual <EOG> tokens. As a consequence, it implicitly learns an alignment between the boundaries of corresponding textual and visual segments. At inference time, the model receives the visual <SOG> already aligned with the textual <SOG>, and is encouraged to produce the appropriate number of visual tokens to ensure that the textual <EOG> provided to the encoder corresponds meaningfully to the visual <EOG> generated by the decoder. In the rest of the chapter, where it is clear from the context, we use <SOG> and <EOG> for both the textual and visual tokens for simplicity of notation.

The generation process continues until either a visual <EOG> token is predicted. At the end of the process, the sequence of $\hat{v}_{g,i}$'s, $\mathbf{v}_g$ is passed to the VAE Decoder to obtain the final generated image I_g .

TEXT CLASSIFIER-FREE GUIDANCE. In order to improve Eruku's ability to correctly render the desired text T_g within the output image, we use an inference-time strategy based on Classifier-Free Guidance (CFG). CFG was developed in the context of conditional Diffusion Models [150] to increase prompt-image alignment by sharpening the sampling distribution towards the conditioning. To this

end, the model is tasked to generate with the conditioning signal (conditional generation) and with a null condition $\emptyset$ (unconditional generation), and the results are combined, scaled with a parameter γ to regulate the intensity of this operation. Recently, this technique has been successfully applied in Autoregressive image generation models [207] and in Autoregressive image editing models [254, 71].

Recall that our proposed Eruku is conditioned on three inputs: the sequence of embeddings representing the style image, $\mathbf{e}_s$, the sequence of style text tokens $\mathbf{t}_s$, and that of the generation text tokens, $\mathbf{t}_g$. The last two are gathered in a single sequence $\mathbf{t}$. To enforce content adherence in the generated image, we apply the CFG formula to the $\hat{e}_{g,i}$'s, but we keep the style image conditioning in the unconditional generation to retain style consistency also in the unconditional branch, *i.e.*

$$\begin{aligned} p(\hat{e}_{g,i} | \hat{e}_{g,<i}, \mathbf{e}_s, \mathbf{t}) = \\ p(\hat{e}_{g,i} | \hat{e}_{g,<i}, \mathbf{e}_s, \emptyset) + \\ \gamma \cdot (p(\hat{e}_{g,i} | \hat{e}_{g,<i}, \mathbf{e}_s, \mathbf{t}) - (p(\hat{e}_{g,i} | \hat{e}_{g,<i}, \mathbf{e}_s, \emptyset))). \end{aligned}$$

8.2.2 ERUKU TRAINING

Note that the training samples for our model consist of tuples containing the style text T_s , the generation text T_g , the style image I_s , and the target text image I_g . We build a dataset by synthesizing such samples as detailed in Section 8.2.2, and we train our model in two phases, as depicted in Fig. 8.7 and described in the following.

TRAINING STRATEGY

At training time, the input to the Transformer Encoder is the same as what is given at inference, *i.e.* a sequence of textual tokens $\mathbf{t}$, computed from T_s and T_g as explained in Section 8.2.1. For the training input to the Transformer Decoder, I_s and I_g are converted to sequences of vectors in the VAE's latent space, *i.e.* $\mathbf{v}_s$ and $\mathbf{v}_g$, and projected into sequences of image embeddings $\mathbf{e}_s$ and $\mathbf{e}_g$. By adding

the embeddings of the special visual tokens, we obtain

$$\mathbf{e} = [e_{\text{SOS}}, e_{s,1}, \dots, e_{s,w_s}, e_{\text{SOG}}, e_{g,1}, \dots, e_{g,w_g}, e_{\text{EOG}}].$$

Then, the model is tasked to replicate the entire sequence of embeddings $\mathbf{e}$, excluding e_{SOS} , *i.e.* to iteratively output the sequence of vectors

$$\hat{\mathbf{e}} = [\hat{e}_{s,1}, \dots, \hat{e}_{s,w_s}, \hat{e}_{\text{SOG}}, \hat{e}_{g,1}, \dots, \hat{e}_{g,w_g}, \hat{e}_{\text{EOG}}].$$

For the generation during training we apply a teacher-forcing strategy.

From each $\hat{e}_i \in \hat{\mathbf{e}}$, we obtain the corresponding $\hat{s}_i$ and $\hat{v}_i$ vectors. Then, we compute a Cross-Entropy (CE) loss $\mathcal{L}_{CE}$ on the $\hat{s}_i$'s and a Mean Square Error (MSE) loss $\mathcal{L}_{MSE}$ on the $\hat{v}_i$'s. Specifically, the $\mathcal{L}_{CE}$ is computed with respect to the reference values in the ground truth sequence given by:

$$\mathbf{s} = [\langle \text{IMG} \rangle_{\times w_s}, \langle \text{SOG} \rangle, \langle \text{IMG} \rangle_{\times w_g}, \langle \text{EOG} \rangle].$$

This sequence is also used to select the $\hat{v}_i$'s corresponding to the style image, $\hat{v}_{s,i}$, and to the desired text image, $\hat{v}_{g,i}$. Then, the $\mathcal{L}_{MSE}$ is computed respectively between the $\hat{v}_{s,i}$'s and the corresponding vector in $\mathbf{v}_s$, and between the $\hat{v}_{g,i}$'s and the corresponding vector in $\mathbf{v}_g$.

TRAINING FOR CLASSIFIER-FREE-GUIDANCE. To enable the text CFG as described in Section 8.2.1, we need to train the model to generate also without conditioning inputs. In our case, this would mean generating without any textual input but only the conditioning given by the style image I_g . To this end, with a given probability p_{uncond} during training, we replace all text embeddings in the models' textual input $\mathbf{t}$ with a learnable text unconditional embedding, $\langle \text{UNCOND} \rangle$. In this way, we obtain the $\emptyset$ conditioning for the unconditional generation. We refer to this setting as text-unconditional generation.

SECOND TRAINING PHASE. After pre-training, we fine-tune Eruku to enable it to generate also when the style text is not available. To this end, with a given probability p_{drop} during training, we do not feed the model with the tokens cor-

responding to T_s , *i.e.* the input to the Transformer Encoder becomes

$$\mathbf{t} = [\langle \text{SOG} \rangle, t_{g,1}, \dots, t_{g,l_g}, \langle \text{EOG} \rangle].$$

Moreover, in this phase, we use samples whose length is more varied compared to those used in the first training phase. This allows the model to handle longer sequences, which is overall beneficial also in terms of performance [6]. Text-unconditional training is also performed during this second phase in order to preserve unconditional generation capabilities, as suggested by [275].

TRAINING DATA

8

Commonly, HTG models are trained on a single dataset, hindering their generalization capabilities on out-of-distribution styles, words, and languages. Therefore, we train our model on a specifically-prepared massive and varied synthetic dataset[‡]. To obtain the dataset, we collect over 100k typewritten and calligraphic fonts available online and use them to render text words as greyscale ink over a white background. Then, we apply random geometric transformations and split the resulting image to obtain the style image and the generated image, as in [277]. The rendered words are picked from a large corpus of English and random words as the one used in [6]. Specifically, For the pre-training stage, we generate samples consisting of 2 to 3 words for the style images and 2 to 3 words for the target text image, and obtain 23M samples. For the fine-tuning stage, we synthesize 10M samples whose style image contains 1 to 8 words and the target text image contains 1 to 32 words.

8.2.3 ERUKU EXPERIMENTS

Following [379][6], we re-run previous approaches, using their publicly released weights, under a unified setup to enable a fair comparison with the State-of-the-Art. For all test datasets considered, we maintain a fixed set of reference style images and target texts to guide generation, ensuring consistency across methods.

[‡]<https://hf.co/datasets/blowing-up-groundhogs/font-square-pretrain-20M>

IMPLEMENTATION DETAILS. We use the pretrained weights from the Emuru VAE [6], which has 4 Encoder layers and 4 Decoder layers, a downscaling factor $f = 8$ and one output channel ($c = 1$). The Autoregressive Transformer architecture or Eruku is the same as T5-Large [299], with $\mathcal{E}_{dim} = \mathcal{D}_{dim} = 1024$. During training, we pad the target images of samples within the same batch so that they have all the same, inter-batch maximum length. For padding, we use the visual $\langle \text{EOG} \rangle$ to teach the model that, once all the visual tokens for I_g are generated, the $\langle \text{EOG} \rangle$ is the only possible output. The first training phase is performed with a batch size of 128 over 65000 iterations, whereas the second phase lasts 5000 iterations with a batch size of 2. In both phases, we use gradient accumulation with a virtual batch size of 256, AdamW as optimizer, with a learning rate of $1e-4$, and weight decay $1e-2$.

EVALUATION SCORES. To comprehensively assess the performance of our model, we employ multiple scores capturing different aspects of HTG. These include the task-specific HWD [1], to capture style fidelity, the Absolute Character Error Rate Difference (ΔCER) [6], which quantifies readability relative to the reference style, the standard image quality evaluation FID [147], and the binarized version of the FID, (**BFID**) [6], which focuses on font fidelity disregarding the background color and texture [4].

DATASETS. Our proposed Eruku is trained only on a large synthetic dataset of images containing English text rendered in calligraphy and typewritten fonts, the same as the one use in [6]. To evaluate its generalization performance, we apply Eruku directly, *i.e.* without fine-tuning, on multiple multi-writer datasets. These include the **IAM** [239] dataset (both word- and line-level), and the line-level **CVL** [190] and **RIMES** [22] datasets. Moreover, we consider the line-level **Karaoke** [6] dataset, consisting of song lyrics in English, French, German, and Italian rendered using 100 publicly available fonts^{||}, encompassing both calligraphy and typewritten styles on a white background.

COMPARED METHODS. We compare Eruku against State-of-the-Art HTG methods with publicly released code and pretrained weights. Specifically, we include

^{||}<https://fonts.google.com/>

Convolutional GAN-based models **HiGAN+** [121] and **TS-GAN** [95], Transformer GAN-based methods **HWT** [42], **VATr** [278], and **VATr++** [379], as well as diffusion-based approaches **DiffPen** [262] and **One-DM** [90]. Finally, we consider the Autoregressive Transformer-based **Emuru** [6], which is the closest to our approach.

ERUKU RESULTS

First, we perform ablation analyses of Eruku’s main characteristics. To this end, we consider the line-level version of the IAM dataset, since it is the most commonly adopted in HTG literature.

STYLE TEXT DROP. We validate the effect of dropping the style text T_s in the Eruku input. In Tab. 8.5, we report the results obtained by applying varying style text drop probability p_{drop} in the second phase of training. For this comparison, we generate both with and without T_s and use $\gamma = 1.25$ for the text CFG. Note that the baseline model is the first line in the same Tab. 8.5, *i.e.* our Eruku architecture trained with style text T_s , which is also provided during inference. Observing the results, we can see that the baseline model, which was not trained with style text dropout ($p_{drop} = 0$), does not work well when the style text is not provided during inference. This is expected, as it has never been trained in this setting. As intended, the ability of the model to generate without style text is significantly improved with increasing values of p_{drop} . In particular, moving from $p_{drop} = 0$ to $p_{drop} = 1$ yields significant performance improvements when style text is not provided. An added benefit of this style text dropout strategy is that the model trained with $p_{drop} = 0.1$ improves its HWD and FID significantly also when provided with T_s , without any decrease in ΔCER . This leads us to believe that this form of style text dropout makes the model more robust to style text that it would otherwise fail to match to the style image. Therefore, the model is able to leverage the provided style text T_s better than to the baseline model.

With increasing values of p_{drop} , we can see that the model’s performance increases without the T_s input and decreases when T_s is provided. We attribute this to the model losing the ability to match T_s and I_g if no longer provided with

p_{drop}	T_s	HWD↓	Δ CER↓	FID↓	BFID↓
0.0	✓	1.81	0.48	14.20	3.88
	✗	3.03	1.02	86.97	77.46
0.1	✓	1.75	0.48	13.49	4.45
	✗	2.15	0.47	16.71	8.10
0.3	✓	1.79	0.50	16.67	4.62
	✗	2.17	0.46	18.73	5.00
0.6	✓	1.74	0.51	15.34	5.47
	✗	2.05	0.46	16.91	6.11
1.0	✓	2.02	0.75	21.41	15.63
	✗	2.03	0.38	15.97	4.25

Table 8.5: Effect of the style text drop probability applied in training, p_{drop} , on the performance of Eruku on IAM Lines, both when the style text T_s is given or not at inference time. For reference, we also report the result of Eruku fed with T_s obtained by running TrOCR on the style image I_s (dubbed T_s^*).

sufficient paired samples in the second phase of training. To show this, we also report the results of a model finetuned with $p_{drop} = 1$, which learns quite well how to generate without style text input. When we input T_s to this model, its performance drops. In light of these results, we find $p_{drop} = 0.1$ to be the best tradeoff between non-style text-conditioned generation and style text-conditioned generation and use this value for our final model.

TEXT-CFG. To validate the effect of the proposed text-CFG on the model’s performance, we consider its effect at different values of the scale γ at inference time. As a reference, we also consider the performance obtained by a variant of our approach not trained to perform text-unconditional generation, which is therefore ran without CFG at inference time. The results of this analysis are reported in Tab. 8.6. We can observe that increasing the CFG scale directly impacts the Δ CER, and therefore the correctness and readability of the text within the output image. For values of γ up to 1.25, this improvement does not hinder the style adherence measured by HWD. When increasing beyond that value, there are diminishing advantages in terms of Δ CER, and worsening HWD values, in-

dicating that style adherence decreases as γ increases beyond 1.25. For this reason, we select $\gamma = 1.25$ as the default CFG scale value for Eruku. Nonetheless, a user could change this CFG scale value to obtain a different style-text adherence trade-off.

Note that an alternative, popular way in the HTG literature to improve the readability of the text output is to fine-tune the models using an auxiliary HTR network [278, 379, 262]. As an ablation, we also try the same approach on Eruku by introducing, in the second stage of training, supervision from an OCR model trained [417] on the same synthetic data used for Eruku in that phase of training. The results obtained are reported in Tab. 8.6 by adding the suffix “+ OCR”. We observe that OCR fine-tuning proves to be less style-preserving than CFG at obtaining the same Δ CER values, as reflected in the HWD scores. Therefore, we do not perform OCR fine-tuning for the final Eruku model, which, as mentioned in Section 8.2.2, is trained without any need for auxiliary networks. To further isolate the effect of CFG training, we also train Eruku’s variants without conditioning dropout, and then fine-tune using the OCR model [417]. The results, reported in Tab. 8.6, show that these variants are worse in both style preservation and text correctness. Finally, we qualitatively show the effect of γ in Fig. 8.8.

COMPARISON WITH THE STATE-OF-THE-ART. In Tabs. 8.7 to 8.9 and Fig. 8.9, we report a quantitative and qualitative comparison between Eruku and other State-of-the-Art HTG approaches. The comparison is performed both on the dataset that most of the competitors have seen in training (IAM) and on unseen datasets. Recall that, instead, for Emuru and our Eruku approach, all the datasets are unseen. Our model exhibits strong generalization capabilities, maintaining solid performance on all line-level datasets. This is underscored by the fact that it is the best-performing model in terms of HWD on all datasets except for IAM Words. The lower performance on this word-level dataset could be attributed to the particular emphasis given during training to long-context generation in the second stage. A potential mitigation that could be implemented in future approaches is to supplement the dataset used for long-context training with a larger sample of short images consisting of one or very few words.

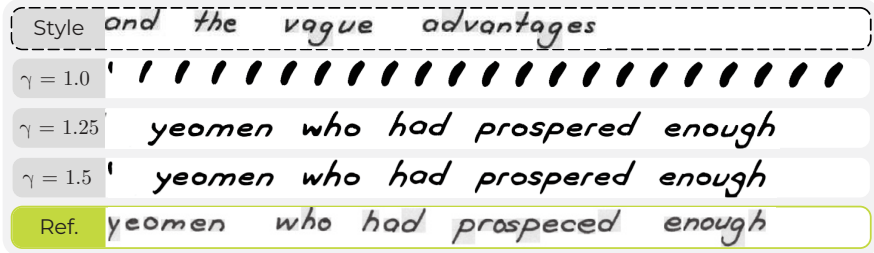


Figure 8.8: Qualitative analysis of the CFG effect on generation at varying γ 's. We generate by giving as I_s the top image in the figure (Style) and as T_s the text contained in the bottom image (Ref.).

8



Figure 8.9: Qualitative results between our proposed Eruku, and the State-of-the-Art Emuru and DiffPen models on the considered datasets. We task the models to generate a replica of the reported reference image (Ref.) by giving them as input the text contained in Ref. and the reported style image (Style).

	HWD↓	Δ CER↓	FID↓	BFID↓
Eruku ($\gamma = 1$)	1.83	0.18	20.12	12.52
Eruku ($\gamma = 1.125$)	1.73	0.10	17.44	7.71
Eruku ($\gamma = 1.25$)	1.70	0.06	16.40	4.88
Eruku ($\gamma = 1.375$)	1.73	0.04	16.76	4.01
Eruku ($\gamma = 1.5$)	1.80	0.04	17.51	4.07
Eruku ($\gamma = 1$) + OCR	1.82	0.14	16.02	7.99
Eruku ($\gamma = 1.125$) + OCR	1.75	0.08	15.27	4.99
Eruku ($\gamma = 1.25$) + OCR	1.78	0.05	16.45	4.86
Eruku ($\gamma = 1.375$) + OCR	1.84	0.04	18.04	5.18
Eruku ($\gamma = 1.5$) + OCR	1.91	0.03	19.66	6.02
Eruku*	1.83	0.21	21.10	17.11
Eruku* + OCR	1.82	0.18	18.48	11.25

Table 8.6: Effect of the text CFG scale on the performance of Eruku. $\gamma = 1$ means that the text CFG is not performed. We also report the results of a variant not trained to support the text CFG (dubbed Eruku*) and variants trained with the additional supervision of an auxiliary OCR network (noted ‘+ OCR’).

	IAM Words				IAM Lines			
	HWD↓	Δ CER↓	FID↓	BFID↓	HWD↓	Δ CER↓	FID↓	BFID↓
TS-GAN	4.22	0.28	129.57	86.45	3.21	0.02	44.17	19.45
HiGAN+	3.12	0.20	50.19	21.92	3.25	0.00	74.41	34.18
HWT	2.01	0.15	27.83	15.09	2.97	0.33	44.72	30.26
VATr	2.19	0.00	30.26	15.81	2.37	0.02	35.32	27.97
VATr++	2.54	0.07	31.91	17.15	2.38	0.03	34.00	21.67
One-DM	2.28	0.10	27.54	10.73	2.83	0.13	43.89	21.54
DiffPen	1.78	0.06	15.54	6.06	2.13	0.03	12.89	6.87
Emuru	3.03	0.19	63.61	37.73	1.87	0.14	13.89	6.19
Eruku	3.23	0.77	79.66	63.31	1.70	0.06	16.40	4.88

Table 8.7: Comparison on the word-level and line-level IAM datasets. Note that Eruku and Emuru have not been trained on IAM.

	CVL Lines				RIMES Lines			
	HWD↓	Δ CER↓	FID↓	BFID↓	HWD↓	Δ CER↓	FID↓	BFID↓
TS-GAN	3.07	0.13	42.12	31.97	3.26	0.12	109.04	36.39
HiGAN+	3.07	0.12	78.44	39.47	3.39	0.14	160.57	47.38
HWT	2.59	0.38	31.22	16.73	3.36	0.45	118.21	35.26
VATr	2.36	0.06	34.40	24.64	3.09	0.07	113.76	30.21
VATr++	2.18	0.12	35.53	19.87	2.83	0.10	110.04	35.61
One-DM	2.66	0.06	60.45	26.58	3.36	0.20	121.18	36.07
DiffPen	2.99	0.01	40.40	17.50	2.58	0.04	89.79	18.25
Emuru	1.82	0.13	14.39	10.77	2.18	0.25	26.93	13.26
Eruku	1.72	0.04	12.32	6.62	1.81	0.11	27.51	10.15

Table 8.8: Comparison on the CVL and RIMES datasets. Note that none of the approaches has been trained on these datasets.

	Karaoke Calligraphy				Karaoke Typewritten			
	HWD↓	Δ CER↓	FID↓	BFID↓	HWD↓	Δ CER↓	FID↓	BFID↓
TS-GAN	4.59	0.23	60.30	12.68	4.70	0.32	141.41	75.78
HiGAN+	4.90	0.08	125.75	69.41	5.19	0.07	135.34	63.39
HWT	4.50	0.32	62.69	43.03	4.57	0.37	72.78	37.40
VATr	3.89	0.05	72.22	47.66	4.14	0.05	80.38	41.02
VATr++	3.96	0.01	67.16	46.53	4.15	0.01	76.03	41.69
One-DM	4.31	0.04	59.73	38.30	4.80	0.05	70.75	44.06
DiffPen	4.18	0.16	34.19	25.78	4.71	0.14	78.07	61.16
Emuru	2.24	0.13	13.87	7.99	1.28	0.11	9.85	4.33
Eruku	2.04	0.13	12.39	7.30	1.21	0.11	10.29	5.07

Table 8.9: Comparison on the Karaoke dataset. Note that none of the approaches has been trained on these datasets.

8.3 CONCLUSION

In this chapter, we have presented two contributions to the autoregressive paradigm for Offline styled text image generation. First, we have introduced Emuru, a novel autoregressive model that relies on a powerful text images representation model (a VAE) combined with an autoregressive Transformer, both trained on a diverse, synthetic dataset of English text rendered in over 100,000 typewritten and calligraphy fonts. Emuru effectively addresses the limitations of previous approaches, such as generalization to novel styles, limited or fixed output length, and entanglement between text style and background style. Then, building upon the Emuru framework, we have proposed Eruku, which improves the autoregressive formulation by tackling its key drawbacks: the dependency on accurate style image transcriptions, the inefficient and error-prone stopping mechanism, and the poor adherence to the target text. Eruku achieves these improvements through the introduction of special visual and textual tokens and a novel CFG-inspired mechanism acting on the text inputs. Through extensive experiments on a variety of handwritten and typewritten datasets, we have validated the effectiveness of both approaches, demonstrating strong zero-shot generalization to unseen styles, arbitrary output length flexibility, and, with Eruku, the ability to generate images whose content closely adheres to the desired target text even without requiring the style image transcription. As shown in Chapter C, Eruku remains robust when the style transcription is missing or obtained via Optical Character Recognition (OCR), whereas Emuru-style conditioning can be destabilized by such noisy inputs.

9

Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas

DIFFUSION models have become the State-of-the-Art for text-to-image generation, and increasing research effort has been dedicated to adapting the inference process of pretrained diffusion models to achieve zero-shot capabilities. An example is the generation of panorama images, which has been tackled in recent works by combining independent diffusion paths over overlapping latent features, which is referred to as *joint diffusion*, obtaining perceptually aligned panoramas. However, these methods often yield semantically incoherent outputs and trade-off diversity for uniformity. To overcome this limitation, we propose the **Merge-Attend-Diffuse** operator, which can be plugged into different types of pretrained diffusion models used in a joint diffusion setting to improve the perceptual and semantic coherence of generated panorama images. Specifically, we merge the diffusion paths, reprogramming self- and cross-

This Chapter is related to the publication “F. Quattrini *et al.*, Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas, ECCV 2024” [5]. See the list of Publications on page 315 for more details.

attention to operate on the aggregated latent space. Extensive quantitative and qualitative experimental analysis, together with a user study, demonstrate that our method maintains compatibility with the input prompt and visual quality of the generated images while increasing their semantic coherence. We release the code of our approach ^{*}.

9.1 INTRODUCTION

Diffusion models [333, 336, 338, 149] have achieved impressive performance in various domains, including image generation [102, 300, 315, 308, 28], video generation [395, 152, 107], and audio synthesis [192, 69, 283], especially when guided by a natural language prompt. Once trained on large-scale datasets, these models can be exploited for zero-shot adaptation in downstream tasks, such as inpainting [230], image editing [144, 25, 23], 3D modeling [212, 404], and video manipulation [186, 387].

The most popular text-to-image diffusion models are trained to generate fixed-size images. Although some methods are trained with different image resolutions and aspect ratios [281, 67], not all image dimensions can be provided during training. This is especially true for oddly-shaped images, such as panoramas. Note that obtaining large or oddly-shaped images by simply applying rescaling or super-resolution techniques to a generated square image can lead to unsatisfactory results due to the artifacts that would be introduced, the lack of detail, and the uneven enlargement that long images require. In light of these considerations, and as the size of the models (and the associated training cost) continues to increase, there is a growing research effort toward inference-time strategies to exploit pre-trained models for generating images with unseen shapes and sizes.

Among those, *iterative diffusion* methods [23, 424] involve generating an initial image and then generating the whole panorama image with a series of out-painting steps. However, this strategy struggles to maintain global coherence and tends to repeat similar patterns [198]. Another family of approaches, referred to as *joint diffusion* methods [30, 198], entails considering the panorama image

^{*}<https://github.com/aimagelab/MAD>



Figure 9.1: MultiDiffusion [30] blends the mountains with the clouds, lacking semantic coherence (top). Applying the proposed MAD leads to semantic coherence (bottom).

canvas, splitting it into overlapping views, performing the diffusion denoising process on each view independently, and combining the resulting scores or latent features at each timestep. This strategy achieves better-blended transitions between adjacent views. However, these are still processed independently, and thus, the resulting image still lacks global coherence and sometimes presents evident artifacts (*e.g.* a mountain smoothly turning into a cloud as in Fig. 9.1-top). Note that, by coherence, one can mean *perceptual coherence*, *i.e.* exhibiting similar colors, illumination conditions, and texture throughout the image, or *semantic coherence*, *i.e.* containing the right amount of elements (objects or parts) with a consistent layout and arranged in a meaningful position in the image. Achieving perceptual coherence has been approached in [198] by performing gradient descent during the denoising process with a perceptual loss. Instead, obtaining semantic coherence is still an open challenge, which entails modeling long-range relations among the elements (for example, the images in Fig. 9.1 are both perceptually coherent, but the first one, which is obtained with the State-of-the-Art MultiDiffusion [30] strategy, contains semantic inconsistencies).

In this work, we take a step towards the generation of long images that are both perceptually and semantically coherent while maintaining variability and realism. To this end, we devise a joint diffusion pipeline to exploit pretrained diffusion models by manipulating their attention operators. Similar to previous approaches [30, 198], we combine the predictions over multiple, strided views. Unlike those approaches, which process the views independently and combine them *at the end* of each denoising step, we devise a strategy to provide additional

interaction points *within* the denoising step. This way, our approach can both model the long image as a whole and focus on the details in each view. In particular, our strategy is based on our proposed Merge-Attend-Diffuse (MAD) operator, which can be selectively applied to the attention layers inside the noise prediction network of a pretrained diffusion model. MAD takes as input the latent features corresponding to the different panorama views, merges them into a single tensor by averaging the overlapped regions, and feeds it to the self- or cross-attention layer it is applied to. Finally, the obtained latent tensor is split back into views for the next layers, which can follow the standard joint diffusion paradigm. Our proposed strategy based on the MAD operator can be applied to different types of pretrained diffusion models to generate perceptually and semantically coherent montages. In this work, we utilize as backbone a pretrained Latent Diffusion Model (LDM), namely, the popular Stable Diffusion [308], and a Latent Consistency Model (LCM) [233] and demonstrate its effectiveness via qualitative and quantitative evaluations, as well as a user study.

9.2 PRELIMINARIES

In this section, we briefly review diffusion and consistency models to provide relevant preliminaries to our pipeline.

DIFFUSION MODELS. Diffusion models [333, 149, 338, 334] are a family of probabilistic models that learn to gradually transform a Gaussian noise input $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ into a data sample $\mathbf{x}_0$ by approximating the data distribution q , $\mathbf{x}_0 \sim q$ in T steps. In particular, the diffusion models framework defines a forward process that gradually injects noise into the data, transforming the data distribution $q(\mathbf{x}_0)$ into the marginal distribution via the transition kernel

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}),$$

where (α_t, σ_t) specifies a differentiable noise schedule such that $q(\mathbf{x}_t) \approx \mathcal{N}(0, \mathbf{I})$. The model is trained to reverse this process by learning to denoise $\mathbf{x}_T \sim q(\mathbf{x}_t | \mathbf{x}_0)$. Predicting $\mathbf{x}_0$ is done iteratively in the reverse diffusion process by estimating

$\mathbf{x}_{t-1}$ from $\mathbf{x}_t$. A possible strategy is to use the ε -prediction parameterization for sampling, as done in Denoising Diffusion Probabilistic Models (DDPMs) [149]. This entails parametrizing both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ as a combination of $\mathbf{x}_0$ and the noise ε , scheduled according to the noise schedule. In this way, the model can be trained to estimate ε directly, which is obtained by optimizing

$$\mathbb{E}_{q(\mathbf{x}_0)}[\|\varepsilon_\theta(\mathbf{x}_t, t) - \varepsilon\|_2^2].$$

In our pipeline for long image generation, we exploit a diffusion model that has been trained with this strategy. However, in our inference-time adaptation, we use the sampler presented in Denoising Diffusion Implicit Models [334]. This is also adopted by other works tackling the long image generation task since it is more efficient than DDPM for requiring fewer sampling steps (usually 25 or 50).

LATENT DIFFUSION MODELS. Training diffusion models in the high-resolution pixel-space can be computationally prohibitive. Latent diffusion models (LDMs) [308, 281] tackle this issue by operating in the latent space of an autoencoder, typically a VQ-GAN [109] or a VQ-VAE [375]. These models achieve good performance while using a fraction of the GFlops required by pixel space diffusion models [308, 281, 271, 67]. Here, we rely on the pretrained Stable Diffusion [308] LDM.

CONDITIONAL GENERATION. Text-to-image diffusion models are able to generate images whose content is specified via a natural language prompt. This is represented by an embedding $\mathbf{e}$ obtained from the text encoder of a pretrained multimodal model (CLIP [297], in our case). In this work, we exploit diffusion models trained with the Classifier-Free Guidance [151] strategy for conditioning. This entails obtaining the noise prediction as a linear combination of conditional and unconditional predictions, with weight s , *i.e.*

$$\hat{\varepsilon}_\theta(\mathbf{x}_t, \mathbf{e}, t) = \varepsilon_\theta(\mathbf{x}_t, \emptyset, t) + s(\varepsilon_\theta(\mathbf{x}_t, \mathbf{e}, t) - \varepsilon_\theta(\mathbf{x}_t, \emptyset, t)),$$

where $\emptyset$ is the embedding of the null prompt. Note that the conditional generation is implemented by performing cross-attention with the embedding $\mathbf{e}$ inside

the noise estimation network. In light of this, by applying our devised MAD operator to the cross-attention layers, we enforce all the latents of the views to jointly attend the prompt, thus favoring semantic coherence.

CONSISTENCY MODELS. Our proposed MAD operator can be applied to different kinds of diffusion models. As an example, in this work, we apply it also to the recently proposed Consistency Models (CMs) [335], which can achieve impressive generation results. In particular, a consistency model f_θ estimates a function that can directly map any intermediate noisy sample $\mathbf{x}_t$ in the forward process to the origin $\mathbf{x}_0$, which is parameterized as

$$f_\theta(\mathbf{x}, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t)F_\theta(\mathbf{x}, t),$$

where $F_\theta(\mathbf{x}, t)$ is a neural network and $c_{\text{skip}}(t), c_{\text{out}}(t)$ are differentiable functions. This way, given an arbitrarily small timestep t_δ , the model respects self-consistency $f_\theta(\mathbf{x}_t, t) = f_\theta(\mathbf{x}_{t'}, t') \forall t, t' \in [t_\delta, T]$ and boundary condition $f(x_{t_\delta}, \delta) = x_{t_\delta}$, allowing CMs to generate high-quality images in a few steps. CMs can be either trained from scratch or distilled from pretrained diffusion models. Moreover, Latent Consistency Models (LCMs) [233], which work on the latent space, are also available. In this work, we use an LCM distilled from Stable Diffusion [308].

9.3 PROPOSED APPROACH

In this section, we describe our pipeline for generating large images by exploiting a pretrained diffusion model at inference-time (see Fig. 9.2). Then, we give the details of our proposed MAD operator, which is the key component of the pipeline for obtaining perceptual and semantic coherence.

JOINT DIFFUSION FOR PANORAMA GENERATION. Diffusion models do not scale well to resolutions and aspect ratios not seen in training [30, 198]. As a result, generating panorama images by simply applying a model trained on squared ones [308] leads to excessive, unrealistic uniformity in the final output and a lack of variability between images generated from the same prompt, as shown in Sec. 9.4.1 and in [198]. Thus, for generating larger images, we average the

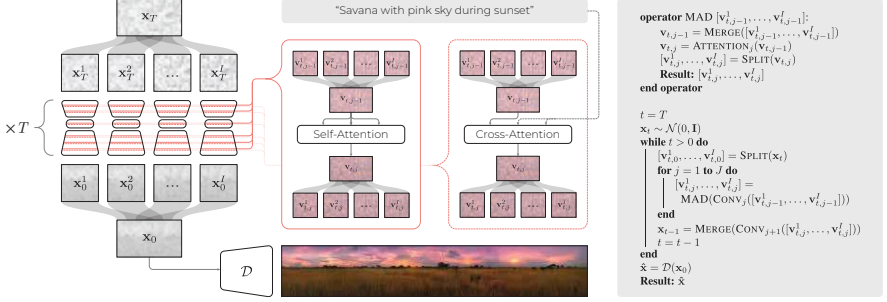


Figure 9.2: Overview of our inference-time pipeline (left) and its pseudo-code (right). During the diffusion process, the image is split into overlapping views, and each is fed to the model separately. Within the attention layers, MAD provides interaction points between the views, enforcing global coherence in the generated panorama.

predictions on multiple views at every reverse process step, similar to [30]. Formally, to generate an image $\hat{\mathbf{x}} \in \mathbb{R}^{H \times W \times C}$, we produce a denoised latent vector $\mathbf{x}_t \in \mathbb{R}^{\frac{H}{n} \times \frac{W}{n} \times C'}$ where n and C' are, respectively, the scaling factor and the channels input of the decoder $\mathcal{D}$ and $t \in [0, T]$, is the timestep in the denoising process of the latent vector. The final image is obtained as $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{x}_0)$. To generate panorama images, we split $\mathbf{x}_t$ into I squared overlapped views $\mathbf{x}_t^i \in \mathbb{R}^{L \times L \times C'}$ where $i \in [1, I]$ denotes the i -th view, and L is its width and height. We set $L = \frac{H}{n}$ for the horizontal panoramas and $L = \frac{W}{n}$ for the vertical ones. For simplicity, we consider the noise predictor module as an alternating sequence of convolutional blocks and attention blocks, plus a final convolutional block. Each of the convolutional blocks, dubbed $\text{CONV}_{j \in [1, J+1]}$, contains one or more convolutional layers, while each of the attention blocks, dubbed $\text{ATTENTION}_{j \in [1, J]}$, contains one or more cross-attention or self-attention layers. In the following, we use the notation $\text{CONV}_j([\dots])$ and $\text{ATTENTION}_j([\dots])$ to indicate that the block is applied to a list of vectors independently. We define as $\mathbf{v}_{t,j} \in \mathbb{R}^{H_j \times W_j \times C_j}$ the feature vector output of the j -th block and as $\mathbf{v}_{t,j}^i \in \mathbb{R}^{L_j \times L_j \times C_j}$ its i -th view. We remark that H_j , W_j , L_j , and C_j depend on depth and type of the j -th block. Finally, we define the SPLIT and MERGE functions. For simplicity, we give their definition by using the latent $\mathbf{x}_t$, but they can be applied also to the $\mathbf{v}_{t,j}$ features. $\text{SPLIT} : \mathbf{x}_t \rightarrow [\mathbf{x}_t^1, \dots, \mathbf{x}_t^I]$ divides its input tensor into multiple overlapping

views. $\text{MERGE} : [\mathbf{x}_t^1, \dots, \mathbf{x}_t^I] \rightarrow \mathbf{x}_t$, merges its input sequence of overlapping views in a single tensor by averaging the overlapped regions. The number of views depends on the extent of the overlap. The more the views overlap, the more seamless the transition between them will be in the final image, at the cost of increasing the inference-time. As a trade-off, we use an overlap of $\frac{3}{4}L$ as in [198].

Note that in previous diffusion approaches for large image generation [30, 198], each of the views is processed independently by the noise prediction model. Then, the outputs of the prediction model at the t -th step, $[\mathbf{x}_t^1, \dots, \mathbf{x}_t^I]$, are combined back into $\mathbf{x}_t$ by averaging the overlapping regions of the views. This averaging procedure allows sharing information between the views. However, in this way, the interactions between the views are infrequent and may be insufficient, which may lead to inconsistent results. Different from these approaches, we introduce interaction points between the views also inside the noise prediction model, which are implemented via our proposed MAD operator.

MERGE-ATTEND-DIFFUSE OPERATOR. We design the MAD operator to be an alternative to the attention layers in the noise prediction module of a pretrained diffusion model when this is used in a joint diffusion setting to generate long images. Through MAD, we enforce consistency between the views while keeping the variability introduced by the convolutional layers, which operate on the views independently. Moreover, note that the MAD operator can be integrated at different stages of the noise prediction network and for a varying number of denoising steps starting from T , defined by a threshold τ . Note that the views are processed jointly by the attention layers where MAD is applied and independently otherwise. In this way, it is possible to choose the prompt-adequate trade-off between uniformity and variability. Specifically, MAD takes as input the feature vectors coming from the $(j-1)$ -th block run over the views, $[\mathbf{v}_{t,j-1}^1, \dots, \mathbf{v}_{t,j-1}^I]$, and feeds the j -th attention block with the merged latent vector $\mathbf{v}_{t,j-1}$. When the operator is integrated with a cross-attention layer, the cross-attention is performed between the merged tensor and the prompt embedding. When the operator is integrated with a self-attention layer, the self-attention is performed on the merged tensor. In this way, the noise estimation over each view is influenced by the information coming from the other views, thus enforcing global coherence.

Finally, the obtained latent vector $\mathbf{v}_{t,j}$ is split back into the views $[\mathbf{v}_{t,j}^1, \dots, \mathbf{v}_{t,j}^I]$. The pseudocode of our approach is in Fig. 9.2.

9.4 EXPERIMENTS

IMPLEMENTATION DETAILS. In our experiments, as the pretrained diffusion models, we use Stable Diffusion 2.0 [308] and Latent Consistency Dreamshaper v7 (distilled from Stable Diffusion 1.5) [233] from HuggingFace [383]. Both models are based on a U-Net [311] architecture, and we apply MAD to all its attention layers. For Stable Diffusion, we perform 50 denoising steps and apply MAD for the first $\tau=15$ steps. For Latent Consistency Dreamshaper, we consider 1, 2, and 4 generation steps and apply MAD with $\tau=2$ over 4 steps for the qualitatives. We empirically found that these values of τ are a good cutoff point, as the global semantic coherence mainly depends on high-level features, defined during the first generation steps. For generating the panoramas, we consider image views of size 512×512 , which correspond to 64×64 latents, and a stride of 16 on such latents, following [198]. For the quantitative evaluation, we use the same prompts as for SyncDiffusion [198] and MultiDiffusion [30], which also exploit our LDM backbone, and report the mean and standard deviation of the performance scores obtained on each prompt. For the comparative analyses, we generate the competitors' images with their official codebases.

EVALUATION SCORES. We adopt the following scores for quantitative evaluation. Mean-CLIP-S (mCLIP) [146], to evaluate the adherence with the textual prompt; Intra-LPIPS (I-LPIPS) [425] and Intra-Style Loss (I-StyleL) [124], to measure the intra-image perceptual coherence; FID [147], KID [44], and Mean-GIQA (mGIQA) [130], to estimate the perceptual similarity and variability with respect to a target distribution. The I-LPIPS and I-StyleL are computed by splitting each image into non-overlapping squared views and then calculating the average LPIPS and Style Loss of all possible combinations of views (*e.g.* for a 512×3072 image, the pairs are 15). The FID, KID, and mGIQA are computed by considering, for each prompt, 500 squared images obtained with the baseline LDM or LCM model as reference and single-view-sized random crops from



Figure 9.3: Long images generated by the considered LDM with MAD, with $\tau=15$, applied in different blocks of the noise prediction model for the prompt *A snowy winter landscape with frosted trees and a frozen lake*. **None** is the setting where MAD is never applied.



Figure 9.4: Long images generated by the considered LDM with MAD applied, in all the U-Net attention layers, up to different numbers of inference steps for the prompt *A herd of Mustang horses crossing a river at sunset*. $\tau=0$ is the setting where MAD is never applied.

9

the generated images, one per image. These crops are also used to compute the mCLIP with the input textual prompt. For comparison, we compute the FID, mGIQA, and KID scores for the baseline LDM and LCM by comparing two random halves of the generated images and the average LPIPS and Style Loss by comparing 1000 random pairs of images for each prompt. Note that these scores capture different, complementary, and somewhat contrasting characteristics of the generated images. As a result, methods that perform well in terms of some scores might achieve worse values for other scores.

9.4.1 RESULTS

VARIETY-UNIFORMITY TRADE-OFF. The noise prediction model of the considered backbone diffusion models is a U-Net, which features multiple attention layers in the downsampling blocks (referred to as *Down* blocks), bottleneck block (referred to as *Mid* block), and upsampling blocks (referred to as *Up* Blocks). As stated above, we apply MAD in all the U-Net blocks (referred to as *All* blocks). Nonetheless, we design our approach to be modular so that it can be applied to selected layers and for a desired number of timesteps, thus obtaining the de-

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$	Runtime
SD	31.63 $\pm$ 1.89	0.74 $\pm$ 0.07	0.74 $\pm$ 0.07	28.31 $\pm$ 10.89	<0.01 $\pm$ 0.13	26.70 $\pm$ 6.90	–
SD-L	32.01 $\pm$ 1.67	0.50 $\pm$ 0.11	0.58 $\pm$ 0.40	87.64 $\pm$ 30.25	76.83 $\pm$ 30.52	27.72 $\pm$ 7.83	88.2 $\pm$ 0.1s
MD	31.77 $\pm$ 2.32	0.69 $\pm$ 0.09	2.98 $\pm$ 2.41	33.52 $\pm$ 12.43	9.04 $\pm$ 4.23	28.54 $\pm$ 7.99	41.8 $\pm$ 0.2s
SyncD	31.84 $\pm$ 2.19	0.56 $\pm$ 0.06	1.39 $\pm$ 1.15	44.60 $\pm$ 18.45	21.00 $\pm$ 11.06	27.17 $\pm$ 6.66	392.4 $\pm$ 0.6s
MAD ($\tau=0$)	31.65 $\pm$ 2.17	0.64 $\pm$ 0.10	2.65 $\pm$ 2.33	34.51 $\pm$ 13.92	9.19 $\pm$ 4.51	27.59 $\pm$ 6.83	41.9 $\pm$ 0.1s
MAD ($\tau=5$)	31.86 $\pm$ 2.22	0.59 $\pm$ 0.10	2.07 $\pm$ 1.31	38.10 $\pm$ 13.71	13.75 $\pm$ 4.00	28.32 $\pm$ 7.64	42.9 $\pm$ 0.2s
MAD ($\tau=15$)	32.03 $\pm$ 2.29	0.56 $\pm$ 0.10	1.90 $\pm$ 1.32	48.52 $\pm$ 17.14	27.15 $\pm$ 9.10	28.32 $\pm$ 7.76	44.3 $\pm$ 0.2s
MAD ($\tau=25$)	32.15 $\pm$ 2.25	0.53 $\pm$ 0.10	1.43 $\pm$ 1.04	61.76 $\pm$ 23.73	43.31 $\pm$ 17.41	28.10 $\pm$ 7.84	45.6 $\pm$ 0.2s
MAD ($\tau=40$)	32.16 $\pm$ 1.95	0.50 $\pm$ 0.10	0.88 $\pm$ 0.61	86.20 $\pm$ 37.23	76.15 $\pm$ 37.73	27.59 $\pm$ 7.61	48.3 $\pm$ 0.1s
MAD ($\tau=50$)	32.14 $\pm$ 1.72	0.49 $\pm$ 0.10	0.71 $\pm$ 0.47	98.01 $\pm$ 43.64	91.51 $\pm$ 49.95	27.05 $\pm$ 7.28	49.8 $\pm$ 0.1s

Table 9.1: Quantitative comparison on 512×3072 panorama generation using the LDM. I-StyleL, KID, and mGIQA values are scaled by 10^3 .

sired variety-uniformity trade-off. To explore this aspect, in Fig. 9.3, we show 512×3072 images generated by applying MAD in all the attention layers of different blocks in the LDM backbone, thus varying the number of interaction points and the stage at which these are performed. Fig. 9.3 shows different levels of global coherence, ranging from a very coherent scene when MAD is applied on all the blocks (which is our selected setting) to a more varied image when MAD is applied only in the Mid block. Quantitative results on the application of MAD at different stages of the U-Net are reported in Chapter D. Moreover, in Fig. 9.4, we show 512×3072 images generated by applying MAD up to a different number of inference steps, defined by the threshold τ (recall that $\tau=15$ in our selected setting). With a low threshold, the images are not globally coherent because the views interactions are insufficient. With increasing thresholds, the images become more and more coherent. We give a quantitative analysis in Sec. 9.4.1 and further substantiate this claim in Chapter D. Both Figs. 9.3 and 9.4 reflect the trade-off between variety and uniformity. When the views interactions are reduced, either by applying MAD to a few attention layers or for a few timesteps, the resulting image contains several perceptually varying views. When such interactions between the views increase, the global semantic layout and the perceptual patterns are more coherent.

QUANTITATIVE COMPARISON. We compare the performance of our method on 512×3072 images with different values of τ against applying Stable Diffu-

sion to generate the whole panorama image directly (which we refer to as SD-L) and the joint diffusion methods MultiDiffusion (MD) [30] and SyncDiffusion (SyncD) [198], which are the closest to our approach. For numerical reference, we also compute the scores on the squared images from Stable Diffusion (SD). The results of this analysis are reported in Sec. 9.4.1. From the table, we observe that MAD achieves the best or second-best performance on all the scores, depending on the parameter τ , which determines how many times it is applied in the denoising process. In particular, with respect to perceptual and semantic coherence of the generated images, we observe that the more MAD is applied, the more the images are compatible with the textual prompt, as measured by the mCLIP. The obtained values are always superior to those obtained with the competitors and slightly inferior w.r.t. SD-L only in the case of a few application steps. Moreover, the I-LPIPS and I-StyleL values indicate that MAD also leads to increasingly perceptually coherent images the more it is applied. It is worth noting that MAD performs on par or better than SyncD in terms of the I-LPIPS, although SyncD specifically entails optimizing this score. As for FID, KID, and mGIQA, note that these measure different aspects of the generated images w.r.t. the other three considered scores. In particular, I-LPIPS and I-StyleL evaluate the perceptual similarity within each image, while FID, KID, and mGIQA the distance of the distributions of random crops from the images generated by the approach under evaluation and SD. If a long image is composed of many perceptually varying views (as those by MD), it is more likely that its crops resemble the squared SD reference images. However, this characteristic does not ensure perceptual coherence between the views in the same long image (measured by I-LPIPS and I-StyleL) nor adherence with the textual prompt (measured by the mCLIP). In summary, the results in Sec. 9.4.1 suggest that our approach is able to generate semantically and perceptually coherent panoramas (which is the goal of our work), regardless of the similarity of their single crops to SD squared images.

QUALITATIVE COMPARISON. Fig. 9.5 shows a qualitative comparison among the joint diffusion methods on two of the six evaluation prompts (more qualitatives are provided in Chapter D). We generate the images using the same seed for each comparison to ensure consistency. As we can see, MD generates images

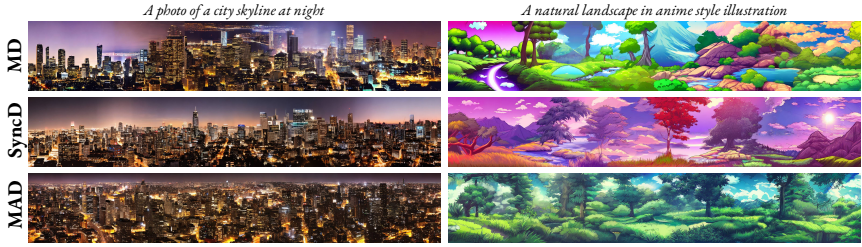


Figure 9.5: Qualitative results of our approach, MD, and SyncD. MD generates panoramas with smooth but incoherent transitions. SyncD increases perceptual coherence but lacks semantic coherence. MAD generates perceptually and semantically coherent images.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$
512$\times$1024						
MD	31.73 $\pm$ 2.22	0.66 $\pm$ 0.09	2.57 $\pm$ 1.97	30.66 $\pm$ 11.79	5.24 $\pm$ 3.04	28.17 $\pm$ 7.54
SyncD	31.71 $\pm$ 2.01	0.53 $\pm$ 0.06	1.09 $\pm$ 0.77	40.35 $\pm$ 16.43	13.20 $\pm$ 7.61	26.41 $\pm$ 6.38
MAD	31.82 $\pm$ 2.09	0.55 $\pm$ 0.10	1.47 $\pm$ 0.89	37.53 $\pm$ 14.72	12.63 $\pm$ 5.74	27.58 $\pm$ 7.04
512$\times$2048						
MD	31.77 $\pm$ 2.14	0.69 $\pm$ 0.09	2.96 $\pm$ 2.41	33.07 $\pm$ 12.38	8.58 $\pm$ 3.99	28.33 $\pm$ 7.79
SyncD	31.77 $\pm$ 2.14	0.55 $\pm$ 0.06	1.39 $\pm$ 1.19	43.33 $\pm$ 17.98	18.77 $\pm$ 10.19	27.08 $\pm$ 6.65
MAD	31.97 $\pm$ 2.23	0.56 $\pm$ 0.10	1.79 $\pm$ 1.21	44.26 $\pm$ 16.26	21.80 $\pm$ 7.54	28.13 $\pm$ 7.61
512$\times$4096						
MD	31.64 $\pm$ 2.18	0.64 $\pm$ 0.10	2.18 $\pm$ 1.16	34.87 $\pm$ 13.56	9.12 $\pm$ 3.72	27.89 $\pm$ 7.08
MAD	32.05 $\pm$ 2.30	0.56 $\pm$ 0.10	1.91 $\pm$ 1.35	49.61 $\pm$ 16.98	28.34 $\pm$ 8.05	28.33 $\pm$ 7.94

Table 9.2: Quantitative comparison on the generation of images at different aspect ratios with the LDM. The I-StyleL, KID, and mGIQA are scaled by 10^3 .

with smooth but incoherent transitions. SyncD generally enhances the perceptual coherence of the panoramas but does not address semantic coherence. For instance, in the anime-style landscape of the last row, we can see a lake morphing into clouds. Conversely, our approach generates perceptually and semantically coherent panoramas with these prompts.

DIFFERENT ASPECT RATIOS. We also evaluate our approach for the generation of images with different aspect ratios. Tab. 9.2 shows the performance of our proposed method and the joint diffusion competitors when generating 512×1024 , 512×2048 , and 512×4096 images. As we can see, MAD is consistently better

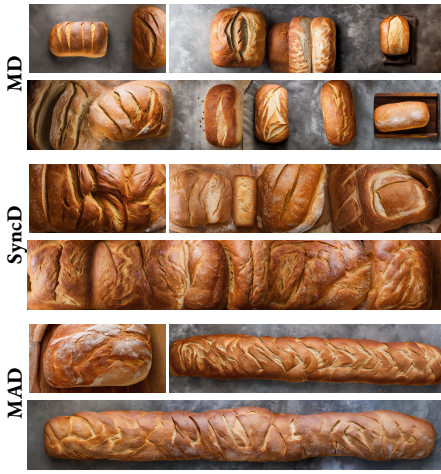


Figure 9.6: Qualitative comparisons between our approach with LDM, MD, and SyncD at different aspect ratios for the prompt *Top view of a loaf of bread*.

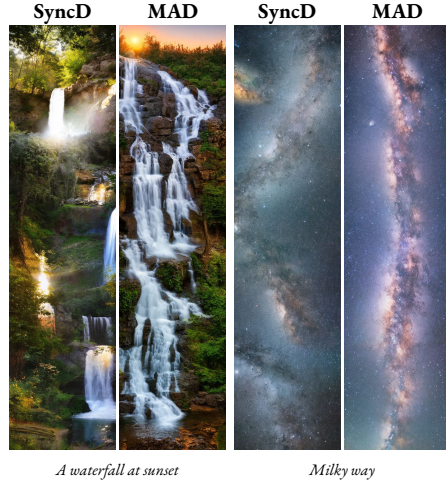


Figure 9.7: Qualitative comparison on vertical 512×2048 images between our approach applied to the LDM and SyncD for prompts requiring semantic coherence.

in terms of mCLIP and competitive in I-LPIPS. The visual quality of the generated images can also be appreciated in Fig. 9.6, where we report 512×1024 , 512×2048 , and 512×3072 images for the same prompt and seed. It can be observed that our approach can generate perceptually and semantically coherent images at different aspect ratios, while the competitors struggle to achieve this capability even in shorter images. We also compare our method with SyncD on long vertical images in Fig. 9.7 by generating 2048×512 images. We chose to compare only with SyncD because, compared to MD, it enforces perceptual coherence. Nonetheless, it struggles to maintain semantic coherence and to generate long images with subjects that should have a beginning and an end, such as the waterfall. Please refer to Chapter D for more qualitative results at different aspect ratios.

JOINT DIFFUSION *VS* DIRECT INFERENCE. When generating long images at inference-time directly with models trained on squared images, a train-test discrepancy can occur, which is more evident as the desired output image width increases. This can be seen in Fig. 9.8, where we compare the FID and KID on long

image generation for MAD and SD-L. As the width increases, the generation performance of SD-L is increasingly impacted. This effect is reduced with our approach. Another issue with directly generating long images by simply applying a model trained on squared images is that this can lead to excessive, unrealistic uniformity in the final output and a lack of variability between images generated from the same prompt as their size increases [198]. To explore this aspect, we run the t-SNE analysis [376] reported in Fig. 9.9 and in Chapter D. Specifically, we extract inception features from square images by SD and compute their t-SNE representations [376]. Then, we consider all the squared crops obtainable from 512×5120 images by SD-L and MAD and represent all their inception features in the same space using openTSNE [216, 282]. As we can see, images generated using MAD (represented by the red dots) cover a bigger portion of the SD distribution (the grey circles) compared to SD-L (the blue dots). Note that in Fig. 9.9, the dark areas represent an overlap between MAD and SD-L. The t-SNE analysis and the perceptual and semantic scores in Sec. 9.4.1 suggest that MAD leads to both intra-image coherence and inter-image variability.

RUNTIME. We also analyze the computation times of our approach and the competitors, all run on the same NVIDIA RTX A5000-24GB for generating 512×3072 images with the LDM backbone (reported in Sec. 9.4.1). MAD runtime ranges from ~ 43 s to ~ 50 s, depending on the threshold τ , which is comparable to MD (41.9s). The small computational overhead is due to the multiple merging and splitting operations and the larger tensors on which self- and cross-attention are performed. Compared to SyncD, which runs in 391.9 ± 0.1 s, MAD is more than $8\times$ faster for not requiring the expensive gradient computation. Moreover, in Fig. 9.10, we report the runtime analysis at increasing output width w.r.t. performing direct inference (SD-L) and never applying MAD in joint diffusion ($\tau=0$). Note that convolutions and attentions are linear w.r.t. the output width when applied to split views and quadratic when applied to the merged latent. MAD entails convolutions on split views and attentions on merged latents only for $t < \tau$ (on splits afterward) and thus the computation grows slower than for SD-L.

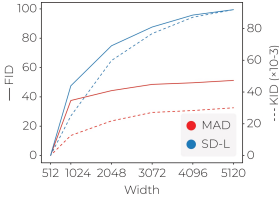


Figure 9.8: Mean FID and KID on images of different widths by MAD and SD-L for the prompts used in the quantitative analysis.

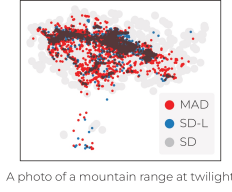


Figure 9.9: T-SNE analysis of crops from images by MAD and SD-L w.r.t. squared images from SD for the same prompt.

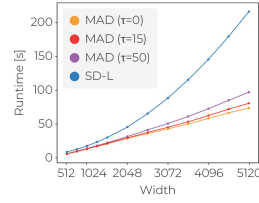


Figure 9.10: Inference time at increasing output widths for SD-L and MAD: for $\tau=0$, MAD is not applied, $\tau=50$ is the upper bound.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$
SD	32.45	0.67	12.76	49.32	<0.01	12.98
SD-L	31.89	0.52	0.73	68.03	6.61	12.58
MD	32.46	0.63	4.81	49.41	0.42	13.22
SyncD	32.34	0.55	4.24	53.52	1.11	12.98
MAD	32.47	0.58	3.89	54.44	1.28	13.03

Table 9.3: Quantitative comparison on 512×3072 GPT-1k panorama generation using the LDM. I-StyleL, KID, and mGIQA values are scaled by 10^3 .

COMPLEX SCENES. Moreover, we evaluate on the generation of 512×3072 images from a set of 1000 prompts describing complex scenes, which we automatically obtain with ChatGPT-4 (referred to as GPT1k). We argue that such scenes allow to better highlight the challenges of the panorama generation task. The results are reported in Table 9.3 and highlight the superiority of our approach over the competitors. They also demonstrate that the task is far from being solved, and thus, we release the GPT1k set for fostering its exploration. Note that the qualitative in this chapter (excluding those in Fig. 9.5) are obtained with prompts in GPT1k. For more qualitative comparisons, see Chapter D.

USER STUDY. Considering the sometimes subjective differences in image quality and the partial nature of quantitative evaluations, we also conduct a user study to evaluate the coherence of the generated images. Participants, presented with pairs of images generated with our approach, MD, and SyncD, were asked which was most coherent and best matched the textual prompt.

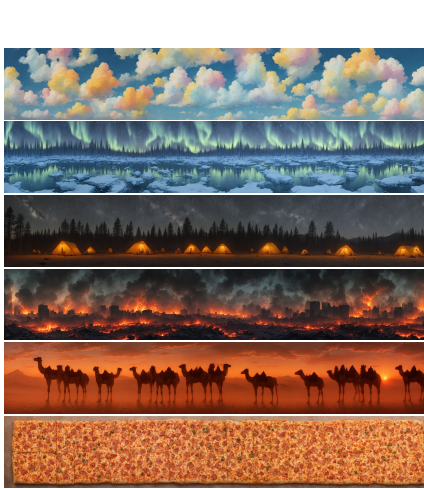


Figure 9.11: Qualitative results on different prompts of our approach by exploiting LCD as backbone.

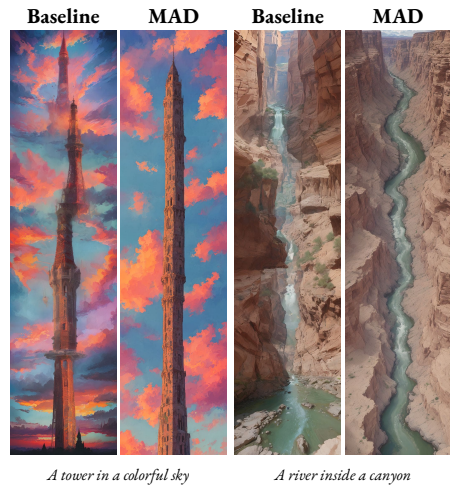


Figure 9.12: Qualitative comparison on vertical images with LCD. **Baseline** entails never applying MAD ($\tau=0$).

RESULTS WITH LCMs. Finally, we apply our method with different τ to Latent Consistency Dreamshaper (LCD) at varying generation steps with resolution 512×3072 . The quantitative results are in Sec. 9.4.1, where we also report the performance obtained by applying LCD to generate the whole panorama image directly (referred to as LCD-L) and the scores on the squared images from LCD, for reference. Applying MAD improves the mCLIP, generating more coherent images (see also the qualitatifs in Figs. 9.11 and 9.12 and in Chapter D).

PLUG&PLAY APPLICATIONS. Note that our MAD operator can be used as it is in settings such as tight and rough region-based generation, and conditional image generation (with Canny edge map guidance). since it is applied to the attention layers of the noise prediction UNet. The visual quality of the resulting images, shown in Fig. 9.13, is on par with that of the generated panoramas (which is our main focus), even for guidance spanning multiple views of the large image.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$
1 Inference Step						
LCD	28.10 $\pm$ 1.65	0.42 $\pm$ 0.06	0.23 $\pm$ 0.12	26.95 $\pm$ 3.92	<0.01 $\pm$ 0.33	42.52 $\pm$ 3.97
LCD-L	29.50 $\pm$ 1.48	0.40 $\pm$ 0.08	0.21 $\pm$ 0.18	67.54 $\pm$ 15.53	66.86 $\pm$ 21.22	32.21 $\pm$ 4.50
MAD ($\tau=0$)	28.42 $\pm$ 1.75	0.45 $\pm$ 0.06	0.25 $\pm$ 0.17	54.65 $\pm$ 24.64	50.56 $\pm$ 33.30	32.27 $\pm$ 4.38
MAD ($\tau=1$)	29.01 $\pm$ 1.66	0.40 $\pm$ 0.07	0.37 $\pm$ 0.28	72.43 $\pm$ 23.75	75.47 $\pm$ 33.24	32.23 $\pm$ 4.95
2 Inference Steps						
LCD	30.57 $\pm$ 1.81	0.53 $\pm$ 0.06	1.43 $\pm$ 0.48	26.82 $\pm$ 11.51	<0.01 $\pm$ 0.32	44.30 $\pm$ 14.31
LCD-L	30.77 $\pm$ 2.09	0.47 $\pm$ 0.06	0.56 $\pm$ 0.29	49.98 $\pm$ 25.20	44.29 $\pm$ 26.88	35.45 $\pm$ 11.94
MAD ($\tau=0$)	30.75 $\pm$ 2.16	0.52 $\pm$ 0.05	0.84 $\pm$ 0.35	27.95 $\pm$ 12.07	13.39 $\pm$ 7.40	35.78 $\pm$ 10.70
MAD ($\tau=1$)	30.97 $\pm$ 2.15	0.50 $\pm$ 0.06	0.85 $\pm$ 0.34	35.69 $\pm$ 17.88	23.74 $\pm$ 15.82	35.32 $\pm$ 11.49
MAD ($\tau=2$)	30.87 $\pm$ 2.18	0.47 $\pm$ 0.06	0.60 $\pm$ 0.23	52.88 $\pm$ 26.12	48.03 $\pm$ 27.75	35.48 $\pm$ 11.77
4 Inference Steps						
LCD	31.37 $\pm$ 1.60	0.55 $\pm$ 0.05	1.84 $\pm$ 0.81	29.05 $\pm$ 13.59	<0.01 $\pm$ 0.23	45.00 $\pm$ 12.97
LCD-L	31.30 $\pm$ 1.63	0.50 $\pm$ 0.06	0.58 $\pm$ 0.24	55.52 $\pm$ 32.82	51.71 $\pm$ 37.79	35.44 $\pm$ 12.94
MAD ($\tau=0$)	31.36 $\pm$ 1.83	0.55 $\pm$ 0.05	0.90 $\pm$ 0.35	31.35 $\pm$ 15.42	13.27 $\pm$ 10.05	35.44 $\pm$ 12.38
MAD ($\tau=2$)	31.48 $\pm$ 1.87	0.52 $\pm$ 0.05	0.71 $\pm$ 0.26	38.77 $\pm$ 19.85	27.41 $\pm$ 16.19	35.23 $\pm$ 13.15
MAD ($\tau=4$)	31.41 $\pm$ 1.94	0.48 $\pm$ 0.06	0.46 $\pm$ 0.19	62.82 $\pm$ 36.24	61.69 $\pm$ 42.07	35.06 $\pm$ 13.61

Table 9.4: Quantitative comparison on 512×3072 panorama generation with the LCM for different numbers of inference steps. I-StyleL, KID, and mGIQA are scaled by 10^3 .



Figure 9.13: Plug&Play applications of MAD

9.5 CONCLUSIONS AND DISCUSSION

In this work, we have presented the **Merge-Attend-Diffuse** operator, which can be applied to the attention layers of a pretrained convolutional-attentive diffusion model for zero-shot generation of long images that are both perceptually and semantically coherent. We have conducted extensive experimental analysis, whose results demonstrate the effectiveness of the proposed approach especially in terms of visual quality and adherence to the input prompt.

Similar to other inference-time joint diffusion approaches [30, 198], a limitation of our method is that it relies on the diffusion paths of the base model for realistic results and can be hindered by a bad seed or inappropriate text prompts. Performing the attention operations on the merged tensor mitigates this issue. For example, in Figs. 9.1, 9.3 and 9.4, the incoherent horizon lines of independent diffusion paths (in the top images) are corrected by applying MAD. Further discussion can be found in Chapter D.

10

Alfie: Democratising RGBA Image Generation With No \$\$\$

DESIGNS and artworks are ubiquitous across various creative fields, requiring graphic design skills and dedicated software to create compositions that include many graphical elements, such as logos, icons, symbols, and art scenes, which are integral to visual storytelling. Automating the generation of such visual elements improves graphic designers' productivity, democratizes and innovates the creative industry, and helps generate more realistic synthetic data for related tasks. These illustration elements are mostly RGBA images with irregular shapes and cutouts, facilitating blending and scene composition. However, most image generation models are incapable of generating such images and achieving this capability requires expensive computational resources, specific training recipes, or post-processing solutions. In this work, we propose a fully-automated approach for obtaining RGBA illustrations by modifying the inference-time behavior of a pre-trained Diffusion Transformer model, exploit-

This Chapter is related to the publication “F. Quattrini *et al.*, Alfie: RGBA Image Generation With No \$\$\$, ECCVW 2024” [3]. See the list of Publications on page 315 for more details.

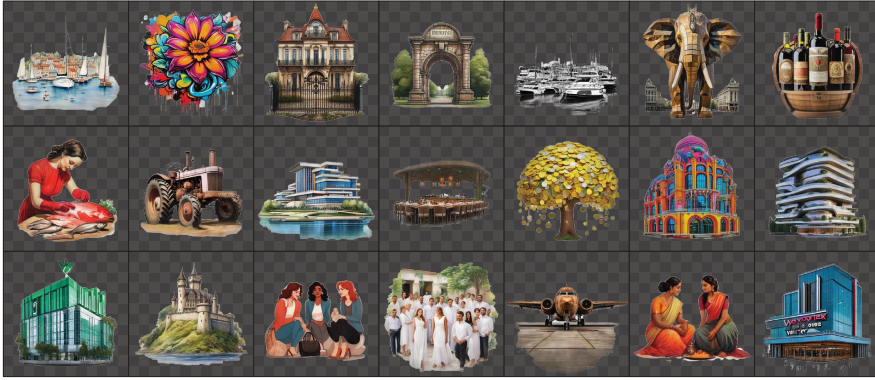


Figure 10.1: We propose a fully-automated pipeline to generate RGBA illustrations by adapting the inference-time behavior of a Diffusion Transformer model.

ing the prompt-guided controllability and visual quality offered by such models with no additional computational cost. We force the generation of entire subjects without sharp croppings, whose background is easily removed for seamless integration into design projects or artistic scenes. We show with a user study that, in most cases, users prefer our solution over generating and then matting an image, and we show that our generated illustrations yield good results when used as inputs for composite scene generation pipelines. We release the code of our approach ^{*}.

10.1 INTRODUCTION

We are surrounded by a lot of visual content, including visually-rich documents in everyday life, such as pamphlets and banners, technical or scientific papers, business reports, storybooks, or novels with intricate and meaningful illustrations. Much effort has been dedicated to generating and integrating multiple design elements into visually-rich documents [432, 202, 409, 156]. Note that, when composing visually-rich documents (*e.g.* slides for a presentation, banners, cards, pamphlets, social media posts), it is not uncommon that designers, content creators, and even non-expert users would like to insert images that should inte-

^{*}<https://github.com/aimagelab/Alfie>

grate and blend nicely with the rest. Those users rely more and more often on AI-generated images for their designs. Generative models, especially prompt-driven diffusion models, have achieved impressive performance, and the increasing availability of user-friendly interfaces also eases non-experts' use of them. Moreover, these models allow for obtaining control over the generated content and producing images free from copyright. Nonetheless, when integrating an illustration inside a design, the designer often wants just the subject of interest to be superimposed on an existing background (possibly patterned or consisting of another image).

Large-scale image generation models have been widely adopted and deeply impacted visual arts. Still, layered image generation has not been extensively tackled, mainly because of the lack of training data. While RGB image generation can leverage open datasets of billions of samples [321], RGBA image datasets are of limited size [307, 405, 353], the biggest being the recently proposed MAG-ICK [55] which contains only 150k elements. These datasets are commonly used to train matting models to estimate the alpha channel of a given image [416, 353, 405, 155, 415]. Another solution to obtain RGBA images consists in directly predicting all four channels at once by fine-tuning a pre-trained generative model [422]. However, these solutions are costly for the required training resources.

To overcome these limitations, in this work, we propose *Alfie*, a pipeline to obtain high-quality, prompt-driven illustrations to be seamlessly integrated into any design or document with limited to no editing effort by the user (see Fig. 10.1). We exploit a pre-trained Diffusion Transformer model [271, 67], *i.e.* the freely-available PixArt- Σ [66], and modify its diffusion process in two ways. First, we mask and combine two latents (one for the subject and one for the background) in order to force the generation of the subject in the center of the canvas and with no sharp crops so that these can be realistic in a design or document. Second, we use the informative cross- and self-attention maps [358, 78] computed during the generation process to extract the foreground regions and estimate the alpha channel values. We opt for a Diffusion Transformer for the greater flexibility of such models in generating images with different sizes and aspect ratios compared to

standard U-Net-based diffusion models. This arguably makes them a more suitable starting point for the task at hand. Moreover, they are currently state-of-the-art both in generation quality and computational requirements [271, 67, 87, 66]. PixArt- Σ [66] (our baseline model) is composed of 0.6B parameters, compared to the 2.6B of Stable Diffusion XL [281]. We quantitatively and qualitatively assess the capability of our approach to generate subjects that are fully contained within the image and its capability to adhere to the prompt. We analyze the cross- and self-attention maps computed during the generation process, exploring various alternatives for estimating the α (transparency) channel. Furthermore, we present a user study, which shows that users mostly prefer our method over generating and then matting (63% of the time). Finally, we demonstrate the versatility of our generated illustrations by integrating them into a scene composition pipeline, achieving results of similar quality to those obtained by using illustrations from Adobe Stock. To facilitate further research on this task, we release the code of our approach and the evaluation setup[†].

10.2 PRELIMINARIES

DIFFUSION MODELS. By diffusion models [333, 149, 338, 334] we refer to a class of generative probabilistic models trained to transform Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ into samples belonging to a certain data distribution $\mathbf{x}_0 \sim q$ in T steps. This is obtained by learning to approximate the data distribution q . To this end, during the so-defined forward process, Gaussian noise is injected into the data to transform the data distribution into the marginal distribution, *i.e.*

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}),$$

where the parameters (α_t, σ_t) define a differentiable noise schedule ensuring that $q(\mathbf{x}_t) \approx \mathcal{N}(0, \mathbf{I})$. Then, the diffusion models framework entails a reverse process in which the model is trained to denoise $\mathbf{x}_T \sim q(\mathbf{x}_t | \mathbf{x}_0)$. To this end, $\mathbf{x}_0$ is predicted iteratively by estimating $\mathbf{x}_{t-1}$ starting from $\mathbf{x}_t$. The most common approach, proposed in [149], entails using a parameterization based on the predic-

[†]<https://github.com/aimagelab/Alfie>

tion of the noise ε for sampling: both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are parametrized as a combination of $\mathbf{x}_0$ and ε , scheduled according to the noise schedule. As a result, the model is trained to estimate the noise ε by optimizing

$$\mathbb{E}_{q(\mathbf{x}_0)} [\|\varepsilon_\theta(\mathbf{x}_t, t) - \varepsilon\|_2^2].$$

LATENT DIFFUSION MODELS. When working with high-resolution image data, training diffusion models in the pixel space is very computationally heavy. To face this issue, latent diffusion models [308, 281] have been introduced. These models work with latent vectors of an autoencoder (*e.g.* a VQ-VAE [375] or a VQ-GAN [109]) and thus are more efficient than models working in the pixel-space in terms of GFlops [308, 281, 271, 67]. Here, we rely on the pre-trained, latent Diffusion Transformer model PixArt- Σ [66].

CONDITIONAL GENERATION. The generation process of diffusion models can be conditioned on a guidance signal. The most popular diffusion models to date are text-to-image models that can be conditioned on a natural language prompt, embedded by the text encoder of a pre-trained multimodal model (*e.g.* CLIP [297] or T5 [299]) in a vector $\mathbf{e}$. The conditioning is achieved by performing cross-attention with $\mathbf{e}$ inside the noise estimation network. Note that, in this work, we use as backbone a diffusion model trained with the Classifier-Free Guidance [151] conditioning strategy, in which the noise prediction $\hat{\varepsilon}_\theta$ is obtained by combining the conditional prediction $\varepsilon_\theta(\mathbf{x}_t, \mathbf{e}, t)$ with the unconditional prediction $\varepsilon_\theta(\mathbf{x}_t, \emptyset, t)$, where $\emptyset$ is the embedding vector of the null prompt, with weight s as

$$\hat{\varepsilon}_\theta(\mathbf{x}_t, \mathbf{e}, t) = \varepsilon_\theta(\mathbf{x}_t, \emptyset, t) + s(\varepsilon_\theta(\mathbf{x}_t, \mathbf{e}, t) - \varepsilon_\theta(\mathbf{x}_t, \emptyset, t)).$$

DIFFUSION TRANSFORMERS. Diffusion models originally featured a U-Net-like model as a noise estimator, combining convolutions and attention. The seminal work [271] proposes to replace the U-Net with a multi-block Diffusion Transformer model (DiT) to achieve better performance and scalability. The noise estimation Transformer takes as input a sequence of tokens obtained from squared patches of the latent vectors added to a positional embedding. At each block of the Transformer, the generation is conditioned on the timestep and the class. After the last block, the tokens are decoded and rearranged into the final image

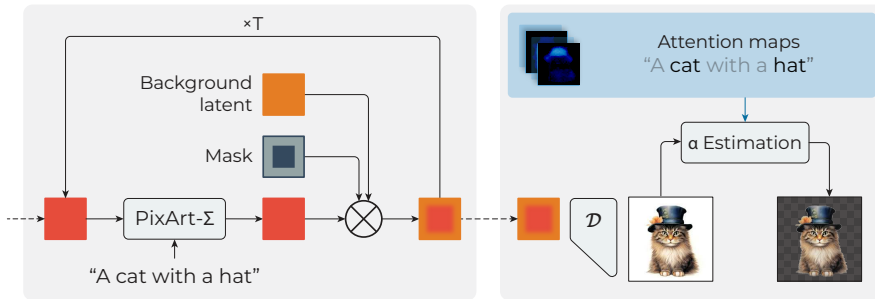


Figure 10.2: Schematic representation of our fully-automated prompt-guided pipeline to obtain RGBA illustrations. The core element is a diffusion model, for which we devise an inference-time adaptation strategy aimed at making the generated images illustration-like. Then, we process the generation attention maps to estimate the α channel.

spatial dimension. Subsequent works [67, 66] adapt the class-guided Diffusion Transformer to work with text guidance. In this work, we exploit the recently-proposed PixArt- Σ Diffusion Transformer [66], which is directly conditioned on the textual prompt.

10.3 INFERENCE-TIME ILLUSTRATION GENERATION

We propose a fully automated pipeline to obtain a high-quality RGBA illustration from a textual prompt. We argue that such illustration should have the following characteristics:

1. Contain the subjects specified in the prompt;
2. Be fully contained in the canvas without crops;
3. Have a precise α channel.

In our proposed inference-time method, we tackle all three aspects. In particular, we exploit the generative power of a pre-trained DiT (1), altering its inference-time behavior so that it always generates entire subjects (2) and extracting and cleaning an α channel estimate (3). Our approach is summarized in Fig. 10.2. We denote the RGB image as $\hat{x} \in \mathbb{R}^{h \times w \times 3}$, with values in range $[-1, 1]$. The transparency channel is denoted as $\alpha \in \mathbb{R}^{h \times w \times 1}$, with values in range $[0, 1]$. The complete RGBA image is obtained by concatenating the RGB image with its α channel and has shape $h \times w \times 4$.

In particular, we propose to obtain the α channel by exploiting the relevant foreground subject’s cross-attention maps and the pixels’ self-attention map relative to the foreground subject. The rationale is that the cross-attention map will have higher values for the pixels of the subject than those of the background, providing a coarse localization of the subject. On the other hand, the self-attention values will help render the texture and appearance of the foreground subject. In fact, when a foreground pixel is relative to an object of solid material, it will attend mainly to neighboring pixels of the same object, resulting in a cohesive appearance. In this case, all the self-attention weights for the pixels of that object will be high. Conversely, the appearance of foreground pixels of objects in a translucent, see-through material (*e.g.*, water, fire, glass) will also depend on the background. Therefore, that pixel will attend to both the pixels of the translucent object and the background. As a result, the self-attention weights for the pixels of that object will be lower than in the case of solid objects. This behavior is in line with what the α channel represents. For this reason, we propose to use the cross- and self-attention maps to define the α channel of our generated RGBA illustrations.

CENTERING THE SUBJECT. Diffusion models are usually trained on billions of images [321] and learn to generate all kinds of styles, from natural to artworks. While some works analyze the disentanglement capability of such models [396], recent works argue that the low quality of the captions in most large-scale datasets negatively affects the generation capabilities and train on LLaVA [220]-refined captions [67, 66]. Still, complete disentangling has yet to be achieved and the models have difficulties with generating meta-descriptions or combinations of subjects and styles not seen in the training data. Note that by meta-descriptions, we define indications on the characteristics of the image, such as the background color or the fact that the full subject should be contained.

When generating illustrations, it is important that the subjects are wholly contained within the image canvas and are not cut or cropped. To achieve this, we chose to employ a method inspired by [23, 30], where we jointly generate the illustration (described by the textual prompt) and a uniform background, and we blend them together at each denoising step. Formally, to generate an image $\mathbf{x}_0$ with the text prompt $\mathbf{e}$, we first define a squared mask $\mathbf{m}$, covering the inner

area of the canvas. Then, we start from the Gaussian noise $\mathbf{x}_t$ at the timestep $t \in [0, T]$ and duplicate it into a foreground latent $\mathbf{x}_{t,fg}$ and a background latent $\mathbf{x}_{t,bg}$, assigning a background textual prompt $\mathbf{e}_{bg}$ to the respective latent. After the denoising step, we merge the obtained predictions into

$$\mathbf{x}_{t-1} = \mathbf{x}_{t-1,fg} \cdot \mathbf{m} + \mathbf{x}_{t-1,bg} \cdot (1 - \mathbf{m}).$$

This process constrains the illustration subjects to be generated inside the mask while ensuring that they are not cropped. After the denoising chain, the image is obtained from the final latent $\mathbf{x}_0$ by using the decoder $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{x}_0)$.

ISOLATING THE SUBJECT. During each forward process generation step t , each layer l performs cross-attention between the latent $\mathbf{z}_{t,l}$ and the encoded prompt $\mathbf{e}$ to guide the subject creation and self-attention to correlate image features. Following DAAM [358] and DAAM-I2I [78], we extract the cross- and self-attention maps. Formally, for a given t and l , the cross-attention map $\mathcal{A}_C^{t,l} \in [0, 1]^{b \times w \times N}$ will represent the relation between each pixel and each of the N prompt tokens, while the self-attention map $\mathcal{A}_S^{t,l} \in [0, 1]^{bw \times bw}$ will represent the relation between each pixel to all the other latent pixels:

$$\begin{aligned} \mathcal{A}_C^{t,l} &:= \text{softmax} \left(Q_{\mathbf{z}_{t,l}} K_{\mathbf{e}}^T / \sqrt{d} \right), \\ \mathcal{A}_S^{t,l} &:= \text{softmax} \left(Q_{\mathbf{z}_{t,l}} K_{\mathbf{z}_{t,l}}^T / \sqrt{d} \right), \end{aligned}$$

where d is the feature size of the layer l .

Note that by performing subject-centering and classifier-free guidance, we have a batch size of four, respectively representing the latent for the null ($\emptyset$) and text prompts ($\mathbf{e}$) for the background and the subject. We discard all the attention maps except those corresponding to the prompt-guided subject latent. We reserve further investigation on integrating the background and null-prompt guided attention maps for future work. Since diffusion models define the coarse image layout during the first steps of the denoising chain and later define the details, we keep the maps of the last ten out of thirty timesteps to obtain more precise localization. Finally, we average across timesteps, layers, and attention heads to obtain the global maps $\mathcal{A}_C \in [0, 1]^{b \times w \times N}$ and $\mathcal{A}_S \in [0, 1]^{bw \times bw}$.

Once the final RGB image $\mathbf{x}_0$ is obtained, we isolate the cross-attention maps

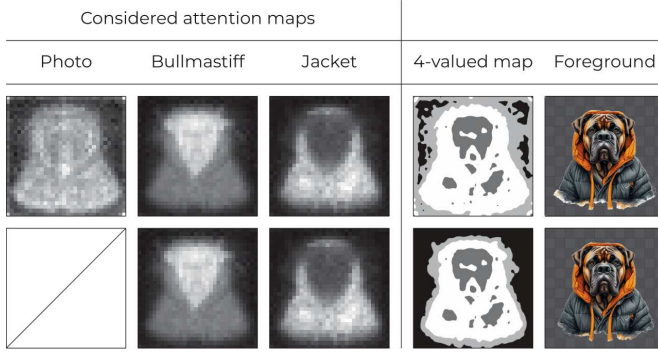


Figure 10.3: Cross-attention map analysis of the prompt *A photo of a bullmastiff with a jacket*. Left to right: maps of the three prompt nouns, candidate region mask computed w/ and w/o the generic noun *photo*, and foreground extraction results.

of the noise estimator relative to the nouns in the input prompt $\mathbf{e}$. The list of nouns $\mathbf{n} = [\mathbf{n}_1, \dots, \mathbf{n}_N]$ is obtained by analyzing the prompt with the standard tokenizer from the NLTK library [45] and excluding generic, uninformative nouns such as *image*, *photo*, and *picture*. We release the complete list of excluded nouns in the provided repository. Empirically, we found that the cross-attention maps associated with these nouns refer to generic areas of the image and not always to the subject, hindering the final quality. This is in line with what is done in [181, 358], where the authors performed analyses on the cross-attention maps of text-to-image convolutional-attentive diffusion models and avoided spurious correlations between text tokens and feature space by removing this type of nouns. To showcase the effect of the generic nouns exclusion, in Fig. 10.3 we visualize the attention maps of different nouns and the effect on the obtained illustration of excluding generic nouns for the prompt *A photo of a bullmastiff with a jacket*. As we can see, the attention maps related to meaningful object nouns (e.g., *dog*, *jacket*) have high activation values in the spatial locations aligned with the specified subjects. In contrast, the attention map of the generic noun *photo* has high activation values also in border regions. In the second-last column, we display the mask of the candidate foreground regions that we use to obtain the guidance masks for a foreground extraction algorithm, namely GrabCut[313]. In particular, we average the maps of each subject and normalize the values be-

tween 0 and 1, obtaining the foreground cross-attention map $\overline{\mathcal{CA}}_{fg}$ and quantizing it into a mask. For foreground extraction, this mask has four values: background (black), probable background (light gray), probable foreground (white), and sure foreground (dark gray). As we can see, this mask is much cleaner when not considering generic, non-descriptive nouns, and the final illustration is of higher quality. In summary, as also shown in [358, 144] for U-Net-based diffusion models, cross-attention maps associated with prompt subjects contain high activation values in the foreground regions. We argue that the attention maps of a Transformer-based model exhibit the same behavior. Therefore, we use them to compute the candidate foreground pixels.

Then, we extract the subject information from the self-attention maps. Inspired by DAAM-I2I [78], we use the foreground map $\overline{\mathcal{CA}}_{fg}$ to merge all the foreground pixels self-attention maps by performing a weighted average, where the weights are the activation values of the foreground maps, normalized in the range $[0, 1]$. This map, denoted as $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{fg})}$, provides us with the attention scores of each foreground pixel with respect to the other foreground pixels. As explained before, we want to combine the different aspects captured by cross- and self-attention and use both for our coarse α channel estimation, which is given by the normalized sum of the two maps, *i.e.*,

$$\hat{\alpha} = \overline{\overline{\mathcal{CA}}_{fg} + \overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{fg})}}.$$

OBTAINING THE RGBA ILLUSTRATION. Once the reverse diffusion process is over, we obtain an image containing the desired subject with no sharp croppings and an estimation of the coarse transparency mask $\hat{\alpha}$. At this point, we need to remove the background to obtain the final illustration. To this end, we exploit the candidate foreground map $\overline{\mathcal{CA}}_{fg}$ obtained with the cross-attention maps of the noise estimation Transformer. Specifically, we quantize it into four values to obtain a map whose values indicate whether the corresponding pixel is in the sure background, the probable background, the probable foreground, or the sure foreground. We use the so-obtained mask as input to the perceptual GrabCut [313] algorithm, which performs graph optimization by combining color and border information to remove the background from the generated image. We use this

algorithm’s output foreground mask to clean $\hat{\alpha}$, zeroing the values outside the subject and obtaining our transparency channel α .

As a side note, we remark that one could alter the intensity of the α channel, *i.e.* augmenting or reducing the opacity of the generated illustration in a semantically meaningful way, by simply introducing a hyperparameter k to obtain

$$\hat{\alpha}' = \min(1, (1 + k)\hat{\alpha}),$$

which the user can tune depending on their needs. We set it to 0.5 in all experiments, leaving further exploration for future work.

10.4 EXPERIMENTS AND RESULTS

IMPLEMENTATION DETAILS. In our experiments, we use the pre-trained PixArt- Σ -512 [66] from HuggingFace [383]. PixArt- Σ is based on DiT [271] and comprises 28 self- and cross-attention layers, each with 16 attention heads. We use the Euler Discrete scheduler [182] with 30 denoising steps. The subject-centering mask has a border of 64 in the pixel space on all image sides. For the attention map quantization into a 4-valued map, we use the percentiles 0.8, 0.3, and 0.1, respectively: for sure foreground, probable foreground, and probable background.

EVALUATION SETUP. For quantitative evaluation, we collect 3000 of the multi-sentence prompts used to train PixArt- Σ [67, 66] and consider only the first sentence of the prompt, which is the one describing the general visual content of the image. These have been obtained by running the LLaVA Multimodal Large Language Model [221] on images from the SAM dataset [189][‡].

We quantitatively evaluate how reliably the pipeline generates illustrations of subjects wholly contained within the image, without crops or cuts. To do so, we consider the image canvas’s left, right, top, and bottom borders and, for each one, a margin of 4 pixels. If all the pixels in the margin area have a value greater than 0.8, we consider that border empty. We choose this threshold value because we aim to keep the background uniform and light rather than perfectly white. Then, we compute the percentage of generated images having each of the borders empty

[‡]We release the list of prompts at <https://github.com/aimagelab/Alfie>.

(denoted as **empty**- $\{\mathbf{l}, \mathbf{r}, \mathbf{t}, \mathbf{b}\}$, respectively), as well as the aggregate percentage of images having all four borders empty (denoted as **empty-a**). Moreover, we consider the CLIP score [146] (denoted as **CLIP-S**) to quantify how much the generated RGB images still respect the original prompt. Regarding the RGBA images, we remark that we do not evaluate in terms of the FID and KID scores, which are commonly adopted in image generation evaluation. This is because these scores need a large reference set (ground truth RGBA images) possibly compatible with the prompts fed to our model and with the characteristics that we enforce in our images. To the best of our knowledge, such a set with the required characteristics is not available.

Moreover, in line with RGBA image generation works [422], we perform a user study to compare the perceived quality of our illustrations compared to the best-performing alternative method. For a fair comparison, we employ the state-of-the-art ViT-B-based ViTMatte [415] and perform matting over our generation results. This matting method requires an input trimap for sure foreground, unsure regions, and sure background. We provide it by quantizing the same map $\overline{\mathcal{CA}}_{fg}$ that we obtain in our pipeline with the percentiles 0.8 and 0.3 for sure foreground and unsure regions, respectively. We consider the values under 0.3 as sure background. We also provide qualitative results showcasing the effects of centering, α estimation, and comparison with matting.

CONSIDERED VARIANTS AND BASELINES. As baselines, we consider the base PixArt- Σ [66] with the same prompts as we use for our pipeline and a version to which we always append the phrase *on a white background* to the input prompt (we refer to this as **PixArt- Σ +suffix**). As for the variants of our approach, we isolate the role of each component by considering the subject centering (which we refer to as **PixArt- Σ +centering**) and different methods for obtaining the α channel, namely: the normalized cross-attention heatmap $\overline{\mathcal{CA}}_{fg}$, the normalized self-attention heatmap $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{fg})}$, the normalized combination of cross- and self-attention $\hat{\alpha}$, and the GrabCut-cleaned α (which we refer to as *Alfie*).

RUNTIME. We run all experiments on a 24GB Nvidia RTX 4090 GPU with half precision. Generating an image with the base model (PixArt- Σ [66]) takes

~ 15 GB of VRAM and ~ 3.15 seconds, increased to ~ 4.13 seconds and ~ 16.7 GB of VRAM when we perform the centering. Cleaning the estimate transparency channel $\hat{z}$ with GrabCut [313] requires additional ~ 0.36 seconds.

10.4.1 RESULTS

CENTERING. We consider the background generation statistics and report it in Tab. 10.1. We can observe that the PixArt- Σ +suffix variant is not entirely reliable, as it leads to $\sim 53\%$ probability of generating the image inside the canvas. Conversely, our approach provides a much higher probability of generating the whole subject, consistently reaching more than 95% . In Tab. 10.1, we also report the CLIP-S between the generated images and the textual prompt. As we can see, the base model obtains the highest value, our reference score, as the generated images are fully based on the textual prompt, and the generation process is unconstrained. On the other hand, our target images have implicit characteristics that are not part of the textual caption but rather represent meta-descriptions, *e.g.*, whole-object generation. As expected, the second-best CLIP-S is obtained by the base model with the suffix. In fact, this baseline sometimes (almost half of the times) ignores the suffix, and thus, it is less penalized. While respecting the desired characteristics defined in Section 10.3, our generated images achieve CLIP-S values very close to the reference ones. Thus, we can conclude that our inference-time adaptation strategy guides the generation toward the desired characteristics while maintaining adherence to the textual prompt. In Fig. 10.4, we showcase the impact of this constraint on the generated images. While the simple appending of the suffix cannot always guarantee whole-subject generation, our method consistently steers the generation towards a correct result and enforces this meta-description on the image.

α CHANNEL ESTIMATION. In Fig. 10.5, we show alternative α estimations obtained using different combinations and processing of the attention maps. As we can see, the normalized cross-attention map $\overline{\mathcal{CA}}_{\text{fg}}$ provides coarse localization of the foreground subjects with uniform transparency values. The normalized self-attention map $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{\text{fg}})}$ is more spatially precise and has a broader range of values,

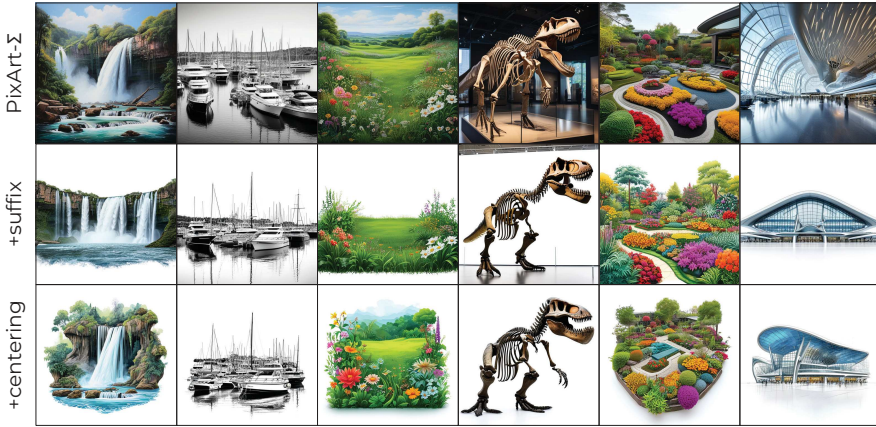


Figure 10.4: Qualitative comparison on the effect of our constrained whole-subject generation compared to meta-descriptions in the input prompt.

	empty-a	empty-l	empty-r	empty-t	empty-b	CLIP-S
PixArt-Σ	3.33	4.33	4.13	8.53	6.06	31.29
PixArt-Σ + suffix	53.10	62.63	62.10	92.53	84.50	30.79
PixArt-Σ + centering	96.50	99.27	99.03	99.53	98.00	30.08

Table 10.1: Quantitative comparison of our approach variants in terms of average probability with which the subjects are generated in the center of the image with no sharp croppings and adherence to the prompt.

reflecting how much the objects appearance depends on a localized or sparse image region. By combining the two, the resulting map $\hat{\alpha}$ gives a balance between absolute and relative transparency values. Finally, using Grabcut [313] for cleanup ensures precise border localization, leading to the best overall results.

COMPARISON WITH MATTING. We compare our method against matting by considering the generated RGB images with the foreground cross-attention map $\overline{\mathcal{A}}_{fg}$ and using it to compute the input trimap for ViTMatte [415]. We run a user study with ten individuals and 100 images randomly sampled from our dataset and ask them “Which of the following RGBA illustrations do you prefer?”. We find that users indicated the images from our method $\sim 63\%$ of the time, demonstrating that we are slightly better than state-of-the-art matting methods while



Figure 10.5: Qualitative comparison on different α estimates. Combining self- and cross-attention maps provides the best balance of spatial localization and transparency values, and their cleanup using GrabCut [313] (*Alfie*) further increases border precision.

requiring no costs for training. In Fig. 10.6, we provide samples obtained by using the two methods. As we can see, our method provides reasonable α channels, comparable to or better than matting, mainly because using the attention maps reduces prediction errors, discarded image parts, and preserved background areas.

COMPOSITIONAL ARTWORK GENERATION. We use the images generated by our method as input for Collage Diffusion [319]. This method allows compositional scene generation by taking sequences of layers as input, which define the spatial arrangement and attributes of objects in the scene. The authors combine



Figure 10.6: Qualitative comparison of our method against ViTMatte [415]. Without training an additional network, we produce reasonable transparency channels.

SDEdit [244], cross-attention manipulation [28], textual inversion [119], and a proposed extension of ControlNet [423] to find a balance between harmonization (of the input subjects and the scene) and fidelity (of the produced collage w.r.t. the different original subjects). The pipeline runs on Stable Diffusion v2-1-base [308]. In Fig. 10.7, we show the results on the pipeline with the same input background and three different subjects: an RGBA illustration obtained from Adobe Stock, an image obtained with the base PixArt- Σ model, and an image obtained with Alfie. The outputs of Collage Diffusion showcase that the quality obtained by using our method is similar to that achieved with a commercial RGBA illustration. Conversely, when the input is the base PixArt- Σ squared illustration, the pipeline tries to incorporate it into the scene by rendering it as a big painting, a plausible but unacceptable result.

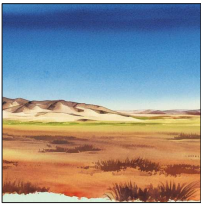
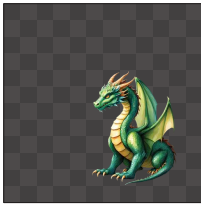
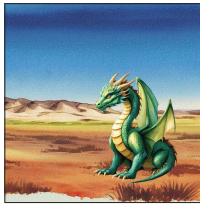
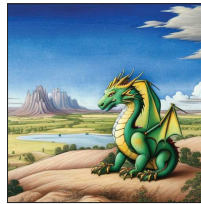
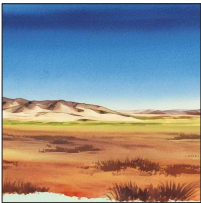
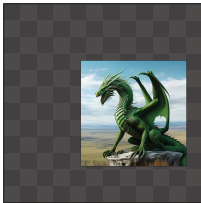
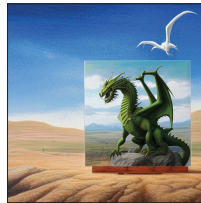
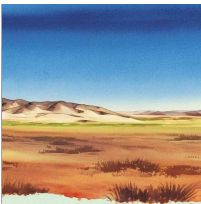
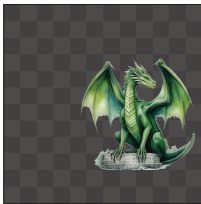
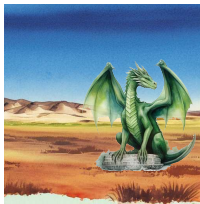
Inputs		Scene Composition	
Background	Foreground	Superimposed	Collage Diffusion
			
			
			

Figure 10.7: Results of Collage Diffusion [319] with different input foreground images. Top to bottom: a commercial RGBA image, an image generated by the baseline PixArt- Σ [66], and an RGBA illustration obtained with our approach. With our generated illustration, the output is comparable to that obtained with the commercial image, differently from the pipeline run with the image from the baseline model. To generate the background and the foreground, we use the prompts *A steppe landscape* and *A green dragon*, combined for the composite scene into *A green dragon in a steppe landscape*.

10.5 CONCLUSIONS

In this work, we have tackled the task of generating illustrations for visual content, artworks, and visually-rich documents. We have devised a fully automated, prompt-guided pipeline for obtaining high-quality RGBA images that can be used by users or integrated into scene compositions with good results at zero additional training cost. As the main component of our pipeline, we have exploited the recently proposed Diffusion Transformer paradigm and explored inference-time adaptation strategies for these models. We have quantitatively and qualitatively validated the effectiveness of the proposed approach, which can serve as a starting point for the community to continue working on the tackled task. We hope that the promising results obtained can encourage the research towards the development of similar low-cost strategies for such models.

Part III

Editing VRD

11

Literature Survey

THIS part of the thesis focuses on editing approaches for visually-rich documents, with particular emphasis on generative image editing and multi-instance manipulation. Editing naturally sits at the intersection of the two previous parts of this thesis: it requires understanding capabilities to localize the regions of interest and interpret what needs to be modified, as well as generation capabilities to render the desired changes while preserving the surrounding context. We review the state of the art in generative image editing, with particular focus on the challenges of multi-instance editing and the main paradigms proposed to address them.

11.1 GENERATIVE EDITING MODELS

GENERATIVE IMAGE EDITING. State-of-the-art editing has consolidated around Autoregressive models [365, 208, 135, 361] and Diffusion Transformers [222, 394, 357, 386, 194], with Rectified Flow Matching (Rectified Flow Matching (RFM)) [218] emerging as a dominant framework for efficient, straight-path sampling [368]. The conditioning signal is generally multimodal, comprising guid-

ance [431], text instructions [108], and context images [72, 194, 394]. However, these techniques rely on globally conditioned velocity fields and struggle with local, fine-grained localization. Recent works address this via attention regularization [330, 421] or constrained cross-token interactions [89, 157], but primarily target single-region or global edits and often fail to maintain coherence when applied to multiple disjoint regions simultaneously. While some methods investigate region-aware generation [73, 106] and causal masking [140] to inject spatial control, most RFM editors [194, 108] still rely on globally conditioned velocity fields, causing attribute leakage between regions during multi-instance editing [255, 436].

To address the challenges of multi-instance editing, current approaches primarily cluster into three categories: scaling inference time (*multi-turn*), scaling the architecture (*multi-branch*), or regularizing the internal generation process.

MULTI-TURN. State-of-the-art editing models are trained to reduce visual drift and show promise in iterative editing workflows [194, 394], where edits are performed sequentially. While serving as a functional baseline, multi-turn editing differs fundamentally from concurrent editing. Technically, visual degradation accumulates as the number of iterations exceeds a few cycles (around six [194]), and inference time scales linearly with the number of instructions, making this approach infeasible as instance edit counts increase. Methodologically, the final result is path-dependent, especially with overlapping regions.

MULTI-BRANCH. To enforce disentanglement, some approaches split the generation process into multiple parallel branches, typically one per edit or region. MultiDiffusion [30], the seminal method in this paradigm, binds multiple generation paths by merging independent denoising steps into a least-squares optimal solution. LoMOE [62] adopts this framework to apply localized influence on the N target regions using foreground masks, while other approaches further refine this by using regional branches [73], performing independent inversion searches [411] or grouping edit “aspects” into specialized branches [157]. Finally, LayerEdit [117] takes this paradigm to its extreme by decomposing the scene into $N + 2$ independent layers. While multiple generation paths allow instance isola-

tion, these methods suffer from severe scaling bottlenecks in time and memory usage. Moreover, computational overhead is worsened by the latent optimization often required: LayerEdit and LoMOE rely on iterative gradient descent and perform a backward pass over N parallel branches. Furthermore, these methods rely on heuristic reassembly to smooth visual boundaries, which often fails to maintain global coherence in dense, overlapping scenes.

REGULARIZATION. Another class of methods focuses on end-to-end, single-pass editing by regularizing the internal generation process. One prominent direction involves inference-time optimization: MDE-Edit [437] performs gradient descent on the latents at each denoising step to minimize specialized losses. While these methods maintain a single-pass memory footprint, they rely on aggregating $O(N)$ loss terms, introducing significant computational overhead and numerical instability during backpropagation, particularly when handling multiple instances of varying spatial scales, where large-scale objects can dominate the gradient signal and suppress smaller targets. Other approaches perform fine-tuning or train additional modules to target quantity perception [209] or personalization [241].

CONDITIONAL FLOW MATCHING. Rectified Flow Matching (RFM) [218] offers efficient inference and exact invertibility, serving as a powerful alternative to diffusion. Recent theoretical advances have further optimized these dynamics via optimal transport paths [368] and guidance mechanisms [431]. While other works investigate region-aware generation [73, 106] and causal masking [140] to inject spatial control, most RFM editors [194, 108] still rely on globally conditioned velocity fields. This global conditioning causes attribute leakage between regions during multi-instance editing [436], a limitation we address via Instance-Disentangled Attention (IDAttn).

Beyond generic image editing, the multi-instance setting becomes especially demanding in visually rich documents such as posters and infographics, where many small textual regions must be modified without disturbing the global layout, background appearance, or visual hierarchy. In these settings, the challenge is not only semantic faithfulness, but also edit locality, text readability, and scal-

ability with the number of edited regions. Unlike object-centric natural images, these domains also admit large-scale annotation through text detection and OCR pipelines, making them a practical and revealing testbed for whether an editing model can truly disentangle concurrent instructions. This document-centric perspective motivates the benchmark introduced in Chapter 12.

12

Shifting the Breaking Point of Flow Matching for Multi-Instance Editing

FLOW matching models have recently emerged as an efficient alternative to diffusion, especially for text-guided image generation and editing, offering faster inference through continuous-time dynamics. However, existing flow-based editors predominantly support global or single-instruction edits and struggle with multi-instance scenarios, where multiple parts of a reference input must be edited independently without semantic interference. We identify this limitation as a consequence of globally conditioned velocity fields and joint attention mechanisms, which entangle concurrent edits. To address this issue, we introduce IDAttn (Instance-Disentangled Attention), a mechanism that partitions joint attention operations, enforcing binding between instance-specific textual instructions and spatial regions during velocity field estimation. We evaluate our approach on both natural image editing and a newly introduced benchmark

This Chapter is related to the publication “C. Zaccagnino*, F. Quattrini* *et al.*, Shifting the Breaking Point of Flow Matching for Multi-Instance Editing, ICML 2026” [9]. *Equal contribution. See the list of Publications on page 315 for more details.

of text-dense infographics with region-level editing instructions. Experimental results demonstrate that our approach promotes edit disentanglement and locality while preserving global output coherence, enabling single-pass, instance-level editing. We release our dataset and code ^{*}.

12.1 INTRODUCTION

High-fidelity text-guided image editing has traditionally relied on diffusion models with U-Net architectures [333, 149, 338] for global and localized edits. These models iteratively denoise a sample in latent or pixel space, enabling a wide range of operations such as attribute modification, object insertion or removal, and inpainting. In particular, given an input image and a text condition, editing capabilities are achieved with techniques such as masked denoising, cross-attention manipulation, and inversion, enabling object-level control and compositional editing [145, 250, 25, 125, 86]. Despite these successes, most works and benchmarks focus on single or at most few-instance edits in the natural image domain [411, 241, 117, 62], offering limited support for *multi-instance editing*. This task requires the model to yield *edits disentanglement*, *i.e.* multiple edits should be independently controllable and combinable, *edits locality*, *i.e.* non-edited regions should remain unchanged, and *global coherence*, *i.e.* the whole image should still be visually coherent. Moreover, when dealing with an increasingly high number of edits N , computational constraints of some methods increasingly bottleneck inference pipelines.

Recently, image generation community is shifting towards Multimodal Diffusion Transformers (MMDiT) [108] trained with Rectified Flow Matching [218]. Compared to diffusing stochastic processes, this framework offers higher visual quality and faster inference by modeling generation as an Ordinary Differential Equation (ODE) and learning to predict velocity fields v_θ over straight paths between data and noise distribution. The MMDiT architecture iteratively refines image tokens from the random noise distribution into data, conditioned on a textual prompt. To allow self- and cross-modality interaction, *joint attention* op-

^{*}<https://github.com/aimagelab/IDAtnn>

erates on concatenated text and image tokens. For editing, the condition image is concatenated to the noisy image tokens and processed through the same blocks as the noisy image [357, 194, 394]. In this setting, the attention operator jointly processes *prompt*, *noisy latents*, and *context image* tokens. While this approach allows great global visual quality, allowing all tokens to interact with all other tokens causes inevitable semantic leakage as N increases.

Formally, Flow matching learns a single global velocity field v_θ governing the evolution of all pixels or latents simultaneously. Conditioning is typically injected as a global signal, $v_\theta(x, t \mid c)$, without explicit mechanisms to enforce instance-level separation. As a consequence, when multiple semantic instructions $\{c_n\}_{n=1}^N$ are applied simultaneously, their effects may interfere within the shared vector field, leading to semantic entanglement and unintended edits. This phenomenon, referred to as *attribute leakage* [302, 65, 255], stems mainly from semantic interference in the prompt encoding and from the vanilla joint attention mechanism, which permits disjoint concepts to interact.

To address this, we enforce instance independence *architecturally* by introducing **Instance-Disentangled Attention (IDAttn)**, a mechanism that partitions the joint attention operations within the MMDiT blocks [108] of the flow matching-based FLUX.1 Kontext model [194]. By partitioning the attention, we enforce locality constraints without disrupting the global flow matching objective. Moreover, we can perform disentangled multi-edits in a single pass, greatly improving computational performance.

We propose to consider both natural image editing and text image editing on Infographic images. In fact, current natural image multi-instance editing benchmarks [62] suffer from low-dimensionality and a low number of edits for each sample, with tens of images and just a few edits requested for each image. To address this, we chose to evaluate models on infographic text replacement. This domain has many practical applications and is particularly challenging. Particularly, infographics are text-dense and aesthetically-pleasing information visualizations. Therefore, uncontrolled edits of a single instance can severely alter the semantics of the whole image. Moreover, this domain facilitates semi-automatic high-quality annotation. Therefore, we curate a benchmark consisting of two

sets of infographic images, paired with paragraph-level text region localization annotations and editing instructions in the form of *Change* ‘SRC’ in ‘TGT’, where SRC is the original text (generally in English), and TGT is the desired text in another language. Our proposed benchmark contains infographic images of notable complexity, both in terms of the number of edits and the small area of the edited regions, opening up many challenges. The benchmark will be made available to the community for fostering the exploration of multi-instance editing in this complex setting. Experimental analysis on these two domains demonstrates the effectiveness and efficiency of our approach.

12.2 PROPOSED APPROACH

In this work, we devise an approach to address the problem of *multi-instance editing* under localized, potentially conflicting conditioning signals. The task translates into obtaining a conditional velocity field that transforms a source distribution (noise) into a target edited sample x_1 , subject to a set of localized instructions $\{(s_n, b_n)\}_{n=1}^N$, where s_n is a text prompt and b_n is the localization information (*e.g.* a bounding box) for the n -th instance. Without explicit structural constraints, the current generative models struggle to disentangle simultaneous, conflicting edits. This phenomenon, often referred to as *attribute leakage*, stems from the entangled nature of the cross-attention mechanism [302, 65]. Standard attention operations allow semantic tokens associated with one subject (*e.g.* “red”) to attend to spatial regions corresponding to another (*e.g.* “car”), causing concepts to bleed across boundaries [255]. While previous works have attempted to mitigate this via inference-time attention guidance [145], these methods often require costly iterative optimization. Our approach also works at inference time and addresses the root cause: the architectural permission for disjoint concepts to interact.

12.2.1 PRELIMINARIES

CONDITIONAL FLOW MATCHING. We focus on conditional rectified flow matching [218, 368] as the generative framework. Specifically, let $x_1 \sim p_{\text{data}}$ denote a target data sample and $x_0 \sim p_0$ a source sample drawn from a simple base distribution (*e.g.* Gaussian noise). Flow matching learns a time-dependent velocity field conditioned on signals $\mathcal{C}$

$$v_\theta : \mathbb{R}^d \times [0, 1] \times \mathcal{C} \rightarrow \mathbb{R}^d$$

such that the induced ODE

$$\frac{dx_t}{dt} = v_\theta(x_t, t \mid c)$$

transports x_0 to x_1 , conditioned on the external information c . In the conditional rectified flow formulation [223], the training objective minimizes the expected flow matching error over time $t \in [0, 1]$, *i.e.*

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, x_1, x_0} [\|v_\theta(x_t, t \mid c) - (x_1 - x_0)\|^2], \quad (12.1)$$

where $x_t = (1-t)x_0 + tx_1$. At inference time, samples are generated by integrating the learned ODE between 0 and 1.

JOINT ATTENTION. State-of-the-art conditional generative models based on flow matching [108, 194] employ an MMDiT as the parameterization of v_θ . MMDiT performs only self-attention operations on a sequence of tokens obtained by concatenating embeddings from multiple modalities. We denote the input tokens with Z , and with Q, K, V the query, key, and value matrices linearly projected from Z through learnable matrices. Specifically, let $Z = Z^{\text{text}} \parallel Z^{\text{latent}} \parallel Z^{\text{context}}$ denote the joint token sequence, where: (i) Z^{text} encodes textual instructions, (ii) Z^{latent} encodes the evolving conditional flow matching latent representation x_t , (iii) Z^{context} encodes auxiliary conditioning signals (*e.g.* a reference image, depth maps, or segmentation mask cues). Each attention layer within MMDiT applies *joint attention* [108], where queries, keys, and values are computed inde-

pends for each modality and concatenated:

$$\begin{aligned} Q &= Q^{\text{text}} \parallel Q^{\text{latent}} \parallel Q^{\text{context}}, \\ K &= K^{\text{text}} \parallel K^{\text{latent}} \parallel K^{\text{context}}, \\ V &= V^{\text{text}} \parallel V^{\text{latent}} \parallel V^{\text{context}}. \end{aligned}$$

The attention output is then given by

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \quad (12.2)$$

12.2.2 INSTANCE-DISENTANGLED ATTENTION

We introduce **IDAttn** (Instance-Disentangled Attention), an architectural intervention that enforces structured conditional dependencies within joint attention, enabling independent control of multiple instances within each single flow field estimation step.

TOKEN SPACE PARTITIONING. First, consider N instances to be edited within the same sample, *e.g.* regions of a reference input image that the model should modify based on some textual prompt. Our approach entails logically partitioning the joint token sequence Z into the following sets, according to both modality and instance association information (*e.g.* bounding boxes b_n in the case of image editing). Note that the localization information b_n can also be provided as a segmentation mask rather than a bounding box; this is particularly useful for irregular shapes, as masks avoid including excessive background in the instance’s region.

- $\mathbb{T}_g$: Global prompt tokens, *i.e.* text tokens Z^{text} relative to style, overall scene description, or a null prompt.
- $\mathbb{T}_n$: Local prompt tokens, *i.e.* text tokens Z^{text} specifically describing the target content for instance n .
- $\mathbb{L}_u$: Latent tokens that do not need to be edited *i.e.* latent tokens Z^{latent} not pertaining any instance (*e.g.* relative to the background).
- $\mathbb{L}_n$: Local latent tokens, *i.e.* latent tokens Z^{latent} corresponding to the n -th instance.

- $\mathbb{C}_u$: Context tokens that do not need to be edited, *i.e.* auxiliary conditioning tokens Z^{context} relative to no specific instance.
- $\mathbb{C}_n$: Local context tokens, *i.e.* auxiliary conditioning tokens Z^{context} relative to instance n .

Note that, when multiple instances overlap (*e.g.* when localizing them via bounding boxes in the case of image editing), their respective tokens end up in multiple partitions.

DISENTANGLEMENT ATTENTION MASKS. We govern the information flow via an additive attention mask $M \in \mathbb{R}^{|Z| \times |Z|}$, which we combine with the standard joint attention matrix. The attention operation for query Q , key K , and value V in Equation (12.2) becomes:

$$\text{IDAttn}(Q, K, V, M) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d}} + M \right) V, \quad (12.3)$$

where $M_{ij} \in \{0, -\infty\}$. We define two distinct masking regimes to control the entanglement of the latent space: one for instance disentanglement (M^{dis}), and the other for context-aware harmonization (M^{har}). Please refer to Fig. 12.1 for a pictorial representation of the defined masks.

DISENTANGLEMENT REGIME. This regime enforces instance isolation. For any two distinct instances n, m (where $n \neq m$), the mask is defined as:

$$M_{ij}^{\text{dis}} = \begin{cases} 0 & Z_i \in \mathbb{T}_g \cup \mathbb{L}_u \cup \mathbb{C}_u \wedge Z_j \notin \bigcup_{n=1}^N \mathbb{P}_n \\ 0 & Z_i, Z_j \in \mathbb{P}_n \cup \mathbb{L}_n \cup \mathbb{C}_n \quad \forall n \in [1, N] \\ -\infty & \text{otherwise.} \end{cases} \quad (12.4)$$

In other words, this mask lets the global prompt tokens $\mathbb{T}_g$ attend to all the image tokens ($\mathbb{L}_u \cup \{\mathbb{L}_n\}_{n=1}^N$) and auxiliary context tokens ($\mathbb{C}_u \cup \{\mathbb{C}_n\}_{n=1}^N$); the image tokens $\mathbb{L}_u$, and context tokens $\mathbb{C}_u$ relative to the background can attend to the global prompt $\mathbb{T}_g$ and to all the image tokens ($\mathbb{L}_u \cup \{\mathbb{L}_n\}_{n=1}^N$) and context tokens ($\mathbb{C}_u \cup \{\mathbb{C}_n\}_{n=1}^N$); the prompt tokens, image tokens, and context tokens relative to the same instance n (*i.e.* $\mathbb{T}_n$, $\mathbb{L}_n$, and $\mathbb{C}_n$) can only attend to each other. As a result, instance-specific conditioning tokens can influence only their corresponding image latents, while global tokens remain self-consistent.

For multi-instance image editing, the main effect of this masking strategy is enabling attention within the token subset corresponding to the region within b_n and its textual editing prompt s_n , while preventing cross-instance interference between regions and editing prompts being ignored by the model. Moreover, by letting the global prompt and the background visual tokens attend to both each other and the instances' tokens, this masking allows maintaining visual consistency.

HARMONIZATION REGIME. To preserve global coherence, we relax the constraints defined in $\mathcal{M}^{\text{dis}}$ by allowing the image and the auxiliary context tokens relative to each instance ($\mathbb{I}_n$, and $\mathbb{C}_n$) attend to also all the other image and context tokens ($\mathbb{I}_u \cup \{\mathbb{I}_n\}_{n=1}^N$, and $\mathbb{C}_u \cup \{\mathbb{C}_n\}_{n=1}^N$), further enforcing global coherence. In other words, the mask in this regime is defined as:

$$\mathcal{M}_{ij}^{\text{har}} = \begin{cases} 0 & Z_i, Z_j \notin \bigcup_{n=1}^N \mathbb{P}_n \\ 0 & Z_j \in \mathbb{P}_n \cup \mathbb{I}_n \cup \mathbb{C}_n \wedge \\ & Z_i \in \mathbb{P}_n \quad \forall n \in [1, N] \\ -\infty & \text{otherwise.} \end{cases} \quad (12.5)$$

MASKS APPLICATION STRATEGY. To obtain the edited sample, within each refinement step, we apply $\mathcal{M}^{\text{har}}$ in the early and late layers of the velocity field estimation Transformer (denoted as L_{early} and L_{late} , respectively), and $\mathcal{M}^{\text{dis}}$ in the central layers (L_{mid}). The resulting instance-disentangled flow matching process is described in Appendix E.1. This design is motivated by empirical and theoretical analyses of Transformer representations [398, 70, 419], which suggest that early layers focus on coarse feature extraction, middle layers encode semantic bindings, and late layers refine and globally harmonize the output. Similar layer-wise intervention strategies have proven effective in controlled generation [436]. Following [436], we set the cut-off between L_{early} and L_{mid} at layer 10, and between L_{mid} and L_{late} at layer 47, using this same configuration across all experiments without any tuning.

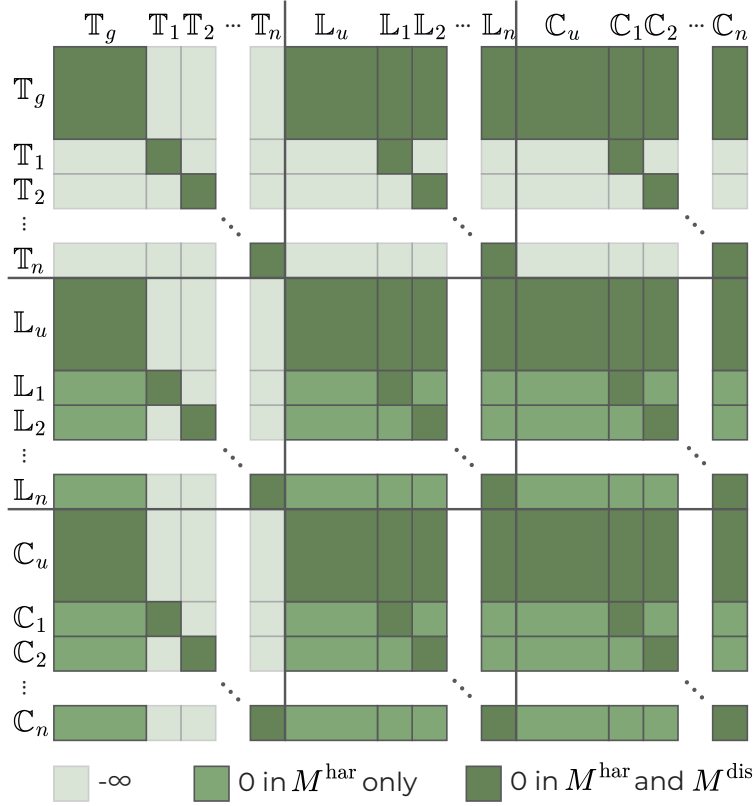


Figure 12.1: Logic visualization of the proposed joint attention masks.

12.2.3 EFFICIENT MULTI-PROMPT INDEPENDENT ENCODING

Effective multi-instance editing requires ensuring that textual instructions for different instances remain semantically isolated. Standard approaches typically concatenate all instructions into a single long prompt. However, this allows the text encoder’s self-attention mechanism to mix unrelated concepts before they even reach the generative model [65, 302]. Existing solutions generally fall into the following two categories. (i) *Post-encoding single prompt masking* [399], *i.e.* encoding a single prompt and enforcing spatial constraints by masking parts of the encoded prompt within the attention layers of the generative architecture. This approach is efficient in terms of token usage, but instance-level disentanglement

is not guaranteed and often worsens as the prompt complexity increases. (ii) *Multi-prompt independent encoding* [436], *i.e.* dividing the prompt into instance-specific subprompts before encoding, each of which is padded or truncated to a fixed length. This enforces instance-level separation by construction, but its computational cost scales linearly with the number of instances.

We propose a multi-prompt encoding strategy that ensures semantic isolation by construction while maintaining a sequence length proportional to the semantic content, thereby optimizing the computational cost of the subsequent attention operations. First, we decompose the input prompt into a global component s_g (in practice, we use a null prompt) and independent instance-specific components $\{s_n\}_{n=1}^N$. Then, we encode these components separately to obtain variable-length embeddings, and concatenate them to construct the final conditioning sequence. In our FLUX.1 Kontext implementation, the instance-specific prompt tokens are encoded independently, while the latent and context partitions are obtained by mapping each annotated box to the corresponding token regions of the VAE representation. As global prompt $\mathbb{T}_g$ we use a **null prompt** made up of 200 padding tokens, which we call a *utility prompt*: it is the only prompt that interacts with $\mathbb{L}_u$ and $\mathbb{C}_u$ in the layers using $\mathcal{M}^{\text{dis}}$, so it is needed to make sure the model preserves the background effectively.

12.2.4 DOMAIN-SPECIFIC FINE-TUNING

All of the modifications proposed so far can be applied at inference time on top of any pre-trained model, but state-of-the-art flow matching models for image generation and editing are typically trained on natural images with long prompts acting on relatively large areas. As a result, they may underutilize fine-grained conditioning signals corresponding to small spatial regions or short textual instructions. The core contribution of this work — IDAttn masking combined with multi-prompt independent encoding — is an inference-time strategy that can be applied on top of any pre-trained MMDiT-based model without additional training. It consistently improves over the base backbone on both natural images and infographics. As an optional extension, we further explore lightweight domain

adaptation for the particularly challenging infographic setting, where the backbone is significantly out-of-domain (images with many small, text-dense regions). We perform Low-Rank Adaptation [327] with rank $r = 32$ on all layers of the MMDiT equipped with our masking strategy, minimizing the conditional flow matching loss $\mathcal{L}_{\text{CFM}}$ defined in Equation (12.1) with multi-prompt independent encoding as the text input. This fine-tuning trains only 0.90% of the model’s parameters, incurring modest computational overhead, but gives the model a better understanding of complex text-heavy editing prompts in this domain.

12.3 INFOGRAPHICS EDITING BENCHMARK

While recent benchmarks have explored object-level and region-based image editing, none explicitly address scenarios requiring a large number of fine-grained edits to be applied simultaneously within a single image. Current multi-instance benchmarks remain limited in both the number of editable regions per image and overall scale [62, 411], arguably due to the cost of accurate annotation in natural images. As a result, they fail to stress-test the ability of editing models to maintain edit disentanglement and locality as the number of concurrent instructions increases. In this respect, we argue that infographics constitute a particularly suitable domain for studying this setting, due to the high density of visual and textual elements within a coherent layout that characterizes these images. In this work, we focus on textual elements, which constitute a demanding and practically relevant testbed for evaluating multi-instance image editing methods. In fact, since infographics often contain a large number of text boxes ($N \gg 5$), when attempting end-to-end editing of all boxes in one go, or in multi-turn, existing image editing models tend to produce inconsistent or collapsed edits and to ignore most of the editing prompts. Moreover, text boxes are usually small relative to the full image, implying that the model must modify high-frequency, localized details. Nonetheless, this setting allows for scalable data collection. In fact, text regions can be automatically detected and annotated with standard Text Detection and Optical Character Recognition (OCR) pipelines, making it feasible to construct large datasets efficiently. Therefore, in this work, we devise a large collection of

densely-annotated infographic text editing samples. This will be made publicly available to favor reproducibility and exploration of the task at hand.

Formally, let $I^{\text{ref}} \in \mathbb{R}^{H \times W \times 3}$ denote the high-resolution input infographic image and consider a set of N bounding boxes $B = \{b_n = (x_n, y_n, w_n, h_n)\}_{n=1}^N$, each enclosing a text element, and a set of corresponding editing instructions $S = \{s_n\}_{n=1}^N$, e.g. *Change ‘Hello’ to ‘Bonjour’*. In real scenarios, these latter two inputs can either be provided by a user or by auxiliary modules for text detection, recognition, and translation. The task entails generating an edited image I^{edit} such that each text element within b_n is replaced according to s_n while maintaining visual coherence of I^{ref} and text style within each local edit. We devise the following subsets of increasing level of difficulty.

CRELLO EDIT. To obtain the most manageable subset, we start from the layer-wise annotated graphic designs of Crello [409]. These are mostly images with English text. Crello comes with precise annotations of each text box, which we translate into four languages (French, German, Italian, and Spanish) by using Seed-X [74]. Then, we re-render the text layer of the original sample so that it contains the translated text, by using the same font metadata accompanying it. In this way, we obtain a paired editing dataset featuring samples where the reference image contains English text and the target image contains text in another language, which can be used for supervised training. To reduce artifacts, we only re-render text boxes where the length of the translated text is up to 25% longer than the original one, and inspect the samples for visual quality. The obtained set contains 1512 samples that can be used in the training phase and 4367 samples used for test, accompanied by metadata to obtain 1 to 25 editing instructions on boxes whose area is on average $4.23 \pm 5.49\%$ of the entire image.

INFOEDIT. To explore a more challenging and realistic infographic editing scenario, we start from the images collected by the authors of [438] and retain only the real, professionally-made, and open-source English infographics. We consider images whose total area is at least of 1024^2 px and further refine the selection via visual inspection. Then, we run LayoutParser [328] to detect the text elements and PaddleOCR [88] to transcribe them. Finally, we aggregate the lines

into paragraphs when relevant. Also in this case, we resort to Seed-X [74] to obtain the translations of the text elements into French, German, Italian, and Spanish, and obtain the located editing prompts from those. The resulting dataset contains 4960 samples, where images are paired with 1 to 285 editing instructions relative to boxes whose area is on average $0.62 \pm 1.64\%$ of the entire image.

12.4 EXPERIMENTAL ANALYSIS

CONSIDERED DATASETS. We evaluate our approach on multi-instance editing of both natural images and infographics. For the former case, we resort to the recently-proposed LoMOE-Bench [62], featuring 80 images paired with 2 to 7 editing instructions. For the latter, we use the Crello Edit and InfoEdit test sets featured in our proposed Infographic Editing Benchmark (Section 12.3).

COMPETITORS AND BASELINES. To assess the benefits of our proposed approach for multi-instance image editing, we apply it to the State-of-the-Art FLUX.1 Kontext RFM-based generative model [194] and compare its performance against the following variants, officially proposed by the respective authors: (i) we apply FLUX.1 Kontext end-to-end by prompting it with all the text editing prompts in one go (**FLUX**); (ii) we apply it in a multi-turn scenario, where we provide it with one editing prompt at a time (**FLUX μ T**). Moreover, for natural image editing, we consider also **LoMOE** [62] and **LayerEdit** [117] and an additional variant of FLUX.1 Kontext, described in [194] that entails drawing boxes over the editing instances as visual cues, and referring to them by color in the prompt (**FLUX w/ v.c.**). As for infographics text editing, we include a FLUX.1 Kontext-based baseline that we design for ensuring instance disentanglement and isolation. We use the model for editing each text box separately, and we stitch the box back on the original image, after having run the ViTEraser [272] inpainting model for removing the original text (**FLUX st.**). Moreover, we consider **Calligrapher** [235], which is a finetuned version of FLUX-Fill [194] for text image editing. Since it can handle one editing instruction at a time, we use it in a multi-turn scenario (**Calligrapher μ T**) and in the same box editing and re-stitching pipeline as for FLUX.1 Kontext (referred to as **Calligrapher st.**).

CONSIDERED SCORES. Here we mention the scores we use to evaluate the performance of our approach and the competitors.

For multi-instance text content editing in the infographics benchmark datasets, we use the following scores:

- We compute the average Character Error Rate (**CER**) obtained by PaddleOCR over all edited text boxes. This score directly quantifies the correctness of the rendered text with respect to the editing prompt, thereby evaluating adherence to the requested textual content. Because of the programmatically generated nature of the Crello Edit subset, we first generate ground-truth renderings of images containing the target text and compute CER on these. We then subtract this ground-truth CER from the one computed on the samples generated by the evaluated methods and refer to the score thus obtained as $\Delta\mathbf{CER}$, which we only report for the Crello Edit subset, as it is the only programmatically generated dataset in our benchmark.
- We report the **MAE** and **MSE** differences between each edited infographic and its reference image, restricted to pixels outside the specified text-box regions. These metrics, dubbed **MAE_B** and **MSE_B**, penalize unintended changes to the original image.
- As customary for assessing the realism of the edited images compared to the original distribution, we report Fréchet Inception Distance (**FID**) [147] between the edited infographics and the corresponding reference images.
- We introduce **Attempt Rate (AR_%)** in order to evaluate the models' ability to attempt to perform edits instead of ignoring prompts. This is quantified as the percentage of target bounding boxes where the Mean Absolute Error (MAE) between the original and the edited image (both images as RGB, with each channel represented by an integer in the range [0, 255]) crop exceeds a fixed threshold (set to 10.0). This score effectively serves as a "sanity check" to distinguish between failed edits (where the model modifies the region but produces poor quality) and ignored instructions

(where the model simply reconstructs the original content). As with many heuristic approaches, we acknowledge that this score has some limitations, but empirically serves its purpose to distinguish areas where significant edits were applied and those in which they were not. A sensitivity analysis varying the threshold across values $\{5, 10, 15, 20, 25\}$ confirms that model rankings remain stable (standard deviation $\approx 5\text{--}8\%$ on Crello Edit and InfoEdit, where the higher number of edits per image ensures statistical significance), validating the choice of 10.0 as the default threshold.

For LoMOE-Bench [62], we use the scores proposed by its authors, computed on the entire image unless otherwise specified, *i.e.*:

- **Target CLIP Score (Tgt C)** [146]. To evaluate the semantic fidelity of the edits, we measure the alignment between the edited image and the target text prompt. Following the LoMOE benchmark protocol, this is computed as the cosine similarity between their feature embeddings extracted by a pre-trained CLIP ViT-B/16 model. A higher score indicates visual content accurately reflecting the requested text description.
- **Structural Distance (SD)**. To assess the preservation of the scene’s geometric layout, we compute the Structural Distance between the **original input image** and the **edited output image**. Following the official implementation of LoMOE-Bench [62], this is calculated as the Mean Squared Error (MSE) between the self-similarity maps of the keys extracted from the last layer of a DINO-trained ViT-B/8 model. This score effectively captures high-level structural correspondence between the source and the edit, ensuring the preservation of object poses and scene composition.
- **Image Reward (IR)** [403] and **Human Preference Score (HPS)** [397]. Following the evaluation protocol proposed in LoMOE-Bench [62], we employ both IR and HPS to quantify perceptual quality. Unlike standard metrics that focus on semantic alignment or pixel-level fidelity, these models are trained on human preference data, serving as robust proxies for visual realism and aesthetic appeal in editing tasks.

- **Background LPIPS ($LPIPS_B$)** [425] and **Background SSIM ($SSIM_B$)** [388].

To verify that the editing process remains localized and does not introduce unintended artifacts in the surrounding scene, we compute LPIPS and SSIM exclusively on the background regions (masked using the inverted edit mask). Background LPIPS utilizes deep features (AlexNet) to measure perceptual deviations, while Background SSIM assesses pixel-level structural consistency between the original and edited images.

Moreover, we use **FID** and **AR_%**, and introduce the **Localized CLIP Score (Loc C)** To compute it, we isolate each target object by masking the surrounding context with a white background and extracting a square crop centered on its bounding box (scaled by a factor of 1.2). We then measure the cosine similarity between the CLIP embedding of this isolated view and the corresponding local editing prompt, averaging the scores across all instances.

We conduct a **user study** on 40 users, asking them to answer 108 binary choice questions on the three datasets, comparing our approach, FLUX-K and FLUX-K μ T. Specifically, we request to pick the best output between two displayed images from two random models and samples, based on the given source image and edit instruction. Additionally, we prompt **Gemini 3 Flash** in an LLM-as-a-judge setting on the same set of questions. We report results for both these evaluations in terms of Elo score.

12.4.1 RESULTS

In Tab. 12.1, we report an ablation study, performed on LoMOE-Bench, on the effect of applying the proposed masks at different stages of the MMDiT. CLIP scores and $LPIPS_B$ indicate that our proposal, balancing harmonization in the early and late layers with disentanglement in the middle layers, better adheres to the prompt and preserves the background. Moreover, Tab. 12.2 contains an ablation analysis on the effect of IDAttn and the efficient prompt encoding strategy. While the former improves all the scores, efficient prompt encoding slightly lowers them. However, the main benefit of this strategy is in terms of efficiency, and is the only feasible option when editing images with tens or hundreds of instances

L_{early}	L_{mid}	L_{late}	Tgt C $\uparrow$	SD $\downarrow$	IR $\uparrow$	HPS $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	Loc C $\uparrow$	AR [%] $\uparrow$
M^{dis}	M^{dis}	M^{dis}	25.51	<u>0.068</u>	0.566	0.276	0.108	0.910	29.14	94.27
M^{dis}	M^{har}	M^{har}	25.01	0.069	0.145	0.270	0.099	0.928	28.53	82.29
M^{dis}	M^{dis}	M^{har}	25.63	0.070	<u>0.598</u>	<u>0.275</u>	0.100	<u>0.927</u>	29.21	91.15
M^{dis}	M^{har}	M^{dis}	25.03	0.076	0.161	0.272	<u>0.097</u>	<u>0.927</u>	28.62	80.21
M^{har}	M^{har}	M^{dis}	25.11	0.073	0.205	0.272	0.101	0.922	28.63	83.33
M^{har}	M^{dis}	M^{dis}	25.59	0.072	0.602	0.275	0.103	0.916	<u>29.23</u>	93.23
M^{har}	M^{dis}	M^{har}	25.67	0.064	0.278	0.657	0.091	0.926	29.26	92.19
M^{har}	M^{har}	M^{har}	25.05	0.073	0.188	0.270	0.103	0.924	28.54	80.21

Table 12.1: Ablation study on layer scheduling on LoMOE-Bench.

Efficient prompt enc.	IDAttn	Tgt C $\uparrow$	SD $\downarrow$	IR $\uparrow$	HPS $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	Loc C $\uparrow$	AR $\%$ $\uparrow$
		25.22	0.080	0.270	0.247	0.126	0.908	27.93	86.46
✓		25.16	0.079	0.270	0.287	0.129	0.903	27.90	86.46
	✓	25.67	0.064	0.278	0.657	0.091	0.926	29.26	92.19
✓	✓	<u>25.60</u>	<u>0.073</u>	<u>0.277</u>	<u>0.574</u>	<u>0.099</u>	<u>0.919</u>	<u>29.08</u>	89.06

Table 12.2: Ablation study on prompt encoding and application of IDAttn on LoMOE-Bench.

(see Fig. 12.3).

A quantitative comparison of the considered models on LoMOE-Bench is reported in Tab. 12.3. We observe that the end-to-end FLUX baseline has the poorest performance, improved by the multi-turn variant. This indicates that, even with 2-7 edits, FLUX is unable to follow all of the prompts, but is comparatively better at following all of the individual prompts. The addition of visual cues in the form of bounding boxes superimposed on the image does not help in this regard. On the other hand, our approach exhibits solid performance, often ranking best or second-best.

The quantitative results on the Infographics datasets are reported in Tab. 12.4. We can see how the single-pass FLUX baseline shows high text rendering error and cannot follow the majority of the editing instructions, with just 40% AR $\%$ on InfoEdit. On the other hand, our approach, and its finetuned version (**ft**), performs better on most scores, with a notable improvement in terms of CER and AR $\%$ indicating how the model is attempting many more edits which are, on av-

	Tgt C $\uparrow$	SD $\downarrow$	IR $\uparrow$	HPS $\uparrow$	LPIPS $\downarrow$	SSIM $\uparrow$	Loc C $\uparrow$	AR $\%$ $\uparrow$
LoMOE	26.00	0.067	0.266	0.546	0.090	0.834	29.40	98.96
LayerEdit	25.61	<u>0.071</u>	0.252	0.186	0.147	0.864	29.07	100.00
FLUX	24.71	0.078	0.266	-0.059	0.206	0.830	27.58	94.79
FLUX μT	<u>25.71</u>	0.074	0.282	<u>0.550</u>	0.150	0.873	28.27	94.27
FLUX w/ v.c.	24.49	0.072	0.267	0.265	0.170	<u>0.893</u>	27.60	92.71
Ours	25.60	0.073	<u>0.277</u>	0.574	<u>0.099</u>	0.919	<u>29.08</u>	89.06

Table 12.3: Quantitative results on LoMOE-Bench.

	Crello Edit						InfoEdit				
	FID $\downarrow$	MAE $\downarrow$	MSE $\downarrow$	CER $\downarrow$	Δ CER $\downarrow$	AR [%] $\uparrow$	FID $\downarrow$	MAE $\downarrow$	MSE $\downarrow$	CER $\downarrow$	AR [%] $\uparrow$
Calligrapher μ T	10.15	<u>3.82</u>	<u>8.58</u>	0.73	0.29	51.21	113.23	40.34	69.80	0.92	99.98
Calligrapher st.	14.40	2.06	8.46	0.69	0.24	99.47	13.37	39.09	70.72	0.82	99.31
FLUX	<u>10.06</u>	61.83	91.73	0.65	0.20	68.72	4.36	30.20	64.26	0.77	39.94
FLUX μT	12.10	62.13	91.32	0.63	0.19	90.44	65.73	35.17	66.76	0.90	99.81
FLUX st.	15.69	61.33	90.90	<u>0.59</u>	0.15	73.49	10.48	29.46	62.67	0.66	63.25
Ours	9.45	61.72	91.58	0.61	<u>0.16</u>	52.00	2.41	<u>3.41</u>	<u>15.33</u>	<u>0.64</u>	52.61
Ours + ft	10.85	61.35	90.86	0.52	0.07	92.16	<u>2.80</u>	3.22	13.41	0.56	80.90

Table 12.4: Quantitative results on the Infographic Editing Benchmark.

erage, more correct. Moreover, the lower background scores indicate the absence of artifacts in those regions. Interestingly, these scores are comparable with those of FLUX st, which does not have background artifacts by design. Note that, by design, the stitching models also have a higher AR%, as each edit is performed separately and stitched upon the inpainted background, resulting in a false-positive.

To further evaluate visual style fidelity on the edited regions, we compute the HWD [1], SSIM, and FID scores restricted to the boxes that each model attempted to edit. HWD and SSIM require paired ground-truth edited images, so they are reported only for Crello Edit; FID is reported for both datasets. The results are shown in Tab. 12.5. Our fine-tuned variant achieves the best performance across all metrics, confirming that LoRA adaptation improves text-rendering style quality without degrading background reconstruction.

Although HWD was originally designed for handwritten style fidelity, we also use it here to quantify the consistency of rendered text appearance across edited infographic boxes, capturing stroke character, spacing, and overall writing-like

	Crello Edit			InfoEdit
	HWD↓	SSIM↑	FID↓	FID↓
Calligrapher μT	0.69	0.30	41.35	183.21
Calligrapher st.	0.69	0.33	25.65	28.90
FLUX	0.43	0.40	13.10	29.52
FLUX μT	0.28	0.45	13.12	109.72
FLUX st.	0.39	0.41	11.62	17.25
Ours	0.37	0.39	13.56	21.28
Ours + ft	0.20	0.56	7.82	7.10

Table 12.5: Style fidelity scores computed on the boxes attempted by each model. HWD and SSIM are computed on Crello Edit only (paired ground-truth available); FID is computed on both datasets.

visual similarity in a feature space specialized for text style.

Additionally, we analyze how editing capabilities scale when increasing the number of edits in Fig. 12.2. We group samples in our Infographics benchmark by number of edits, with up to 20 randomly selected samples per group, and compute the average CER and $AR_{\%}$. We can see that our model scales better even on samples with tens of editing instances, while other methods ignore the editing instructions and commit more text rendering mistakes.

The results of our user study and the LLM-as-Judge evaluation are reported in Tab. 12.6. We observe that both the users and Gemini strongly favor our method, with the FLUX end-to-end baseline being the second-best. Noting the strong correlation between the automated LLM-as-a-judge results and the user study, we report results of an extended LLM-as-a-judge evaluation with more evaluated baselines and samples in Appendix E.4. This extended evaluation is important because some competitors, especially LoMOE, attain strong CLIP-based scores while being judged less favorably by both users and Gemini, suggesting that CLIP-style metrics can overestimate the perceptual quality of multi-instance edits.

Finally, in Fig. 12.4, we present qualitative results on samples from the three datasets. In the first image, which has 6 editing instructions, the end-to-end

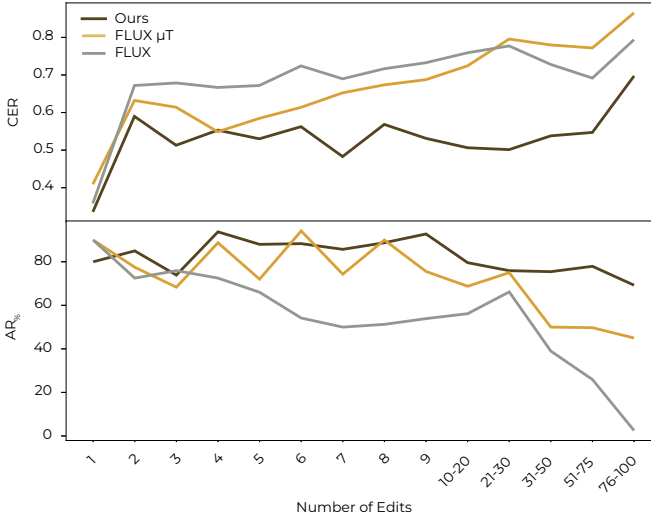


Figure 12.2: CER and AR w.r.t. the number of edits.

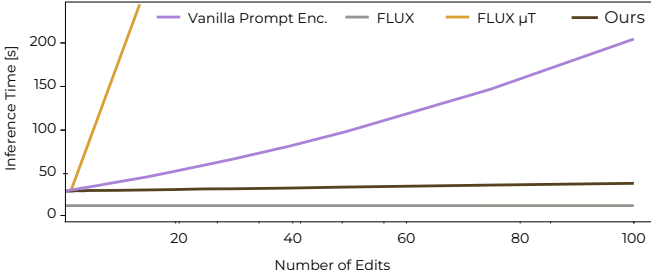


Figure 12.3: Inference time comparison w.r.t. the number of edits.

FLUX baseline ignores half the prompts and incorrectly edits an area (where the word *AND* appears) despite the editing instruction entailing rewriting *AND* there. The multi-turn and stitching variants attempt all edits but make several mistakes, while our approach offers a good balance between editing quantity and quality. On the sample from InfoEdit, the FLUX only attempts one edit in the large title area with some rendering mistakes. Due to the high number of iterations, FLUX μ T severely degrades image quality. FLUX st. fares somewhat better by attempting all edits with higher success, but introduces many artifacts and character errors. Our method is capable of editing almost all of the instances,



Figure 12.4: Qualitative results on **Crello Edit**, **InfoEdit** and **LoMOE-Bench**. The leftmost column is the source image, the second one is the source image with overlaid bounding boxes indicating areas in which to perform editing operations, and the other columns contain qualitative results on that same sample. For readability, we list the prompts for these samples in Appendix E.5.

introducing minor artifacts, and only missing a few characters for one of the text rewriting prompts. On LoMOE-Bench, FLUX performs the requested edits but renders slightly unrealistic hats with no shadows. The multi-turn variant introduces artifacts and extra textures, which are even more evident for the variant exploiting visual cues (it draws colored rectangles, disrupting visual coherence and fidelity to the source image). Our approach successfully performs all of the requested edits. The prompts used to generate these samples are listed in Appendix E.5, for readability, and more qualitative results for all three datasets are shown in Appendix E.6.

	LoMOE-Bench		Infographics	
	Users	Gemini	Users	Gemini
FLUX	<u>1331</u>	<u>1319</u>	<u>1102</u>	<u>1109</u>
FLUX μT	680	695	1095	1096
Ours	1589	1585	1404	1394

Table 12.6: Elo scores of our approach on natural images and Infographics with a User Study and LLM-as-a-judge.

12.5 CONCLUSION

In this work, we have explored the challenges of multi-instance editing within the framework of Flow Matching and MMDiT. We have identified that the reliance on global velocity fields and joint attention mechanisms can lead to semantic entanglement when multiple instructions are processed concurrently. To address this, we have proposed IDAttn, a modification that seeks to localize attention operations by associating specific textual tokens with their corresponding spatial regions. Moreover, we have devised a multi-prompt encoding strategy that makes inference more efficient. Our experimental results on natural images and our devised benchmark of challenging infographic text editing suggest that our approach can help mitigate attribute leakage and improve prompt following ability while maintaining global coherence. Furthermore, by performing edits in a single inference pass, the method offers a more efficient alternative to multi-step or iterative editing pipelines as the number of instances increases. Finally, it is worth noting that, while our implementation was integrated into the FLUX architecture, the concept of attention partitioning applies straightforwardly to other Transformer-based flow models, particularly those based on MMDiT.

LIMITATIONS. Like prior multi-instance editing approaches [62, 117], our method relies on external inputs to localize instances (*i.e.* bounding boxes or segmentation masks), and is therefore not end-to-end. Devising an approach that jointly localizes and edits instances remains an interesting future direction, which could be addressed, for example, by orchestrating agentic pipelines that combine promptable detection or segmentation for natural images, and text detection with OCR

for infographics, with an MMDiT-based editor such as ours. Our method is robust to imprecise localization (*e.g.* boxes that are too tight or too broad), but can fail when the provided boxes are entirely incorrect. Artifacts arising from the hard attention mask boundaries are mitigated by the harmonization regime applied in the early and late layers, which blends instance-level features back into a globally coherent representation.

Part IV

Takeaways

13

Conclusions

THIS dissertation has presented a comprehensive body of work addressing the analysis, generation, and editing of Visually Rich Documents (VRD). We pursued this goal by exploring novel architectures, training strategies, and evaluation protocols tailored to the unique challenges of structured layouts, severely degraded historical manuscripts, and complex multi-modal interactions. The contributions spanned three main themes corresponding to the parts of this thesis: (i) analyzing and understanding VRD, from 3D tomographic scans to low-resource languages; (ii) generating realistic handwritten and graphical content with increased control and zero-shot style adaptation; and (iii) editing multi-instance compositions while maintaining global structural coherence. In the following, we summarize the main findings by answering the interconnected research objectives posed at the beginning of this journey, and outline possible directions for future research.

13.1 SUMMARY OF CONTRIBUTIONS

Recall that we began this thesis by identifying a set of interconnected research objectives in Chapter 1. These can be distilled into the following key questions:

- **Q1:** How can we effectively analyze and parse complex documents, particularly in scenarios characterized by severe degradation, complex layouts, or limited annotated data?
- **Q2:** How can we design generative models for handwriting that seamlessly adapt to unseen styles zero-shot, and how can we systematically evaluate their true style fidelity?
- **Q3:** How can we adapt powerful generative paradigms, such as diffusion models, to synthesize visually rich graphical content with diverse structures, variable aspect ratios, and transparency?
- **Q4:** How can we achieve disentangled, multi-instance control for fine-grained document editing, preventing unintended attribute leakage across different regions?

Through the research work presented in this thesis, we answered these questions with our main contributions, detailed below.

DOCUMENT ANALYSIS IN DEGRADED AND LOW-RESOURCE SCENARIOS (Q1). We demonstrated that frequency-domain processing offers a powerful mechanism for document restoration. By proposing FourBi for 2D binarization (Chapter 3), we showed that capturing long-range dependencies via Fast Fourier Convolutions is crucial for handling complex, severe degradation patterns. As a further proof of the generality of this spectral approach, we extended it to the volumetric domain with vFFC, demonstrating preliminary applicability to 3D tomographic ink detection on carbonized ancient scrolls. Moreover, to parse multi-page documents effectively, we introduced μ gat (Chapter 4), proving that incorporating global cross-page context via cross-attention significantly reduces character error rates in structured historical and academic publications. Finally, to address the profound data scarcity in ancient scripts, we introduced the SCAM dataset (Chapter 5) for Sahidic Coptic manuscripts, establishing a new baseline and highlighting the architectural strategies necessary for effective low-resource Handwritten Text Recognition (HTR). These experiments also revealed that standard Character Error Rate underestimates recognition difficulty on continuous scripts with

no word boundaries, motivating the joint use of Sentence Error Rate as a more faithful measure of practical transcription quality.

ZERO-SHOT GENERATION AND EVALUATION OF STYLED HANDWRITING (Q2). We devised novel architectures for Styled Handwritten Text Generation (HTG) that adapt to unseen writing styles without costly per-style fine-tuning. We introduced Emuru and Eruku (Chapter 8), two autoregressive Transformer frameworks. Emuru bridges textual content with visual appearance using continuous latent representations with a fully-synthetic training strategy, a robust vision autoencoder, and automatic prediction of generation termination. Eruku leverages CFG, a better training recipe, and a combination of continuous and discrete token prediction to remove the dependency on style image transcriptions. Both models can generate arbitrary-length text with strong zero-shot generalization to diverse, real-world scripts. Crucially, recognizing the limitations of generic metrics like FID, which easily entangle background artifacts with calligraphic features and whose square-crop approximation discards the rightmost part of text images, we proposed the Handwriting Distance (HWD) (Chapter 7). HWD computes distributional distances in the feature space of a handwriting-specific backbone, resulting in an evaluation protocol that is demonstrably more robust, numerically stable, and aligned with human judgment. This contribution exposes a flaw that affected all prior HTG evaluation, not just the models proposed in this thesis. Notably, HWD was subsequently adopted as an evaluation score within the IDAttn infographic editing experiments, demonstrating that evaluation tools developed in the generation part of this thesis directly inform the editing part. Supplementary analyses further indicate that this autoregressive framework can be extended beyond Latin scripts with suitable synthetic pretraining data, and that reducing or removing reliance on noisy style-text transcriptions is crucial for robust deployment in realistic settings.

CONTROLLABLE GENERATION OF GRAPHICAL CONTENT (Q3). Moving beyond text synthesis, we tackled the structural constraints inherent in generating graphical documents. We proposed the Merge-Attend-Diffuse (MAD) operator (Chapter 9), a training-free attention synchronization mechanism that coordi-

nates multiple diffusion views, enabling the generation of structurally coherent, artifact-free panoramas of arbitrary aspect ratios. Additionally, to address challenges of modern graphic design and printing workflows, we introduced Alfie (Chapter 10). By precisely modulating the attention maps and Classifier-Free Guidance within pre-trained Diffusion Transformers, Alfie enables the automated generation of detailed RGBA illustrations with native transparency, without requiring any expensive architectural retraining or specialized matting models.

DISENTANGLED MULTI-INSTANCE DOCUMENT EDITING (Q4). Finally, we addressed simultaneous multi-instance editing within complex layouts (Chapter 12). We introduced IDAttn (Instance-Disentangled Attention), an approach that prevents semantic interference and attribute leakage when modifying multiple localized regions in flow matching models. Combined with our newly introduced Infographic Editing Benchmark (comprising Crello Edit and InfoEdit), IDAttn demonstrated superior disentanglement and precise local text editing capabilities, setting the first baseline for concurrent multi-instance editing at infographic scale. Crucially, by operating in a single inference pass, IDAttn also addresses the scalability bottleneck that makes sequential or multi-branch methods impractical when tens or hundreds of instances must be edited simultaneously. This ensures that modifications respect both the semantic meaning and the visual style of individual document elements without inducing any unintended global changes across the structured composition.

13.2 OPEN CHALLENGES AND FUTURE DIRECTIONS

While this thesis has advanced the state of the art in VRD processing, the contributions presented here open several avenues for future research.

UNIFYING UNDERSTANDING AND GENERATION. A recurring theme across this thesis is the interplay between understanding and generation of VRD. While our works have addressed these capabilities through specialized models, future research could explore unified architectures that jointly perform document understanding and generation within a single framework. Recent advances in large

multimodal models suggest that such unification is increasingly feasible, potentially leading to more versatile and capable document processing systems.

SCALING TO DIVERSE DOCUMENT TYPES. Our contributions have focused on specific document types, including historical manuscripts, academic papers, and infographics. An important direction for future work is to extend these methods to a broader range of visually rich documents, such as forms, invoices, and complex multi-modal layouts. This would require developing more generalizable models and richer datasets that capture the diversity of real-world documents.

ADVANCING LOW-RESOURCE AND HISTORICAL DOCUMENT PROCESSING. The work on SCAM and on the Herculaneum papyri highlights the potential of computational methods for preserving and accessing cultural heritage. Future research could investigate the combination of state-of-the-art recognition models with generative data augmentation techniques, such as styled handwritten text generation, to improve HTR performance in low-resource settings. In particular, the generation of synthetic training data in the style of rare scripts could help bridge the data scarcity gap that currently limits the applicability of deep learning to endangered and extinct languages. A concrete next step would be to use synthetic rare-script generation as an additional data source for low-resource HTR settings such as SCAM, where annotation scarcity remains a primary bottleneck.

IMPROVING CONTROLLABILITY AND EDITABILITY. The methods proposed in this thesis for controllable generation (MAD, Alfie) and editing (IDAttn) represent steps towards more fine-grained user control over generative models. Future work could focus on developing richer conditioning mechanisms, such as natural language instructions, spatial constraints, or interactive feedback loops, allowing users to iteratively refine generated content. Additionally, extending multi-instance editing to handle more complex relationships between edited regions, such as preserving visual harmony and stylistic consistency across concurrent modifications, remains an open challenge.

EVALUATION METHODOLOGY. The HWD score demonstrated the importance of domain-specific evaluation metrics. As generative models continue to diversify across document types and modalities, the development of tailored evaluation

protocols that align with human perception and task-specific requirements will become increasingly important. A recurring observation in this thesis is that standard metrics systematically fail in specialized settings: FID entangles background and calligraphic style in HTG, standard CER underestimates difficulty in scriptio continua HTR, and heuristic pixel-level scores cannot always distinguish attempted edits from ignored instructions in multi-instance editing. Future work could formalize this pattern into a unified framework for benchmark design in Document AI, extending the HWD paradigm to layout generation, document design synthesis, and multi-modal editing tasks.

13.3 FINAL REMARKS

This thesis has addressed the full lifecycle of Visually Rich Document processing, and a number of its results stand as concrete demonstrations of what is now possible. Viewed as a whole, the dissertation progresses from recovering difficult visual evidence, to structuring and recognizing it, to controlling generation, and finally to composing precise edits in dense document-centric settings. A model trained on spectral convolutions can detect ink in multi-resolution 2D documents and in 3D tomographic scans of 2000-year-old carbonized scrolls that have been physically inaccessible for centuries. A generative model trained entirely on synthetic data can reproduce the handwriting style of a writer it has never seen, in a language it has never encountered, automatically adapting the output resolution depending on the textual content and the style. An attention-partitioning mechanism applied at inference time, requiring no architectural changes and no mandatory retraining, can simultaneously apply dozens of independent edits to a complex infographic without instances interfering with each other.

Each of these results reflects a broader lesson: that the unique challenges of Visually Rich Documents, namely structured layouts, rare scripts, heterogeneous degradation, and compositional complexity, are not obstacles but design constraints that lead to better-principled and more general methods. The evaluation tools, datasets, and benchmarks introduced alongside the models are, in this sense, contributions of equal importance: they define what good performance

means.

We hope that the methodologies, datasets, and benchmarks introduced in this thesis serve as stepping stones for future researchers, helping make Document AI more capable and accessible, and that the problems identified here, from low-resource ancient scripts to dense multi-instance infographic editing, continue to challenge and inspire the community.

13.3.1 PUBLICATIONS AND ACHIEVEMENTS

The majority of the works presented in this thesis have been published in prestigious international conferences and workshops:

1. **Document Image Binarization** (Chapter 3): published as an oral presentation at ICDAR 2024 [4]. The companion work on volumetric Fast Fourier Convolutions for ink detection on the Herculaneum papyri appeared as an oral presentation at the 4th ICCV Workshop on e-Heritage 2023 [2].
2. **Multi-page Document Parsing** (Chapter 4): published at the AI4DH: Artificial Intelligence for Digital Humanities workshop at ECCV 2024 [8].
3. **Sahidic Coptic Ancient Manuscripts dataset** (Chapter 5): submitted to ICDAR 2026, under review [7].
4. **Handwritten Text Generation evaluation** (Chapter 7): published at BMVC 2024 [1].
5. **Autoregressive styled text image generation** (Chapter 8): Emuru was published at CVPR 2025 [6] and Eruku at WACV 2026 [10].
6. **Semantically coherent panorama generation** (Chapter 9): published at ECCV 2024 [5].
7. **Automatic RGBA image generation** (Chapter 10): published at the AI for Visual Arts Workshop at ECCV 2024, receiving an oral spotlight presentation (Top-4) [3].
8. **Multi-instance Infographic Editing** (Chapter 12): accepted at ICML 2026 [9].

13.4 PH.D. ACTIVITIES

In this final section, I present a list of the main activities I carried out during my Ph.D. program in Information and Communication Technologies (ICT).

13.4.1 EXCHANGE PERIODS

15 September 2024 - 14 March 2025: Applied Scientist Intern at Amazon in the “Computer Vision North America - CVNA” team, supervised by Dr. Bojan Pepik and Dr. Michael Opitz.

13.4.2 RESEARCH COLLABORATIONS

International Collaborations: Joint research with Google and ETH Zürich on autoregressive generation and image editing.

13.4.3 PARTICIPATIONS TO NATIONAL AND INTERNATIONAL PROJECTS

ITSERR PhD Student: My research activity within the ITSERR project (Italian Strengthening of the ESFRI RI RESILIENCE) focused on developing multimodal AI models for the analysis and generation of historical documents.

13.4.4 CONFERENCE ATTENDANCES

02 - 06 october 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, *Paris, France*.

29 September - 4 October 2024: European Conference on Computer Vision (ECCV), 2024, *Milan, Italy*.

30 August - 4 September 2024 IAPR International Conference on Document Analysis and Recognition (ICDAR), 2024, *Athens, Greece*.

11 - 15 June 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, *Nashville, US*.

13.4.5 SEMINARS AND WORKSHOPS

17 - 18 July 2023: 1st ITSERR in-house training meeting at Fondazione per le Scienze Religiose (FSCIRE), *Bologna, Italy*;

09 - 10 June 2024: 2nd ITSERR in-house training meeting at DISMI (Uni-MoRe), *Reggio Emilia, Italy*;

10 - 11 April 2025: 6th ITSERR in-house training meeting at ISTI - CNR, *Pisa, Italy*;

11 - 12 September 2025: 8th ITSERR in-house training meeting at the University of Palermo, *Palermo, Italy*.

13.4.6 SCHOOLS

4 - 8 September 2023: Attendance and completion of the “International Summer School on Machine Vision - VISMAL 2023” summer school.

7 - 13 July 2024: Attendance and completion of the “International Computer Vision Summer School - ICVSS 2024” summer school.

18 - 22 September 2023: Attendance and completion (Best Project Award) of the “ELLIS Summer School on Large-Scale AI for Research and Industry” summer school.

13

13.4.7 CONFERENCE SERVICE

OUTSTANDING REVIEWER.

British Machine Vision Conference (BMVC), 2024, *Glasgow, UK*;

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, *Nashville, US*.

REVIEWER.

British Machine Vision Conference (BMVC), 2023, *Aberdeen, UK*;

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, *Seattle, US*;

European Conference on Computer Vision (ECCV), 2024, *Milan, Italy*;

British Machine Vision Conference (BMVC), 2025, *Sheffield, UK*;

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2026, *Denver, US*;

International Conference on Document Analysis and Recognition (ICDAR), 2026, *Vienna, Austria*;

European Conference on Computer Vision (ECCV), 2026, *Malmö, Sweden*.

WORKSHOP ORGANIZER.

Automatically Domain-Adapted and Personalized Document Analysis (ADAPDA) at 20th International Conference on Document Analysis and Recognition (ICDAR), 2026, *Vienna, Austria*.

13.4.8 MASTER THESIS CO-ADVISING

Carmine Zaccagnino, “Improving Document Parsing for Historical Documents with Multi-Page Context”, 2024.

Appendices



HWD: A Novel Evaluation Score for Styled Handwritten Text Generation

A.1 ADDITIONAL RESULTS ON THE SENSITIVITY TO HAND- WRITING

In this section, we report the results on the sensitivity of the HWD and the FID to the handwriting style obtained on the Norhand [236] and BanglaWriting [253] multi-author datasets. We consider half of the samples for each featured writer as references and the other half as if they were the output of a perfect Styled HTG model. Then, we compare the distributions of the HWD and the FID values computed on text images of multiple matching and non-matching authors pairs.

The obtained distributions are reported in Figure A.1.

A.2 FURTHER COMPARISON BETWEEN HTG APPROACHES

In Table A.1, we report some qualitative examples of images generated by two HTG approaches being scored with both FID and HWD. These show that HWD better separates cases in which HTG models perform one better than the other,

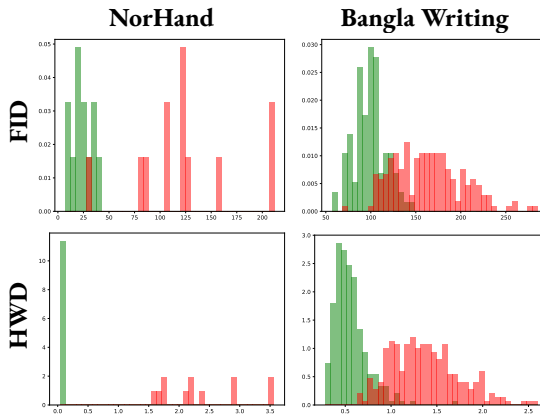


Figure A.1: Distributions of different scores used to evaluate HTG models when applied on same-author (green) or different-author (red) subsets. The overlap area is in dark red.

Reference	VATr		HWT	
	HWD	FID	HWD	FID
<i>you as the</i>	0.84	128.4	1.09	128.7
<i>air is so</i>	0.70	111.7	0.93	111.9
<i>the IF UN</i>	0.60	86.6	0.92	86.7

Table A.1: Qualitatives, FID, and HWD of HTG models.

compared to the FID, which has similar values both for good cases and failure cases.

A.3 ALTERNATIVE METRIC DISTANCES FOR HWD

In Table A.2 we report the results obtained on the IAM dataset when using Mahalanobis and the Hamming distances in the final step of the HWD computation. It emerges that the Euclidean distance works best for HWD, leading to the smaller Overlap and EER.

Distance	Overlap	EER
Mahalanobis	7.4	3.6
Hamming	4.3	2.1
Euclidean	0.7	0.3

Table A.2: Ablation results when changing the distance metric at the final step of HWD.

A.4 ADDITIONAL RESULTS ON THE SENSITIVITY TO THE NUMBER OF SAMPLES

In this section, we report further results on the numerical stability of the proposed HWD, compared to the FID and two baseline scores, namely the **FID w/ Euclidean** (obtained by computing the Euclidean distance on the Inception-v3 features) and **HWD w/ Fréchet** (obtained by computing the Fréchet distance on the VGG16). In particular, we use the images from the considered single-author datasets (ICFHR14 [317], Saint Gall [113], Leopardi [58], Rodrigo [326], Washington [114], and LAM [59]). The results are expressed as mean and range between the 25th and 75th percentiles of the values obtained over multiple runs by varying the number of samples. These are reported in Figure A.2. For the plots in each row, we consider the whole indicated dataset as the set of reference images and compute the score when comparing it with a variable number of samples from the other datasets and from the dataset itself for reference.

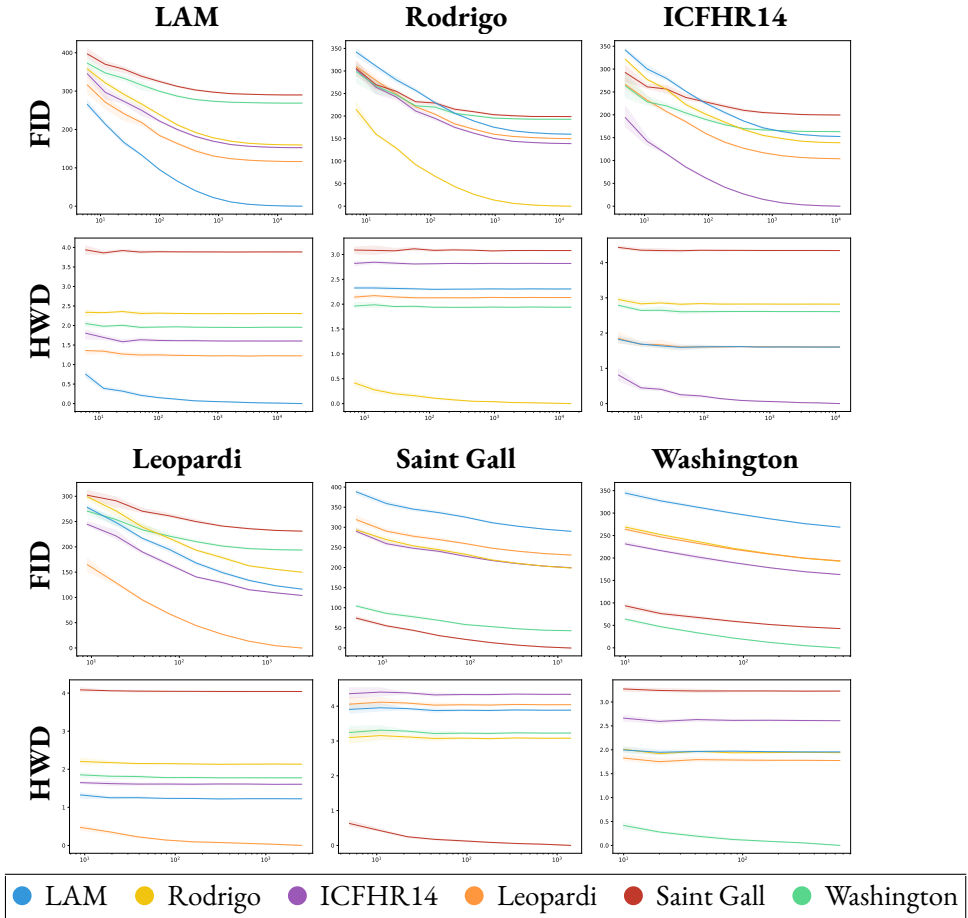


Figure A.2: Comparison between FID and HWD with varying number of samples on different single-author datasets. The lines denote the mean, and the transparent bands represent the range between the 25th and 75th percentiles, obtained with 10 calculation runs.

B

Zero-Shot Styled Text Image Generation with Emuru

B.1 FURTHER DETAILS ABOUT THE TRAINING DATASET

The proposed synthetic dataset* comprises RGB images that feature text lines superimposed on diverse background patterns (Fig. B.2). The process to create a text image is divided into two macro-steps: Text Content Sampling and Text Line Rendering.

TEXT CONTENT SAMPLING. First, we formed the text lines to render in the samples of the dataset. To this end, we consider the following English corpora available on the NLTK[†] library: `abc`, `brown`, `genesis`, `inaugural`, `state_union`, and `webtext`. To ensure a balanced character distribution and to address the learning of rare characters, we employ a rarity-based weighting strategy. Specifically, each word in the corpora is assigned a weight, allowing for weighted sampling from the resulting distribution. The weight of each word is computed based on the frequency of unigrams and bigrams contained within it. This ap-

*<https://huggingface.co/datasets/blowing-up-groundhogs/font-square-v2>

[†]<https://www.nltk.org/>

proach prioritizes words that include infrequent character patterns, enabling the HTG model to learn from the long-tail distribution of characters more effectively. The Python implementation of the algorithm for computing the weight of each word is reported in Listing B.1. The function `word_weight(word)` calculates a score for each word based on the unigram and bigram frequencies, stored in `u_counts` and `b_counts`, respectively. These counts represent the frequencies of unigrams and bigrams across the entire corpus. The resulting word weight is the average of the unigram and bigram-based scores. In Fig. B.1, we show the character’s distribution with the custom weights per word and the naive sampling (original). Each bar of the distribution represents a character in our charset (which contains 157 characters) in this order: [, e, a, n, s, i, r, t, o, l, A, ., m, c, d, u, 0, S, h, E, p, M, g, C, R, 1, I, -, D, b, T, 2, N, 3, O, y, P, B, L, ', F, k, f, H, 4, 5, G, U, v, 7, 8, ", 6, 9, V, K, x, w, Y,), ä, W, ,, *, z, :, J, (, X, !, j, /, _, é, >, Z, ?, Q, ;, \$, å,], ö, q, %, <, [, á, ã, #, i, &, }, +, ü, â, ó, Å, ê, =, {, ô, \, í, è, », ¨, É, ú, @, À, à, ` , ç, ù, ~, ", ' , \$, ", µ, ", Å, ¢, °, Å, ®, ´, -, ¢, Ä, û, Ó, ", f, ¤, ^, ¨, ~, |, î, ë, |, Š, ö, Ö, Á, °, Å, ¢, ¤, Ü, ß, ž, ñ, ĺ, d].

Note that, our current Emuru is trained on a dataset derived from fonts covering primarily Latin characters, there is nothing in the approach itself that prohibits extension to non-Latin scripts. Specifically, to extend Emuru to non-Latin scripts, its VAE should be re-trained on a similarly-built synthetic dataset of non-Latin strings, rendered with fonts that support the desired characters. Given that there are a number of such fonts available online, we argue that there is no obvious method or model limitation as to why the results of Emuru would not translate also to non-Latin scripts.

TEXT LINE RENDERING. Once we sample a sentence, we select a font from a list of 100,000 fonts and render the text image on a white background, obtaining $\hat{I}_T$. Then, we sample a background $\hat{I}_B$ from a set of images depicting reasonable supports for writings (*e.g.* paper-like textures, wood, walls), a transparency value $\alpha \in [0.5, 1]$ and process $\hat{I}_T$ and $\hat{I}_B$ in parallel. To obtain I_T (Fig. B.2, center), we use the following random transforms: Rotation, Warping, Gaussian Blur, Dilation, and Color Jitter (with probabilities respectively set to 0.5, 1.0, 0.5, 0.1, and 0.5) and apply the transparency α . To obtain I_B (Fig. B.2, right), we use the fol-

```
def word_weight(word):
    # Compute the score
    # based on the unigrams
    u_score = 0
    for c in word:
        u_score += u_counts[c]
    u_score /= len(word)

    # Compute the score
    # based on the bigrams
    bigrams = pairwise(f' {word} ')
    bigrams = list(bigrams)
    b_score = 0
    for b in bigrams:
        b_score += b_counts[' '.join(b)]
    b_score /= len(bigrams)

    # Average the two scores
    return (u_score + b_score) / 2
```

Listing B.1: Python code for computing the words weight, needed for the character frequency balancing procedure adopted to obtain the text in our synthetic training dataset.

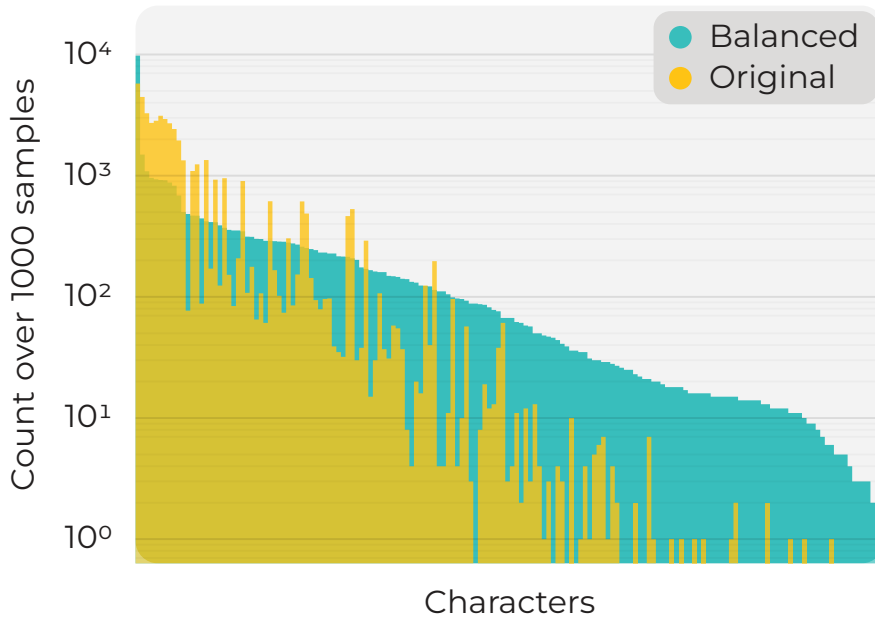


Figure B.1: Characters frequency in text from the original corpora used to generate the synthetic training dataset, and from the balanced version of the corpora.

lowing random transforms: Dilation, Color Jitter, and Random Inversion of the values (with probabilities respectively set to 0.1, 0.5, 0.2). The image I (Fig. B.2, left) is the result of superimposing I_T on I_B .

B.2 DETAILS ABOUT THE CONSIDERED DATASETS

In this section, we give further details about the datasets we use for evaluation.

IAM. The IAM Handwriting Database is a collection of greyscale document scans written in English by multiple writers. The dataset features free-layout modern English text lines from the Lancaster-Oslo/Bergen (LOB) corpus. The number of pages per writer varies significantly, ranging from 1 to 10. The dataset comes with both line-level and word-level segmentation information. The HTG community commonly used this dataset for training and evaluation of the proposed approaches and has adopted a split defined in [178], in which the samples



Figure B.2: Exemplar sample images I from the synthetic dataset used for training Emuru (left). These samples consist of a rendered greyscale text image I_T (center) superimposed to RGB paper-like background images I_B (right).

from 339 writers are used for training, and those from 161 other writers are used for test. In this work, we adopt the same split. However, while most of the competitors are trained on IAM (and, therefore, see the test set as in-distribution samples), we never use any training data from it to train our model (therefore, test samples are out-of-distribution for us).

CVL. The CVL Database is a collection of RGB scans of English and German manuscripts written with ink on white paper by 310 writers. In this work, we consider the line-level annotation of the dataset and perform the evaluation on its test set, which contains lines from 283 different authors. We include this dataset since it is mostly in the same language as the commonly-adopted IAM, but the handwriting styles are out-of-distribution with respect to IAM.

RIMES. The RIMES Database is a collection of binary images of customer service-themed letters in French, written by multiple authors. For our experiments, we consider the lines in the official test split of the dataset, which have been written by 100 authors. We include this dataset as a more challenging collection of out-of-distribution generation cases. In fact, both the language and the styles differ from those in IAM.

KARAOKE.[‡] To assess the capabilities of Emuru and the existing handwriting style imitation approaches, we devise a dataset of text lines images obtained by rendering song lyrics with 100 publicly available fonts[§], 50 calligraphy, and 50 typewritten, on a white background. This dataset, especially its typewritten split, serves as a challenging test to measure the performance of HTG models when dealing with significantly out-of-distribution, font-like styles.

B.3 DETAILS ABOUT THE CONSIDERED SCORES

Measuring the performance of styled HTG models is a challenging task per se. Early works on HTG simply employ the Fréchet Inception Distance (FID) to measure the realism of the generated images, and sometimes the CER to measure their readability. Recent works [263, 379][1] have highlighted the limitations of such scores (especially the FID) and thus proposed to assess the performance of HTG models both by using multiple scores or introducing novel, task-specific scores. In line with these works, for our evaluation, we adopt the following scores (some of which have been reported in Chapter 8).

FRÉCHET INCEPTION DISTANCE (FID). The FID is widely employed in HTG literature and captures the realism/similarity between the distribution of the real images and the generated ones. In the context of HTG, it somehow captures the texture-wise style similarity. Note that in most HTG works, the reported FID is computed only on the initial square crop of the real and generated images. In this work, we compute it on non-overlapping crops obtained from the entire images.

KERNEL INCEPTION DISTANCE (KID). Similar to the FID, also the KID captures the realism of the generated images by comparing their distribution to the real images distribution. Despite being less used in image generation evaluation, it is more numerically stable than the FID.

BINARIZED FID (BFID). To reduce the impact of the texture (both of the strokes and the background) on the similarity between generated and real images, we in-

[‡]<https://huggingface.co/datasets/blowing-up-groundhogs/karaoke>,
<https://open.spotify.com/playlist/5GTPvzVNedUIkYkV7x0tgI>

[§]<https://fonts.google.com/>

<https://open.spotify.com/playlist/5GTPvzVNedUIkYkV7x0tgI>

troduce this variant of the FID score by computing it on binarized images. Intuitively this measure should capture mostly the perceptual handwriting style and disregard color or textures.

BINARIZED KID (BKID). Similar to the BFID, this score is obtained by computing the KID on the binarized reference and generated images.

ABSOLUTE CHARACTER ERROR RATE DIFFERENCE (Δ CER). We consider the State-of-the-Art TrOCR-Base model, which can handle both handwritten and typewritten images and compute its CER on the reference and generated images to capture their readability while taking into consideration the style. Intuitively, we want the generated images to be as readable as the reference ones (which suggest that they actually contain text) but not significantly more, which could be that the HTG model has collapsed to a very readable style that does not necessarily resemble the reference.

HANDWRITING DISTANCE (HWD). The HWD is a perceptual score that has been recently proposed specifically for HTG and captures the similarity in terms of calligraphic style between the real and generated images pairs. For this reason, we include it in our evaluation setup.

ABSOLUTE INTRA LEARNED PERCEPTUAL IMAGE PATCH SIMILARITY DIFFERENCE (Δ I-LPIPS). To measure the style preservation over long generated images, we split each image into non-overlapping crops and compute the Learned Perceptual Image Patch Similarity (LPIPS) (which is based on the similarity of feature maps) between all possible combinations of crops of the same image. We repeat this process for both the real and the generated image separately and then compute the absolute difference between these values. We want its value on the generated images to be similar to what is obtained on the real ones (to ensure mimicking the degree of variability of the reference style) and not significantly lower, which could indicate mode collapse to a repeated stroke.

B.4 ANALYSIS ON THE STYLE INPUT LENGTH

As mentioned in Section 6.2, existing HTG methods have a variety of settings for the style reference image, ranging from single words to entire paragraphs. These approaches offer different advantages depending on the use case, with some models demonstrating some adaptability to a style from minimal input and others leveraging more context for improved style fidelity. We argue that an HTG model should be flexible to the number of input words available, handling both data-scarce scenarios with limited reference samples and situations with more style images available. To deepen the analysis on this aspect, we evaluate the performance of Emuru when generating IAM lines given increasingly longer reference style images, which we build starting from the word-level annotations of the IAM dataset. Note that the IAM dataset contains all the information needed to determine which sequence of words appears within each line and where to cut the line image to obtain a new line image with the desired width. Therefore, to generate the set of style samples within a specific width range, we exploit the IAM information of the bounding boxes around the words inside the lines. Specifically, we group images whose width is in a range of ± 25 px around average width values that vary from a minimum of 125px to a maximum of 1025px with steps of 50px. In Fig. B.3, we report the results of this analysis. As expected, the performance improves with increasing input length, showing that our model is flexible to different scenarios and more applicable to real-world handwriting generation tasks.

B.5 ANALYSIS ON THE STYLED OUTPUT LENGTH

To assess the optimal output text length manageable by Emuru, we perform experiments by increasing the number of output characters to render in the styles from the Karaoke dataset. We report FID and CER scores in Fig. B.4, representing style fidelity and text readability. We observe that the model performs best when generating text lines between 25 and 75 characters, which aligns with the training sequence length. Outside this range, Emuru exhibits performance degradation in both style fidelity and text readability. Note that some workarounds for

improving the quality of generated long sequences can be applied, as mentioned in the following. In Fig. B.5, we report qualitative results of image generation with Emuru for increasingly longer text lines. In the top part of the figure, we feed Emuru with the reference style image and the entire text line to be generated all at once. As we can see, the model maintains style consistency across long outputs but struggles to correctly render all the words as their number increases. This discrepancy between the input and the output text is reflected in the Δ CER score, as observed in the quantitative analysis in Section 8.1.2. To improve this aspect, in the bottom part of Fig. B.5, we also show a simple yet effective solution for generating longer lines with increased text fidelity. Specifically, we let Emuru iteratively generate one word at a time by using the previously generated line as a style reference. This incremental approach improves the alignment of the output image with the input text with a computational overhead of 29.5% (2.24s), mainly due to the additional 10 padding embeddings that the model has to generate as a stop signal after each new word generation.

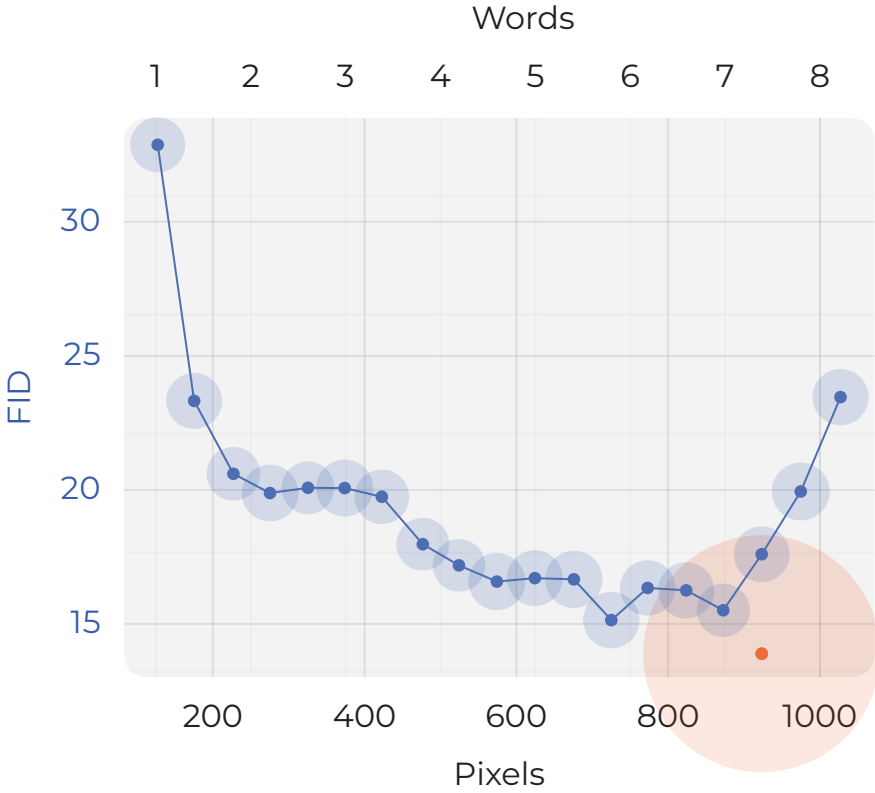


Figure B.3: Emuru FID scores for IAM line generation using style inputs of varying length (in blue), compared to those generated with the original input style (in red). The length of the style input is expressed as the average pixel width of the style samples, with standard deviations represented by semi-transparent circles behind the data points. Additionally, the average number of words in each style sample is indicated, where an “average word” corresponds to approximately 125 pixels in width.



Figure B.4: Emuru FID and CER when generating increasingly longer lines in unseen styles from the Karaoke dataset, separated into calligraphy and typewritten styles. We express the generated lines length in terms of the number of characters contained.

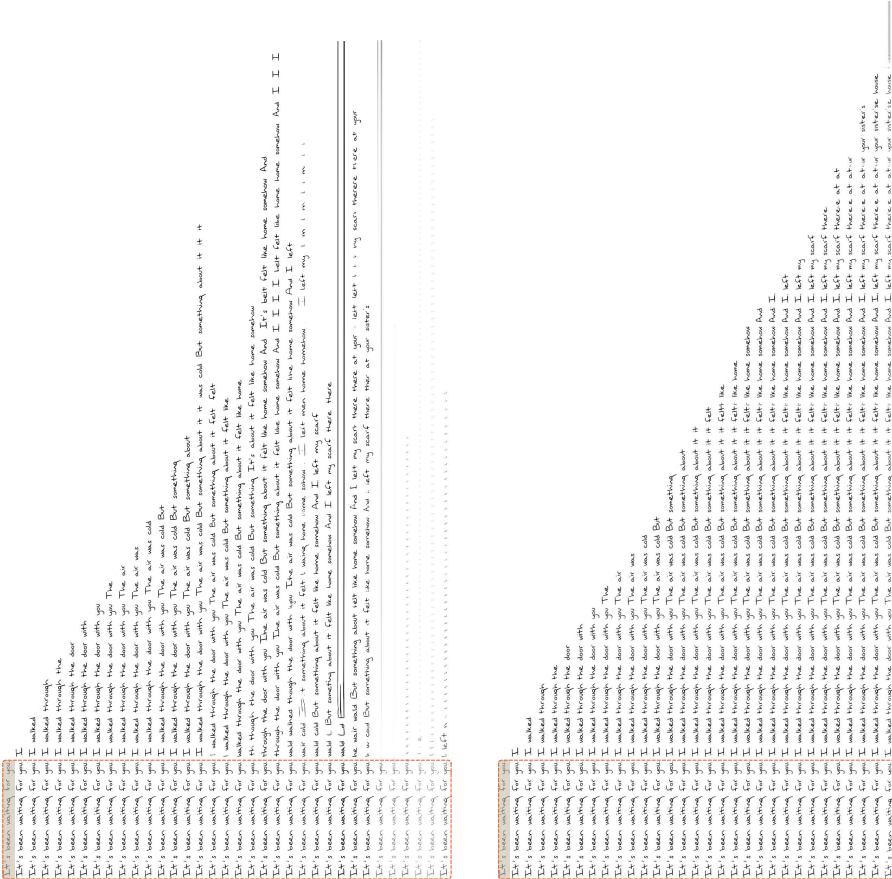


Figure B.5: Increasingly long text image generation with Emuru. We use the top, colored image as style input and task the model to generate an incrementally longer sentence in each line. In the top figure, we task the model to generate the whole sentence in one go, while, in the bottom one, each line is used as the style sample of the next one and the model generates only the last word.

B.6 DETAILED COMPARISON BETWEEN VAEs

In this section we expand the analysis on the performance of Emuru VAE on multiple datasets reported in Section 8.1.2. As the performance of the autoregressive Transformer in Emuru is tied to the quality of the latent vectors it has been trained on, this detailed analysis helps determine a soft bound on the reconstruction quality. In Tab. B.1, we report the results of the reconstruction performance comparison on different datasets, which we compute on images reconstructed with different VAEs: Emuru VAE, SD1.5 VAE (used by DiffPen and OneDM), and SD3 VAE (the current State-of-the-Art in image reconstruction). As we can see, Emuru VAE consistently gives the best or second-best BFID and BKID due to its ability to closely capture and reproduce the handwritten style. Looking at FID and KID, in some datasets the performance is a bit lower with respect to the alternatives considered, likely due to the fact that Emuru VAE is trained to reconstruct text on a clean background (and not the original input image). Moreover, the Δ I-LPIPS is consistently better, which indicates that the images reproduced by our VAE have a consistent style in all their parts. In general, we can observe either better or comparable performance to the other considered VAEs. The tailored training of our VAE allows us to obtain these results with a much more lightweight model and a more compressed latent space ($\sim 16\%$ of the parameters of the other VAEs and 1 latent channel instead of 4 and 16, respectively for SD1.5 VAE and SD3 VAE).

B.7 MULTI-STAGE TRAINING ABLATION

In Tab. B.2, we report a detailed analysis of the performance obtained with Emuru across different training stages. As explained in Section 8.1.1, we train Emuru with the MSE loss in two stages: pretraining with text lines containing 4 to 7 words (**Pretraining** in Tab. B.2) and fine-tuning on text lines containing 1 to 32 words (+ **Var. len. ft** in Tab. B.2). Predictably, there is a significant performance improvement across all metrics and datasets, which contain variable-length text lines. These results indicate the benefit of the two-stage training approach. Dur-

ing the first stage, the autoregressive Transformer learns to correlate the text input to the VAE embeddings extracted from the reference style image in a simpler setup. Then, in the second stage, the model learns to deal with variable-length inputs and can also focus on understanding when to stop generating. In fact, although the pertaining stage is longer, the samples used in this stage have similar text length and reference image width, and thus, batches of such samples are more informative since they contain less padding. On the other hand, the images used in the variable length fine-tuning stage need to be padded more to be batched, which makes this stage less efficient. Nonetheless, thanks to pertaining, few iterations are sufficient to achieve good performance.

B.8 ADDITIONAL QUANTITATIVE RESULTS

In Tabs. B.3 to B.5, we report the performance of Emuru compared to State-of-the-Art HTG solutions on the considered datasets, expressed in terms of all the scores in the extended set of considered scores, as described in Section B.3.

B.9 ADDITIONAL QUALITATIVE RESULTS

In Figs. B.6 to B.11, we report additional qualitative results comparing Emuru, the GAN-based VATr++, and the Diffusion Model-based DiffPen when generating images from the considered datasets. For each sample, we report the input style image used for guiding the generation and another reference image in the same style. We let the models generate the same text as in the reference to better observe the style imitation capabilities of the models.

IAM words							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	21.15	2.42	20.72	1.65	0.01	0.92	2.34
SD3 VAE	15.28	2.09	13.62	1.35	0.01	0.84	2.72
Emuru VAE	27.74	2.60	23.41	1.32	0.08	0.64	0.02
IAM lines							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	15.87	1.64	16.47	1.27	0.01	0.68	74.04
SD3 VAE	9.80	1.77	9.66	1.53	0.00	0.59	82.95
Emuru VAE	16.74	1.11	13.76	0.53	0.01	0.76	6.24
CVL							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	83.33	34.82	101.89	34.74	0.01	0.63	45.74
SD3 VAE	70.84	8.77	84.68	9.30	0.00	0.57	52.81
Emuru VAE	11.25	1.10	9.77	0.53	0.05	0.79	5.28
RIMES							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	20.35	2.77	19.57	2.14	0.02	0.79	0.50
SD3 VAE	19.07	5.10	17.65	4.95	0.03	0.90	3.52
Emuru VAE	47.79	2.62	44.27	1.45	0.01	1.24	39.72
Karaoke Typewritten							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	12.28	0.87	9.54	0.31	0.01	0.55	20.60
SD3 VAE	5.29	1.13	2.89	0.50	0.01	0.65	32.76
Emuru VAE	6.55	0.75	3.92	0.09	0.01	0.56	7.43
Karaoke Calligraphy							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
SD1.5 VAE	23.38	1.63	24.65	1.16	0.00	1.04	9.37
SD3 VAE	11.09	2.80	9.58	2.37	0.00	0.87	13.84
Emuru VAE	5.28	1.55	2.98	0.51	0.00	1.13	10.63

Table B.1: Comparison between the VAE featured in Emuru and other VAEs used in popular Diffusion Models when reproducing the images in the test set of different HTG datasets. Note that none of these VAEs has been trained on the considered datasets. KID, BKID, and Δ I-LPIPS are multiplied by 10^3 and the best performance is in bold for readability.

IAM words							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	85.15	53.28	83.53	52.54	0.56	3.51	0.39
+ Var. len. ft	63.61	37.73	62.34	37.22	0.19	3.03	0.16
IAM lines							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	44.90	42.36	28.19	29.75	0.83	2.70	11.39
+ Var. len. ft	13.89	6.19	11.30	5.36	0.14	1.87	38.27
CVL							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	77.23	70.52	69.28	54.47	0.87	2.52	6.69
+ Var. len. ft	14.39	10.77	12.34	10.54	0.13	1.82	0.75
RIMES							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	100.34	75.90	76.32	56.07	0.69	2.96	87.25
+ Var. len. ft	26.93	13.36	21.19	9.40	0.25	2.18	47.96
Karaoke Typewritten							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	47.06	39.28	22.76	14.69	0.67	2.03	45.52
+ Var. len. ft	9.85	4.33	5.60	1.24	0.11	1.28	5.07
Karaoke Calligraphy							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
Pretraining	37.26	29.65	18.54	16.81	0.54	2.99	4.90
+ Var. len. ft	13.87	7.99	9.24	5.37	0.13	2.24	0.73

Table B.2: Ablation analysis on the multi-stage training strategy for the autoregressive Transformer (‘ft’ stands for ‘fine-tuning’). The scores are computed on all the considered datasets. KID, BKID, and ΔI-LPIPS are multiplied by 10^3 and the best performance is in bold for readability.

	IAM words						
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
TS-GAN	129.57	86.45	141.08	92.48	0.28	4.22	1.15
HiGAN+	50.19	21.92	43.39	14.21	0.20	3.12	0.17
HWT	27.83	15.09	19.64	11.95	0.15	2.01	0.06
VATr	30.26	15.81	22.31	13.37	0.00	2.19	0.56
VATr++	31.91	17.15	23.05	15.20	0.07	2.54	0.55
One-DM	27.54	10.73	21.39	6.72	0.10	2.28	0.00
DiffPen	15.54	6.06	11.55	3.93	0.06	1.78	0.53
Emuru	63.61	37.73	62.34	37.22	0.19	3.03	0.16

	IAM lines						
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
TS-GAN	44.17	19.45	45.42	18.17	0.02	3.21	25.66
HiGAN+	74.41	34.18	77.27	31.24	0.00	3.25	63.43
HWT	44.72	30.26	43.49	31.14	0.33	2.97	34.31
VATr	35.32	27.97	33.61	27.81	0.02	2.37	28.17
VATr++	34.00	21.67	29.68	19.04	0.03	2.38	35.86
One-DM	43.89	21.54	44.48	20.94	0.13	2.83	78.42
DiffPen	12.89	6.87	9.73	4.98	0.03	2.13	3.27
Emuru	13.89	6.19	1.30	5.36	0.14	1.87	38.27

Table B.3: Comparison on the word-level and line-level IAM datasets between Emuru and State-of-the-Art approaches trained on IAM. KID, BKID, and Δ I-LPIPS are multiplied by 10^3 and the best performance is in bold for readability.

CVL							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
TS-GAN	42.12	31.97	43.15	32.44	0.13	3.07	2.02
HiGAN+	78.44	39.47	80.39	36.50	0.12	3.07	53.91
HWT	31.22	16.73	26.14	14.44	0.38	2.59	10.23
VATr	34.40	24.64	32.21	25.01	0.06	2.36	8.77
VATr++	35.53	19.87	34.15	16.08	0.12	2.18	13.30
One-DM	60.45	26.58	64.13	26.76	0.06	2.66	88.94
DiffPen	40.40	17.50	38.21	18.30	0.01	2.99	51.58
Emuru	14.39	10.77	12.34	10.54	0.13	1.82	0.75

RIMES							
	FID↓	BFID↓	KID↓	BKID↓	Δ CER↓	HWD↓	Δ I-LPIPS↓
TS-GAN	109.04	36.39	132.90	41.64	0.12	3.26	93.87
HiGAN+	160.57	47.38	183.82	46.23	0.14	3.39	48.48
HWT	118.21	35.26	128.66	35.60	0.45	3.36	1.89
VATr	113.76	30.21	114.21	27.88	0.07	3.09	7.88
VATr++	110.04	35.61	104.05	31.90	0.10	2.83	26.84
One-DM	121.18	36.07	121.67	34.68	0.20	3.36	86.49
DiffPen	89.79	18.25	94.78	18.49	0.04	2.58	78.41
Emuru	26.93	13.36	21.19	9.40	0.25	2.18	47.96

Table B.4: Comparison on the line-level CVL and RIMES datasets between Emuru and State-of-the-Art approaches trained on IAM. KID, BKID, and Δ I-LPIPS are multiplied by 10^3 and the best performance is in bold for readability.

Karaoke Handwritten							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
TS-GAN	60.30	12.68	64.80	6.75	0.05	4.59	75.30
HiGAN+	125.75	69.41	136.75	72.14	0.04	4.90	49.49
HWT	62.69	43.03	59.35	43.98	0.34	4.50	33.89
VATr	72.22	47.66	67.70	46.23	0.04	3.89	60.46
VATr++	67.16	46.53	58.57	42.59	0.01	3.96	84.57
One-DM	59.73	38.30	56.55	37.93	0.24	4.31	47.65
DiffPen	34.19	25.78	28.91	24.03	0.16	4.18	33.33
Emuru	13.87	7.99	9.24	5.37	0.13	2.24	0.73

Karaoke Typewritten							
	FID↓	BFID↓	KID↓	BKID↓	ΔCER↓	HWD↓	ΔI-LPIPS↓
TS-GAN	141.41	75.78	157.33	80.14	0.02	4.70	235.45
HiGAN+	135.34	63.39	146.34	65.81	0.03	5.19	85.45
HWT	72.78	37.40	62.77	31.51	0.39	4.57	138.76
VATr	80.38	41.02	70.46	37.26	0.04	4.14	115.80
VATr++	76.03	41.69	63.17	36.50	0.01	4.15	89.74
One-DM	70.75	44.06	60.90	42.78	0.25	4.80	119.63
DiffPen	78.07	61.16	67.17	61.05	0.14	4.71	187.68
Emuru	9.85	4.33	5.60	1.24	0.11	1.28	5.07

Table B.5: Comparison on the line-level Karaoke dataset, separating calligraphy and typewritten styles, between Emuru and State-of-the-Art approaches trained on IAM. KID, BKID, and ΔI-LPIPS are multiplied by 10^3 and the best performance is in bold for readability.

B.9. ADDITIONAL QUALITATIVE RESULTS

IAM Words Style <i>revolutionarily with an effort.</i> VATr++ <i>two</i> DiffPen <i>two</i> Emuru <i>two</i> Ref. <i>two</i>	IAM Words Style <i>fell for her but couldn't say, so he let</i> VATr++ <i>arithmetic</i> DiffPen <i>arithmetie</i> Emuru <i>arithmetie</i> Ref. <i>arithmetie</i>
IAM Words Style <i>contraction of the ventricles. Going on in Chapter Three</i> VATr++ <i>apex</i> DiffPen <i>apex</i> Emuru <i>apex</i> Ref. <i>apex</i>	IAM Words Style <i>how to live and accept all the bird</i> VATr++ <i>Mortar</i> DiffPen <i>Mortar</i> Emuru <i>Mortar</i> Ref. <i>Mortar</i>
IAM Words Style <i>The others were all dancing and</i> VATr++ <i>doesn't</i> DiffPen <i>doesn't</i> Emuru <i>doesn't</i> Ref. <i>doesn't</i>	IAM Words Style <i>in Paris' direction. What a first for the</i> VATr++ <i>error</i> DiffPen <i>error</i> Emuru <i>error</i> Ref. <i>error</i>

Figure B.6: Further qualitative results on IAM words.

IAM Lines Style <i>As it is, with so much of our life already</i> VATr++ <i>In many cases, their homes know little</i> DiffPen <i>In many cases, their homes know little of</i> Emuru <i>In many cases, their homes know little of</i> Ref. <i>In many cases, their homes know little of</i>	IAM Lines Style <i>To this true world, as you call it, we come to die.</i> VATr++ <i>had no place here.</i> DiffPen <i>had no place here.</i> Emuru <i>had no place here.</i> Ref. <i>had no place here.</i>
IAM Lines Style <i>It is possible that our saga contains the</i> VATr++ <i>The saga is built around the saying</i> DiffPen <i>The saga is built around the saying of</i> Emuru <i>The saga is built around the saying of</i> Ref. <i>The saga is built around the saying of</i>	IAM Lines Style <i>They had been married for six years, but the</i> VATr++ <i>salary raise, on the expectation of w</i> DiffPen <i>salary raise on the expectation of which they</i> Emuru <i>salary raise, on the expectation of which they</i> Ref. <i>salary raise, on the expectation of which they</i>
IAM Lines Style <i>And we answer, that the Lord our God gave us four freedoms</i> VATr++ <i>"four cups" to thank the Lord our God</i> DiffPen <i>four cups' to thank the Lord our God for the four</i> Emuru <i>four cups" to thank the Lord our God for the four</i> Ref. <i>"four cups" to thank the Lord our God for the four</i>	IAM Lines Style <i>You made some sort of a protest to Amy.</i> VATr++ <i>idyllic surroundings." "How dare you</i> DiffPen <i>idyllic surroundings. "How dare you say</i> Emuru <i>idyllic surroundings. How dare you say</i> Ref. <i>idyllic surroundings. "How dare you say</i>

Figure B.7: Further qualitative results on IAM lines.

CVL Lines Style <i>Marvelous is an Austrian rich-</i> VATr++ <i>When we look to the individuals</i> DiffPen <i>When we look to the individuals</i> Emuru <i>When we look to the individuals</i> Ref. <i>When we look to the individuals</i>	CVL Lines Style <i>When we look to the individuals of the sand variety</i> VATr++ <i>Lines Triangles Squares Pentagons Hexagons and other</i> DiffPen <i>Lines Triangles Squares Pentagons Hexagons and other</i> Emuru <i>Lines Triangles Squares Pentagons Hexagons and other</i> Ref. <i>Lines, Triangles, Squares, Pentagons, Hexagons, and other</i>
CVL Lines Style <i>When we look to the individuals of</i> VATr++ <i>of any one species or variety in a</i> DiffPen <i>of any one species or variety in a</i> Emuru <i>of any one species or variety in a</i> Ref. <i>of any one species or variety in a</i>	CVL Lines Style <i>And fortune, on his damned quavrel</i> VATr++ <i>but now my mind has been opened to</i> DiffPen <i>but now my mind has been opened to</i> Emuru <i>but now my mind has been opened to</i> Ref. <i>but now my mind has been opened to</i>
CVL Lines Style <i>Marvelous is an Austrian nickname for the first</i> VATr++ <i>du bist so schön Dann magst du mich in Fies</i> DiffPen <i>du bist so schön Dann magst du mich in Fies</i> Emuru <i>du bist schon! magst du mich in Fies</i> Ref. <i>du bist so schön! Dann magst du mich in Fies</i>	CVL Lines Style <i>Imagine a wood sheet of paper</i> VATr++ <i>said my universe but my mind has</i> DiffPen <i>said my universe but my mind has</i> Emuru <i>said my universe but my mind has</i> Ref. <i>said "my universe" had my mind has</i>

Figure B.8: Further qualitative results on CVL lines.

Rimes Lines Style <i>J'ai desee change de royaume des vêtements masculins</i> VATr++ <i>Je souhaite désormais verser 45 chaque</i> DiffPen <i>Je souhaite désormais verser 45 chaque mois</i> Emuru <i>Je souhaite désormais verser 45 chaque mois au lieu des</i> Ref. <i>Je souhaite désormais verser 45 chaque mois, au lieu des</i>	Rimes Lines Style <i>mes impôts sur le revenu.</i> VATr++ <i>Dans un souci pratique, j'aimerais du</i> DiffPen <i>Dans un souci pratique j'ai essayé d'échelonner le p</i> Emuru <i>Dans un souci pratique, j'aimerais échelonner le paie</i> Ref. <i>Dans un souci pratique, j'aimerais échelonner le paiement de</i>
Rimes Lines Style <i>Je me poutel de vous être afin d'obtenir de la</i> VATr++ <i>mes sentiments distingués.</i> DiffPen <i>mes sentiments distingués</i> Emuru <i>mes sentiments distingués.</i> Ref. <i>mes sentiments distingués.</i>	Rimes Lines Style <i>En la présente je vous informe que</i> VATr++ <i>Vous trouverez ci-joint le constat.</i> DiffPen <i>Vous trouverez ci-joint le constat.</i> Emuru <i>Vous trouvez ci-joint le constat.</i> Ref. <i>Vous trouverez ci-joint le constat.</i>
Rimes Lines Style <i>Je vous informe que j'ai malencontreusement presque</i> VATr++ <i>uniquement matériel.</i> DiffPen <i>uniquement matériel</i> Emuru <i>uniquement matériel.</i> Ref. <i>uniquement matériel.</i>	Rimes Lines Style <i>Je vous écris afin de vous pour part de mon</i> VATr++ <i>de mes sincères salutations</i> DiffPen <i>de mes sincères salutations</i> Emuru <i>de mes sincères salutations</i> Ref. <i>de mes sincères salutations</i>

Figure B.9: Further qualitative results on RIMES lines.



Figure B.10: Further qualitative results on Karaoke Calligraphy.

B.10 EDITING APPLICATION

Due to our Emuru VAE, there can be a difference in ink color between the reference image, I_{style} , and the Emuru-generated one, I_{out} . However, this difference can be easily recovered. Note that, once passed through the VAE, both I_{style} and I_{out} have a perfectly white background. Therefore, the background can be easily removed from both images via standard chroma keying on white. Then, the average color of the ink pixels can be computed from the reference image and directly applied to the ink pixels in the generated image (see Fig. B.12).

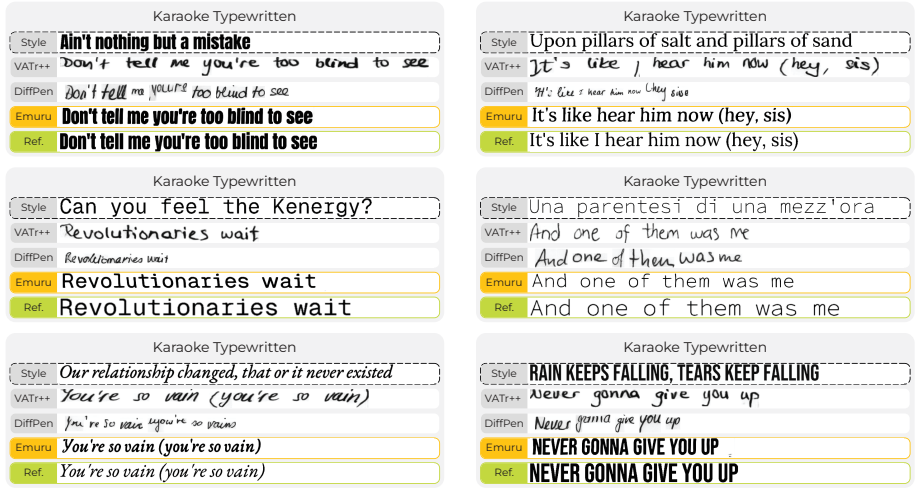


Figure B.11: Further qualitative results on Karaoke Typewritten.

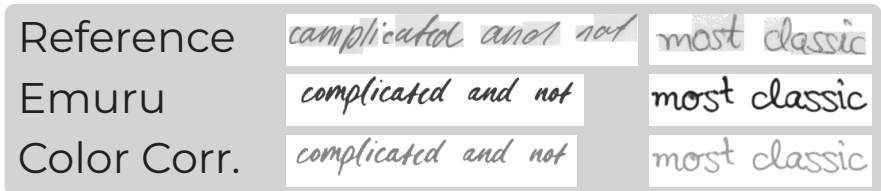


Figure B.12: Thanks to the characteristics of the VAE, ink color can be easily corrected on Emuru-generated images to look more similar to the color in the reference image.



Improving Autoregressive HTG with Eruku

In this document, we report additional analyses on style text independence and the effect of the training strategies adopted in the second phase of training. Moreover, we report results that include color correction of the output.

C.1 STYLE TEXT RELIANCE ANALYSIS

When compared to Emuru [6], Eruku does not need style text input. A possible workaround to use Eruku with a style sample with no known ground-truth textual transcription is to use an OCR model to obtain it. We test both Emuru and Eruku with style text input obtained from running TrOCR-Base [205] as an OCR model and comparing them against each other when using the T_s^* text generated by TrOCR, against Eruku ran with no style text input and against a version of Eruku which has never been trained with style text dropout. The results are displayed in Tab. C.1.

Eruku is (except for FID) better than Emuru even when using the ground truth text T_s . When using T_s^* , Eruku is able to maintain very low $\Delta\text{Character Er-}$

ror Rate (CER), whereas Emuru tends to collapse and/or generate incorrect text more often. Both manage to maintain style consistency. Eruku with no style text gets even better Δ CER scores, but compromises in a significant way on style adherence, as indicated by the high HWD score. The version of Eruku trained with no style text dropout and style text from OCR suffers, just like Emuru, from significantly increased Δ CER from the reliance on this noisy style text. Emuru is incapable of running with no style text input.

C.2 ABLATION ON SECOND STAGE TRAINING

In the second stage of pretraining, as described in Section 8.2.2, two variations are made to the way the model trains: it is trained on the dataset of images with longer context described in Section 8.2.2, and it is trained to randomly drop style text conditioning with a probability of $p_{drop} = 0.1$. We investigate the effects of each of those by running training for the same amount of iterations as the full Eruku second stage of training, but with just one strategy or the other. We then compare those runs on IAM lines to the full second stage of training and to the result of just the first stage of training.

The results, shown in Tab. C.2, highlight how long-context training improves Δ CER significantly. Style text dropout instead, in addition to allowing the model to generate unconditionally as shown in Section 8.2.3, also improves style image adherence, as indicated by the improvement in HWD. The model using both strategies (Eruku) combines the advantages of both and reaches the best HWD values and much-improved Δ CER values when compared to the model resulting from the first stage of training.

C.3 RESULTS INCLUDING COLOR CORRECTION

Since it relies on the same Variational Autoencoder (VAE) as Emuru, Eruku generates images with a white background and usually very dark text strokes. This allows the simple color correction strategy proposed in [6] for Emuru to be applicable also for Eruku. The strategy uses the VAE’s background removal abilities

	HWD↓	Δ CER↓	FID↓	BFID↓
Eruku w/ T_s	1.70	0.06	16.40	4.88
Emuru w/ T_s	1.87	0.14	13.89	6.19
Eruku w/ T_s^*	1.73	0.06	16.59	5.07
Eruku $p_{drop} = 0$ w/ T_s^*	1.72	0.53	15.81	7.68
Emuru w/ T_s^*	1.79	0.42	14.09	6.23
Eruku w/o T_s	2.51	0.04	20.44	9.63
Emuru w/o T_s	-	-	-	-

Table C.1: Emuru and Eruku results on IAM lines when fed with the actual T_s or a T_s obtained by running TrOCR on the style image I_s (dubbed T_s^*). As a reference, we report the results of the generation without T_s .

longer input	$p_{drop} = 0.1$	HWD↓	Δ CER↓	FID↓	BFID↓
$\times$	$\times$	1.81	0.40	14.20	3.38
$\checkmark$	$\times$	1.92	0.04	19.45	5.50
$\times$	$\checkmark$	1.75	0.40	13.49	4.45
$\checkmark$	$\checkmark$	1.70	0.06	16.40	4.88

Table C.2: Ablation analysis on the effect of the second training phase inputs and strategy in terms of performance on IAM Lines.

to isolate the mask containing the text within the style image, then computes the average of the color values among the foreground ink pixels and applies that to those of the generated image. The effect of such color correction post-processing can be observed quantitatively in Tab. C.3 (especially in terms of FID).

	HWD↓	Δ CER↓	FID↓	BFID↓
Eruku	1.70	0.06	16.40	4.88
Emuru	1.87	0.14	13.89	6.19
Eruku w/ c.c.	1.68	0.04	12.21	4.54
Emuru w/ c.c.	1.85	0.14	11.40	6.20

Table C.3: Emuru and Eruku results on IAM lines in the standard setting and when the color correction strategy (c.c.) is applied as post-processing.

D

Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas

In this document, we report additional results of our proposed MAD approach for inference-time adaptation of diffusion models for generating long images. We explore the effect of changing the application stage and varying the number of application steps, present more detailed information regarding the user study, and offer further comparisons with SD-L. Additionally, we report other qualitative results, also at different aspect ratios, and further discuss the limitations.

D.1 EFFECT OF THE APPLICATION STAGE

In Tabs. D.1 and D.2, we report a quantitative analysis of the effects of applying MAD at different stages in the noise prediction model of the considered LDM and LCM. We observe that when MAD is applied only in the bottleneck (Mid blocks), FID, KID, and mGIQA values are comparable to those obtained with the baseline in which MAD is never applied and the perceptual coherence is lower than in the other settings. On the other hand, I-LPIPS, I-StyleL, and mCLIP get better when MAD is applied more (*i.e.* in all blocks) and in later stages of the

U-Net (*i.e.* in the upsampling blocks). The same trend can be observed from the LDM qualitatives in Fig. D.2.

Table D.1: Quantitative results on the generation of images with MAD applied at different stages of the LDM noise prediction model to generate 512×3072 images.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$ ($\times 10^{-3}$)	FID $\downarrow$	KID $\downarrow$ ($\times 10^{-3}$)	mGIQA $\uparrow$ ($\times 10^{-3}$)
No MAD (Baseline)	31.65 $\pm$ 2.17	0.64 $\pm$ 0.10	2.65 $\pm$ 2.33	34.51 $\pm$ 13.92	9.19 $\pm$ 4.51	27.59 $\pm$ 6.83
MAD in Mid Block	31.66 $\pm$ 2.16	0.64 $\pm$ 0.10	2.19 $\pm$ 1.17	34.80 $\pm$ 13.80	9.35 $\pm$ 3.91	28.04 $\pm$ 7.23
MAD in Down Blocks	31.83 $\pm$ 2.21	0.62 $\pm$ 0.10	1.98 $\pm$ 1.03	40.46 $\pm$ 16.58	16.40 $\pm$ 8.83	28.11 $\pm$ 7.42
MAD in Up Blocks	32.09 $\pm$ 2.20	0.53 $\pm$ 0.10	1.19 $\pm$ 0.86	60.84 $\pm$ 24.65	43.56 $\pm$ 19.82	28.60 $\pm$ 7.51
MAD in All Blocks	32.15 $\pm$ 2.25	0.53 $\pm$ 0.10	1.43 $\pm$ 1.04	61.76 $\pm$ 23.73	43.31 $\pm$ 17.41	28.10 $\pm$ 7.84

Table D.2: Quantitative results on the generation of images with MAD applied at different stages of the LCM noise prediction model to generate 512×3072 images.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$ ($\times 10^{-3}$)	FID $\downarrow$	KID $\downarrow$ ($\times 10^{-3}$)	mGIQA $\uparrow$ ($\times 10^{-3}$)
No MAD (Baseline)	31.36 $\pm$ 1.83	0.55 $\pm$ 0.05	0.90 $\pm$ 0.35	31.35 $\pm$ 15.42	13.27 $\pm$ 10.05	35.44 $\pm$ 12.38
MAD in Mid Block	31.44 $\pm$ 1.79	0.55 $\pm$ 0.05	0.87 $\pm$ 0.33	29.35 $\pm$ 15.62	13.85 $\pm$ 11.27	35.49 $\pm$ 12.22
MAD in Down Blocks	31.44 $\pm$ 1.78	0.55 $\pm$ 0.05	0.84 $\pm$ 0.32	29.01 $\pm$ 14.83	13.35 $\pm$ 11.02	35.38 $\pm$ 12.46
MAD in Up Blocks	31.46 $\pm$ 1.84	0.51 $\pm$ 0.05	0.71 $\pm$ 0.26	40.22 $\pm$ 20.99	28.98 $\pm$ 17.43	35.45 $\pm$ 13.08
MAD in All Blocks	31.48 $\pm$ 1.87	0.52 $\pm$ 0.05	0.71 $\pm$ 0.26	38.77 $\pm$ 19.85	27.41 $\pm$ 16.19	35.23 $\pm$ 13.15

D.2 EFFECT OF THE NUMBER OF APPLICATION STEPS

We report a quantitative analysis (Tab. D.3) and a more exhaustive qualitative analysis (Fig. D.1) on the effect of the number of inference timesteps in which the proposed MAD operator is applied (defined by the hyperparameter τ). In this experiment, we use the considered LDM run for 50 reverse process steps and apply MAD at each timestep from the first one up to a certain threshold. As we can see, 15 steps lead to a good trade-off between variety and uniformity.

D.3 DETAILS OF THE USER STUDY

Here, we report further details on the user study conducted to assess human preference between images generated with our approach or with two other State-of-the-Art ones (namely, MultiDiffusion [30] and SyncDiffusion [198]). For each approach, we generated 60 images for each of the six prompts used for the quan-

Table D.3: Results on the generation of 512×3072 images with MAD applied for different numbers of timesteps from the beginning of the reverse diffusion process in the LDM.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$ ($\times 10^{-3}$)	FID $\downarrow$	KID $\downarrow$ ($\times 10^{-3}$)	mGIQA $\uparrow$ ($\times 10^{-3}$)
$\tau=0$	31.65 $\pm$ 2.17	0.64 $\pm$ 0.10	2.65 $\pm$ 2.33	34.51 $\pm$ 13.92	9.19 $\pm$ 4.51	27.59 $\pm$ 6.83
$\tau=5$	31.86 $\pm$ 2.22	0.59 $\pm$ 0.10	2.07 $\pm$ 1.31	38.10 $\pm$ 13.71	13.75 $\pm$ 4.00	28.32 $\pm$ 7.64
$\tau=10$	31.95 $\pm$ 2.24	0.57 $\pm$ 0.10	2.00 $\pm$ 1.34	42.60 $\pm$ 14.94	19.38 $\pm$ 5.56	28.36 $\pm$ 7.83
$\tau=15$	32.03 $\pm$ 2.29	0.56 $\pm$ 0.10	1.90 $\pm$ 1.32	48.52 $\pm$ 17.14	27.15 $\pm$ 9.10	28.32 $\pm$ 7.76
$\tau=20$	32.09 $\pm$ 2.29	0.54 $\pm$ 0.10	1.68 $\pm$ 1.19	54.31 $\pm$ 19.92	34.43 $\pm$ 12.54	28.23 $\pm$ 7.90
$\tau=25$	32.15 $\pm$ 2.25	0.53 $\pm$ 0.10	1.43 $\pm$ 1.04	61.76 $\pm$ 23.73	43.31 $\pm$ 17.41	28.10 $\pm$ 7.84
$\tau=30$	32.16 $\pm$ 2.17	0.52 $\pm$ 0.10	1.22 $\pm$ 0.90	70.29 $\pm$ 28.24	54.40 $\pm$ 23.19	27.94 $\pm$ 7.73
$\tau=35$	32.18 $\pm$ 2.08	0.51 $\pm$ 0.10	1.02 $\pm$ 0.73	78.50 $\pm$ 32.94	65.06 $\pm$ 29.94	27.82 $\pm$ 7.65
$\tau=40$	32.16 $\pm$ 1.95	0.50 $\pm$ 0.10	0.88 $\pm$ 0.61	86.20 $\pm$ 37.23	76.15 $\pm$ 37.73	27.59 $\pm$ 7.61
$\tau=45$	32.14 $\pm$ 1.83	0.50 $\pm$ 0.10	0.78 $\pm$ 0.53	92.69 $\pm$ 40.76	84.13 $\pm$ 44.13	27.35 $\pm$ 7.50
$\tau=50$	32.14 $\pm$ 1.72	0.49 $\pm$ 0.10	0.71 $\pm$ 0.47	98.01 $\pm$ 43.64	91.51 $\pm$ 49.95	27.05 $\pm$ 7.28

titative evaluation (for a total of 360 images per method). We then formed pairs of images with matching prompts and presented 60 random pairs to the users (plus 6 additional pairs as a vigilance task), asking them to select the image that they found to be “*the most coherent, and that conforms the most to the prompt*”. As mentioned in the main chapter, 139 out of 201 participants passed the vigilance task and provided a total of 7807 preference answers. In Fig. D.3, we plot the overall and the per-prompt results. As we can see, our method is consistently preferred over the competitors.

D.4 FURTHER COMPARISON WITH DIRECT INFERENCE

Here, we report further comparison on the variability between the generated images using MAD and SD-L. In particular, for each one of the six evaluation prompts, we exploit the same network used for computing the FID, KID, and mGIQA (an Inception-v3 trained on Imagenet) to extract features from square images generated with SD and embed them using t-SNE [376]. Then, we generate 512×5120 images with SD-L and MAD and embed all the squared crops using openTSNE [216, 282]. In Fig. D.4, we report the embeddings, showing that images generated by MAD have more variability with respect to the images generated by the vanilla SD. Thus, MAD has increased inter-image variability compared to SD-L. For each prompt, we plot the SD embeddings using grey circles, the MAD embeddings using red dots, and the SD-L embeddings using blue

dots. Darker areas indicate an overlap between the embeddings of MAD and SD-L images.

D.5 COMBINATION WITH ATTENTION SCALING

To adapt Stable Diffusion to the generation of larger images, a recent work has proposed to re-scale the attention operations in the U-Net [170] (which we refer to as Attn-S). We evaluate this approach both in combination with a method that generates the whole image directly and MAD (we refer to this combination as MAD-L+Attn-S). To this end, we consider the generation of 512×3072 images. The result of generation with LDMs and Latent Consistency Models (LCMs) are reported in Tab. D.4 and Tab. D.5, respectively. It can be noticed that the combination with the Attn-S has limited to no benefit in terms of generation performance. This can be explained by looking at the small differences in terms of qualitative results, reported in Figs. D.5 and D.6.

Table D.4: Quantitative comparison on 512×3072 panorama generation using the LDM. I-StyleL, KID, and mGIQA values are scaled by 10^3 . For MAD, $\tau=15$.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$
Standard Prompts						
SD-L	32.01 $\pm$ 1.67	0.50 $\pm$ 0.11	0.58 $\pm$ 0.40	87.64 $\pm$ 30.25	76.83 $\pm$ 30.52	27.72 $\pm$ 7.83
SD-L+Attn-S	32.02 $\pm$ 1.66	0.52 $\pm$ 0.11	0.73 $\pm$ 0.47	80.16 $\pm$ 27.11	67.54 $\pm$ 26.00	27.89 $\pm$ 7.67
MAD	32.03 $\pm$ 2.29	0.56 $\pm$ 0.10	1.90 $\pm$ 1.32	48.52 $\pm$ 17.14	27.15 $\pm$ 9.10	28.32 $\pm$ 7.76
MAD+Attn-S	32.17 $\pm$ 2.20	0.53 $\pm$ 0.11	1.19 $\pm$ 0.75	64.08 $\pm$ 28.19	47.33 $\pm$ 23.77	28.30 $\pm$ 0.75
GPT1k Prompts						
SD-L	31.89	0.52	0.73	68.03	6.61	12.58
SD-L+Attn-S	32.03	0.53	0.87	65.08	5.47	12.68
MAD	32.47	0.58	3.89	54.44	1.28	13.03
MAD+Attn-S	32.40	0.57	2.94	55.58	1.82	13.09

Table D.5: Quantitative comparison on 512×3072 panorama generation with the LCM for different numbers of inference steps. I-StyleL, KID, and mGIQA are scaled by 10^3 . For MAD, $\tau=1/1/2$ for 1/2/4 inference steps, respectively.

	mCLIP $\uparrow$	I-LPIPS $\downarrow$	I-StyleL $\downarrow$	FID $\downarrow$	KID $\downarrow$	mGIQA $\uparrow$
1 Inference Step						
LCD-L	29.50 $\pm$ 1.48	0.40 $\pm$ 0.08	0.21 $\pm$ 0.18	67.54 $\pm$ 15.53	66.86 $\pm$ 21.22	32.21 $\pm$ 4.50
LCD-L+Attn-S	29.77 $\pm$ 1.43	0.41 $\pm$ 0.08	0.33 $\pm$ 0.33	77.50 $\pm$ 14.72	79.76 $\pm$ 20.96	31.13 $\pm$ 4.15
MAD	29.01 $\pm$ 1.66	0.40 $\pm$ 0.07	0.37 $\pm$ 0.28	72.43 $\pm$ 23.75	75.47 $\pm$ 33.24	32.23 $\pm$ 4.95
MAD+Attn-S	29.36 $\pm$ 1.46	0.41 $\pm$ 0.07	0.51 $\pm$ 0.42	80.53 $\pm$ 24.19	86.79 $\pm$ 36.77	31.28 $\pm$ 4.75
2 Inference Steps						
LCD-L	30.77 $\pm$ 2.09	0.47 $\pm$ 0.06	0.56 $\pm$ 0.29	49.98 $\pm$ 25.20	44.29 $\pm$ 26.88	35.45 $\pm$ 11.94
LCD-L+Attn-S	31.11 $\pm$ 1.95	0.49 $\pm$ 0.06	0.66 $\pm$ 0.32	56.37 $\pm$ 28.61	54.55 $\pm$ 33.00	34.94 $\pm$ 12.09
MAD	30.97 $\pm$ 2.15	0.50 $\pm$ 0.06	0.85 $\pm$ 0.34	35.69 $\pm$ 17.88	23.74 $\pm$ 15.82	35.32 $\pm$ 11.49
MAD+Attn-S	31.16 $\pm$ 2.07	0.50 $\pm$ 0.06	0.81 $\pm$ 0.31	36.84 $\pm$ 18.15	25.83 $\pm$ 16.26	35.68 $\pm$ 11.60
4 Inference Steps						
LCD-L	31.30 $\pm$ 1.63	0.50 $\pm$ 0.06	0.58 $\pm$ 0.24	55.52 $\pm$ 32.82	51.71 $\pm$ 37.79	35.44 $\pm$ 12.94
LCD-L+Attn-S	31.60 $\pm$ 1.57	0.51 $\pm$ 0.06	0.72 $\pm$ 0.27	59.25 $\pm$ 32.32	57.57 $\pm$ 37.91	34.92 $\pm$ 13.27
MAD	31.48 $\pm$ 1.87	0.52 $\pm$ 0.05	0.71 $\pm$ 0.26	38.77 $\pm$ 19.85	27.41 $\pm$ 16.19	35.23 $\pm$ 13.15
MAD+Attn-S	31.64 $\pm$ 1.87	0.51 $\pm$ 0.06	0.68 $\pm$ 0.23	38.68 $\pm$ 19.10	26.49 $\pm$ 14.77	35.45 $\pm$ 13.27

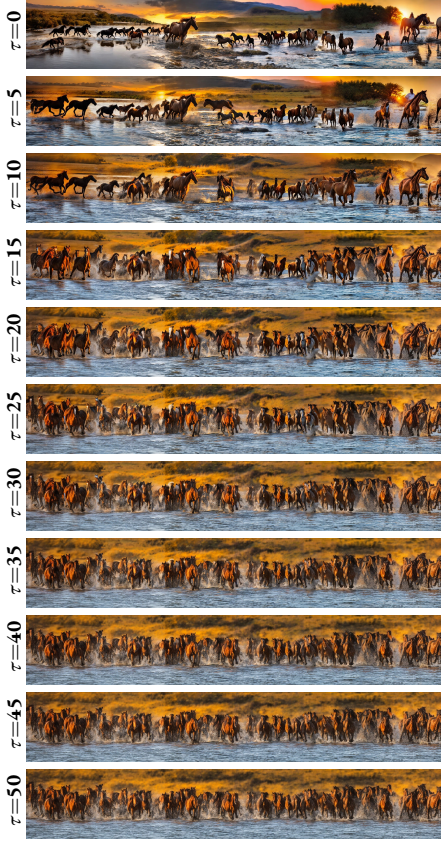


Figure D.1: Long images generated by the considered LDM with MAD applied up to different numbers of inference steps for the prompt *A herd of Mustang horses crossing a river at sunset*. When τ is too low, the view interactions are not enough to produce a globally coherent image. As τ increases, the image becomes more and more coherent, with maximal uniformity when MAD is applied at every timestep.



Figure D.2: Long images generated by the considered LDM with MAD applied in different blocks of the noise prediction model, with $\tau=15$ for the prompt *A herd of Mustang horses crossing a river at sunset*.

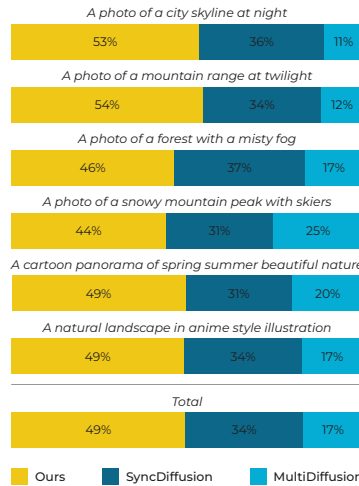


Figure D.3: User study per-prompt results of MAD, SyncDiffusion, and MultiDiffusion.

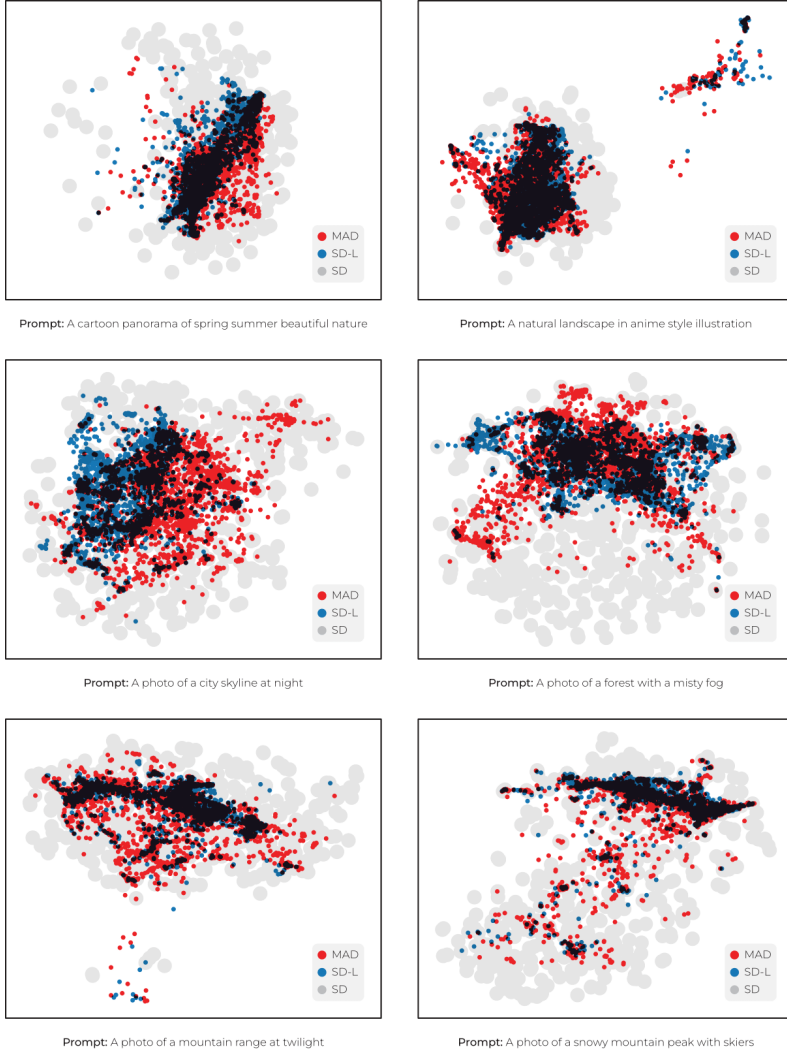


Figure D.4: Comparison on the distribution coverage of panoramas generated by MAD (red) and SD-L (blue) with respect to square images generated by SD (gray). Darker areas indicate an overlap between the embeddings of MAD and SD-L images.

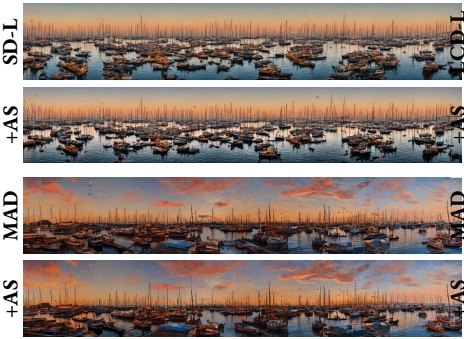


Figure D.5: Effect of attention scaling (AS) on LDMs used directly and with MAD.

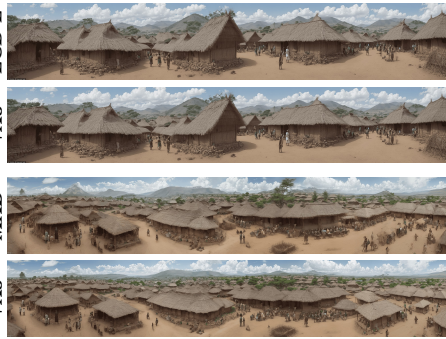


Figure D.6: Effect of attention scaling (AS) on LCMs used directly and with MAD.

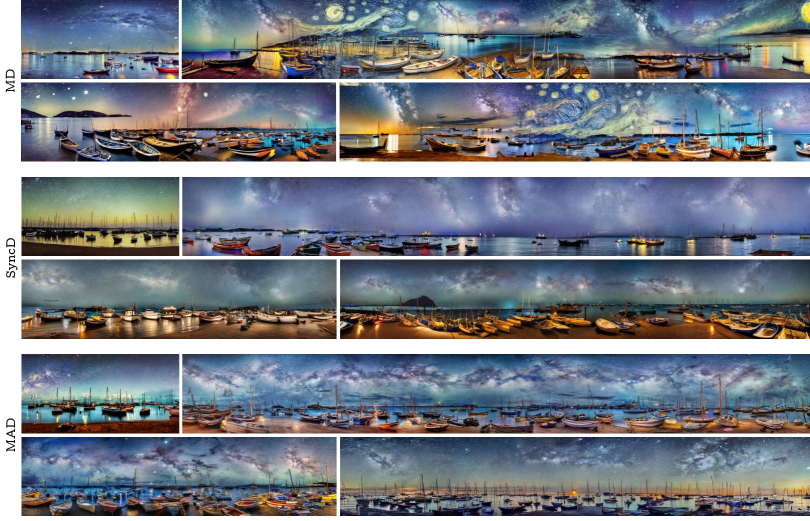


Figure D.7: Qualitative comparisons with MD and SyncD on the generation of horizontal images at different aspect ratios with the LDM model and the prompt *A photo of the seaside with some boats in a starry night*. These images are 512×1024 (top-left), 512×4096 (top-right), 512×2048 (bottom-left), and 512×3072 (bottom-right).

D.6 FURTHER RESULTS AT DIFFERENT ASPECT RATIOS

We report a qualitative comparison with the State-of-the-Art approaches (Multi-Diffusion [30] and SyncDiffusion [198]) on the generation of images with different aspect ratios. In particular, we consider the generation of horizontal images with the LDM (Fig. D.7) and with the LCM (Fig. D.8) and of vertical images with the same models (Fig. D.9 and Fig. D.10, respectively).

D.7 FURTHER QUALITATIVE RESULTS

We report some randomly picked qualitative results on the GPT1k prompts and the six prompts used for the quantitative analysis of the LDM in Figs. D.11, D.12 and D.15 to D.20 and of the LCM in Figs. D.13, D.14 and D.21 to D.26. Moreover, in Figs. D.27 to D.30, we report more qualitative results, both horizontal and vertical, on different GPT1k prompts for the LDM and LCM, respectively.

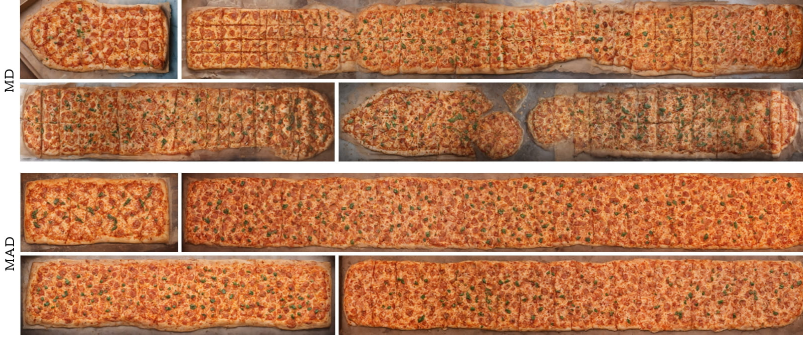


Figure D.8: Qualitative comparisons with respect to a baseline on the generation of horizontal images at different aspect ratios with the LCM model and the prompt *Top-view of a squared long pizza*. The images are 512×1024 (top-left), 512×4096 (top-right), 512×2048 (bottom-left), and 512×3072 (bottom-right).

D.8 FURTHER DISCUSSION OF THE LIMITATIONS

Approaches for panorama image generation that adapt diffusion models pretrained on squared images, including ours, can be limited by the performance and the image distribution learned by the base model. This results in generated images whose quality depends on the backbone, as showcased in Fig. D.31. Moreover, inference-time panorama generation approaches struggle with scenes or objects that do not fit well with the specified aspect ratio and prompts where the base model itself does not produce good-quality results. We provide examples in Figs. D.32 and D.33. As we can see, the prompt *A gothic cathedral nave* (Fig. D.33) does not fit well with the horizontal aspect ratio. However, when asked to generate a vertical image, the results improve. Moreover, the prompt *A fancy living room*, which entails a usually small indoor space, is rendered at an inferior quality when the desired output width is too large.

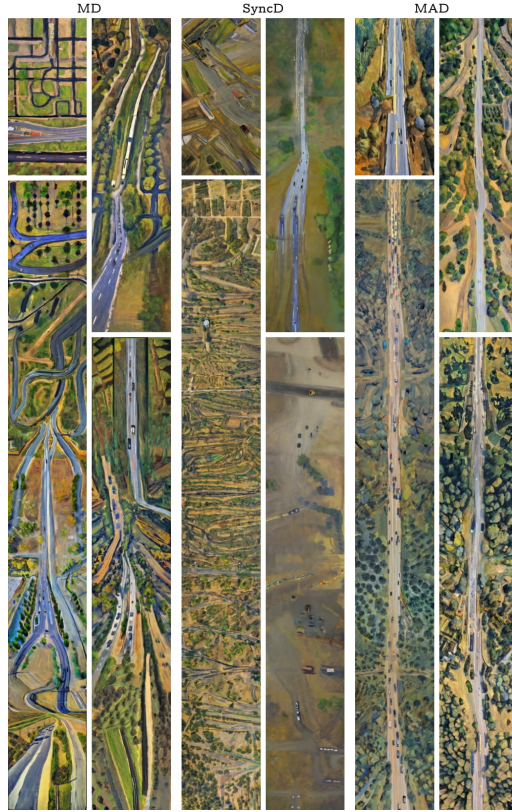


Figure D.9: Qualitative comparisons with MD and SyncD on the generation of vertical images at different aspect ratios with the LDM model and the prompt *Top view of a long road with the Manet style*. The images are 1024×512 (top-left), 2048×512 (top-right), 4096×512 (bottom-left), and 3072×512 (bottom-right).

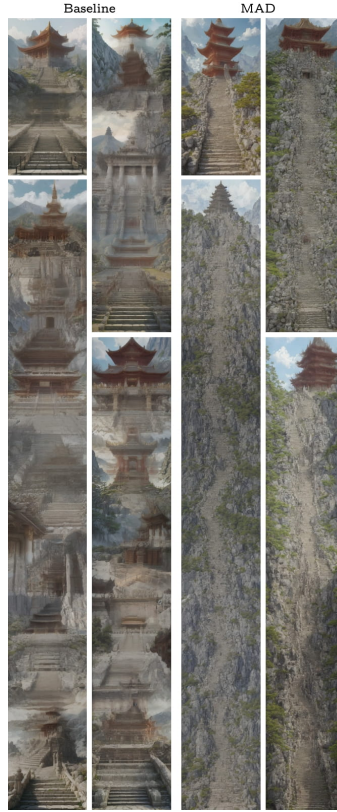


Figure D.10: Qualitative comparisons with respect to a baseline on the generation of vertical images at different aspect ratios with the LCM model and the prompt *Stairs on a mountain with a temple on top*. The images are 1024×512 (top-left), 2048×512 (top-right), 4096×512 (bottom-left), and 3072×512 (bottom-right).

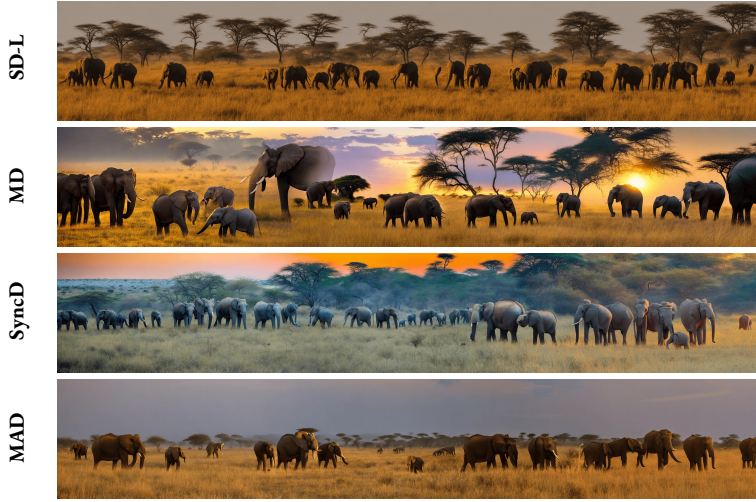


Figure D.11: Qualitative comparisons with respect to SD-L, MD, and SyncD for the prompt *A group of elephants grazing on the Savannah at sunset* from GPT1k, using the LDM model.



Figure D.12: Qualitative comparisons with respect to SD-L, MD, and SyncD for the prompt *A line of beach chairs facing the ocean* from GPT1k, using the LDM model.

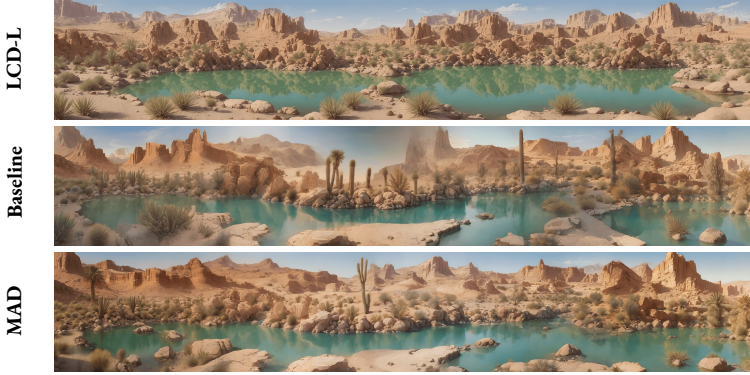


Figure D.13: Qualitative comparisons with respect to LCD-L, MD, and SyncD for the prompt *A tranquil desert oasis* from GPT1k, using the LCM model.

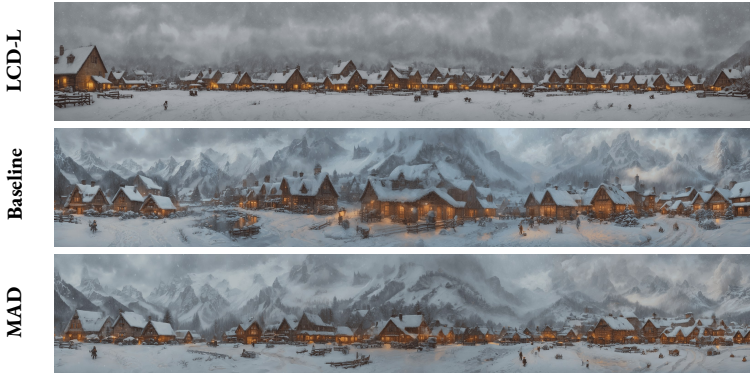


Figure D.14: Qualitative comparisons with respect to LCD-L and a baseline for the prompt *A snowstorm blanketing a small village* from GPT1k, using the LCM model.

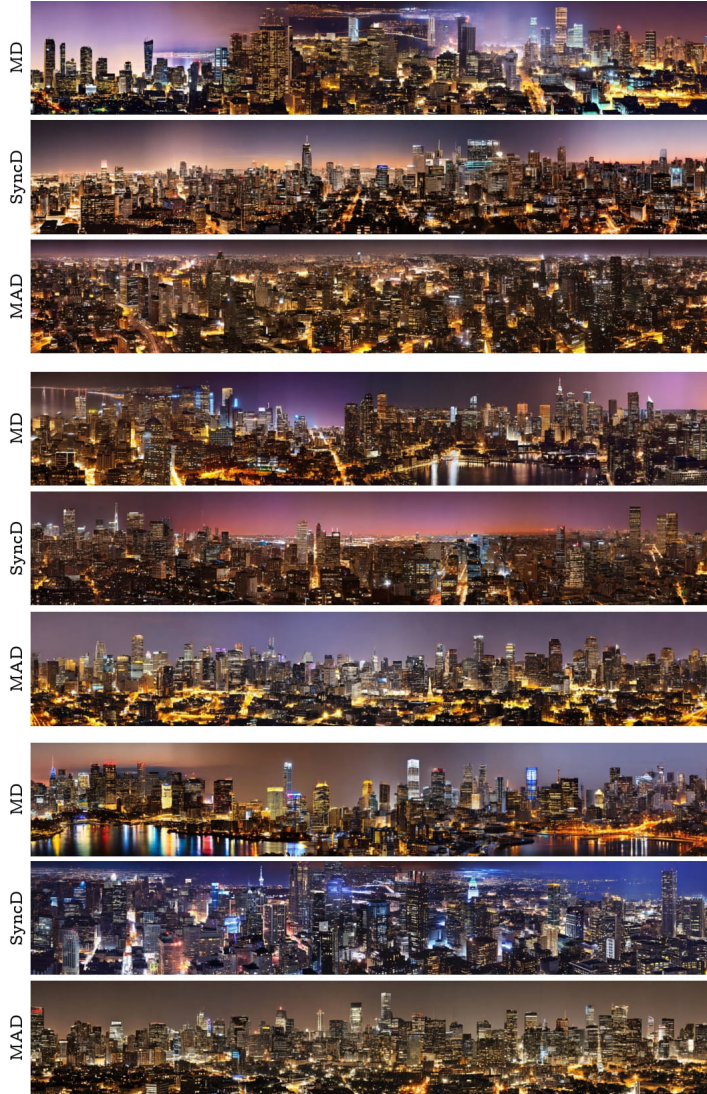


Figure D.15: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a city skyline at night*, using the LDM model.



Figure D.16: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a mountain range at twilight*, using the LDM model.



Figure D.17: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a forest with a misty fog*, using the LDM model.



Figure D.18: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a snowy mountain peak with skiers*, using the LDM model.



Figure D.19: Qualitative comparisons with respect to MD and SyncD for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LDM model.

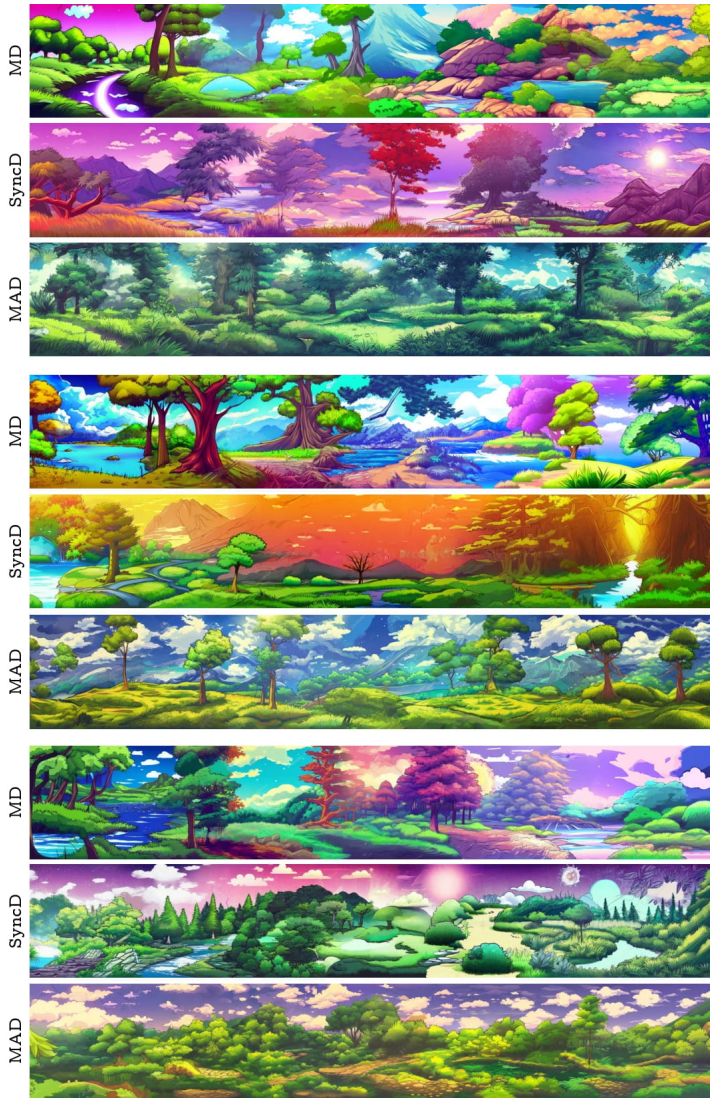


Figure D.20: Qualitative comparisons with respect to MD and SyncD for the prompt *A natural landscape in anime style illustration*, using the LDM model.



Figure D.21: Qualitative comparisons with respect to a baseline for the prompt *A photo of a city skyline at night*, using the LCM model.



Figure D.22: Qualitative comparisons with respect to a baseline for the prompt *A photo of a mountain range at twilight*, using the LCM model.



Figure D.23: Qualitative comparisons with respect to a baseline for the prompt *A photo of a forest with a misty fog*, using the LCM model.



Figure D.24: Qualitative comparisons with respect to a baseline for the prompt *A photo of a snowy mountain peak with skiers*, using the LCM model.



Figure D.25: Qualitative comparisons with respect to a baseline for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LCM model.

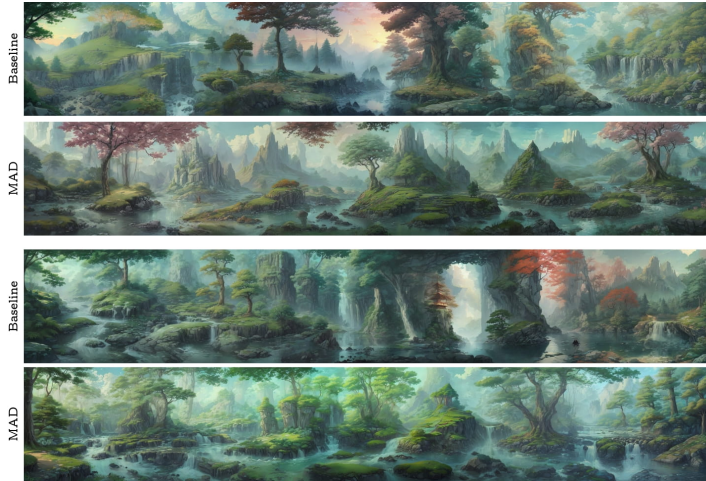


Figure D.26: Qualitative comparisons with respect to a baseline for the prompt *A natural landscape in anime style illustration*, using the LCM model.

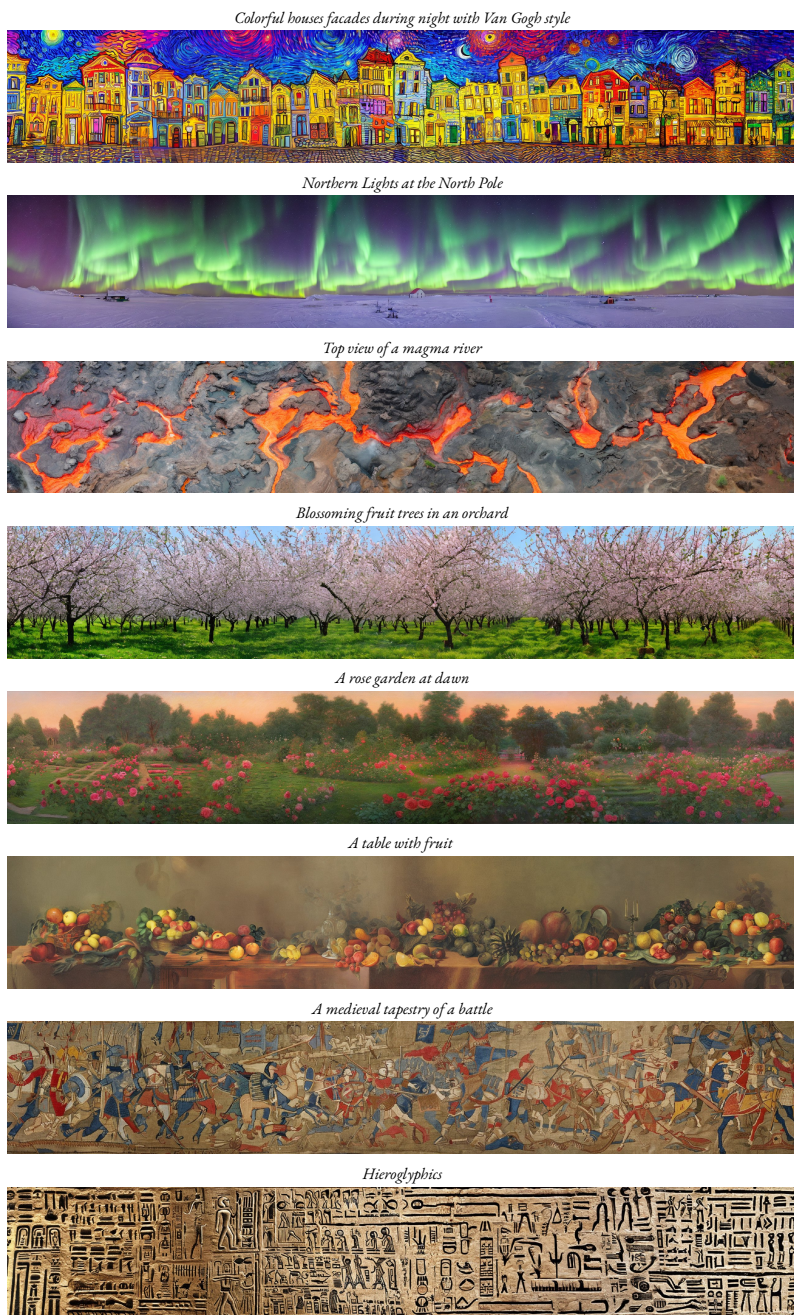


Figure D.27: Horizontal panorama images with various prompts and the LDM model.

A blooming lavender field, with the sun setting on the horizon, creating a peaceful and colorful landscape



A pink panorama at sunset with mountains and a river



A cove with a sandy beach and clear, calm sea under a sunset sky



A serene sunrise over a misty lake, with soft colors reflecting on the water's surface



Raging waves crashing against a rocky coast, under gray and stormy skies



Majestic icebergs floating in the Arctic Ocean, under a bright polar sun



Gentle rolling hills covered in vibrant wildflowers, under a clear blue sky



The park of a royal palace



Figure D.28: Horizontal panorama images with various prompts and the LCM model.



Figure D.29: Vertical panorama images with various prompts and the LDM model.

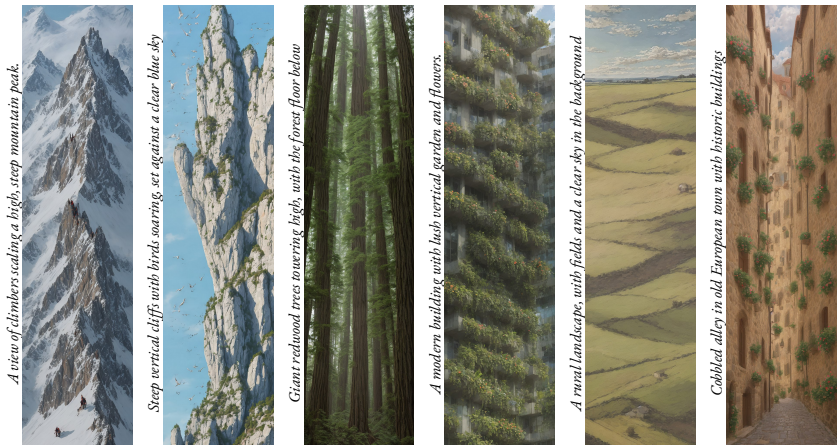


Figure D.30: Vertical panorama images with various prompts and the LCM model.

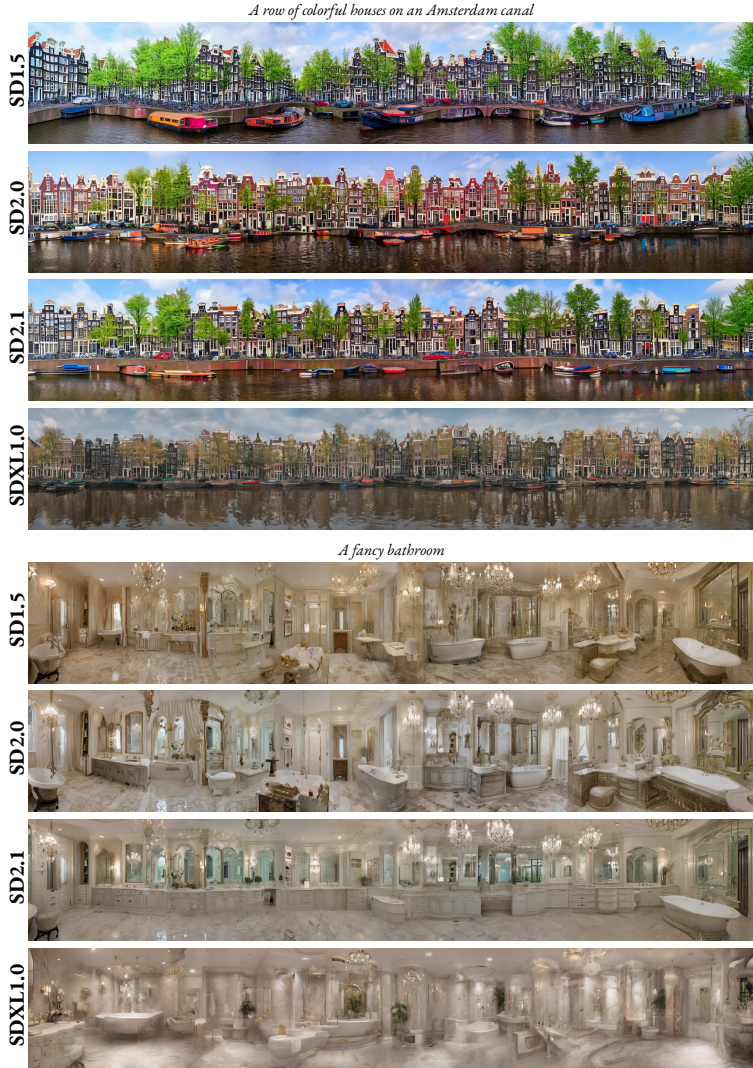


Figure D.31: Images generated with MAD applied to different LDM backbones and resolution 512×3072 . For SDXL1.0 [281], the resolution is 1024×6144 .



Figure D.32: Qualitative comparison between SD-L, SD-L+AttnS, MD, SyncD, and MAD on the generation of images with a prompt that does not fit well into all the specified aspect ratios (top: 512×3072 , bottom-left: 512×2048 , bottom-right: 512×1024). We use the LDM as backbone, the same seed, and the prompt *A fancy living room*.

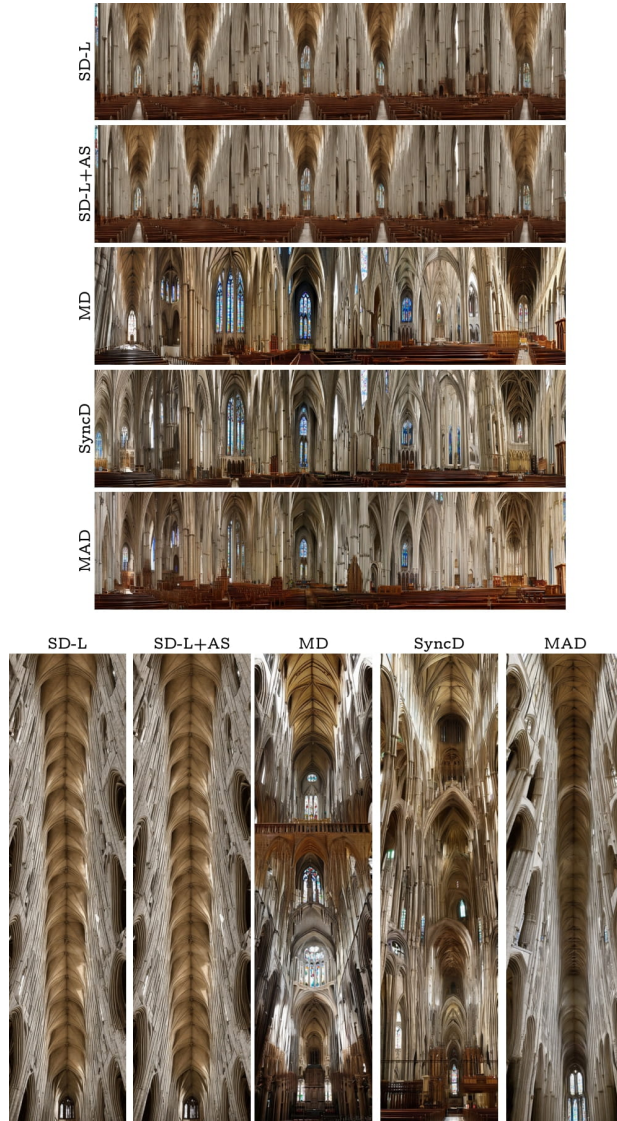


Figure D.33: Horizontal and vertical images generated with the same seed and for the prompt *A gothic cathedral nave* using different methods applied on the LDM model: direct inference (SD-L), Attention Scaling (SD-L+AS), MD, SyncD, and MAD.



Shifting the Breaking Point of Flow Matching for Multi-Instance Editing

E.1 MASKING APPLICATION STRATEGY

In Algorithm 1 we share the full algorithm of our multi-instance editing method. In particular, it shows how and where standard rectified flow inference is adapted to implement Instance-Disentangled Attention.

E.2 IMPLEMENTATION DETAILS

We implement and analyze our proposed approach on the FLUX.1 Kontext [194] image editing framework. It uses a custom VAE as an image encoder, and a custom MMDiT [108] as a generative model.

The instance prompt tokens $\{\mathbb{T}_n\}_{n=1}^N$ are obtained by encoding the instance text instructions $\{s_n\}_{n=1}^N$ using its finetuned T5 [299] text encoder, where the effective tokens are those obtained until the EOS token is generated. The entire Z^{context} sequence is obtained by encoding the input image with FLUX’s custom VAE, whereas the Z^{latent} is initialized to Gaussian noise. The instance tokens

Algorithm 1 Flow Matching-based Editing with Instance-Disentangled Attention

Require: Initial latent $x_0 \sim p_0$; **Conditioning** $Z^{\text{text}}, Z^{\text{context}}$; **Partitions** $\{\mathbb{T}_g, \mathbb{T}_n, \mathbb{L}_u, \mathbb{L}_n, \mathbb{C}_u, \mathbb{C}_n\}_{n=1}^N$; Steps T

Ensure: Edited sample x_1

- 1: # Construct masks using definitions in Equation (12.4) and Equation (12.5)
 - 2: $\mathcal{M}^{\text{dis}}, \mathcal{M}^{\text{har}} \leftarrow \text{ConstructMasks}(\{\mathbb{T}, \mathbb{L}, \mathbb{C}\})$
 - 3: Initialize $x \leftarrow x_0, \Delta t \leftarrow 1/T$
 - 4: **for** $k = 0$ to $T - 1$ **do**
 - 5: $t \leftarrow k \cdot \Delta t$
 - 6: $Z^{\text{latent}} \leftarrow \text{ImageEncoder}(x)$
 - 7: $Z \leftarrow Z^{\text{text}} \parallel Z^{\text{latent}} \parallel Z^{\text{context}}$
 - 8: $b \leftarrow \text{Embed}(Z, t)$
 - 9: **for** $\ell = 1$ to L **do**
 - 10: **if** $\ell \in L_{\text{early}} \cup L_{\text{late}}$ **then**
 - 11: $M \leftarrow \mathcal{M}^{\text{har}}$
 - 12: **else**
 - 13: $M \leftarrow \mathcal{M}^{\text{dis}}$
 - 14: $b \leftarrow \text{TransformerBlock}(b, \text{mask} = M)$
 - 15: $v_\theta \leftarrow \text{OutputHead}(b)$
 - 16: $x \leftarrow x + v_\theta \cdot \Delta t$
 - 17: **return** x
-

$\{\mathbb{L}_n\}_{n=1}^N$ and $\{\mathbb{C}_n\}_{n=1}^N$ are obtained from Z^{latent} and Z^{context} , respectively, by converting the $\{b_n\}_{n=1}^N$ bounding boxes to masks. Specifically, the instance latent and context tokens retain the place they have within the context image’s token sequence Z^{context} as encoded by the VAEs, and the mask acts on those areas of the sequence.

As global prompt $\mathbb{T}_g$ we use a **null prompt** made up of 200 padding tokens, which we call a *utility prompt*: it is the only prompt that interacts with $\mathbb{L}_u$ and $\mathbb{C}_u$ in the layers using $\mathcal{M}^{\text{dis}}$, so it is needed to make sure the model preserves the background effectively.

E.3 ATTENTION MASKING IMPLEMENTATION ANALYSIS

Here we report a detailed analysis of the attention masking approach when applied to FLUX.1 Kontext as described in Section E.2. In particular, considering the way the inputs are encoded by the Flux VAE, and our efficient text encoding scheme on top of T5, we discuss the attention mask values as they appear when running the model on an image from our benchmark datasets.

In this section we use the following definitions for simplicity of notation and as a direct reference to our specific implementation on FLUX.1 Kontext for multi-instance image editing:

$$\mathbb{T} := Z^{\text{text}} = \mathbb{T}_g \parallel \mathbb{T}_1 \parallel \dots \parallel \mathbb{T}_N$$

$$\mathbb{L} := Z^{\text{latent}} = \mathbb{L}_u \parallel \mathbb{L}_1 \parallel \dots \parallel \mathbb{L}_N$$

$$\mathbb{C} := Z^{\text{context}} = \mathbb{C}_u \parallel \mathbb{C}_1 \parallel \dots \parallel \mathbb{C}_N$$

In Fig. E.1 and Fig. E.2 we respectively show the M^{dis} and M^{har} attention masks constructed for a sample from LoMOE-Bench, where each pixel is a query-key combination in the attention matrix, each row in the image is a different query, each column is a key, so by looking at a row we can see to which tokens of the sequence each query attends to.

In Fig. E.1 we can see, from top to bottom, the relatively large $\mathbb{T}_g$ (200px high) being allowed to attend to everything, except for the thin strip where the keys of the two instance prompts $\mathbb{T}_1$ and $\mathbb{T}_2$ are placed within the sequence. Below that, the horizontal strip corresponding to those two instance prompts should be analyzed in detail: the first 12 rows of pixels are those of the queries derived from $\mathbb{T}_1$, which do not attend to the global prompt, attend to the 12 columns of pixels referring to $\mathbb{T}_1$'s keys. This interactions among tokens in $\mathbb{T}$ is shown within Fig. E.1 in the zoomed-in section on the right called *A: Prompt-to-Prompt mask*.

Proceeding along the $\mathbb{T}_1$ row, the queries are allowed to attend also to the regions within the latent $\mathbb{L}$ and the context $\mathbb{C}$ that fall within the bounding boxes as-

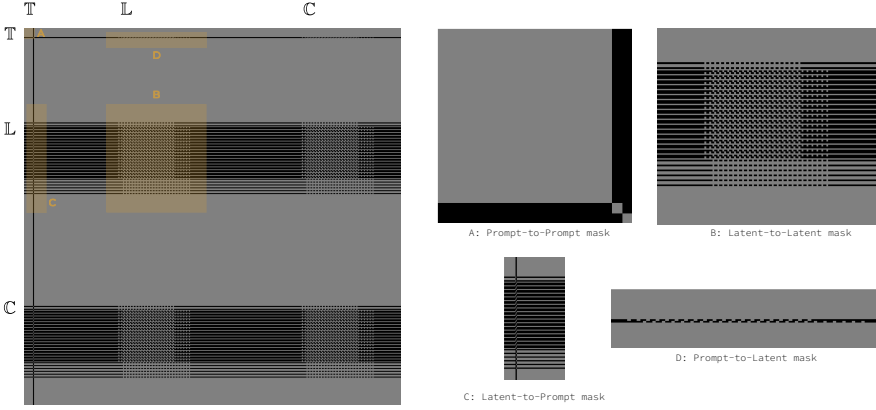


Figure E.1: Detailed representation and analysis of the construction of M^{dis} when using FLUX Kontext as a baseline on a sample from LoMOE-Bench. We use the color black to indicate areas which are blocked from attending each other and the color grey to indicate areas where attention is allowed.

sociated with the first instance prompt: this is the upper row of rectangles within *D: Prompt-to-Latent mask*. They look like isolated regions, because the tokens are encoded in raster order, and the box is not as wide as the entire image, but much smaller. The next 11 rows of pixels are those controlling attention masking for the second prompt, for which the same considerations apply, but in reference to a different instance.

The areas corresponding to the $\mathbb{L}$ and $\mathbb{C}$ tokens are exactly the same as each other because the positions of the bounding boxes within the latents and within the context are the same for this task. In *C: Latent-to-Prompt mask* the interaction between the queries of the latent and the keys of the prompt is highlighted: the tokens attending to the global prompt and all of the latents and context tokens are those corresponding to the background $\mathbb{L}_u$ and $\mathbb{C}_u$, whereas those corresponding to the latent instances $\mathbb{L}_n$ and context instances $\mathbb{C}_n$ only attend to their corresponding $\mathbb{P}_n$, $\mathbb{L}_n$ and $\mathbb{C}_n$, as shown in particular in *B: Latent-to-Latent mask*.

M^{har} , shown in Fig. E.2, is a relaxation of M^{dis} allowing all instance tokens $\mathbb{L}$ and context tokens $\mathbb{C}$ to attend to all instance and context tokens.

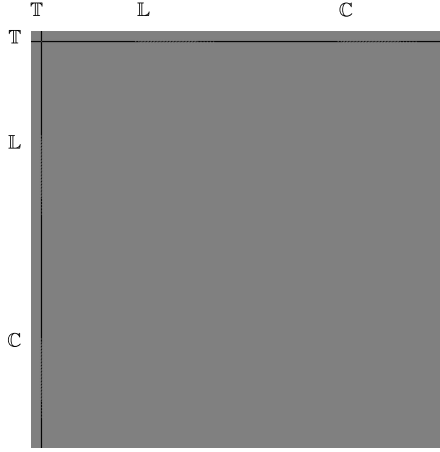


Figure E.2: Representation of $\mathcal{M}^{\text{har}}$ when running FLUX Kontext as a baseline on a sample from LoMOE-Bench. We use the color black to indicate areas which are blocked from attending each other and the color grey to indicate areas where attention is allowed.

E.4 EXTENDED LLM-AS-A-JUDGE EVALUATION

We repeat the same experiments performed on the LLM-as-a-judge setting based on Gemini 3 Flash described in Section 12.4 on an extended set of questions including all binary comparisons of 6 models on 37 images from LoMOE-Bench and 119 images from the infographic datasets and we report the results in Fig. E.3. In this extended validation our model obtains the highest score, particularly on infographics and we find out the following: (i) on LoMOE-Bench LoMOE underperforms when compared to the scores proposed in [62] and reported in Tab. 12.3: the Elo score computed from the LLM-as-a-judge evaluation only ranks it higher than the FLUX Kontext baseline with visual cues, whereas the end-to-end FLUX Kontext baseline gets ranked as second-best, with a score that is close to our method, but significantly higher than all others; (ii) on infographics, Calligrapher in particular severely underperforms when compared to the other models, and our method leads by a significant margin over all competitors and baselines.

Especially on LoMOE-Bench, this underlines a lack of reliable scores to measure this task in a consistent way: on that dataset, LoMOE appears from the scores to be among the best models. The generation quality, though, is low, so

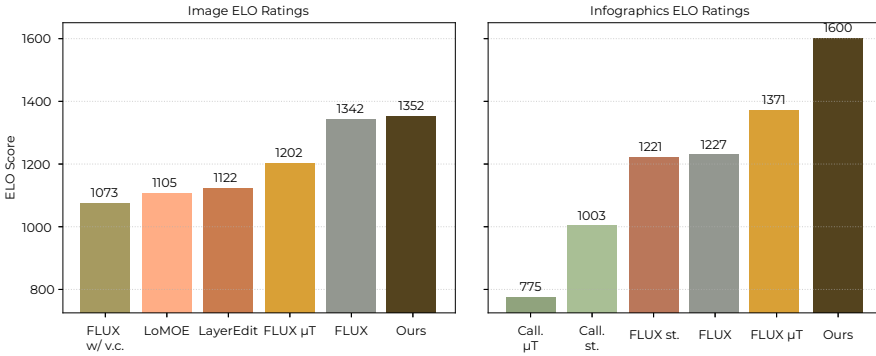


Figure E.3: Elo score of LLM-as-a-judge on the image benchmark (LOMOE) and the Infographics (Crello and Infodet). We use the abbreviation *Call.* for the Calligrapher baselines.

the LLM-as-a-judge model prefers FLUX-based models, which generate images that look much better. This is qualitatively evident by looking at samples like those in Fig. E.4, but is not captured by CLIP score, for example.

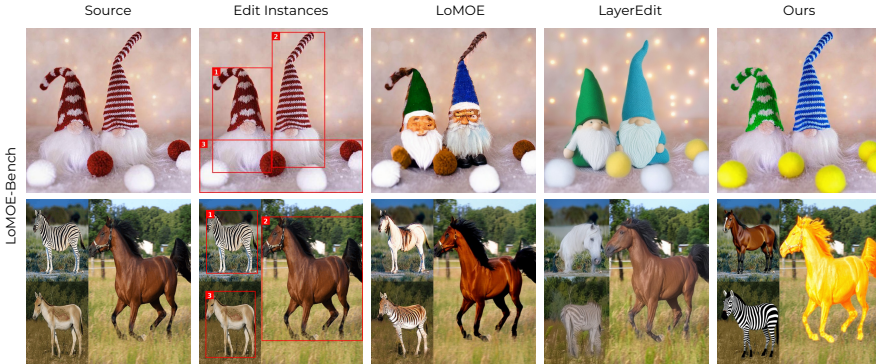


Figure E.4: Further qualitative comparison on **LOMOE-Bench**.

E.5 PROMPTS FOR QUALITATIVE RESULTS IN THE MAIN CHAPTER

The prompts for the qualitative results are presented here in the format Instance ID. Source $\rightarrow$ Target, which are passed to all FLUX Kontext-based models

as Replace Source with Target instead. For other approaches, we follow their standard pipelines, including prompt generation. The prompts for the samples shown in Fig. 12.4 are as follows.

For the Crello Edit sample:

1. FOR YOU -> POUR TOI
2. JUST -> SEULEMENT
3. AND -> AND
4. with care -> avec précaution
5. WITH LOVE -> AVEC AMour
6. MADE -> FABRIQUÉ

For the InfoEdit sample:

1. Who's Getting The Most Sleep?. -> Chi sta dormendo di più?
2. Netherlands -> Paesi Bassi
3. 5.22 -> 5.22
4. New Zealand -> Nuova Zelanda.
5. 4.00 -> 4.00
6. France -> Francia
7. 3.19 -> 3.19
8. Australia -> Australia
9. 1.30 -> 1.30
10. -1.73 -> -1,73
11. Canada -> Canada
12. -3.16 -> -3.16
13. United Arab Emirates -> Emirati Arabi Uniti.
14. -6.12 -> -6.12
15. United Kingdom -> Regno Unito.
16. Italy -> Italia.
17. -7.32 -> -7.32
18. -7.88 -> -7.88
19. United States -> Stati Uniti

- 20. -8.51 -> -8.51
- 21. China -> Cina
- 22. -9.85 -> -9,85
- 23. Switzerland -> Svizzera
- 24. -12.50 -> -12,50
- 25. -18.26 -> -18,26
- 26. Germany -> Germania
- 27. -25.88 -> -25,88
- 28. Brazil -> Brasile
- 29. -29.49 -> -29,49
- 30. Japan -> Giappone
- 31. -36.23 -> -36,23
- 32. Singapore -> Singapore
- 33. Latest available data:2016 -> Ultimi dati disponibili:
2016
- 34. Source: Science Advances -> Fonte: Science Advances.
- 35. statista -> Statista.
- 36. Minutes above and below eight hours of sleep in selected
countries -> Minuti sopra e sotto le otto ore di sonno in
paesi selezionati.

For the LoMOE-Bench sample:

- 1. a gnome with a red hat -> a gnome with a green hat
- 2. a gnome with a red hat -> a gnome with a blue hat
- 3. white and red cotton balls -> yellow cotton balls

E.6 ADDITIONAL QUALITATIVE RESULTS

In this appendix, we report further qualitative results for all three datasets.

The prompts for the Crello Edit qualitatives in Fig. E.5 are as follows.

For the sample in the first row:

1. DON'T MISS THE UPDATES! -> NE MANQUEZ PAS LES MISES À JOUR !
2. PHOTOS -> PHOTOS
3. STUNNING -> ÉTONNANT
4. THIS ACCOUNT CONTAINS -> CE COMPTE CONTIENT

For the sample in the second row:

1. not visible -> Non visible.
2. Some things are -> Alcune cose sono
3. from the shore -> dalla riva

For the sample in the third row:

1. GO AT ALL? -> TU PEUX Y ALLER ?
2. WHY -> POUQUOI ?
3. www.greendaleshop.com -> www.greendaleshop.com
4. All the way -> Tout du long.
5. GOING -> PARTIR.
6. Aren't -> N'est-ce pas ?
7. IF YOU -> SI VOUS

The prompts for the InfoEdit qualitatifs in Fig. E.6 are as follows.

For the sample in the first row:

1. China -> La Chine.
2. \$229.29 -> 229,29 \$
3. Thailand -> La Thaïlande
4. \$215.77 -> 215,77 \$
5. \$186.52 -> 186,52 \$
6. Singapore -> Singapour
7. \$181.85 -> 181,85 \$
8. Greece -> La Grèce.
9. \$164.26 -> 164,26 \$
10. France -> La France.

E.6. ADDITIONAL QUALITATIVE RESULTS



Figure E.5: Qualitative results on the **Crello Edit** dataset.

11. \$155.94 -> 155,94 \$
12. Germany -> L'Allemagne.
13. \$154.77 -> 154,77 \$
14. Turkey -> La Turquie.
15. \$150.09 -> 150,09 \$
16. United Kingdom -> Royaume-Uni
17. \$148.60 -> 148,60 \$
18. United States -> États-Unis.
19. \$135.00 -> 135,00 \$
20. Source:Cosmetify -> Source : Cosmetify.
21. statistaS -> StatistaS.
22. Where Posh Perfume Has A Prohibitive Price Tag -> Où le parfum Posh a un prix prohibitif.

23. Average cost of Chanel No 5 in selected countries in 2021 (in U.S. dollars) -> Coût moyen de Chanel N°5 dans certains pays en 2021 (en dollars américains)

24. United Arab Emirates -> Émirats arabes unis.

For the sample in the second row:

1. by State -> Par État.

2. Savings needed for a comfortable retirement (USD) -> Épargne nécessaire pour une retraite confortable (en USD)

3. \$0 -> 0 \$

4. \$200K -> 200 000 dollars.

5. \$400K -> 400 000 dollars.

6. \$600K -> 600 000 dollars.

7. \$800K -> 800 000 dollars.

8. \$1.0M -> 1,0 million de dollars.

9. \$1.2M -> 1,2 million de dollars.

10. Difference -> La différence.

11. Hawaii -> Hawaï.

12. \$423K -> 423 000 dollars.

13. New York -> New York.

14. \$396K -> 396 000 dollars.

15. California -> Californie

16. \$396K -> 396 000 dollars.

17. New Jersey -> New Jersey.

18. \$365K -> 365 000 dollars.

19. Maryland -> Maryland

20. \$364K -> 364 000 dollars.

21. Arizona -> L'Arizona.

22. \$356K -> 356 000 dollars.

23. Massachusetts -> Le Massachusetts.

24. \$355K -> 355 000 dollars.

25. Connecticut -> Le Connecticut.

26. \$343K -> 343 000 dollars.
27. Illinois -> L'Illinois.
28. \$342K -> 342 000 dollars.
29. South Carolina -> La Caroline du Sud.
30. \$334K -> 334 000 dollars.
31. Florida -> La Floride.
32. \$330K -> 330 000 dollars.
33. Mississippi -> Le Mississippi.
34. \$328K -> 328 000 dollars.
35. Rhode Island -> Rhode Island.
36. \$327K -> 327 000 dollars.
37. Tennessee -> Le Tennessee.
38. \$327K -> 327 000 dollars.
39. New Mexico -> Le Nouveau-Mexique.
40. \$326K -> 326 000 dollars.
41. Louisiana -> Louisiane.
42. \$322K -> 322 000 dollars.
43. Georgia -> Géorgie.
44. \$320K -> 320 000 dollars.
45. Nevada -> Nevada.
46. \$320K -> 320 000 dollars.
47. Maine -> Le Maine.
48. \$319K -> 319 000 dollars.
49. Alabama -> L'Alabama.
50. \$317K -> 317 000 dollars.
51. Delaware -> Delaware.
52. \$316K -> 316 000 dollars.
53. How Much Men & Women Need to Save to Retire Comfortably
-> Combien les hommes et les femmes doivent-ils épargner
pour prendre une retraite confortable
54. Visualizing how life expectancy differences and local
living costs affect the retirement savings men and women

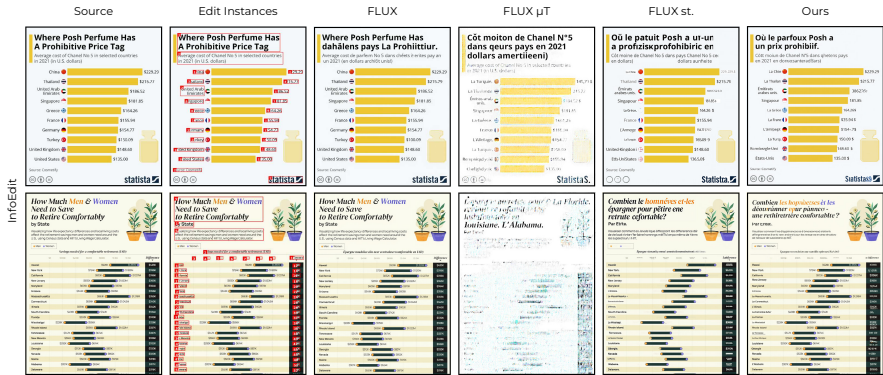


Figure E.6: Qualitative results on the **InfoEdit** dataset.

need around the U.S., using Census data and MIT's Living Wage Calculator. -> Visualiser comment les différences d'espérance de vie et les coûts de la vie locaux affectent l'épargne-retraite dont les hommes et les femmes ont besoin aux États-Unis, en utilisant les données du recensement et le calculateur de salaire de subsistance de MIT.

The prompts used to generate the LoMOE-Bench qualitative results in Fig. E.7 are as follows.

For the sample in the first row:

1. a barstool with a black cushion -> a barstool with a green cushion
2. a barstool with a black cushion -> a barstool with a blue cushion
3. a barstool with a black cushion -> a barstool with a blue cushion
4. a barstool with a black cushion -> a barstool with a green cushion
5. a wooden basket -> a steel basket with fruits

For the sample in the second row:

1. a cotton ball -> a violet coloured cotton ball
2. a cotton ball -> an indigo coloured cotton ball
3. a cotton ball -> a blue coloured cotton ball
4. a cotton ball -> a dark green coloured cotton ball
5. a cotton ball -> a light yellow coloured cotton ball
6. a cotton ball -> an orange coloured cotton ball
7. a cotton ball -> a red coloured cotton ball

For the sample in the third row:

1. a red flag -> a pink flag
2. a red flag -> a purple flag
3. a red flag -> a blue flag
4. a red flag -> an orange flag

For the sample in the fourth row:

1. a zebra -> a horse
2. a horse -> a golden fire horse
3. a mule -> a zebra

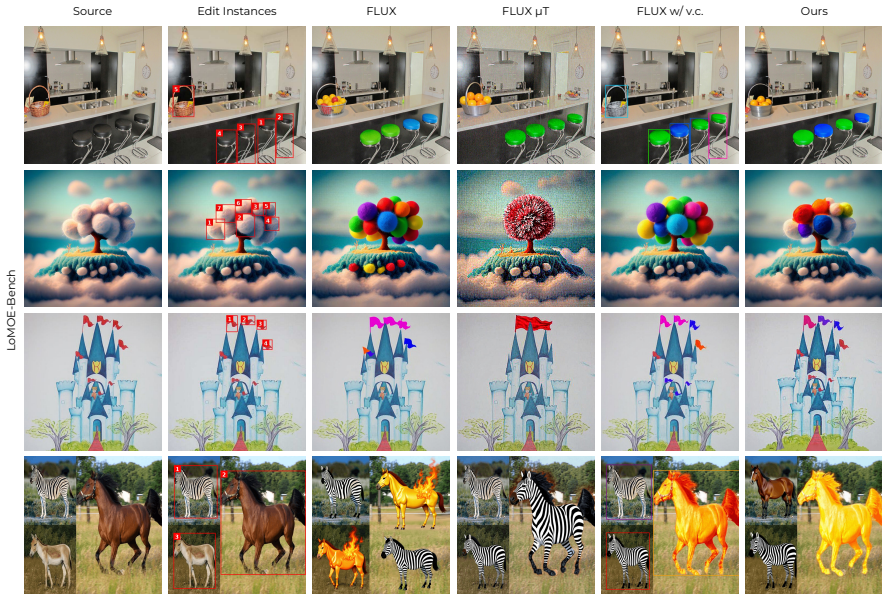


Figure E.7: Qualitative results on the LoMOE-Bench.

F

Color Palettes

The visual design of figures, diagrams, and plots throughout this thesis follows color palettes derived from a curated selection of artworks. Each palette was sampled directly from the source image and discretized into a set of representative colors. This appendix documents the source artworks and their corresponding palettes, one per chapter. Artworks are reproduced in reduced format solely for illustrative purposes (Art. 70, L. 633/1941).

F.1 μ GAT — *ALEXANDER MOSAIC*, c. 100 BCE

The palette used in Chapter 4 was derived from the *Alexander Mosaic*, a Roman floor mosaic discovered in the House of the Faun, Pompeii, dating to approximately 100 BCE. The work depicts the Battle of Issus between Alexander the Great and Darius III, and is now preserved at the Museo Archeologico Nazionale di Napoli.



Figure F.1: *Alexander Mosaic* (c. 100 BCE), attributed to Philoxenus of Eretria. Originally from the House of the Faun, Pompeii; now at the Museo Archeologico Nazionale di Napoli, Naples. Public domain. Reproduced for illustrative purposes (Art. 70, L. 633/1941).

F.2 SCAM — *REST ON THE FLIGHT INTO EGYPT*, LUC-OLIVIER MERSON, 1879

The palette used in Chapter 5 was derived from *Rest on the Flight into Egypt* (1879) by Luc-Olivier Merson, now at the Museum of Fine Arts, Boston. The work depicts the Holy Family sheltering at the base of the Great Sphinx of Giza, bathed in cold moonlight against a vast desert night.



Figure F.2: Luc-Olivier Merson, *Rest on the Flight into Egypt* (1879). Oil on canvas, 72.4 × 128.3 cm. Museum of Fine Arts, Boston. Public domain. Reproduced for illustrative purposes (Art. 70, L. 633/1941).

F.3 IDATTN — *LE SACRE DE NAPOLÉON*, JACQUES-LOUIS DAVID, 1807

The palette used in Chapter 12 was derived from *Le Sacre de Napoléon* (1807) by Jacques-Louis David, one of the largest and most elaborate paintings in the Musée du Louvre, Paris. The work depicts the coronation ceremony of Napoleon Bonaparte at Notre-Dame de Paris on 2 December 1804.



Figure F.3: Jacques-Louis David, *Le Sacre de Napoléon* (1807). Oil on canvas, 610 × 931 cm. Musée du Louvre, Paris. Public domain. Reproduced for illustrative purposes (Art. 70, L. 633/1941).

List of Publications

- [1] Pippi, V., Quattrini, F., Cascianelli, S., and Cucchiara, R. (2024). HWD: A Novel Evaluation Score for Styled Handwritten Text Generation. In *Proceedings of the British Machine Vision Conference*.
- [2] Quattrini, F., Pippi, V., Cascianelli, S., and Cucchiara, R. (2023). Volumetric Fast Fourier Convolution for Detecting Ink on the Carbonized Herculaneum Papyri. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- [3] Quattrini, F., Pippi, V., Cascianelli, S., and Cucchiara, R. (2024a). Alfie: Democratising RGBA Image Generation with No \$\$\$\$. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 38–55. Springer.
- [4] Quattrini, F., Pippi, V., Cascianelli, S., and Cucchiara, R. (2024b). Binarizing Documents by Leveraging Both Space and Frequency. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [5] Quattrini, F., Pippi, V., Cascianelli, S., and Cucchiara, R. (2024c). Merging and splitting diffusion paths for semantically coherent panoramas. In *Proceedings of the European Conference on Computer Vision*, pages 234–251. Springer.
- [6] Quattrini, F., Pippi, V., Cascianelli, S., and Cucchiara, R. (2025a). Zero-Shot Styled Text Image Generation, but Make It Autoregressive. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *Equal contribution.
- [7] Quattrini, F., Zaccagnino, C., Cascianelli, S., and Cucchiara, R. (2025b). A Text Recognition Dataset from Sahidic Coptic Ancient Manuscripts. In *Proceedings of the International Conference on Document Analysis and Recognition*. *Equal contribution.
- [8] Quattrini, F., Zaccagnino, C., Cascianelli, S., Righi, L., and Cucchiara, R. (2024d). μ gat: Improving Single-Page Document Parsing by Providing Multi-Page Context. In *Proceedings of the European Conference on Computer Vision Workshops*, pages 212–228. Springer. *Equal contribution.

- [9] Quattrini, F., Zaccagnino, C., Simsar, E., Gazulla, M. T., Cucchiara, R., Tonioni, A., and Cascianelli, S. (2026). Shifting the Breaking Point of Flow Matching for Multi-Instance Editing. In *Proceedings of the International Conference on Machine Learning*. *Equal contribution.
- [10] Zaccagnino, C., Quattrini, F., Pippi, V., Cascianelli, S., Tonioni, A., and Cucchiara, R. (2026). Autoregressive Styled Text Image Generation, but Make it Reliable. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

Bibliography

- [11] Aberdam, A., Litman, R., Tsiper, S., Anshel, O., Slossberg, R., Mazor, S., Manmatha, R., and Perona, P. (2021). Sequence-to-Sequence Contrastive Learning for Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [12] Akbari, Y., Al-Maadeed, S., and Adam, K. (2020). Binarization of degraded document images using convolutional neural networks and wavelet-based multichannel images. *IEEE Access*, 8:153517–153534.
- [13] Akbari, Y., Britto, A. S., Al-Maadeed, S., and Oliveira, L. S. (2019). Binarization of degraded document images using convolutional neural networks based on predicted two-channel images. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [14] Aksan, E. and Hilliges, O. (2018). STCN: Stochastic Temporal Convolutional Networks. In *Proceedings of the International Conference on Learning Representations*.
- [15] Aksan, E., Pece, F., and Hilliges, O. (2018). DeepWriting: Making digital ink editable via deep generative modeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.
- [16] AlKendi, W., Gechter, F., Heyberger, L., and Guyeux, C. (2024). Advancements and challenges in handwritten text recognition: A comprehensive survey. *J. Imaging*.
- [17] Allegra, D., Ciliberto, E., Ciliberto, P., Milotta, F. L. M., Petrillo, G., Stanco, F., and Trombato, C. (2015). Virtual unrolling using X-ray computed tomography. In *European Signal Processing Conference*.
- [18] Allegra, D., Ciliberto, E., Ciliberto, P., Petrillo, G., Stanco, F., and Trombatore, C. (2016). X-ray computed tomography for virtually unrolling damaged papyri. *Applied Physics A*, 122:1–7.
- [19] Alonso, E., Moysset, B., and Messina, R. (2019). Adversarial Generation of Handwritten Text Images Conditioned on Sequences. In *Proceedings of the International Conference on Document Analysis and Recognition*.

- [20] Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., and Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [21] Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., and Manmatha, R. (2024). Docformerv2: Local features for document understanding. In *Proceedings of the Conference on Artificial Intelligence*.
- [22] Augustin, E., Carré, M., Grosicki, E., Brodin, J.-M., Geoffrois, E., and Prêteux, F. (2006). RIMES evaluation campaign for handwritten mail processing. In *IWFHR*.
- [23] Avrahami, O., Fried, O., and Lischinski, D. (2023a). Blended Latent Diffusion. *ACM Trans. Graphics*, 42(4):1–11.
- [24] Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., and Yin, X. (2023b). SpáText: Spatio-Textual Representation for Controllable ImageGgeneration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [25] Avrahami, O., Lischinski, D., and Fried, O. (2022). Blended Diffusion for Text-driven Editing of Natural Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [26] Ayatollahi, S. M. and Nafchi, H. Z. (2013). Persian heritage image binarization competition (PHIBC 2012). In *Pattern Recognition and Image Analysis*.
- [27] Baena, R., Kalleli, S., and Aubry, M. (2024). General Detection-based Text Line Recognition. *Advances in Neural Information Processing Systems*.
- [28] Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al. (2022). eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*.
- [29] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- [30] Bar-Tal, O., Yariv, L., Lipman, Y., and Dekel, T. (2023). MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. *Proceedings of the International Conference on Machine Learning*.

- [31] Baró, A., Chen, J., Fornés, A., and Megyesi, B. (2019). Towards a generic unsupervised method for transcription of encoded manuscripts. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 73–78.
- [32] Barrere, K., Soullard, Y., Lemaitre, A., and Coüasnon, B. (2022). A Light Transformer-Based Architecture for Handwritten Text Recognition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [33] Barrere, K., Soullard, Y., Lemaitre, A., and Coüasnon, B. (2024). Training transformer architectures on few annotated data: an application to historical handwritten text recognition. *International Journal on Document Analysis and Recognition*.
- [34] Barron, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [35] Barron, J. T. (2020). A generalization of Otsu’s method and minimum error thresholding. In *Proceedings of the European Conference on Computer Vision*.
- [36] Baum, D., Herter, F., Larsen, J. M., Lichtenberger, A., and Raja, R. (2021). Revisiting the Jerash Silver Scroll: A new visual data analysis approach. *Digital applications in archaeology and cultural heritage*.
- [37] Ben-Ezra, M. (2000). Segmentation with Invisible Keying Signal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [38] Berenguel-Baeta, B., Bermudez-Cameo, J., and Guerrero, J. J. (2022). FreDSNet: Joint Monocular Depth and Semantic Segmentation with Fast Fourier Convolutions. *Proceedings of the IEEE International Conference on Robotics and Automation*.
- [39] Bhunia, A. K., Bhunia, A. K., Sain, A., and Roy, P. P. (2019a). Improving document binarization via adversarial noise-texture augmentation. In *Proceedings of the IEEE International Conference on Image Processing*.

- [40] Bhunia, A. K., Das, A., Bhunia, A. K., Kishore, P. S. R., and Roy, P. P. (2019b). Handwriting Recognition in Low-Resource Scripts Using Adversarial Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [41] Bhunia, A. K., Ghose, S., Kumar, A., Chowdhury, P. N., Sain, A., and Song, Y.-Z. (2021a). MetaHTR: Towards Writer-Adaptive Handwritten Text Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [42] Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Khan, F. S., and Shah, M. (2021b). Handwriting Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [43] Bhunia, A. K., Sain, A., Chowdhury, P. N., and Song, Y.-Z. (2021c). Text is Text, No Matter What: Unifying Text Recognition using Knowledge Distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [44] Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
- [45] Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.
- [46] Biten, A. F., Litman, R., Xie, Y., Appalaraju, S., and Manmatha, R. (2022). Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [47] Blau, T., Fogel, S., Ronen, R., Golts, A., Ganz, R., Ben Avraham, E., Aberdam, A., Tsiper, S., and Litman, R. (2024). GRAM: Global Reasoning for Multi-Page VQA. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [48] Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. (2024). Nougat: Neural optical understanding for academic documents. In *Proceedings of the International Conference on Learning Representations*.
- [49] Bluche, T. (2016). Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In *Advances in Neural Information Processing Systems*, volume 29, pages 838–846.

- [50] Bluche, T., Louradour, J., and Messina, R. (2017). Scan, Attend and Read: End-to-End Handwritten Paragraph Recognition with MDLSTM Attention. In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, pages 1050–1055. IEEE.
- [51] Bluche, T. and Messina, R. (2017). Gated convolutional recurrent neural networks for multilingual handwriting recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [52] Brock, A., Donahue, J., and Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [53] Brooks, T., Holynski, A., and Efros, A. A. (2023). InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [54] Brown, M. S. and Seales, W. B. (2001). Document restoration using 3D shape: a general deskewing algorithm for arbitrarily warped documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [55] Burgert, R. D., Price, B. L., Kuen, J., Li, Y., and Ryoo, M. S. (2024). MAG-ICK: A Large-scale Captioned Dataset from Matting Generated Images using Chroma Keying. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [56] Cao, H., Ding, X., and Liu, C. (2003). A cylindrical surface model to rectify the bound document image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [57] Cascianelli, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2022a). Boosting Modern and Historical Handwritten Text Recognition with Deformable Convolutions. *International Journal on Document Analysis and Recognition*, 25(3):207–217.
- [58] Cascianelli, S., Cornia, M., Baraldi, L., Piazzzi, M. L., Schiuma, R., and Cucchiara, R. (2021). Learning to Read L’Infinito: Handwritten Text Recognition with Synthetic Training Data. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns*.
- [59] Cascianelli, S., Pippi, V., Martin, M., Cornia, M., Baraldi, L., Christopher, K., and Cucchiara, R. (2022b). The LAM Dataset: A Novel Benchmark

for Line-Level Handwritten Text Recognition. In *Proceedings of the International Conference on Pattern Recognition*.

- [60] Causer, T. and Wallace, V. (2012). Building a volunteer community: results and findings from Transcribe Bentham. *Digital Humanities Quarterly*, 6(2).
- [61] Chai, S., Zhuang, L., and Yan, F. (2023). LayoutDM: Transformer-based Diffusion Model for Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [62] Chakrabarty, G., Chandrasekar, A., Hebbalaguppe, R., and AP, P. (2024). LoMOE: Localized Multi-Object Editing via Multi-Diffusion. In *Proceedings of the ACM International Conference on Multimedia*, pages 3342–3351.
- [63] Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. (2022). Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [64] Charbonnier, P., Blanc-Feraud, L., Aubert, G., and Barlaud, M. (1994). Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the IEEE International Conference on Image Processing*.
- [65] Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., and Cohen-Or, D. (2023). Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. *ACM Trans. Graph.*
- [66] Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., and Li, Z. (2024a). PixArt- Σ : Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. *arXiv preprint arXiv:2403.04692*.
- [67] Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., and Li, Z. (2024b). PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [68] Chen, J., Zhang, R., Zhou, Y., and Chen, C. (2023). Towards Aligned Layout Generation via Diffusion Model with Aesthetic Constraints. In *Proceedings of the International Conference on Learning Representations*.

- [69] Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. (2020). WaveGrad: Estimating Gradients for Waveform Generation. In *Proceedings of the International Conference on Learning Representations*.
- [70] Chen, X., Liu, Z., Xie, S., and He, K. (2025a). Deconstructing Denoising Diffusion Models for Self-Supervised Learning. In *Proceedings of the International Conference on Learning Representations*.
- [71] Chen, Y., Jha, A., Shah, H., Anand, S., Kumar, A., Chari, V., Prasad, S., De Sa, C., and Xu, Z. (2025b). Context Diffusion: In-Context Aware Image Generation. *arXiv preprint arXiv:2501.01697*.
- [72] Chen, Y., Ma, Z., Jia, G., Jiang, C., Li, J., and Zhou, B. (2025c). Context-Aware Autoregressive Models for Multi-Conditional Image Generation. *arXiv preprint arXiv:2505.12274*.
- [73] Chen, Z., Li, Y., Wang, H., Chen, Z., Jiang, Z., Li, J., Wang, Q., Yang, J., and Tai, Y. (2025d). Region-aware text-to-image generation via hard binding and soft refinement. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [74] Cheng, S., Bao, Y., Cao, Q., Huang, L., Kang, L., Liu, Z., Lu, Y., Zhu, W., Chen, J., Huang, Z., et al. (2025). Seed-X: Building Strong Multilingual Translation LLM with 7B Parameters. *arXiv preprint arXiv:2507.13618*.
- [75] Chi, L., Jiang, B., and Mu, Y. (2020). Fast fourier convolution. *Advances in Neural Information Processing Systems*.
- [76] Chong, M. J. and Forsyth, D. (2020). Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [77] Chowdhury, A. and Vig, L. (2018). An efficient end-to-end neural model for handwritten text recognition. In *Proceedings of the British Machine Vision Conference*.
- [78] Chowdhury, R. D. (2024). DAAM-Image2Image: Extension of DAAM for Image Self-Attention in Diffusion Models. <https://github.com/RishiDarkDevil/daam-i2i>. Accessed: 2024-01-22.
- [79] Cilia, N. D., De Stefano, C., Fontanella, F., and di Freca, A. S. (2019). A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognition Letters*, 121:77–86.

- [80] Clanuwat, T., Lamb, A., and Kitamoto, A. (2019). KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 607–614. IEEE.
- [81] Cojocaru, I., Cascianelli, S., Baraldi, L., Corsini, M., and Cucchiara, R. (2020). Watch Your Strokes: Improving Handwritten Text Recognition with Deformable Convolutions. In *Proceedings of the International Conference on Pattern Recognition*.
- [82] Cojocaru, I., Cascianelli, S., Baraldi, L., Corsini, M., and Cucchiara, R. (2021). Watch your strokes: improving handwritten text recognition with deformable convolutions. In *Proceedings of the International Conference on Pattern Recognition*.
- [83] Cooley, J. W. and Tukey, J. W. (1965). An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*.
- [84] Coquenet, D., Chatelain, C., and Paquet, T. (2023). End-to-End Handwritten Paragraph Text Recognition Using a Vertical Attention Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):508–524.
- [85] Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. (2022). DiffEdit: Diffusion-based semantic image editing with mask guidance. In *Proceedings of the International Conference on Learning Representations*.
- [86] Couairon, G., Verbeek, J., Schwenk, H., and Cord, M. (2023). DiffEdit: Diffusion-based Semantic Image Editing with Mask Guidance. In *Proceedings of the International Conference on Learning Representations*.
- [87] Crowson, K., Baumann, S. A., Birch, A., Abraham, T. M., Kaplan, D. Z., and Shippole, E. (2024). Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Proceedings of the International Conference on Machine Learning*.
- [88] Cui, C., Sun, T., Lin, M., Gao, T., Zhang, Y., Liu, J., Wang, X., Zhang, Z., Zhou, C., Liu, H., et al. (2025). PaddleOCR 3.0 Technical Report. *arXiv preprint arXiv:2507.05595*.
- [89] Dahary, O., Patashnik, O., Aberman, K., and Cohen-Or, D. (2024). Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation.

- In *Proceedings of the European Conference on Computer Vision*, pages 432–448. Springer.
- [90] Dai, G., Zhang, Y., Ke, Q., Guo, Q., and Huang, S. (2024). One-DM: One-Shot Diffusion Mimicker for Handwritten Text Generation. In *Proceedings of the European Conference on Computer Vision*.
 - [91] Dai, G., Zhang, Y., Wang, Q., Du, Q., Yu, Z., Liu, Z., and Huang, S. (2023). Disentangling Writer and Character Styles for Handwriting Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
 - [92] Dang, Q.-V. and Lee, G.-S. (2021). Document image binarization with stroke boundary feature guided network. *IEEE Access*, 9:36924–36936.
 - [93] d’Arce, R., Norton, T., Hannuna, S., and Cristianini, N. (2022). Self-attention networks for non-recurrent handwritten text recognition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
 - [94] Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., and Morariu, V. (2022). End-to-end document recognition and understanding with dessurt. In *Proceedings of the European Conference on Computer Vision*.
 - [95] Davis, B., Tensmeyer, C., Price, B., Wigington, C., Morse, B., and Jain, R. (2020). Text and Style Conditioned GAN for Generation of Offline Handwriting Lines. In *Proceedings of the British Machine Vision Conference*.
 - [96] de Buy Wenniger, G. M., Schomaker, L., and Way, A. (2019). No padding please: Efficient neural handwriting recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*.
 - [97] De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O’Donoghue, B., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*.
 - [98] De Gregorio, G., Biswas, S., Souibgui, M. A., Bensalah, A., Lladós, J., Fornés, A., and Marcelli, A. (2022). A Few Shot Multi-representation Approach for N-Gram Spotting in Historical Manuscripts. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.

- [99] Deng, F., Wu, Z., Lu, Z., and Brown, M. S. (2010). Binarizationshop: a user-assisted software suite for converting old documents to black-and-white. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*.
- [100] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.
- [101] Dey, S. and Jawanpuria, P. (2021). Light-weight document image cleanup using perceptual loss. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [102] Dhariwal, P. and Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*.
- [103] Doncescu, A., Bouju, A., and Quillet, V. (1997). Former books digital processing: image warping. In *Proceedings Workshop on Document Image Analysis*.
- [104] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*.
- [105] Du, Z. and He, C. (2021). Nonlinear diffusion equation with selective source for binarization of degraded document images. *Appl. Math. Model.*, 99:243–259.
- [106] Eijkelboom, F., Zimmermann, H., Vadvaga, S., Bekkers, E. J., Welling, M., Naesseth, C. A., and van de Meent, J.-W. (2025). Controlled Generation with Equivariant Variational Flow Matching. In *Proceedings of the International Conference on Machine Learning*.
- [107] Esser, P., Chiu, J., Atighehchian, P., Granskog, J., and Germanidis, A. (2023). Structure and Content-Guided Video Synthesis with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

- [108] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. (2024). Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*.
- [109] Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [110] Fan, L., Li, K., Li, H., Yang, X., Zhang, P., Wang, L., Duan, Y., and Li, B. (2025a). Fluid: Scaling Autoregressive Continuous Image Generation. In *Proceedings of the International Conference on Learning Representations*.
- [111] Fan, L., Tang, L., Qin, S., Li, T., Yang, X., Qiao, S., Steiner, A., Sun, C., Li, Y., Zhu, T., et al. (2025b). Unified autoregressive visual generation and understanding with continuous tokens. In *Proceedings of the International Conference on Learning Representations*.
- [112] Feng, S. (2022). Effective document image binarization via a convex variational level set model. *Appl. Math. Comput.*, 419:126861.
- [113] Fischer, A., Frinken, V., Fornés, A., and Bunke, H. (2011). Transcription alignment of Latin manuscripts using hidden Markov models. In *HIP*.
- [114] Fischer, A., Keller, A., Frinken, V., and Bunke, H. (2012). Lexicon-free handwritten word spotting using character HMMs. *Pattern Recognition Letters*, 33(7):934–942.
- [115] Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., and Stolz, M. (2009). Automatic transcription of handwritten medieval documents. In *Proceedings of the International Conference on Virtual Systems and Multimedia*, pages 137–142. IEEE.
- [116] Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., and Litman, R. (2020). ScrabbleGAN: Semi-Supervised Varying Length Handwritten Text Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [117] Fu, F., Huang, M., Zhang, L., and Mao, Z. (2026). LayerEdit: Disentangled Multi-Object Editing via Conflict-Aware Multi-Layer Learning. *Proceedings of the Conference on Artificial Intelligence*.

- [118] Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. (2022). Make-a-Scene: Scene-Based Text-to-Image Generation with Human Priors. In *Proceedings of the European Conference on Computer Vision*.
- [119] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- [120] Gan, J. and Wang, W. (2021). HiGAN: Handwriting Imitation Conditioned on Arbitrary-Length Texts and Disentangled Styles. In *Proceedings of the Conference on Artificial Intelligence*.
- [121] Gan, J., Wang, W., Leng, J., and Gao, X. (2022). HiGAN+: Handwriting Imitation GAN with Disentangled Representations. *ACM Trans. Graphics*, 42(1):1–17.
- [122] Gatos, B., Ntirogiannis, K., and Pratikakis, I. (2009). ICDAR 2009 document image binarization contest (DIBCO 2009). In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [123] Gatos, B., Pratikakis, I., and Ntirogiannis, K. (2007). Segmentation based recovery of arbitrarily warped document images. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [124] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [125] Goel, V., Peruzzo, E., Jiang, Y., Xu, D., Xu, X., Sebe, N., Darrell, T., Wang, Z., and Shi, H. (2024). Pair diffusion: A comprehensive multimodal object-level image editor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8609–8618.
- [126] Gonda, F., Wei, D., Parag, T., and Pfister, H. (2018). Parallel separable 3D convolution for video and volumetric data understanding. *arXiv preprint arXiv:1809.04096*.
- [127] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.

- [128] Graves, A. (2013). Generating Sequences with Recurrent Neural Networks. *arXiv preprint arXiv:1308.0850*.
- [129] Graves, A. and Schmidhuber, J. (2008). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Advances in Neural Information Processing Systems*, volume 21, pages 545–552.
- [130] Gu, S., Bao, J., Chen, D., and Wen, F. (2020). GIQA: Generated Image Quality Assessment. In *Proceedings of the European Conference on Computer Vision*.
- [131] Guo, J., He, C., and Wang, Y. (2020). Fourth order indirect diffusion coupled with shock filter and source for text binarization. *Signal Processing*, 171:107478.
- [132] Guo, J., He, C., and Zhang, X. (2019). Nonlinear edge-preserving diffusion with adaptive source for document images binarization. *Appl. Math. Comput.*, 351:8–22.
- [133] Haines, T., Mac Aodha, O., and Brostow, G. (2016). My Text in Your Handwriting. *ACM Trans. Graphics*, 35(3).
- [134] Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., and Liu, X. (2025a). Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [135] Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., and Liu, X. (2025b). Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [136] Han, S., Ji, S., and Rhee, J. (2023). Diffusion-Denoising Process with Gated U-Net for High-Quality Document Binarization. *Applied Sciences*, 13(20):11141.
- [137] Harley, A. W., Ufkes, A., and Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [138] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [139] He, L., Lu, Y., Corring, J., Florencio, D., and Zhang, C. (2023). Diffusion-Based Document Layout Generation. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [140] He, S., Levine, D., Vrkic, I., Bressana, M. F., Zhang, D., Rizvi, S. A., Zhang, Y., Zappala, E., and van Dijk, D. (2024a). CaLMFlow: Volterra Flow Matching using Causal Language Models. *arXiv preprint arXiv:2410.05292*.
- [141] He, S. and Schomaker, L. (2019). DeepOtsu: Document enhancement and binarization using iterative deep learning. *Pattern Recognition*, 91:379–390.
- [142] He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., and Shan, Y. (2024b). ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. In *Proceedings of the International Conference on Learning Representations*.
- [143] Hedjam, R. and Cheriet, M. (2013). Ground-truth estimation in multi-spectral representation space: Application to degraded document image binarization. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [144] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. (2023a). Prompt-to-Prompt Image Editing with Cross-Attention Control. In *Proceedings of the International Conference on Learning Representations*.
- [145] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. (2023b). Prompt-to-Prompt Image Editing with Cross-Attention Control. In *Proceedings of the International Conference on Learning Representations*.
- [146] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., and Choi, Y. (2021). CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [147] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*.

- [148] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. *Proceedings of the International Conference on Learning Representations*.
- [149] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.
- [150] Ho, J. and Salimans, T. (2021a). Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- [151] Ho, J. and Salimans, T. (2021b). Classifier-Free Diffusion Guidance. In *Advances in Neural Information Processing Systems Workshops*.
- [152] Ho, J., Salimans, T., Gritsenko, A. A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video Diffusion Models. In *Proceedings of the International Conference on Learning Representations Workshops*.
- [153] Hoffmann Barfod, G., Larsen, J. M., Lichtenberger, A., and Raja, R. (2015). Revealing text in a complexly rolled silver scroll from Jerash with computed tomography and advanced imaging software. *Scientific Reports*.
- [154] Howe, N. R. (2013). Document binarization with automatic parameter tuning. *International Journal on Document Analysis and Recognition*, 16:247–258.
- [155] Hu, Y., Lin, Y., Wang, W., Zhao, Y., Wei, Y., and Shi, H. (2023). Diffusion for Natural Image Matting. *arXiv preprint arXiv:2312.05915*.
- [156] Huang, D., Guo, J., Sun, S., Tian, H., Lin, J., Hu, Z., Lin, C.-Y., Lou, J.-G., and Zhang, D. (2023). A Survey for Graphic Design Intelligence. *arXiv preprint arXiv:2309.01371*.
- [157] Huang, M., Cai, J., Jia, S., Lokhande, V. S., and Lyu, S. (2024). ParallelEdits: Efficient Multi-object Image Editing. *arXiv preprint arXiv:2406.00985*.
- [158] Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F. (2022). Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the ACM International Conference on Multimedia*.

- [159] Huh, M., Cheung, B., Agrawal, P., and Isola, P. (2023). Straightening out the straight-through estimator: Overcoming optimization challenges in vector quantized networks. In *Proceedings of the International Conference on Machine Learning*.
- [160] Hui, M., Zhang, Z., Zhang, X., Xie, W., Wang, Y., and Lu, Y. (2023). Unifying Layout Generation with a Decoupled Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [161] Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., and Yamaguchi, K. (2023). LayoutDM: Discrete Diffusion Model for Controllable Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [162] Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 2017–2025.
- [163] Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General Perception with Iterative Attention. In *Proceedings of the International Conference on Machine Learning*.
- [164] Jain, A., Xie, A., and Abbeel, P. (2023). VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [165] Jaume, G., Ekenel, H. K., and Thiran, J.-P. (2019). Funsd: A dataset for form understanding in noisy scanned documents. In *Proceedings of the International Conference on Document Analysis and Recognition Workshops*.
- [166] Jemni, S. K., Souibgui, M. A., Kessentini, Y., and Fornés, A. (2022). Enhance to read better: A multi-task adversarial network for handwritten document image enhancement. *Pattern Recognition*, 123:108370.
- [167] Ji, S., Xu, W., Yang, M., and Yu, K. (2012). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [168] Jia, F., Shi, C., He, K., Wang, C., and Xiao, B. (2018). Degraded document image binarization using structural symmetry of strokes. *Pattern Recognition*, 74:225–240.

- [169] Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., and Wei, Y. (2021). Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30:5875–5888.
- [170] Jin, Z., Shen, X., Li, B., and Xue, X. (2023). Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. *Advances in Neural Information Processing Systems*.
- [171] Johansson, S., Leech, G. N., and Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computer*. Department of English, University of Oslo.
- [172] John, S. (2015). The Invisible Library.
- [173] Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Medical image analysis*.
- [174] Kang, L., Riba, P., Rusinol, M., Fornes, A., and Villegas, M. (2020a). Distilling Content from Style for Handwritten Word Recognition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [175] Kang, L., Riba, P., Rusinol, M., Fornes, A., and Villegas, M. (2021a). Content and Style Aware Generation of Text-line Images for Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.
- [176] Kang, L., Riba, P., Rusiñol, M., Fornés, A., and Villegas, M. (2022a). Pay attention to what you read: non-recurrent handwritten text-line recognition. *Pattern Recognition*.
- [177] Kang, L., Riba, P., Rusiñol, M., Fornés, A., and Villegas, M. (2022b). Pay attention to what you read: Non-recurrent handwritten text-line recognition. *Pattern Recognition*.
- [178] Kang, L., Riba, P., Wang, Y., Rusiñol, M., Fornés, A., and Villegas, M. (2020b). GANwriting: Content-Conditioned Generation of Styled Handwritten Word Images. In *Proceedings of the European Conference on Computer Vision*.

- [179] Kang, L., Tito, R., Valveny, E., and Karatzas, D. (2024). Multi-Page Document Visual Question Answering using Self-Attention Scoring Mechanism. *arXiv preprint arXiv:2404.19024*.
- [180] Kang, S., Iwana, B. K., and Uchida, S. (2021b). Complex image processing with less data—Document image binarization by integrating multiple pre-trained U-Net modules. *Pattern Recognition*, 109:107577.
- [181] Karazija, L., Laina, I., Vedaldi, A., and Rupprecht, C. (2023). Diffusion models for zero-shot open-vocabulary segmentation. *arXiv preprint arXiv:2306.09316*.
- [182] Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. *Advances in Neural Information Processing Systems*.
- [183] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [184] Katherine Crowson (2021). CLIP guided diffusion HQ 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj. Accessed: 2026-02-13.
- [185] Kennedy, C. J. and Wess, T. J. (2006). The use of X-ray scattering to analyse parchment structure and degradation. In *Physical techniques in the study of art, archaeology and cultural heritage*. Elsevier.
- [186] Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. (2023). Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [187] Khrulkov, V. and Oseledets, I. (2018). Geometry Score: A Method For Comparing Generative Adversarial Networks. In *Proceedings of the International Conference on Machine Learning*.
- [188] Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. (2022). Ocr-free document understanding transformer. In *Proceedings of the European Conference on Computer Vision*.

- [189] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [190] Kleber, F., Fiel, S., Diem, M., and Sablatnig, R. (2013). CVL-DataBase: An Off-Line Database for Writer Retrieval, Writer Identification and Word Spotting. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [191] Koloda, J. and Wang, J. (2023). Context Aware Document Binarization and Its Application to Information Extraction from Structured Documents. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [192] Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [193] Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., and Hassner, T. (2023). TextStyleBrush: Transfer of Text Aesthetics from a Single Example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [194] Labs, B. F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., et al. (2025). FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.
- [195] Lai, W.-S., Huang, J.-B., Ahuja, N., and Yang, M.-H. (2018). Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [196] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*.
- [197] Lee, K., Joshi, M., Turc, I. R., Hu, H., Liu, F., Eisenschlos, J. M., Khandelwal, U., Shaw, P., Chang, M.-W., and Toutanova, K. (2023a). Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the International Conference on Machine Learning*.

- [198] Lee, Y., Kim, K., Kim, H., and Sung, M. (2023b). Syncdiffusion: Coherent montage via Synchronized Joint Diffusions. In *Advances in Neural Information Processing Systems*.
- [199] Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- [200] Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., and Heard, J. (2006). Building a test collection for complex document information processing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- [201] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [202] Li, D.-W., Huang, D., Ma, T., and Lin, C.-Y. (2021). Towards topic-aware slide generation for academic papers with unsupervised mutual learning. In *Proceedings of the Conference on Artificial Intelligence*.
- [203] Li, F., Liu, A., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z., et al. (2023a). Relation-aware diffusion model for controllable poster layout generation. In *CIKM*.
- [204] Li, J., Xu, Y., Lv, T., Cui, L., Zhang, C., and Wei, F. (2022). Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the ACM International Conference on Multimedia*.
- [205] Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., and Wei, F. (2023b). TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the Conference on Artificial Intelligence*.
- [206] Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., and Zhou, M. (2020). DocBank: A benchmark dataset for document layout analysis. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [207] Li, T., Tian, Y., Li, H., Deng, M., and He, K. (2024a). Autoregressive Image Generation without Vector Quantization. In *Advances in Neural Information Processing Systems*.

- [208] Li, X., Qiu, K., Chen, H., Kuen, J., Lin, Z., Singh, R., and Raj, B. (2024b). ControlVAR: Exploring Controllable Visual Autoregressive Modeling. *arXiv preprint arXiv:2406.09750*.
- [209] Li, Y., Chan, K., Sun, Y., Lam, C., Tong, T., Yu, Z., Fu, K., Liu, X., and Tan, T. (2025a). MoEdit: On Learning Quantity Perception for Multi-object Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2683–2693.
- [210] Li, Y., Chen, D., Tang, T., and Shen, X. (2025b). HTR-VT: Handwritten Text Recognition with Vision Transformer. *Pattern Recognition*.
- [211] Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023c). Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [212] Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., and Lin, T.-Y. (2023). Magic3D: High-Resolution Text-to-3D Content Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [213] Lin, Y. and Seales, W. B. (2005). Opaque document imaging: Building images of inaccessible texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [214] Lin, Y.-S., Ju, R.-Y., Chen, C.-C., Lin, T.-Y., and Chiang, J.-S. (2022). Three-stage binarization of color document images based on discrete wavelet transform and generative adversarial networks. *arXiv preprint arXiv:2211.16098*.
- [215] Lin, Z. and Wan, L. (2007). Style-preserving English handwriting synthesis. *Pattern Recognition*, pages 2097–2109.
- [216] Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S., and Kluger, Y. (2019). Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods*, 16(3):243–245.
- [217] Lins, R. (2011). Nabuco—Two Decades of Processing Historical Documents in Latin America. *J. Univers. Comput. Sci*, 17:151–161.
- [218] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow Matching for Generative Modeling. In *Proceedings of the International Conference on Learning Representations*.

- [219] Liu, C., Wei, H., Chen, J., Kong, L., Ge, Z., Zhu, Z., Zhao, L., Sun, J., Han, C., and Zhang, X. (2024a). Focus Anywhere for Fine-grained Multi-page Document Understanding. *arXiv preprint arXiv:2405.14295*.
- [220] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023a). Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*.
- [221] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2024b). Visual Instruction Tuning. *Advances in Neural Information Processing Systems*.
- [222] Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., et al. (2025). Step1X-Edit: A Practical Framework for General Image Editing. *arXiv preprint arXiv:2504.17761*.
- [223] Liu, X., Gong, C., et al. (2023b). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *Proceedings of the International Conference on Learning Representations*.
- [224] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [225] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *Proceedings of the International Conference on Learning Representations*.
- [226] Loshchilov, I. and Hutter, F. (2019). Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*.
- [227] Lu, H., Kot, A. C., and Shi, Y. Q. (2004). Distance-reciprocal distortion measure for binary document images. *IEEE Signal Process. Lett.*, 11(2):228–231.
- [228] Lu, S., Su, B., and Tan, C. L. (2010). Document image binarization using background estimation and stroke edges. *International Journal on Document Analysis and Recognition*, 13:303–314.
- [229] Lüddecke, T. and Ecker, A. (2022). Image Segmentation Using Text and Image Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [230] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [231] Luhman, T. and Luhman, E. (2020). Diffusion Models for Handwriting Generation. *arXiv preprint arXiv:2011.06704*.
- [232] Luo, C., Zhu, Y., Jin, L., Li, Z., and Peng, D. (2022). SLOGAN: Handwriting Style Synthesis for Arbitrary-Length and Out-of-Vocabulary Text. *IEEE Trans. Neural Netw. Learn. Syst.*
- [233] Luo, S., Tan, Y., Huang, L., Li, J., and Zhao, H. (2023). Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference.
- [234] Lv, T., Huang, Y., Chen, J., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., Dong, L., Luo, W., et al. (2023). Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*.
- [235] Ma, C., Liu, W., Tang, H., Li, J., and Yuan, Y. (2025). Calligrapher: Scalable and Diverse Text Rendering for Document Images. *arXiv preprint arXiv:2501.03456*.
- [236] Maarand, M., Beyer, Y., Kåsen, A., Fosseide, K. T., and Kermorvant, C. (2022). A Comprehensive Comparison of Open-Source Libraries for Handwritten Text Recognition in Norwegian. In *IAPR International Workshop on Document Analysis Systems*.
- [237] Mahmoud, S. A., Ahmad, I., Al-Khatib, W. G., Alshayeb, M., Parvez, M. T., Märgner, V., and Fink, G. A. (2014). KHATT: An open Arabic offline handwritten text database. *Pattern Recognition*, 47:1096–1112.
- [238] Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693.
- [239] Marti, U.-V. and Bunke, H. (2002). The IAM-database: an English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46.

- [240] Mathew, M., Karatzas, D., and Jawahar, C. (2021). Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [241] Matsuda, H., Togo, R., Maeda, K., Ogawa, T., and Haseyama, M. (2024). Multi-object editing in personalized text-to-image diffusion model via segmentation guidance. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 8140–8144. IEEE.
- [242] Mattick, A., Mayr, M., Seuret, M., Maier, A., and Christlein, V. (2021). SmartPatch: Improving Handwritten Word Imitation with Patch Discriminators. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [243] Mayr, M., Dreier, M., Kordon, F., Seuret, M., Zöllner, J., Wu, F., Maier, A., and Christlein, V. (2024). Zero-Shot Paragraph-level Handwriting Imitation with Latent Diffusion Models. *arXiv preprint arXiv:2409.00786*.
- [244] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. (2022). SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *Proceedings of the International Conference on Learning Representations*.
- [245] Mentzer, F., Minnen, D., Agustsson, E., and Tschannen, M. (2024). Finite Scalar Quantization: VQ-VAE Made Simple. In *Proceedings of the International Conference on Learning Representations*.
- [246] Messaoud, I. B., Amiri, H., El Abed, H., and Märgner, V. (2012). Region based local binarization approach for handwritten ancient documents. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [247] Michael, J., Labahn, R., Grüning, T., and Zöllner, J. (2019). Evaluating sequence-to-sequence models for handwritten text recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 1286–1293. IEEE.
- [248] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision*.

- [249] Moghaddam, R. F., Moghaddam, F. F., and Cheriet, M. (2013). Unsupervised ensemble of experts (EoE) framework for automatic binarization of document images. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [250] Mokady, R., Hertz, A., Aberman, K., Pritch, Y., and Cohen-Or, D. (2023). Null-text Inversion for Editing Real Images using Guided Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [251] Moysset, B., Kermorvant, C., and Wolf, C. (2017). Full-page text recognition: Learning where to start and when to stop. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [252] Moysset, B. and Messina, R. (2019). Are 2D-LSTM really dead for offline text recognition? *International Journal on Document Analysis and Recognition*, 22(3):193–208.
- [253] Mridha, M. F., Ohi, A. Q., Ali, M. A., Emon, M. I., and Kabir, M. M. (2021). BanglaWriting: A multi-purpose offline Bangla handwriting dataset. *Data in Brief*, 34:106633.
- [254] Mu, N., Darrell, T., and Wang, X. (2025). EditAR: Controllable Image Editing using Autoregressive Models. *arXiv preprint arXiv:2501.00256*.
- [255] Mun, S., Nam, J., Cho, S., and Ok, J. (2025). Addressing Text Embedding Leakage in Diffusion-based Image Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [256] Nafchi, H. Z., Ayatollahi, S. M., Moghaddam, R. F., and Cheriet, M. (2013). An efficient ground truthing tool for binarization of historical manuscripts. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [257] Nafchi, H. Z., Moghaddam, R. F., and Cheriet, M. (2014). Phase-based binarization of ancient document images: Model and applications. *IEEE Transactions on Image Processing*, 23(7):2916–2930.
- [258] Nair, N. G., Valanarasu, J. M. J., and Patel, V. M. (2024). Diffscaler: Enhancing the Generative Prowess of Diffusion Transformers. *arXiv preprint arXiv:2404.09976*.

- [259] Nguyen, Q., Vu, T., Tran, A., and Nguyen, K. (2024). Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*.
- [260] Nichol, A. Q. and Dhariwal, P. (2021). Improved Denoising Diffusion Probabilistic Models. In *Proceedings of the International Conference on Machine Learning*.
- [261] Nikolaidou, K., Retsinas, G., Christlein, V., Seuret, M., Sfikas, G., Smith, E. B., Mokayed, H., and Liwicki, M. (2023). WordStylist: Styled Verbatim Handwritten Text Generation with Latent Diffusion Models. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [262] Nikolaidou, K., Retsinas, G., Sfikas, G., and Liwicki, M. (2024a). DiffusionPen: Towards Controlling the Style of Handwritten Text Generation. *Proceedings of the European Conference on Computer Vision*.
- [263] Nikolaidou, K., Retsinas, G., Sfikas, G., and Liwicki, M. (2024b). Rethinking HTG Evaluation: Bridging Generation and Recognition. *Proceedings of the European Conference on Computer Vision Workshops*.
- [264] Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2012). Performance evaluation methodology for historical document image binarization. *IEEE Transactions on Image Processing*, 22(2):595–609.
- [265] Ntirogiannis, K., Gatos, B., and Pratikakis, I. (2014). ICFHR2014 competition on handwritten document image binarization (H-DIBCO 2014). In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [266] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*.
- [267] Panariello, A., Porrello, A., Calderara, S., and Cucchiara, R. (2022). Consistency-based Self-supervised Learning for Temporal Anomaly Localization. In *Proceedings of the European Conference on Computer Vision Workshops*.
- [268] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

- [269] Parker, C. S., Parsons, S., Bandy, J., Chapman, C., Coppens, F., and Seales, W. B. (2019). From invisibility to readability: recovering the ink of Herculaneum. *PloS one*.
- [270] Parsons, S., Parker, C. S., Chapman, C., Hayashida, M., and Seales, W. B. (2023). EduceLab-Scrolls: Verifiable Recovery of Text from Herculaneum Papyri using X-ray CT. *arXiv preprint arXiv:2304.02084*.
- [271] Peebles, W. and Xie, S. (2023). Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [272] Peng, D., Liu, C., Jin, L., Guo, J., and Liu, Y. (2024). ViTEraser: Harnessing the Power of Vision Transformers for Scene Text Removal with SegMIM Pretraining. *arXiv preprint arXiv:2306.12106*.
- [273] Pérez, D., Tarazón, L., Serrano, N., Castro, F., Ramos Terrades, O., and Juan, A. (2010). The GERMANA Database. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 19–21.
- [274] Pham, V., Bluche, T., Kermorvant, C., and Louradour, J. (2014). Dropout improves recurrent neural networks for handwriting recognition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 285–290. IEEE.
- [275] Phunyaphibarn, P., Yaman, B., and Yildirim, S. (2025). Unconditional Priors Matter: Improving Diffusion Models with Joint Guidance. *arXiv preprint arXiv:2501.12345*.
- [276] Pilu, M. (2001). Undoing paper curl distortion using applicable surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [277] Pippi, V., Cascianelli, S., Baraldi, L., and Cucchiara, R. (2023a). Evaluating synthetic pre-Training for handwriting processing tasks. *Pattern Recognition Letters*, 172:44–50.
- [278] Pippi, V., Cascianelli, S., and Cucchiara, R. (2023b). Handwritten Text Generation from Visual Archetypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [279] Pippi, V., Cascianelli, S., Kermorvant, C., and Cucchiara, R. (2023c). How to Choose Pretrained Handwriting Recognition Models for Single Writer Fine-Tuning. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [280] Pippi, V., Cascianelli, S., Kermorvant, C., and Cucchiara, R. (2023d). How to Choose Pretrained Handwriting Recognition Models for Single Writer Fine-Tuning. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [281] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. (2024). SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *Proceedings of the International Conference on Learning Representations*.
- [282] Poličar, P. G., Stražar, M., and Zupan, B. (2019). openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv*.
- [283] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. (2021). Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *Proceedings of the International Conference on Machine Learning*.
- [284] Potthast, A. (1874). *Regesta pontificum romanorum inde a 1198 ad 1304*, volume 1. De Decker.
- [285] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2010). H-DIBCO 2010-handwritten document image binarization competition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [286] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2011). ICDAR 2011 Document Image Binarization Contest (DIBCO 2011). In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [287] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2012). ICFHR 2012 competition on handwritten document image binarization (H-DIBCO 2012). In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [288] Pratikakis, I., Gatos, B., and Ntirogiannis, K. (2013). ICDAR 2013 document image binarization contest (DIBCO 2013). In *Proceedings of the International Conference on Document Analysis and Recognition*.

- [289] Pratikakis, I., Zagori, K., Kaddas, P., and Gatos, B. (2018). ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [290] Pratikakis, I., Zagoris, K., Barlas, G., and Gatos, B. (2016). ICFHR2016 handwritten document image binarization contest (H-DIBCO 2016). In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [291] Pratikakis, I., Zagoris, K., Barlas, G., and Gatos, B. (2017). ICDAR2017 competition on document image binarization (DIBCO 2017). In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [292] Pratikakis, I., Zagoris, K., Karagiannis, X., Tsochatzidis, L., Mondal, T., and Marthot-Santaniello, I. (2019). ICDAR 2019 Competition on Document Image Binarization (DIBCO 2019). In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [293] Puigcerver, J. (2017). Are multidimensional recurrent layers really necessary for handwritten text recognition? In *Proceedings of the International Conference on Document Analysis and Recognition*, volume 1, pages 67–72. IEEE.
- [294] Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., and Guibas, L. J. (2016). Volumetric and Multi-view CNNs for Object Classification on 3D Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [295] Qin, X., Dai, H., Hu, X., Fan, D. P., Shao, L., and Van Gool, L. (2022). Highly Accurate Dichotomous Image Segmentation. In *Proceedings of the European Conference on Computer Vision*.
- [296] Qureshi, R., Uzair, M., Khurshid, K., and Yan, H. (2019). Hyperspectral document image processing: Applications, challenges and future prospects. *Pattern Recognition*, 90:12–22.
- [297] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*.

- [298] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, page 9.
- [299] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- [300] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv e-prints*, pages arXiv–2204.
- [301] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*.
- [302] Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., and Chechik, G. (2023). Linguistic Binding in Diffusion Models: Enhancing Attribute Correspondence through Attention Map Alignment. In *Advances in Neural Information Processing Systems*.
- [303] Ray, A., Sharma, M., Upadhyay, A., Makwana, M., Chaudhury, S., Trivedi, A., Singh, A., and Saini, A. (2019). An end-to-end trainable framework for joint optimization of document enhancement and recognition. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [304] Razavi, A., Van den Oord, A., and Vinyals, O. (2019). Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*.
- [305] Ren, M.-S., Zhang, Y.-M., Wang, Q.-F., Yin, F., and Liu, C.-L. (2023). Diff-Writer: A Diffusion Model-Based Stylized Online Handwritten Chinese Character Generator. In *Proceedings of the International Conference on Neural Information Processing*.
- [306] Rezanezhad, V., Baierer, K., and Neudecker, C. (2023). A hybrid CNN-Transformer model for Historical Document Image Binarization. In *HIP*.
- [307] Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., and Rott, P. (2009). A perceptually motivated online benchmark for image matting.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [308] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [309] Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., Vidal, E., and Lladós, J. (2013). The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting recognition. *Pattern Recognition*, 46(6):1658–1669.
- [310] Ronneberger, O., Fischer, P., and Brox, T. (2015a). U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 234–241.
- [311] Ronneberger, O., Fischer, P., and Brox, T. (2015b). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*.
- [312] Rosin, P. L., Lai, Y.-K., Liu, C., Davis, G. R., Mills, D., Tuson, G., and Russell, Y. (2018). Virtual recovery of content from X-ray micro-tomography scans of damaged historic scrolls. *Scientific Reports*, 8(1):11901.
- [313] Rother, C., Kolmogorov, V., and Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graphics*, 23(3):309–314.
- [314] Rowley-Brooke, R., Pitié, F., and Kokaram, A. (2012). A ground truth bleed-through document image database. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*.
- [315] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. (2022). Image Super-Resolution via Iterative Refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726.
- [316] Samko, O., Lai, Y.-K., Marshall, D., and Rosin, P. L. (2014). Virtual unrolling and information recovery from scanned scrolled historical documents. *Pattern Recognition*.

- [317] Sánchez, J. A., Romero, V., Toselli, A. H., and Vidal, E. (2014). ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [318] Sanchez, J. A., Romero, V., Toselli, A. H., and Vidal, E. (2016). ICFHR2016 competition on handwritten text recognition on the READ dataset. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 630–635. IEEE.
- [319] Sarukkai, V., Li, L., Ma, A., Ré, C., and Fatahalian, K. (2024). Collage Diffusion. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [320] Sauvola, J., Seppanen, T., Haapakoski, S., and Pietikainen, M. (1997). Adaptive document binarization. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [321] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*.
- [322] Seales, B. and Delattre, D. (2013). *Virtual unrolling of carbonized Herculanum scrolls: Research Status (2007-2012)*. Macchiaroli editore.
- [323] Seales, W. (2017). Reading the invisible library: A retrospective. *Modern Alchemy: New Technology for Museum Collections*. Gilcrease Museum.
- [324] Seales, W. B. and Lin, Y. (2004). Digital restoration using volumetric scanning. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*.
- [325] Seales, W. B., Parker, C. S., Segal, M., Tov, E., Shor, P., and Porath, Y. (2016). From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi. *Science Advances*.
- [326] Serrano, N., Castro, F., and Juan, A. (2010). The RODRIGO Database. In *Proceedings of the International Conference on Language Resources and Evaluation*.

- [327] Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations*.
- [328] Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., and Li, W. (2021). LayoutParser: A unified toolkit for deep learning based document image analysis. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 131–146. Springer.
- [329] Shi, B., Bai, X., and Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.
- [330] Simsar, E., Tonioni, A., Xian, Y., Hofmann, T., and Tombari, F. (2025). LIME: Localized Image Editing via Attention Regularization in Diffusion Models. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 222–231. IEEE.
- [331] Sinha, A. K., Moorthi, S. M., and Dhar, D. (2022). NL-FFC: Non-Local Fast Fourier Convolution for Image Super Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [332] Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.
- [333] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*.
- [334] Song, J., Meng, C., and Ermon, S. (2021a). Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations*.
- [335] Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. (2023). Consistency Models. *Proceedings of the International Conference on Machine Learning*.
- [336] Song, Y. and Ermon, S. (2019). Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*.

- [337] Song, Y. and Ermon, S. (2020). Improved Techniques for Training Score-Based Generative Models. *Advances in Neural Information Processing Systems*.
- [338] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021b). Score-Based Generative Modeling through Stochastic Differential Equations. In *Proceedings of the International Conference on Learning Representations*.
- [339] Souibgui, M. A., Biswas, S., Jemni, S. K., Kessentini, Y., Fornés, A., Lladós, J., and Pal, U. (2022). DocEnTr: an end-to-end document image enhancement transformer. In *Proceedings of the International Conference on Pattern Recognition*.
- [340] Souibgui, M. A., Biswas, S., Mafla, A., Biten, A. F., Fornés, A., Kessentini, Y., Lladós, J., Gomez, L., and Karatzas, D. (2023). Text-DIAE: Degradation Invariant Autoencoders for Text Recognition and Document Enhancement. *Proceedings of the Conference on Artificial Intelligence*.
- [341] Souibgui, M. A. and Kessentini, Y. (2020). DE-GAN: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1180–1191.
- [342] Souibgui, M. A., Kessentini, Y., and Fornés, A. (2021). A conditional gan based approach for distorted camera captured documents recovery. In *Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence*.
- [343] Stauffer, M., Fischer, A., and Riesen, K. (2018). Keyword spotting in historical handwritten documents based on graph matching. *Pattern Recognition*.
- [344] Strauß, T., Leifert, G., Labahn, R., Hodel, T., and Mühlberger, G. (2018). ICFHR2018 competition on automated text recognition on a READ dataset. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 477–482. IEEE.
- [345] Stromer, D., Christlein, V., Huang, X., Zippert, P., Hausotte, T., and Maier, A. (2019). Virtual cleaning and unwrapping of non-invasively digitized soiled bamboo scrolls. *Scientific Reports*, 9(1):2311.

- [346] Stromer, D., Christlein, V., Maier, A., Zippert, P., Helmecke, E., Hausotte, T., and Huang, X. (2018a). Non-destructive digitization of soiled historical Chinese bamboo scrolls. In *IAPR International Workshop on Document Analysis Systems*.
- [347] Stromer, D., Christlein, V., Martindale, C., Zippert, P., Haltenberger, E., Hausotte, T., and Maier, A. (2018b). Browsing through sealed historical manuscripts by using 3-D computed tomography with low-brilliance X-ray sources. *Scientific Reports*, 8:15335.
- [348] Stromer, D., Christlein, V., Schön, T., Holub, W., and Maier, A. (2017). Browsing Through Closed Books: Evaluation of Preprocessing Methods for Page Extraction of a 3-D CT Book Volume. In *IOP Conference Series: Materials Science and Engineering*.
- [349] Such, F. P., Peri, D., Brockler, F., Paul, H., and Ptucha, R. (2018). Fully convolutional networks for handwriting recognition. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*.
- [350] Sueiras, J., Ruiz, V., Sanchez, A., and Velez, J. F. (2018). Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing*, 289:119–128.
- [351] Suh, S., Kim, J., Lukowicz, P., and Lee, Y. O. (2020). Two-stage generative adversarial networks for document image binarization with color noise and background removal. *arXiv preprint arXiv:2010.10103*.
- [352] Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., and Yuan, Z. (2024a). Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*.
- [353] Sun, Y., Tang, C.-K., and Tai, Y.-W. (2021). Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [354] Sun, Y., Zhou, D., Lin, C., He, C., Ouyang, W., and Zhong, H.-S. (2024b). LOCR: Location-Guided Transformer for Optical Character Recognition. *arXiv preprint arXiv:2403.02127*.
- [355] Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., and Lempitsky, V. (2022).

- Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [356] Tamrin, M. O., El-Amine Ech-Cherif, M., and Cheriet, M. (2021). A two-stage unsupervised deep learning framework for degradation removal in ancient documents. In *Pattern Recognition*, pages 292–303.
- [357] Tan, Z., Liu, S., Yang, X., Xue, Q., and Wang, X. (2025). OminiControl: Minimal and Universal Control for Diffusion Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14940–14950.
- [358] Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., and Ture, F. (2023a). What the daam: Interpreting stable diffusion using cross attention. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 5644–5659.
- [359] Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., and Bansal, M. (2023b). Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [360] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.
- [361] Team, G. D. (2025). Next-Step Prediction for Continuous Image Generation. In *Proceedings of the International Conference on Learning Representations*.
- [362] Tensmeyer, C. and Martinez, T. (2017). Document image binarization with fully convolutional neural networks. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [363] Tensmeyer, C. and Martinez, T. (2020). Historical document image binarization: A review. *SN comput. sci.*, 1(3):173.
- [364] Thomas, A. O., Rusu, A., and Govindaraju, V. (2009). Synthetic Handwritten CAPTCHAs. *Pattern Recognition*, pages 3365–3373.
- [365] Tian, K., Jiang, Y., Yuan, Z., Jia, B., and Chen, B. (2024). Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. In *Advances in Neural Information Processing Systems*.

- [366] Tito, R., Karatzas, D., and Valveny, E. (2023). Hierarchical multimodal transformers for Multipage DocVQA. *Pattern Recognition*, 144:109834.
- [367] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., and Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [368] Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, pages 1–34.
- [369] Toselli, A. H., Juan, A., González, J., Salvador, I., Vidal, E., Casacuberta, F., Keysers, D., and Ney, H. (2004). Integrated handwriting recognition and interpretation using finite-state models. *Int. J. Pattern Recognit. Artif. Intell.*, 18(04):519–539.
- [370] Toselli, A. H. and Vidal, E. (2015). Handwritten text recognition results on the Bentham collection with improved classical n-gram-HMM methods. In *HIP*, pages 15–22.
- [371] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [372] Tschannen, M., Eastwood, C., and Mentzer, F. (2025). Givt: Generative infinite-vocabulary transformers. In *Proceedings of the European Conference on Computer Vision*.
- [373] Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. (2023). Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [374] Turski, M., Stanisławek, T., Kaczmarek, K., Dyda, P., and Graliński, F. (2023). Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [375] Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30.

- [376] Van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11).
- [377] Van Landeghem, J., Biswas, S., Blaschko, M., and Moens, M.-F. (2024). Beyond document page classification: Design, datasets, and challenges. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.
- [378] Van Landeghem, J., Tito, R., Borchmann, L., Pietruszka, M., Joziak, P., Powalski, R., Jurkiewicz, D., Coustaty, M., Anckaert, B., Valveny, E., et al. (2023). Document understanding dataset and evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [379] Vanherle, B., Pippi, V., Cascianelli, S., Michiels, N., Van Reeth, F., and Cucchiara, R. (2024). VATr++: Choose Your Words Wisely for Handwritten Text Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [380] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.
- [381] Vo, Q. N., Kim, S. H., Yang, H. J., and Lee, G. (2018). Binarization of degraded document images based on hierarchical deep supervised network. *Pattern Recognition*, 74:568–586.
- [382] Voigtlaender, P., Doetsch, P., and Ney, H. (2016). Handwriting recognition with large multidimensional long short-term memory recurrent neural networks. In *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pages 228–233. IEEE.
- [383] Von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., and Wolf, T. (2022). Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.
- [384] Wang, J., Wu, C., Xu, Y.-Q., and Shum, H.-Y. (2005). Combining Shape and Physical Models for On-line Cursive Handwriting Synthesis. *International Journal on Document Analysis and Recognition*, 7(4):219–227.
- [385] Wang, P., Li, Y., and Vasconcelos, N. (2021). Rethinking and Improving the Robustness of Image Style Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

- [386] Wang, P., Shi, Y., Lian, X., Zhai, Z., Xia, X., Xiao, X., Huang, W., and Yang, J. (2025). SeedEdit 3.0: Fast and High-Quality Generative Image Editing. *arXiv preprint arXiv:2506.05083*.
- [387] Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., and Shen, C. (2023). Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. *arXiv preprint arXiv:2303.17599*.
- [388] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, pages 600–612.
- [389] Weng, H., Huang, D., Qiao, Y., Hu, Z., Lin, C.-Y., Zhang, T., and Chen, C. (2024). Design: A Pipeline for Controllable Design Template Generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [390] Wick, C., Zöllner, J., and Grüning, T. (2021). Rescoring Sequence-to-Sequence Models for Text Line Recognition with CTC-Prefixes. *arXiv preprint arXiv:2110.05909*.
- [391] Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., and Cohen, S. (2018). Start, Follow, Read: End-to-End Full-Page Handwriting Recognition. In *Proceedings of the European Conference on Computer Vision*, pages 367–383.
- [392] Wolf, C., Jolion, J.-M., and Chassaing, F. (2002). Text localization, enhancement and binarization in multimedia documents. In *Proceedings of the International Conference on Pattern Recognition*.
- [393] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- [394] Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.-m., Bai, S., Xu, X., Chen, Y., et al. (2025). Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*.

- [395] Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. (2023a). Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [396] Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., and Chang, S. (2023b). Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [397] Wu, X., Sun, K., Zhu, F., Zhao, R., and Li, H. (2023c). Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [398] Xiang, W., Yang, H., Huang, D., and Wang, Y. (2023). Denoising Diffusion Autoencoders are Unified Self-supervised Learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [399] Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., and Shou, M. Z. (2023). BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461.
- [400] Xing, X., Zhou, H., Wang, C., Zhang, J., Xu, D., and Yu, Q. (2023a). SVGDreamer: Text Guided SVG Generation with Diffusion Model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [401] Xiong, W., Jia, X., Xu, J., Xiong, Z., Liu, M., and Wang, J. (2018). Historical document image binarization using background estimation and energy minimization. In *Proceedings of the International Conference on Pattern Recognition*.
- [402] Xiong, W., Zhou, L., Yue, L., Li, L., and Wang, S. (2021). An enhanced binarization framework for degraded historical document images. *EURASIP J Image Video Process*, 2021(1):13.
- [403] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., and Dong, Y. (2023a). ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. *Advances in Neural Information Processing Systems*.

- [404] Xu, J., Wang, X., Cheng, W., Cao, Y.-P., Shan, Y., Qie, X., and Gao, S. (2023b). Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [405] Xu, N., Price, B., Cohen, S., and Huang, T. (2017). Deep image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [406] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020a). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- [407] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al. (2020b). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [408] Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.
- [409] Yamaguchi, K. (2021). CanvasVAE: Learning to Generate Vector Graphic Documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE.
- [410] Yang, M. and Xu, S. (2023). A Novel Degraded Document Binarization Model through Vision Transformer Network. *Information Fusion*, 93:159–173.
- [411] Yang, Z., Ding, G., Wang, W., Chen, H., Zhuang, B., and Shen, C. (2024a). Object-Aware Inversion and Reassembly for Image Editing. In *Proceedings of the International Conference on Learning Representations*.
- [412] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., and Wang, L. (2023a). The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421*.

- [413] Yang, Z., Liu, B., Xiong, Y., Yi, L., Wu, G., Tang, X., Liu, Z., Zhou, J., and Zhang, X. (2023b). DocDiff: Document enhancement via residual diffusion models. In *Proceedings of the ACM International Conference on Multimedia*.
- [414] Yang, Z., Xiong, Y., and Wu, G. (2024b). GDB: Gated convolutions-based Document Binarization. *Pattern Recognition*.
- [415] Yao, J., Wang, X., Yang, S., and Wang, B. (2024a). ViTMatte: Boosting image matting with pre-trained plain vision transformers. *Inform. Fusion*, 103:102091.
- [416] Yao, J., Wang, X., Ye, L., and Liu, W. (2024b). Matte anything: Interactive natural image matting with segment anything model. *Image and Vision Computing*, 147:105067.
- [417] Yousef, M. and Bishop, T. E. (2020). OrigamiNet: Weakly-Supervised, Segmentation-Free, One-Step, Full Page Text Recognition by learning to unfold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [418] Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. (2022). Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*.
- [419] Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. (2025). Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *Proceedings of the International Conference on Learning Representations*.
- [420] Zhang, D., Huang, F., Liu, S., Wang, X., and Jin, Z. (2022). SwinFIR: Revisiting the SWINIR with fast Fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*.
- [421] Zhang, H., Duan, Z., Wang, X., Chen, Y., and Zhang, Y. (2025). EliGen: Entity-Level Controlled Image Generation with Regional Attention. *arXiv preprint arXiv:2501.01097*.
- [422] Zhang, L. and Agrawala, M. (2024). Transparent image layer diffusion using latent transparency. *arXiv preprint arXiv:2402.17113*.

- [423] Zhang, L., Rao, A., and Agrawala, M. (2023a). Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [424] Zhang, Q., Song, J., Huang, X., Chen, Y., and Liu, M.-Y. (2023b). Diff-Collage: Parallel Generation of Large Content With Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [425] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [426] Zhang, X., He, C., and Guo, J. (2020). Selective diffusion involving reaction for binarization of bleed-through document images. *Appl. Math. Comput.*, 81:844–854.
- [427] Zhang, Y., Nie, S., Liu, W., Xu, X., Zhang, D., and Shen, H. T. (2019). Sequence-to-sequence domain adaptation network for robust text image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [428] Zhang, Z., Tan, C. L., and Fan, L. (2004). Restoration of curved document images through 3D shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [429] Zhao, G., Liu, J., Jiang, J., Guan, H., and Wen, J.-R. (2018). Skip-connected deep convolutional autoencoder for restoration of document images. In *Proceedings of the International Conference on Pattern Recognition*.
- [430] Zhao, J., Shi, C., Jia, F., Wang, Y., and Xiao, B. (2019). Document image binarization with cascaded generators of conditional generative adversarial networks. *Pattern Recognition*, 96:106968.
- [431] Zheng, Q., Le, M., Shaul, N., Lipman, Y., Grover, A., and Chen, R. T. (2023). Guided flows for generative modeling and decision making. *arXiv preprint arXiv:2311.13443*.
- [432] Zheng, X., Xiaotian, Q., Ying, C., and Rynson, W. (2019). Content-aware Generative Modeling of Graphic Design Layouts. *ACM Trans. Graphics*, 38(4):133.

- [433] Zhong, X., Tang, J., and Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. In *Proceedings of the International Conference on Document Analysis and Recognition*.
- [434] Zhong, Z., Shen, T., Yang, Y., Lin, Z., and Zhang, C. (2018). Joint sub-bands learning with clique structures for wavelet domain super-resolution. *Advances in Neural Information Processing Systems*.
- [435] Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. (2025a). Transfusion: Predict the next token and diffuse images with one multi-modal model. In *Proceedings of the International Conference on Learning Representations*.
- [436] Zhou, D., Li, M., Yang, Z., and Yang, Y. (2025b). DreamRenderer: Taming multi-instance attribute control in large-scale text-to-image models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [437] Zhu, H., Liu, H., Fu, B., and Wang, Y. (2025a). MDE-Edit: Masked Dual-Editing for Multi-Object Image Editing via Diffusion Models. *arXiv preprint arXiv:2505.05101*.
- [438] Zhu, J., Zhou, Y., Wang, Z., Yao, J., Gu, Y., Yuan, Y., and Liu, S. (2025b). InfoDet: A Dataset for Infographic Element Detection. *arXiv preprint arXiv:2505.17473*.
- [439] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- [440] Zhu, Q., Shao, L., Li, X., and Wang, L. (2015). Targeting Accurate Object Extraction From an Image: A Comprehensive Study of Natural Image Matting. *IEEE Trans. Neural Netw. Learn. Syst.*, page 185.
- [441] Zhu, Y., Li, Z., Wang, T., He, M., and Yao, C. (2023). Conditional Text Image Generation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Glossary

- AdamW** Adam with Weight Decay, an extension of the Adam optimizer that decouples weight decay from the gradient update. 29, 48, 129
- AI** branch of computer science that develops machines and software with human-like intelligence. 173
- BLSTM** Bidirectional Long Short-Term Memory, a type of recurrent neural network that processes sequences in both forward and backward directions. 74, 82
- BN** Batch Normalization, a technique to standardize the inputs to a layer for each mini-batch. 24, 25
- CER** Character Error Rate, a common metric for evaluating the performance of text recognition systems. 81, 83, 84, 86, 261, 262
- CFG** Classifier-Free Guidance, a technique to control the generation process of diffusion or autoregressive models by mixing conditional and unconditional scores. 93, 120, 136, 139–141, 144–148, 150, 223
- CNN** convolutional neural network. 12, 14, 17, 23, 74, 82
- Computer Vision** scientific field that deals with how computers can gain high-level understanding from digital images or videos. 13, 14, 21, 22, 42, 119, 120
- CRNN** Convolutional Recurrent Neural Network, a neural network architecture combining convolutional and recurrent layers for sequence recognition tasks. 82, 85, 87, 88

- CTC** Connectionist Temporal Classification, a loss function used for training recurrent neural networks on sequence data without requiring pre-segmented alignments. 18, 74, 82, 83, 87, 88
- DDR** Digital Document Restoration, the process of recovering content from severely damaged documents using digital imaging and computer vision techniques. 13, 14, 40, 42, 54
- Deep Learning** a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. 17, 74
- DFT** Discrete Fourier Transform, a mathematical transform that decomposes a sequence of values into components of different frequencies. 14, 15
- DIBCO** Document Image Binarization Contest, a periodic competition providing benchmark datasets for document enhancement. 29, 30, 34
- Document Analysis** the process of automatically extracting information from document images, involving layout analysis, character recognition, and semantic understanding. 21–23, 42, 119, 120
- Document Parsing** the task of extracting a structured representation of a document in markup language from its image. 55, 56, 58
- DQA** Document Question Answering, the task of answering questions about a document given its image. 16, 58
- DRD** Distance Reciprocal Distortion, a measure of visual distortion in binary document images. 30–33, 35–37
- FFC** Fast Fourier Convolution, an operator that performs convolutions in the spectral domain to achieve a global receptive field. 6, 12–15, 21, 22, 24–28, 30–33, 36, 39, 40, 42–44, 54, 365

- FFT** Fast Fourier Transform, an algorithm that computes the discrete Fourier transform of a sequence, or its inverse. 14, 26, 43, 44, 365
- FID** Fréchet Inception Distance, a score for evaluating the quality of generative models by comparing distributions of features extracted by an Inception network. 79, 86, 96, 98, 104–109, 111–116, 127, 131, 143, 144, 212, 213, 233–236, 242–244, 246, 247, 249, 261, 263, 265–269
- FM** F-Measure, the harmonic mean of precision and recall. 30–33, 35–37
- GAN** Generative Adversarial Network, a class of machine learning frameworks where two neural networks contest with each other in a game. 12, 23, 34, 91, 92, 95, 114, 119, 128, 132, 144, 250
- HTG** Handwritten Text Generation, the task of producing realistic images of arbitrary handwritten text in a desired style. 3, 6, 7, 94–99, 103–109, 111–115, 117, 119–123, 126–128, 130, 133, 136, 142–144, 146, 223, 233, 234, 238, 240, 242–244, 250, 251
- HTR** Handwritten Text Recognition, the task of converting images of handwritten text into machine-encoded text. 2, 6, 17–19, 73–76, 81–83, 88, 104, 125, 129, 132, 137, 146, 222, 225
- HWD** Handwriting Distance, a tailored evaluation score for styled handwritten text generation. 79, 80, 86, 98, 100, 103–109, 111–117, 127, 130, 143–146, 212, 213, 233–236, 243, 262
- IAM** a large dataset of handwritten English text, widely used for training and evaluating handwriting recognition and generation systems. 18, 262, 263
- IRFFT** Inverse Real Fast Fourier Transform, used to convert frequency domain signals back into the spatial domain. 27
- ISOS-BTD** Irish Script on Screen Bleed-Through Database. 30, 35, 37, 38

- KID** Kernel Inception Distance, a metric that measures the dissimilarity between two probability distributions of images. 79, 265–269
- LCM** Latent Consistency Model, a diffusion model distilled for fast high-quality image generation. 154, 265, 266, 268, 269, 272–274, 276, 278, 285–287, 289, 290
- LDM** Latent Diffusion Model, a diffusion model that operates in the latent space of a powerful pretrained autoencoder. 92, 154, 265–268, 270, 272, 273, 275, 277, 279–284, 288, 290–293
- LPIS** Learned Perceptual Image Patch Similarity, a perceptual distance metric based on deep features. 243, 249, 251–255
- LSTM** long short-term memory. 17, 82
- MAD** Merge-Attend-Diffuse, a joint diffusion operator that merges latent features from multiple views, processes them with attention layers, and splits them back to enforce semantic coherence. 167, 168, 265–272, 277, 278, 291–293
- micro-CT** X-ray micro-computed tomography, a non-invasive imaging technique that generates 3D volumetric representations of objects. 13, 14, 41, 48
- MMDiT** Multimodal Diffusion Transformer, a transformer-based architecture that processes both text and image modalities using joint attention mechanisms. 196, 197, 199, 204, 205, 217
- OCR** Optical Character Recognition, the process of converting images of text into machine-encoded text. 16, 57, 59, 71, 146, 148, 150, 216, 261, 262
- pFM** pseudo-F-Measure, a metric for document binarization that incorporates a distance-based weight to emphasize character edges. 30–33, 35–37, 48, 51, 52

- PHIBD** Persian Heritage Image Binarization Dataset. 30, 35, 37, 38
- PSNR** Peak Signal-to-Noise Ratio, the ratio between the maximum possible power of a signal and the power of corrupting noise. 30–37, 48, 51, 52
- ReLU** Rectified Linear Unit, a common activation function used in deep neural networks. 24, 25
- RFFT** Real Fast Fourier Transform, a variant of the FFT optimized for real-valued inputs. 26
- RFM** Rectified Flow Matching, a generative modeling framework that learns to predict velocity fields over straight paths between data and noise distributions. 191–193
- RNN** recurrent neural network. 82
- SER** Sequence Error Rate, a metric for evaluating text recognition systems that measures the percentage of incorrectly recognized sequences or lines. 81–84, 86
- ST** Spectral Transform, a block within the FFC that processes features in the frequency domain. 26, 27, 30
- TConv** Transpose Convolution, an operator used to perform up-sampling of feature maps. 25
- VAE** Variational Autoencoder, a generative model that learns a latent representation of data by combining an encoder-decoder architecture with a probabilistic framework. 93, 119, 121–127, 129, 130, 133, 134, 136–140, 143, 150, 238, 249–251, 258, 259, 262
- vFFC** volumetric Fast Fourier Convolution, an extension of the FFC operator to handle 3D volumetric data using 3D convolutions and 3D discrete Fourier Transform. 15, 21, 41, 43–45, 49, 50, 52, 54

ViT Vision Transformer, a model that applies the transformer architecture to image processing by treating image patches as tokens. 12, 13, 18, 22, 34, 35

VRD Visually-Rich Document. 1–3, 5–8, 55, 221, 224

VrDU Visually-rich Document Understanding, an area of research that aims to analyze and extract information from document images by combining textual, visual, and spatial information. 11, 15, 16, 57–59, 62, 71

Acknowledgments

When a chapter as all-encompassing and important as a Ph.D. comes to a close, it is natural to take a step back and reflect on the journey.

This dissertation would not have been possible without the support and mentorship of many people, whose importance I cannot overstate. I'll try.

To my mother Claudia, my father Loris, my sister Giulia, and my grandmother Anna, who have been a constant source of love throughout my life. You have always encouraged me to pursue my dreams and have been there for me, unconditionally. I also want to mention my young nephews and the joy they have brought into my life: Andrea, Mattia, and Enea.

To my friends, who have had a profound effect on my life and whom I have always counted among my pillars. Thinking of the classic proverb "Tell me who your friends are, and I'll tell you who you are", I am happy if that holds true for me. Simone, Samuele, Nicholas, Eleonora, Davide, Luca, Daniele, Francesco, Marco, Andrea, Riccardo, Valeria, Giorgia, Giovanni, and many others. You have been a fundamental source of joy, and I am grateful that our paths have crossed.

I want to thank all the AImageLab people at my university. Among them, I would mention some whom I consider my friends, such as Vittorio, Carmine, Davide, Luca, Giuseppe, Nicholas, and many more; I'm sure that we will continue building our friendship in the future. I am likewise grateful to my tutor, the Rector of the University of Modena and Reggio Emilia, Rita Cucchiara, for the opportunities she has given me during these three years.

Doing a Ph.D. is a demanding journey, influenced by effort and luck. I was lucky enough to be mentored by Dr. Silvia Cascianelli, who gave me the freedom and support to follow my interests and passions. Always a source of guidance, she has shaped my thinking and approach to research. I know it's sometimes been challenging to mentor me, and I hope many more students in the future will have the same opportunity. I am also happy to be able to call you a friend.

Many asked me if I am happy about the choice to pursue a Ph.D., and my answer has always been “ask me in a few years.” While it’s early to tell, looking back I can confidently say that for now it’s looking good.