This is the peer reviewd version of the followng article:

Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas / Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.. - 15137:(2025), pp. 234-251. (18th European Conference on Computer Vision, ECCV 2024 ita SEP 29-OCT 04, 2024) [10.1007/978-3-031-72986-7_14].

Springer Science and Business Media Deutschland GmbH
Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

06/11/2026 23:55

(Article begins on next page)

arXiv:2408.15660v1 [cs.CV] 28 Aug 2024

Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas

Fabio Quattrini, Vittorio Pippi,
Silvia Cascianelli, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

Abstract. 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 **M**erge-**A**ttend-**D**iffuse operator, which can be plugged into different types of pretrained diffusion models used in a joint diffusion setting to improve the perceptual and semantical coherence of the generated panorama images. Specifically, we merge the diffusion paths, reprogramming self- and cross-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 at https://github.com/aimagelab/MAD.

Keywords: Image Generation · Diffusion Models · Text-to-Image

1 Introduction

Diffusion models [23, 48, 51, 53] have achieved impressive performance in various domains, including image generation [4,12,44,45,47], video generation [13,25,58], and audio synthesis [10, 29, 42], 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 [33], image editing [1, 3, 20], 3D modeling [31, 59], and video manipulation [28, 57].

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 [9,40], 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 superresolution techniques to a generated square image can lead to unsatisfactory results due to the artifacts that would be introduced, the lack of detail,



Fig. 1: MultiDiffusion [5] blends the mountains with the clouds, lacking semantic coherence (top). Applying the proposed MAD leads to semantic coherence (bottom).

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 pretrained models for generating images with unseen shapes and sizes.

Among those, *iterative diffusion* methods [1, 61] involve generating an initial image and then, with a series of outpainting steps, generating the whole panorama image. However, this strategy struggles to maintain global coherence and tends to repeat similar patterns [30]. Another family of approaches, referred to as *joint diffusion* methods [5, 30], entails considering the panorama image 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. 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 [30] 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. 1 are both perceptually coherent, but the first one, which is obtained with the State-of-the-Art MultiDiffusion [5] 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 [5,30], 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 [45], and a Latent Consistency Model (LCM) [34] and demonstrate its effectiveness via qualitative and quantitative evaluations, as well as a user study. The code for our approach is available at `https://github.com/aimagelab/MAD`.

2 Related Work

Diffusion models [12,23,38,48,51–53] are powerful generative probabilistic models that have achieved drastic performance improvements compared to previous methods, such as Generative Adversarial Networks [7, 17, 27], especially in the case of text-to-image generation, where the generation process is conditioned on a natural language prompt [9, 40, 44, 45, 47]. The introduction of LDMs [45] has even 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. Moreover, recent works have introduced strategies for fine-tuning diffusion models over additional guiding signals to gain better controllability [2,8,15,60] and for manipulating the diffusion process at inference-time to obtain zero-shot capabilities [1, 3, 11, 20, 33, 36, 37, 54]. Some of the proposed strategies for inference-time adaptation exploit the attention layers of pretrained models to perform tasks such as image editing [20], varied-size image generation [26], or zero-shot video generation [28]. Inspired by these works, we propose to modify the attention operations of diffusion models pretrained on square images for generating coherent panorama images in a zero-shot fashion.

Inference-Time Panorama 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 [1,5,19,26,30,61]. One of the most commonly considered pretrained models in this context is Stable Diffusion [45], which is adapted at inference-time to generate long images by following either an iterative diffusion paradigm [1,61] or a joint diffusion one [5,30], as we do in this work. 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 [5] 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 [30] entails providing guidance in the form of the gradient from a perceptual similarity loss [62] 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 [52]. In this work, we aim to generate varying and both perceptually and semantically coherent long images. To this end, different from the previous methods tackling this task, we modify the attention operations inside the backbone used to predict the noise at each sampling step.

3 Preliminaries

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

Diffusion Models. Diffusion models [23, 48, 49, 53] 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 $\mathbf{x}_{t-1}$ starting from $\mathbf{x}_t$. A possible strategy is to use the ϵ-prediction parameterization for sampling, as done in Denoising Diffusion Probabilistic Models (DDPMs) [23]. 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

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

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 [49]. 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) [40, 45] tackle this issue by operating in the latent space of an autoencoder, typically a VQ-GAN [14] or a VQ-VAE [55]. These models achieve good performance while using a fraction of the GFlops required by pixel space diffusion models [9, 39, 40, 45]. Here, we rely on the pretrained Stable Diffusion [45] 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 [43], in our case). In this work, we exploit diffusion models trained with the Classifier-Free Guidance [24] strategy for conditioning. This entails obtaining the noise prediction as a linear combination of conditional and unconditional predictions, with weight s, *i.e.*,

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

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) [50], 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) [34], which work on the latent space, are also available. In this work, we use an LCM distilled from Stable Diffusion [45].

4 Proposed Approach

In this section, we describe our pipeline for generating large images by exploiting a pretrained diffusion model at inference-time (see Fig. 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 [5,30]. As a result, generating panorama images by simply applying a model trained on squared ones [45] 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. 5.1 and in [30]. Thus, for generating larger images, we average the predictions on multiple views at every reverse process step, similar to [5]. 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

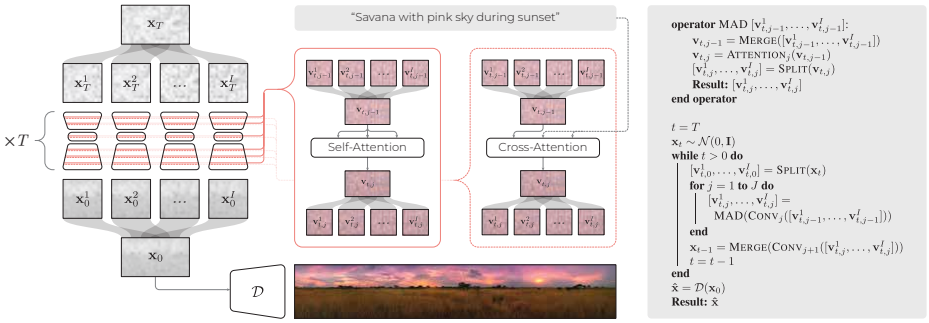


Fig. 2: Overview of our inference-time pipeline (left) and its pseudo-code (right). During the diffusion process, the image is split into overlapping views, and each is fed to the model separately. Within the attention layers, MAD provides interaction points between the views, enforcing global coherence in the generated panorama.

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 $\mathbf{x}^i_t \in \mathbb{R}^{L \times L \times C'}$ 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 $\mathbf{v}^i_{t,j} \in \mathbb{R}^{L_j \times L_j \times C_j}$ 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. $\text{SPLIT} : \mathbf{x}_t \to [\mathbf{x}^1_t, \dots, \mathbf{x}^I_t]$ divides its input tensor into multiple overlapping views. $\text{MERGE} : [\mathbf{x}^1_t, \dots, \mathbf{x}^I_t] \to \mathbf{x}_t$, 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 [30].

Note that in previous diffusion approaches for large image generation [5,30], each of the views is processed independently by the noise prediction model. Then, the outputs of the prediction model at the t-th step, $[\mathbf{x}^1_t, \dots, \mathbf{x}^I_t]$, 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, $[\mathbf{v}^1_{t,j-1}, \dots, \mathbf{v}^I_{t,j-1}]$, 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 $[\mathbf{v}^1_{t,j}, \dots, \mathbf{v}^I_{t,j}]$. The pseudocode of our approach is in Fig. 2.

5 Experiments

Implementation Details. In our experiments, as the pretrained diffusion models, we use Stable Diffusion 2.0 [45] and Latent Consistency Dreamshaper v7 (distilled from Stable Diffusion 1.5) [34] from HuggingFace [56]. Both models are based on a U-Net [46] architecture, and we apply MAD to all its attention layers. For Stable Diffusion, we perform 50 denoising steps and apply MAD for the first τ=15 steps. For Latent Consistency Dreamshaper, we consider 1, 2, and 4 generation steps and apply MAD with τ=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 [30]. For the quantitative evaluation, we use the same prompts as for SyncDiffusion [30] and MultiDiffusion [5], 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) [21], to evaluate the adherence with the textual prompt; Intra-LPIPS (I-LPIPS) [62] and Intra-Style Loss (I-StyleL) [16], to measure the intra-image perceptual coherence; FID [22], KID [6], and Mean-GIQA (mGIQA) [18], to estimate the perceptual similarity and variability with respect to a target distribution. The I-LPIPS and I-StyleL are computed by splitting



Fig. 3: Long images generated by the considered LDM with MAD, with τ=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.



Fig. 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*. τ=**0** is the setting where MAD is never applied.

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 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.

5.1 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 desired variety-uniformity trade-off. To explore this aspect, in Fig. 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. 3 shows different levels of global co-

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

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

herence, 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 the supplementary. Moreover, in Fig. 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 τ=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 Tab. 1 and further substantiate this claim in the supplementary. Both Figs. 3 and 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 Diffusion to generate the whole panorama image directly (which we refer to as SD-L) and the joint diffusion methods MultiDiffusion (MD) [5] and SyncDiffusion (SyncD) [30], 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 Tab. 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



Fig. 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.

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 Tab. 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. 5 shows a qualitative comparison among the joint diffusion methods on two of the six evaluation prompts (more qualitatives are provided in the supplementary). We generate the images using the same seed for each comparison to ensure consistency. As we can see, MD generates images 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. 2 shows the performance of our method and the joint diffusion competitors when generating 512×1024, 512×2048, and 512×4096 images. As we can see, MAD is consistently better in terms of mCLIP and competitive in I-LPIPS. The visual quality of the generated images can also be appreciated in Fig. 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. 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

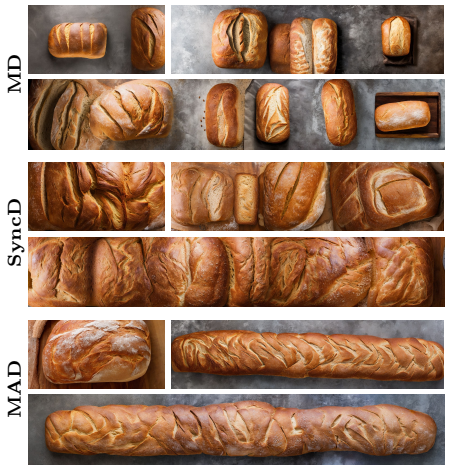


Fig. 6: Qualitative comparisons between our approach with LDM, MD, and SyncD at different aspect ratios for the prompt *Top view of a loaf of bread.*



Fig. 7: Qualitative comparison on vertical 512×2048 images between our approach applied to the LDM and SyncD for prompts requiring semantic coherence.

Table 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.

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

should have a beginning and an end, such as the waterfall. Please refer to the supplementary 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. 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 [30]. To explore this aspect, we run the

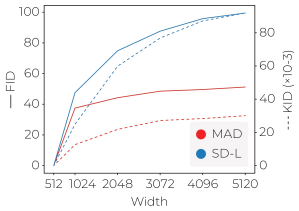


Fig. 8: Mean FID and KID on images of different widths by MAD and SD-L for the prompts used in the quantitative analysis.

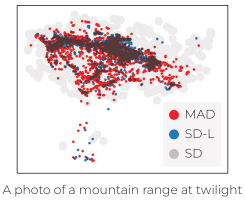


Fig. 9: T-SNE analysis of crops from images by MAD and SD-L *w.r.t.* squared images from SD for the same prompt.

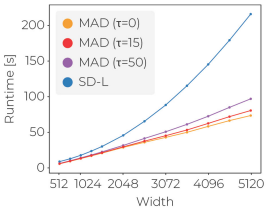


Fig. 10: Inference time at increasing output widths for SD-L and MAD: for τ=0, MAD is not applied, τ=50 is the upper bound.

t-SNE analysis [35] reported in Fig. 9 and in the supplementary. Specifically, we extract inception features from square images by SD and compute their t-SNE representations [35]. 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 [32, 41]. 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, the dark areas represent an overlap between MAD and SD-L. The t-SNE analysis and the perceptual and semantic scores in Tab. 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 RTXA5000-24GB for generating 512×3072 images with the LDM backbone (reported in Tab. 1). MAD runtime ranges from ∼43s to ∼50s, 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±0.1s, MAD is more than 8× faster for not requiring the expensive gradient computation. Moreover, in Fig. 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 (τ=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.

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 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 qualitatives in this paper (excluding those in Fig. 5) are obtained with prompts in GPT1k. For more qualitative comparisons, see the supplementary.

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

	mCLIP ↑	I-LPIPS ↓	I-StyleL ↓	FID ↓	KID ↓	mGIQA ↑
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



Fig. 11: Qualitative results on different prompts of our approach by exploiting LCD as backbone.

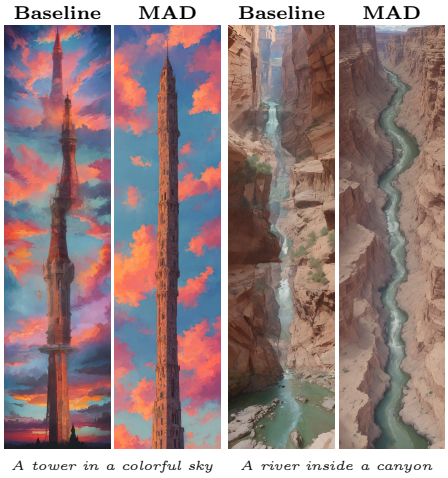


Fig. 12: Qualitative comparison on vertical images with LCD. **Baseline** entails never applying MAD (τ=0).

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. Every user answered six control questions to assess their understanding of the task. We collected responses from 201 users and kept those with at least five correct answers to the control questions (139 users). These provided 7807 answers, from which it has emerged that MAD was preferred 80.92% of the times over MD and 66.54% over SyncD (note that SyncD was preferred over MD 68.19% of the times, which is in line with what found in [30]). For further details, please see the supplementary.

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 Tab. 4, 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 qualitatives in Figs. 11 and 12 and in the supplementary).

Table 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.

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

6 Conclusions and Discussion

In this work, we have presented the **M**erge-**A**ttend-**D**iffuse 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.

Limitations. Similar to other inference-time joint diffusion approaches [5,30], 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. 1, 3 and 4, the incoherent horizon lines of independent diffusion paths (in the top images) are corrected by applying MAD. Further discussion can be found in the supplementary.

Potential Negative Societal Impacts. Image generative models can produce harmful content like deepfakes and NSFW images, violate copyrights, and reflect existing societal biases. Further research is necessary to address these issues.

Acknowledgement

This work was supported by the "AI for Digital Humanities" project funded by "Fondazione di Modena" and the PNRR project Italian Strengthening of ESFRI RI Resilience (ITSERR) funded by the European Union – NextGenerationEU.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended Latent Diffusion. ACM Trans. Graphics **42**(4), 1–11 (2023)
2. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: SpaText: Spatio-Textual Representation for Controllable ImageGgeneration. In: CVPR (2023)
3. Avrahami, O., Lischinski, D., Fried, O.: Blended Diffusion for Text-driven Editing of Natural Images. In: CVPR (2022)
4. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv preprint arXiv:2211.01324 (2022)
5. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. ICML (2023)
6. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: ICLR (2018)
7. Brock, A., Donahue, J., Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis. In: ICLR (2018)
8. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to Follow Image Editing Instructions. In: CVPR (2023)
9. Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In: ICLR (2024)
10. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W.: WaveGrad: Estimating Gradients for Waveform Generation. In: ICLR (2020)
11. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: DiffEdit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2022)
12. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. NeurIPS (2021)
13. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and Content-Guided Video Synthesis with Diffusion Models. In: ICCV (2023)
14. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: CVPR (2021)
15. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-Scene: Scene-Based Text-to-Image Generation with Human Priors. In: ECCV (2022)
16. Gatys, L.A., Ecker, A.S., Bethge, M.: Image Style Transfer Using Convolutional Neural Networks. In: CVPR (2016)
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Nets. NeurIPS (2014)
18. Gu, S., Bao, J., Chen, D., Wen, F.: GIQA: Generated Image Quality Assessment. In: ECCV (2020)
19. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: ScaleCrafter: Tuning-free Higher-Resolution Visual Generation with Diffusion Models. In: ICLR (2024)
20. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control. In: ICLR (2022)
21. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)

22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NeurIPS (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. NeurIPS (2020)
24. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance. In: NeurIPS Workshop (2021)
25. Ho, J., Salimans, T., Gritsenko, A.A., Chan, W., Norouzi, M., Fleet, D.J.: Video Diffusion Models. In: ICLR Workshop (2022)
26. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free Diffusion Model Adaptation for Variable-Sized Text-to-Image Synthesis. NeurIPS (2023)
27. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and Improving the Image Quality of StyleGAN. In: CVPR (2020)
28. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In: ICCV (2023)
29. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: DiffWave: A Versatile Diffusion Model for Audio Synthesis. In: ICLR (2021)
30. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via Synchronized Joint Diffusions. In: NeurIPS (2023)
31. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3D: High-Resolution Text-to-3D Content Creation. In: CVPR (2023)
32. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. Nat. Methods **16**(3), 243–245 (2019)
33. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In: CVPR (2022)
34. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference (2023)
35. Van der Maaten, L., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research **9**(11) (2008)
36. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In: ICLR (2022)
37. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text Inversion for Editing Real Images using Guided Diffusion Models. In: CVPR (2023)
38. Nichol, A.Q., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In: ICML (2021)
39. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: ICCV (2023)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR (2024)
41. Poličar, P.G., Stražar, M., Zupan, B.: openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. BioRxiv (2019)
42. Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., Kudinov, M.: Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In: ICML (2021)
43. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning Transferable Visual Models from Natural Language Supervision. In: ICML (2021)

44. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv e-prints pp. arXiv–2204 (2022)
45. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)
46. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: MICCAI (2015)
47. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image Super-Resolution via Iterative Refinement. IEEE Trans. PAMI **45**(4), 4713–4726 (2022)
48. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: ICML (2015)
49. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: ICLR (2021)
50. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency Models. ICML (2023)
51. Song, Y., Ermon, S.: Generative Modeling by Estimating Gradients of the Data Distribution. NeurIPS (2019)
52. Song, Y., Ermon, S.: Improved Techniques for Training Score-Based Generative Models. NeurIPS (2020)
53. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: ICLR (2021)
54. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In: CVPR (2023)
55. Van Den Oord, A., Vinyals, O., et al.: Neural Discrete Representation Learning. NeurIPS (2017)
56. Von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
57. Wang, W., Xie, K., Liu, Z., Chen, H., Cao, Y., Wang, X., Shen, C.: Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. arXiv preprint arXiv:2303.17599 (2023)
58. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. In: ICCV (2023)
59. Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., Gao, S.: Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models. In: CVPR (2023)
60. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: ICCV (2023)
61. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: DiffCollage: Parallel Generation of Large Content With Diffusion Models. In: CVPR (2023)
62. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: CVPR (2018)

Supplementary Material for: Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas

Fabio Quattrini, Vittorio Pippi,
Silvia Cascianelli, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

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.

1 Effect of the Application Stage

In Tables 1 and 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-LPIPIS, 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. 2.

2 Effect of the Number of Application Steps

We report a quantitative analysis (Tab. 3) and a more exhaustive qualitative analysis (Fig. 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.

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 [5] and SyncDiffusion [30]). For each

approach, we generated 60 images for each of the six prompts used for the quantitative 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 paper, 139 out of 201 participants passed the vigilance task and provided a total of 7807 preference answers. In Fig. 3, we plot the overall and the per-prompt results. As we can see, our method is consistently preferred over the competitors.

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 [35]. Then, we generate 512×5120 images with SD-L and MAD and embed all the squared crops using openTSNE [32, 41]. In Fig. 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.

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 [26] (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 LCMs are reported in Table 4 and Table 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 Figure 5.

6 Further Results at Different Aspect Ratios

We report a qualitative comparison with the State-of-the-Art approaches (MultiDiffusion [5] and SyncDiffusion [30]) on the generation of images with different aspect ratios. In particular, we consider the generation of horizontal images with the LDM (Fig. 7) and with the LCM (Fig. 8) and of vertical images with the same models (Fig. 9 and Fig. 10, respectively).

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. 11, 12 and 15 to 20 and of the LCM in Figs. 13, 14 and 21 to 26. Moreover, in Figs. 27 to 30, we report more qualitative results, both horizontal and vertical, on different GPT1k prompts for the LDM and LCM, respectively.

8 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 Figure 31, is on par with that of the generated panoramas (which is our main focus), even for guidance spanning multiple views of the large image.

9 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. 32. 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. 33 and 34. As we can see, the prompt *A gothic cathedral nave* (Fig. 34) 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.

Table 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↑	I-LPIPS↓	I-StyleL↓ ($\times 10^{-3}$)	FID↓	KID↓ ($\times 10^{-3}$)	mGIQA↑ ($\times 10^{-3}$)
No MAD (Baseline)	31.65±2.17	0.64±0.10	2.65±2.33	34.51±13.92	9.19±4.51	27.59±6.83
MAD in Mid Block	31.66±2.16	0.64±0.10	2.19±1.17	34.80±13.80	9.35±3.91	28.04±7.23
MAD in Down Blocks	31.83±2.21	0.62±0.10	1.98±1.03	40.46±16.58	16.40±8.83	28.11±7.42
MAD in Up Blocks	32.09±2.20	0.53±0.10	1.19±0.86	60.84±24.65	43.56±19.82	28.60±7.51
MAD in All Blocks	32.15±2.25	0.53±0.10	1.43±1.04	61.76±23.73	43.31±17.41	28.10±7.84

Table 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↑	I-LPIPS↓	I-StyleL↓ ($\times 10^{-3}$)	FID↓	KID↓ ($\times 10^{-3}$)	mGIQA↑ ($\times 10^{-3}$)
No MAD (Baseline)	31.36±1.83	0.55±0.05	0.90±0.35	31.35±15.42	13.27±10.05	35.44±12.38
MAD in Mid Block	31.44±1.79	0.55±0.05	0.87±0.33	29.35±15.62	13.85±11.27	35.49±12.22
MAD in Down Blocks	31.44±1.78	0.55±0.05	0.84±0.32	29.01±14.83	13.35±11.02	35.38±12.46
MAD in Up Blocks	31.46±1.84	0.51±0.05	0.71±0.26	40.22±20.99	28.98±17.43	35.45±13.08
MAD in All Blocks	31.48±1.87	0.52±0.05	0.71±0.26	38.77±19.85	27.41±16.19	35.23±13.15

Table 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↑	I-LPIPS↓	I-StyleL↓ ($\times 10^{-3}$)	FID↓	KID↓ ($\times 10^{-3}$)	mGIQA↑ ($\times 10^{-3}$)
τ=**0**	31.65±2.17	0.64±0.10	2.65±2.33	34.51±13.92	9.19±4.51	27.59±6.83
τ=**5**	31.86±2.22	0.59±0.10	2.07±1.31	38.10±13.71	13.75±4.00	28.32±7.64
τ=**10**	31.95±2.24	0.57±0.10	2.00±1.34	42.60±14.94	19.38±5.56	28.36±7.83
τ=**15**	32.03±2.29	0.56±0.10	1.90±1.32	48.52±17.14	27.15±9.10	28.32±7.76
τ=**20**	32.09±2.29	0.54±0.10	1.68±1.19	54.31±19.92	34.43±12.54	28.23±7.90
τ=**25**	32.15±2.25	0.53±0.10	1.43±1.04	61.76±23.73	43.31±17.41	28.10±7.84
τ=**30**	32.16±2.17	0.52±0.10	1.22±0.90	70.29±28.24	54.40±23.19	27.94±7.73
τ=**35**	32.18±2.08	0.51±0.10	1.02±0.73	78.50±32.94	65.06±29.94	27.82±7.65
τ=**40**	32.16±1.95	0.50±0.10	0.88±0.61	86.20±37.23	76.15±37.73	27.59±7.61
τ=**45**	32.14±1.83	0.50±0.10	0.78±0.53	92.69±40.76	84.13±44.13	27.35±7.50
τ=**50**	32.14±1.72	0.49±0.10	0.71±0.47	98.01±43.64	91.51±49.95	27.05±7.28

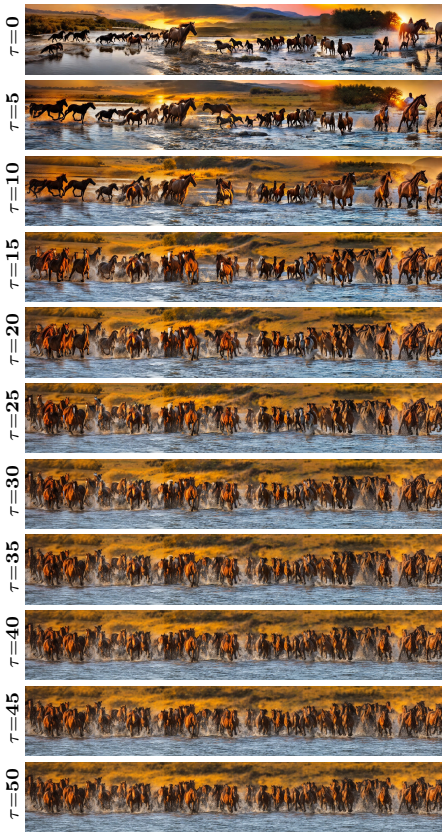


Fig. 1: Long images generated by the considered LDM with MAD applied up to different numbers of inference steps for the prompt *A herd of Mustang horses crossing a river at sunset*. When τ is too low, the view interactions are not enough to produce a globally coherent image. As τ increases, the image becomes more and more coherent, with maximal uniformity when MAD is applied at every timestep.



Fig. 2: Long images generated by the considered LDM with MAD applied in different blocks of the noise prediction model, with τ=15 for the prompt *A herd of Mustang horses crossing a river at sunset*.

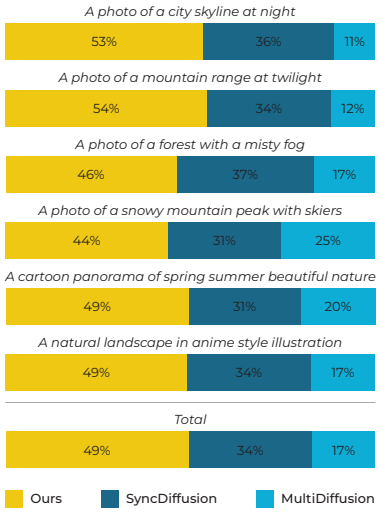


Fig. 3: User study per-prompt results of MAD, SyncDiffusion, and MultiDiffusion.

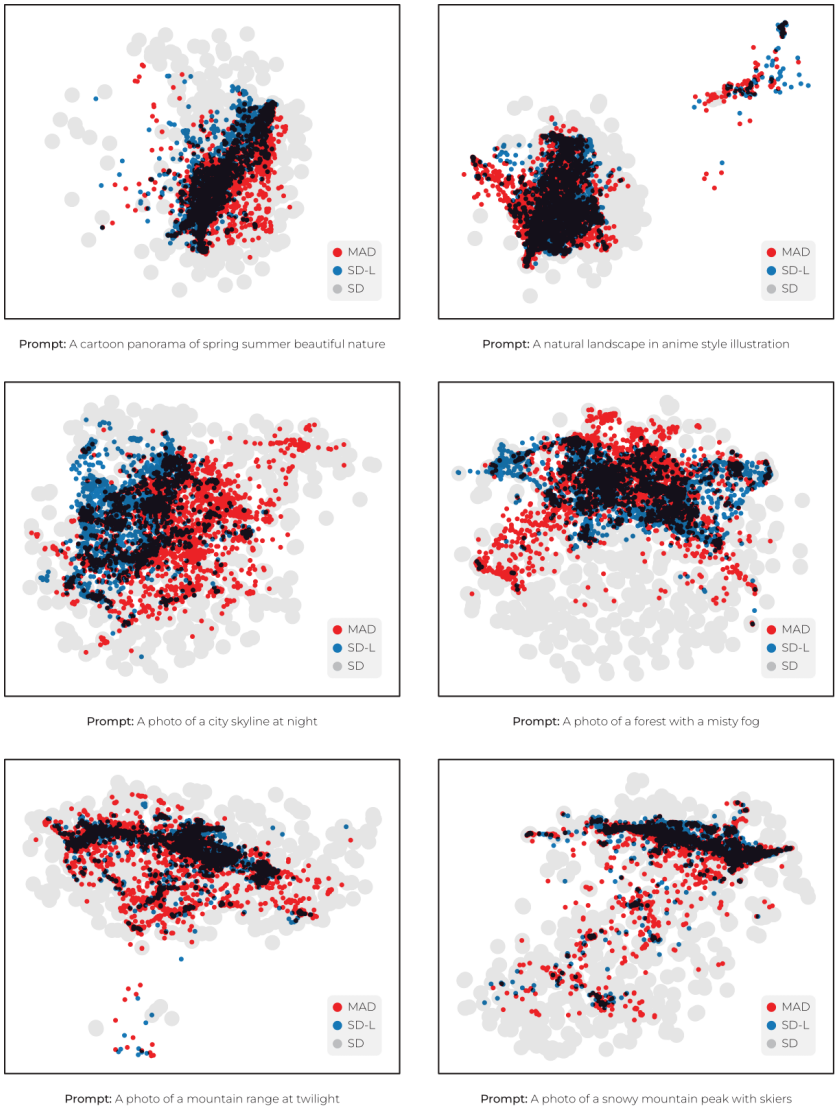


Fig. 4: Comparison on the distribution coverage of panoramas generated by MAD (red) and SD-L (blue) with respect to square images generated by SD (gray). Darker areas indicate an overlap between the embeddings of MAD and SD-L images.

Table 4: Quantitative comparison on 512×3072 panorama generation using the LDM. I-StyleL, KID, and mGIQA values are scaled by 10^3. For MAD, τ=15.

	mCLIP ↑	I-LPIPS ↓	I-StyleL ↓	FID ↓	KID ↓	mGIQA ↑
Standard Prompts						
SD-L	32.01±1.67	0.50±0.11	0.58±0.40	87.64±30.25	76.83±30.52	27.72±7.83
SD-L+Attn-S	32.02±1.66	0.52±0.11	0.73±0.47	80.16±27.11	67.54±26.00	27.89±7.67
MAD	32.03±2.29	0.56±0.10	1.90±1.32	48.52±17.14	27.15±9.10	28.32±7.76
MAD+Attn-S	32.17±2.20	0.53±0.11	1.19±0.75	64.08±28.19	47.33±23.77	28.30±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 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, τ=1/1/2 for 1/2/4 inference steps, respectively.

	mCLIP ↑	I-LPIPS ↓	I-StyleL ↓	FID ↓	KID ↓	mGIQA ↑
1 Inference Step						
LCD-L	29.50±1.48	0.40±0.08	0.21±0.18	67.54±15.53	66.86±21.22	32.21±4.50
LCD-L+Attn-S	29.77±1.43	0.41±0.08	0.33±0.33	77.50±14.72	79.76±20.96	31.13±4.15
MAD	29.01±1.66	0.40±0.07	0.37±0.28	72.43±23.75	75.47±33.24	32.23±4.95
MAD+Attn-S	29.36±1.46	0.41±0.07	0.51±0.42	80.53±24.19	86.79±36.77	31.28±4.75
2 Inference Steps						
LCD-L	30.77±2.09	0.47±0.06	0.56±0.29	49.98±25.20	44.29±26.88	35.45±11.94
LCD-L+Attn-S	31.11±1.95	0.49±0.06	0.66±0.32	56.37±28.61	54.55±33.00	34.94±12.09
MAD	30.97±2.15	0.50±0.06	0.85±0.34	35.69±17.88	23.74±15.82	35.32±11.49
MAD+Attn-S	31.16±2.07	0.50±0.06	0.81±0.31	36.84±18.15	25.83±16.26	35.68±11.60
4 Inference Steps						
LCD-L	31.30±1.63	0.50±0.06	0.58±0.24	55.52±32.82	51.71±37.79	35.44±12.94
LCD-L+Attn-S	31.60±1.57	0.51±0.06	0.72±0.27	59.25±32.32	57.57±37.91	34.92±13.27
MAD	31.48±1.87	0.52±0.05	0.71±0.26	38.77±19.85	27.41±16.19	35.23±13.15
MAD+Attn-S	31.64±1.87	0.51±0.06	0.68±0.23	38.68±19.10	26.49±14.77	35.45±13.27



Fig. 5: Effect of attention scaling (AS) on LDMs used directly and with MAD.

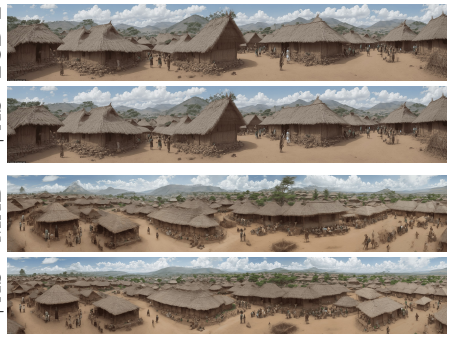


Fig. 6: Effect of attention scaling (AS) on LCMs used directly and with MAD.

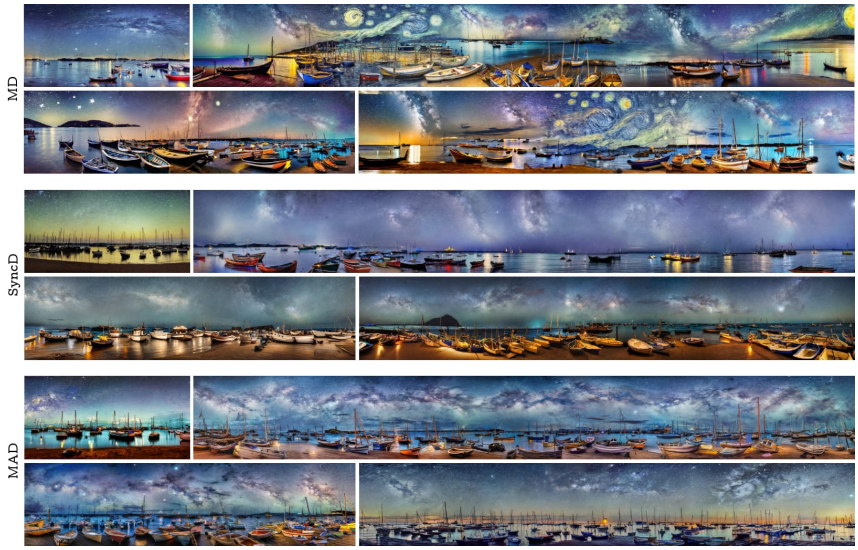


Fig. 7: Qualitative comparisons with MD and SyncD on the generation of horizontal images at different aspect ratios with the LDM model and the prompt *A photo of the seaside with some boats in a starry night*. These images are 512×1024 (top-left), 512×4096 (top-right), 512×2048 (bottom-left), and 512×3072 (bottom-right).



Fig. 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).



Fig. 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).

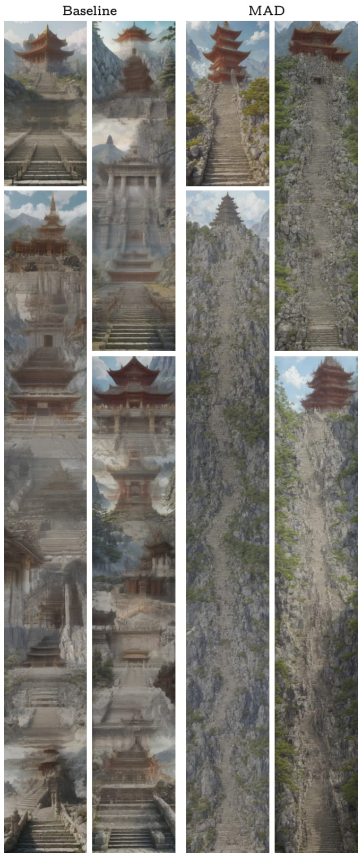


Fig. 10: Qualitative comparisons with respect to a baseline on the generation of vertical images at different aspect ratios with the LCM model and the prompt *Stairs on a mountain with a temple on top*. The images are 1024×512 (top-left), 2048×512 (top-right), 4096×512 (bottom-left), and 3072×512 (bottom-right).



Fig. 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.



Fig. 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.

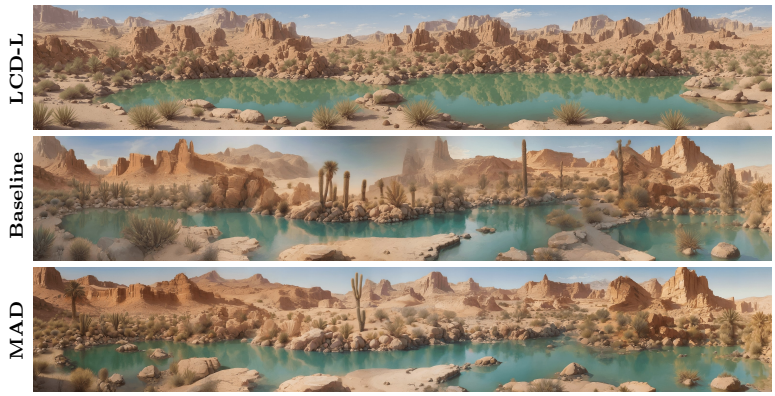


Fig. 13: Qualitative comparisons with respect to LCD-L, MD, and SyncD for the prompt *A tranquil desert oasis* from GPT1k, using the LCM model.

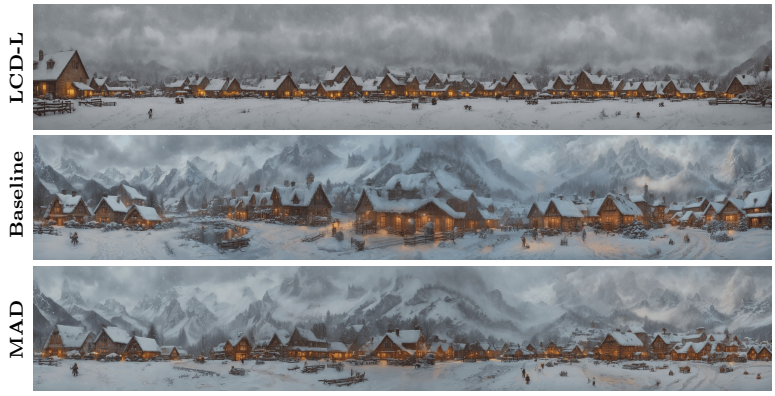


Fig. 14: Qualitative comparisons with respect to LCD-L and a baseline for the prompt *A snowstorm blanketing a small village* from GPT1k, using the LCM model.

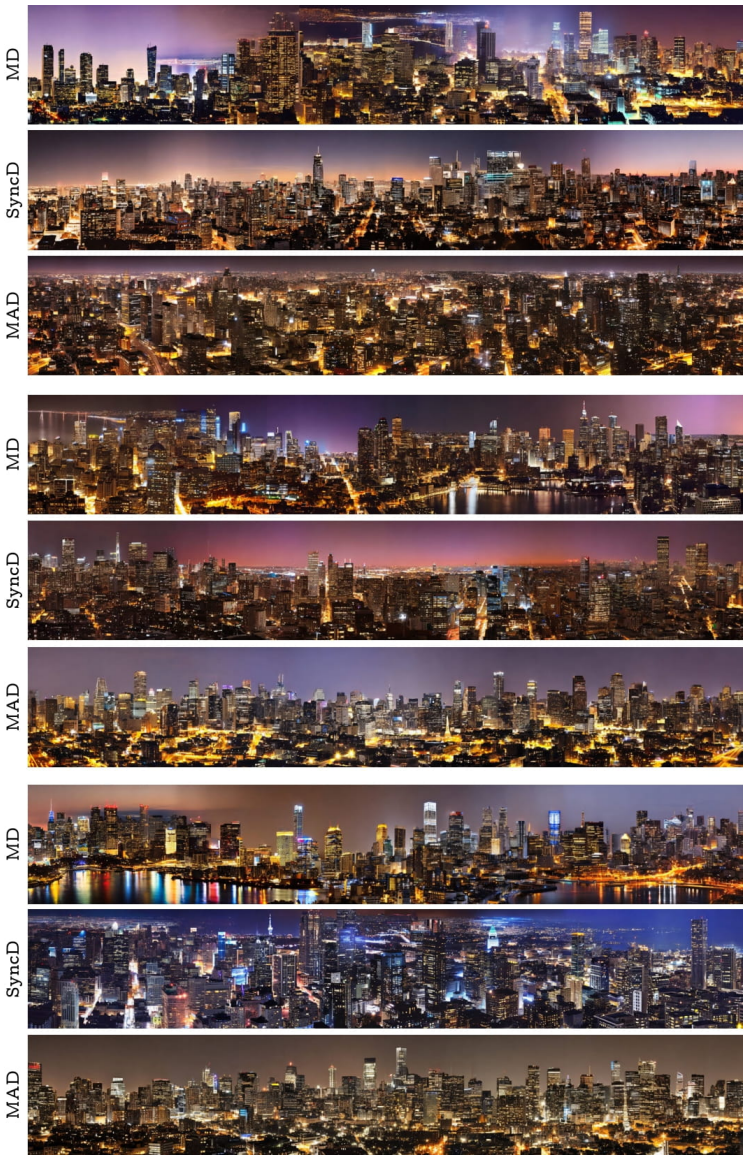


Fig. 15: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a city skyline at night*, using the LDM model.



Fig. 16: Qualitative comparisons with respect to MD and SyncD for the prompt *A photo of a mountain range at twilight*, using the LDM model.



Fig. 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.



Fig. 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.



Fig. 19: Qualitative comparisons with respect to MD and SyncD for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LDM model.

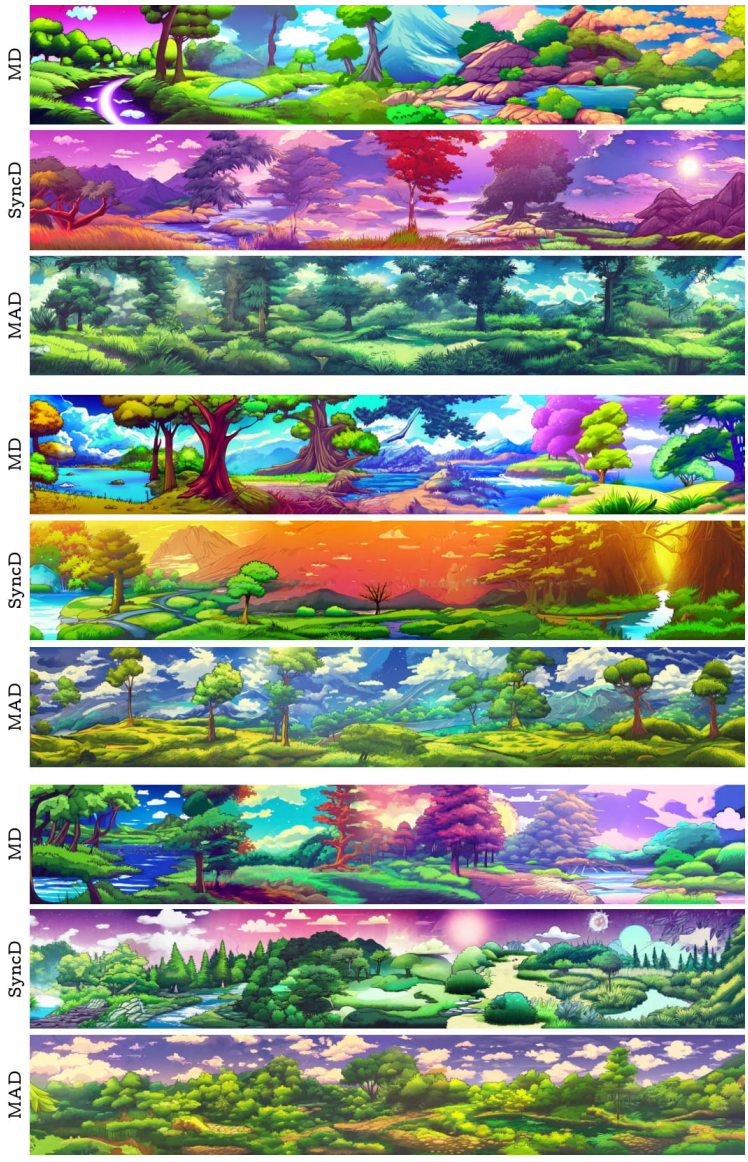


Fig. 20: Qualitative comparisons with respect to MD and SyncD for the prompt *A natural landscape in anime style illustration*, using the LDM model.



Fig. 21: Qualitative comparisons with respect to a baseline for the prompt *A photo of a city skyline at night*, using the LCM model.



Fig. 22: Qualitative comparisons with respect to a baseline for the prompt *A photo of a mountain range at twilight*, using the LCM model.



Fig. 23: Qualitative comparisons with respect to a baseline for the prompt *A photo of a forest with a misty fog*, using the LCM model.



Fig. 24: Qualitative comparisons with respect to a baseline for the prompt *A photo of a snowy mountain peak with skiers*, using the LCM model.



Fig. 25: Qualitative comparisons with respect to a baseline for the prompt *A cartoon panorama of spring summer beautiful nature*, using the LCM model.

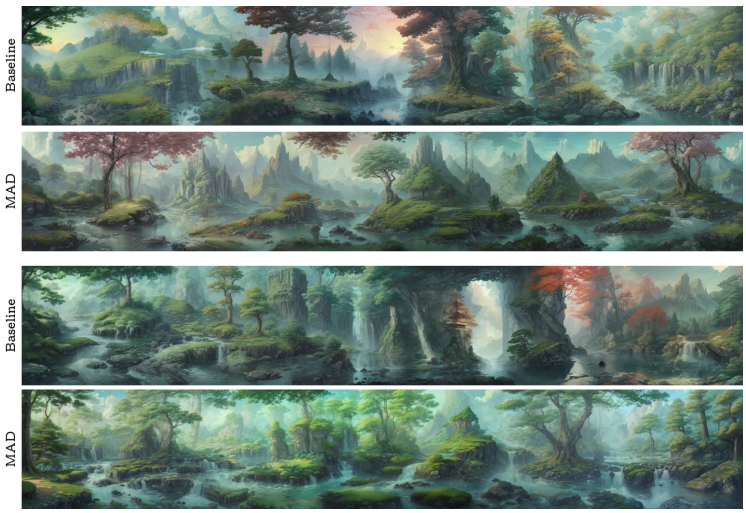


Fig. 26: Qualitative comparisons with respect to a baseline for the prompt *A natural landscape in anime style illustration*, using the LCM model.

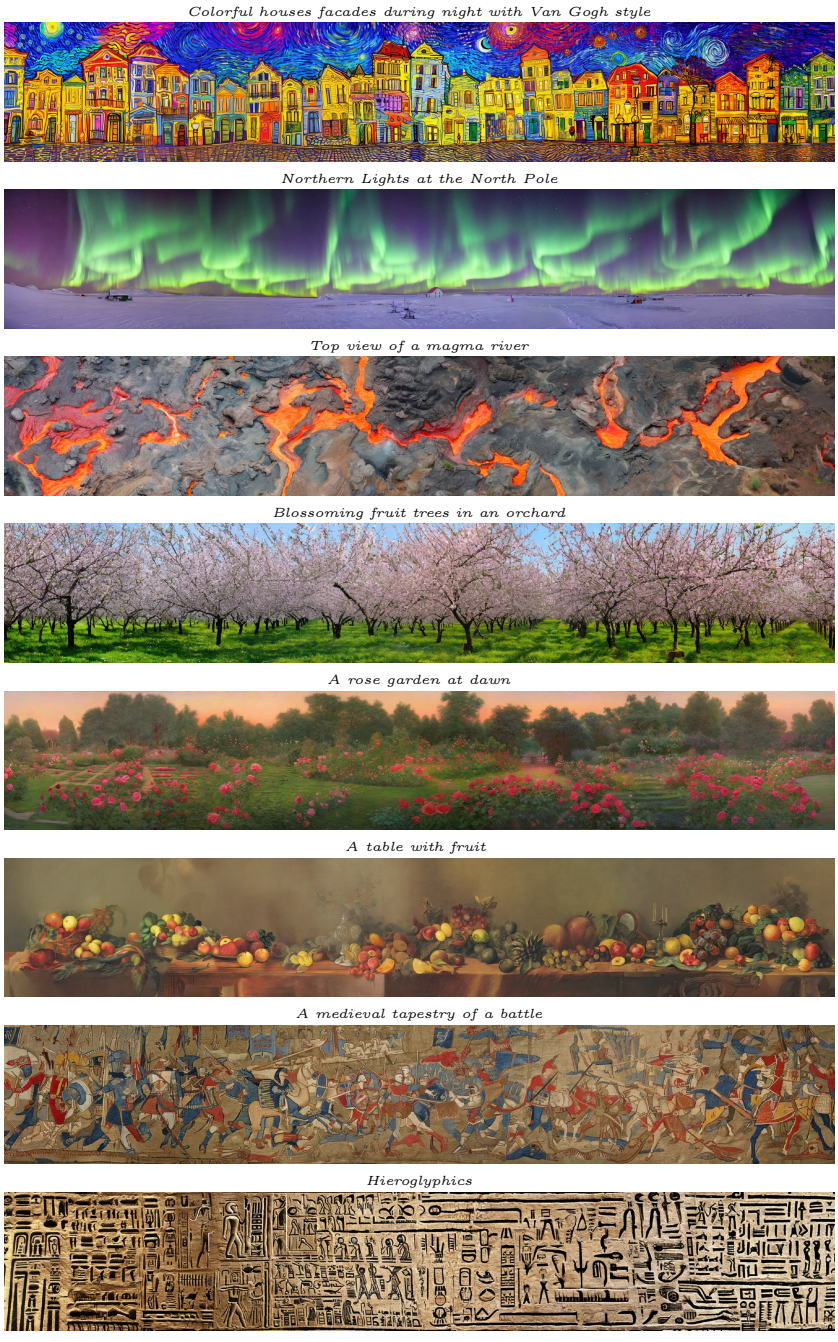


Fig. 27: Horizontal panorama images with various prompts and the LDM model.

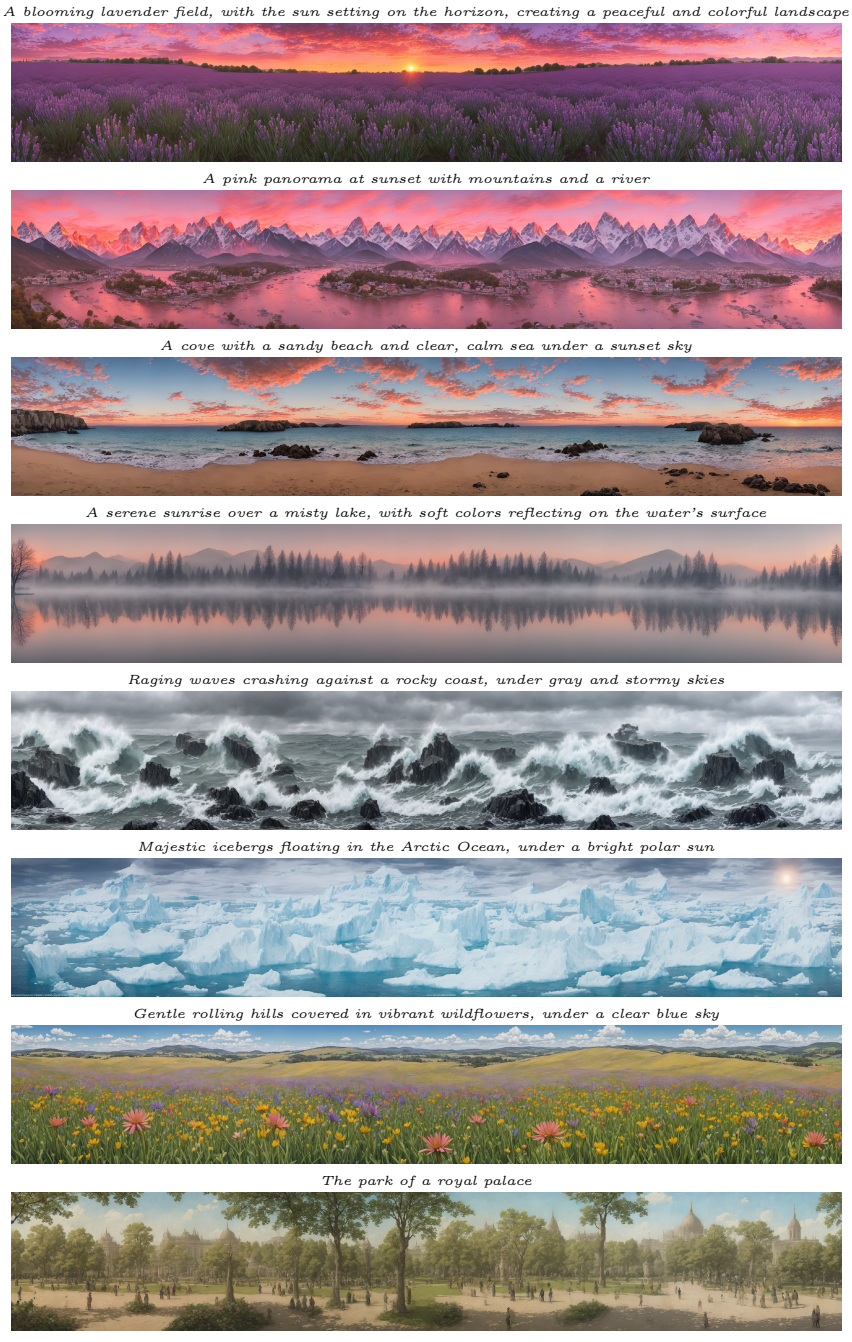


Fig. 28: Horizontal panorama images with various prompts and the LCM model.



Fig. 29: Vertical panorama images with various prompts and the LDM model.

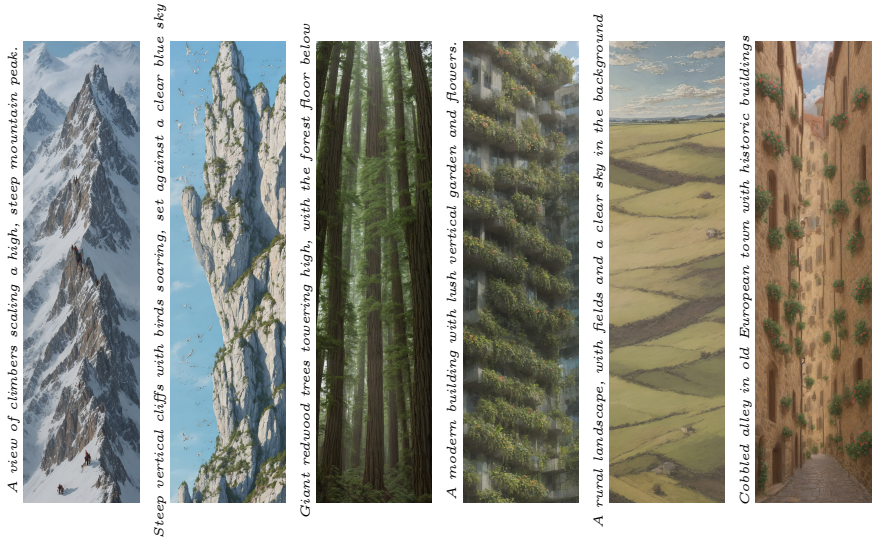


Fig. 30: Vertical panorama images with various prompts and the LCM model.



Fig. 31: Plug&Play applications of MAD.

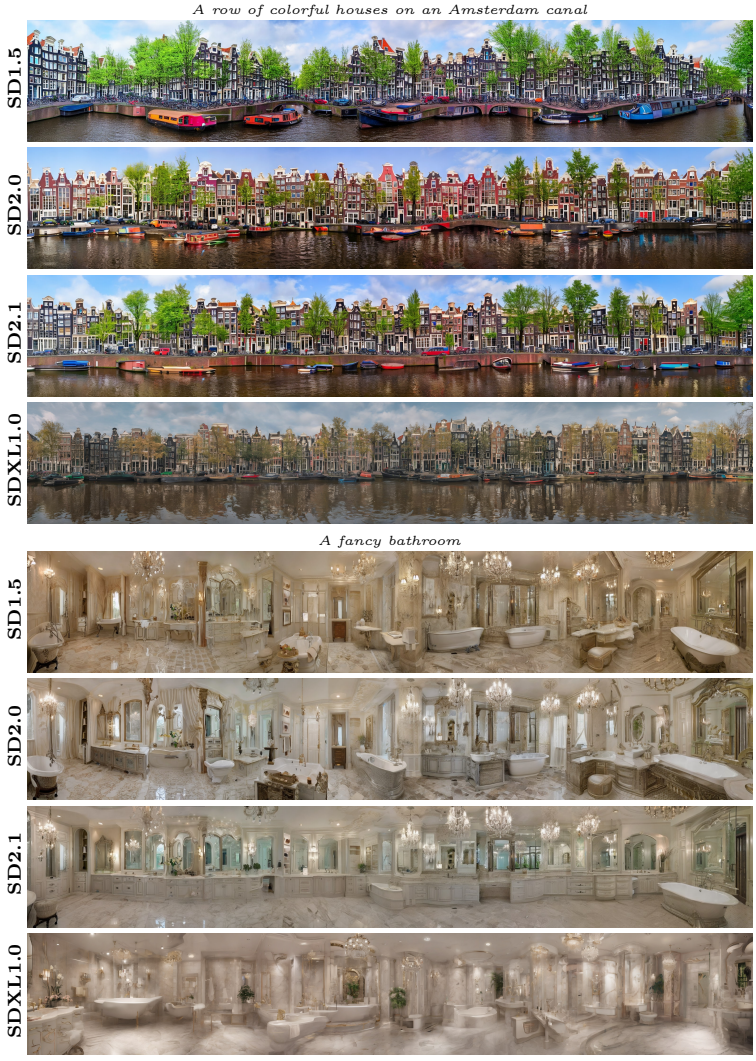


Fig. 32: Images generated with MAD applied to different LDM backbones and resolution 512×3072. For SDXL1.0 [40], the resolution is 1024×6144.



Fig. 33: 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*.

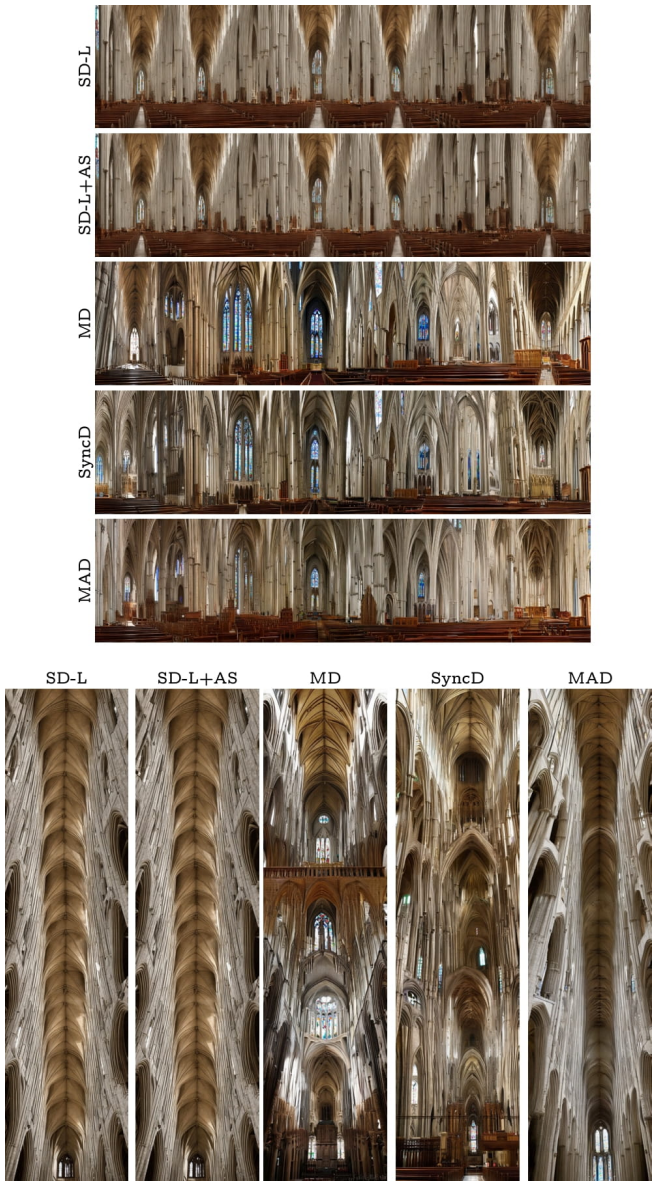


Fig. 34: Horizontal and vertical images generated with the same seed and for the prompt *A gothic cathedral nave* using different methods applied on the LDM model: direct inference (SD-L), Attention Scaling (SD-L+AS), MD, SyncD, and MAD.