This is the peer reviewd version of the followng article:

Dress-ED: Instruction-Guided Editing for Virtual Try-On and Try-Off / Lobba, D., Sanguigni, F., Ren, B., Cornia, M., Cucchiara, R., Sebe, N.. - (2026). (European Conference on Computer Vision Malmö, Sweden September 8-12, 2026).

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.

08/09/2026 05:25

(Article begins on next page)

Dress-ED: Instruction-Guided Editing for Virtual Try-On and Try-Off

Davide Lobba[1,3*], Fulvio Sanguigni[2,3*], Bin Ren[4†], Marcella Cornia[2], Rita Cucchiara[2], and Nicu Sebe[1]

[1] University of Trento, Italy
{davide.lobba, nicu.sebe}@unitn.it
[2] University of Modena and Reggio Emilia, Italy
{fulvio.sanguigni, marcella.cornia, rita.cucchiara}@unimore.it
[3] University of Pisa, Italy
[4] MBZUAI, United Arab Emirates
bin.ren@mbzuai.ac.ae

aimagelab.github.io/Dress-ED/

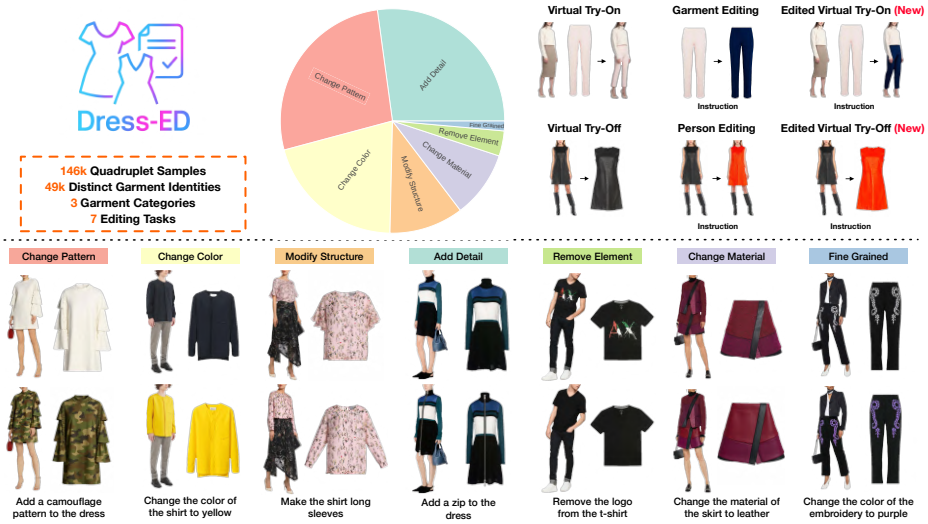

Fig. 1: We propose the **Dress Editing Dataset** (**Dress-ED**), the first benchmark for instruction-driven virtual try-on and try-off with over 146k verified samples across seven editing types, including both appearance and structural modifications.

Abstract. Recent advances in Virtual Try-On (VTON) and Virtual Try-Off (VTOFF) have greatly improved photo-realistic fashion synthesis and garment reconstruction. However, existing datasets remain static, lacking instruction-driven editing for controllable and interactive fashion generation. In this work, we introduce the **Dress Editing Dataset** (**Dress-ED**), the first large-scale benchmark that unifies VTON, VTOFF, and text-guided garment editing within a single framework. Each sample in Dress-ED includes an in-shop garment image, the corresponding

* Equal contribution. † Corresponding author.

person image wearing the garment, their edited counterparts, and a natural-language instruction of the desired modification. Built through a fully automated multimodal pipeline that integrates MLLM-based garment understanding, diffusion-based editing, and LLM-guided verification, Dress-ED comprises over 146k verified quadruplets spanning three garment categories and seven edit types, including both appearance (*e.g.*, color, pattern, material) and structural (*e.g.*, sleeve length, neckline) modifications. Based on this benchmark, we further propose a unified multimodal diffusion framework that jointly reasons over linguistic instructions and visual garment cues, serving as a strong baseline for instruction-driven VTON and VTOFF.

Keywords: Virtual Try-On · Virtual Try-Off · Editing · Fashion

1 Introduction

Digital garment manipulation on human images has emerged as a central challenge at the intersection of computer vision, computer graphics, and fashion intelligence. Tasks such as Virtual Try-On (VTON) [21, 33, 39] and Virtual Try-Off (VTOFF) [36, 50] aim to bridge physical garments and digital humans through realistic cross-modal synthesis and reconstruction. Recent progress in diffusion-based generation [5, 10, 72] and Transformer architectures [25, 28, 32] has achieved remarkable realism. Nevertheless, existing datasets remain largely static, containing fixed person-garment pairs without support for instruction-level control or semantic supervision, thereby constraining progress toward interactive and controllable fashion manipulation.

Beyond conventional VTON and VTOFF, emerging applications demand fine-grained garment editing, such as modifying color, texture, or silhouette via natural-language instructions. Achieving this requires precise alignment between garment semantics, human appearance, and linguistic intent. Yet, the existing efforts remain fragmented: *(i)* VTON datasets [9, 21, 39] focus on person–garment synthesis without editability; *(ii)* instruction-driven datasets [7, 24, 66, 69] support generic image editing but lack consistent garment–person correspondences; and *(iii)* fashion editing approaches [23, 64, 71] tackle narrow sub-tasks without unified benchmarks. As a result, controllable, multimodal fashion editing remains underexplored.

Motivated by these limitations, in this work we propose a unified benchmark for instruction-driven fashion editing that jointly captures garment semantics, human appearance, and editing intent. To this end, we introduce the **Dress Editing Dataset** (**Dress-ED**), the first large-scale, systematically curated benchmark that integrates VTON and VTOFF tasks under a single instruction-based framework. Each sample in Dress-ED includes an in-shop garment image, the corresponding person image wearing the garment, their edited counterparts, and a natural-language instruction describing the desired modification (Fig. 1). The dataset is constructed through a fully automated multimodal pipeline. In particular, the pipeline first extracts structured garment attributes using Qwen3-

VL [3], then synthesizes diverse editing instructions and applies diffusion-based editing via FLUX.2 Klein [29]. Finally, multi-stage quality verification with GPT-5 [40] and a fine-tuned InternVL-3.5 [53] ensures semantic consistency and visual fidelity. The resulting dataset comprises over 146k verified samples spanning three garment categories and seven editing types, covering both *appearance edits* (*e.g.*, color, pattern, material) and *structural modifications* (*e.g.*, sleeve length, neckline, details). Dress-ED thus establishes a scalable and semantically grounded benchmark for controllable, instruction-driven fashion generation.

In addition to dataset construction, Dress-ED contributes along two key dimensions: *benchmarking* and *methodology*. It introduces two complementary benchmarks (*i.e.*, instruction-driven VTON and VTOFF) under unified evaluation protocols assessing both visual realism and semantic edit accuracy. Building on this foundation, we propose a unified multimodal diffusion framework for instruction-driven editing in both try-on and try-off scenarios. The framework integrates a pretrained MLLM, a lightweight connector, and a DiT-based component [45] to jointly reason over textual instructions and visual garment cues. This design enables fine-grained, localized edits through token-level conditioning while preserving global semantic coherence, providing a strong and extensible baseline for future research in multimodal fashion understanding and generation.

Contributions. Our main contributions are threefold:

- We present Dress-ED, the first large-scale dataset that unifies VTON, VTOFF, and instruction-driven garment editing within a single, semantically consistent framework.
- We develop a fully automated multimodal pipeline that combines MLLM-based garment understanding, diffusion-based synthesis, and LLM-guided verification, generating over 146k high-quality samples across diverse edit types.
- We establish comprehensive benchmarks and a unified diffusion-based baseline for instruction-driven fashion editing in both VTON and VTOFF settings.

2 Related Work

Virtual Try-On and Try-Off. VTON aims to synthesize realistic images of a target person wearing a given garment [4, 12, 15, 47], while VTOFF addresses the inverse process of recovering clean garment representations from dressed-person photos for digitization and editing. Early *warping-based* methods [8, 21, 52, 58, 60] deform garments to align with body shapes but often suffer from misaligned textures and ghosting artifacts. Recent *warping-free* models [5, 10, 38, 72] remove explicit deformation by using attention-based conditioning, achieving higher realism though sometimes losing fine texture fidelity. To mitigate this, approaches based on diffusion-based architectures [25, 28, 32, 71] enhance garment details, while conditioning modules such as LOTS [19] and LEFFA [70] improve structural consistency. In parallel, diffusion-based VTOFF models [36, 50, 51, 57] advance garment disentanglement from occlusions and complex poses across multiple categories. More unified frameworks like Voost [30], One-Model-for-All [34], and

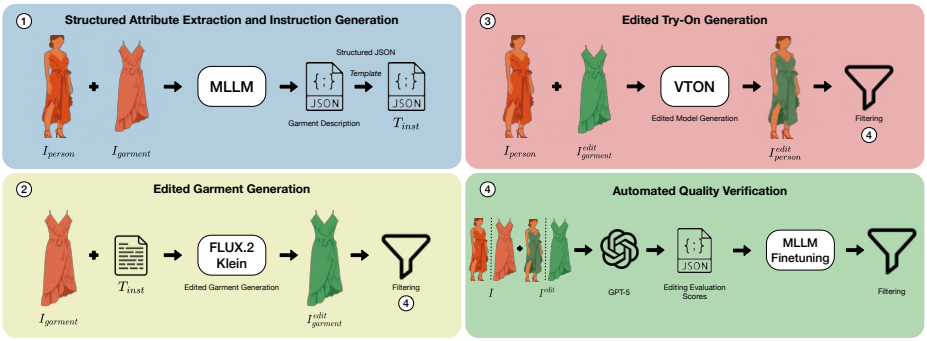

Fig. 2: Overview of the Dress-ED curation pipeline. Starting from Dress Code [39], (i) structured garment attributes are extracted with Qwen3-VL and used to generate natural-language edit instructions, (ii) edited in-shop garments are synthesized via FLUX.2 Klein, (iii) the corresponding edited try-on images are generated using FitDiT, and (iv) all results are validated through a version of InternVL-3.5, fine-tuned with samples annotated using GPT-5 [40], to ensure semantic and visual consistency.

Any2AnyTryon [20] attempt to integrate both VTON and VTOFF within a single model. However, the progress of such unified modeling remains limited by the lack of large-scale datasets that jointly support person-centric synthesis and garment-centric reconstruction with fine-grained correspondence. Motivated by this gap, we introduce a large-scale dataset that unifies both VTON and VTOFF tasks, enabling detailed garment editing and realistic human-garment interaction beyond prior approaches.

General-Purpose Editing. Image editing has become an active research area, enabling flexible and fine-grained manipulation of image content. Existing approaches can be broadly categorized into several types. Finetuning-based methods [7, 16, 35, 49, 55, 61, 65, 66, 68] adapt pretrained generative models using real or synthetic supervision. Some studies [16, 35] further incorporate MLLM guidance to inject richer multimodal cues. Other research lines explore zero-shot editing via noise or flow inversion [17, 26, 43, 48] and probing-based analysis to identify edit-receptive components [2]. However, these general-purpose methods often underperform in fashion scenarios, where garments exhibit complex geometry, layered structure, and fine-grained textures that require precise spatial and semantic control. Recently, foundation models such as FLUX.2 Klein [29] and Qwen-Image-Edit [56] have advanced instruction-based image-to-image editing, supporting diverse manipulation tasks. Yet, these models remain limited when the input and target belong to different domains (*e.g.*, person-to-garment or garment-to-person transformations).

Fashion-Oriented Editing. In the VTON setting, several methods have explored instruction-based editing as an auxiliary task. M&M VTO [71] and HieraFashDiff [59] demonstrate qualitative results on simple edits such as rolled sleeves. However, these works neither provide quantitative benchmarking nor address more challenging edits involving complex structures or patterns. In the

Algorithm 1 Automated Multimodal Data Generation.

Require: Paired samples $\{(I_{\text{garment}}, I_{\text{person}})\}$ from Dress Code [39]
Ensure: Verified quadruplets $\mathcal{S}^* = \{(I_{\text{garment}}, I_{\text{person}}, I^{\text{edit}}_{\text{garment}}, I^{\text{edit}}_{\text{person}}, T_{\text{inst}})\}$
Stage 1: Attribute Extraction and Instruction Generation
$\mathcal{A}_g \leftarrow$ Qwen3-VL$(I_{\text{garment}}, I_{\text{person}})$ ▷ structured garment attributes
$T_{\text{inst}} \leftarrow f_{\text{temp}}(\mathcal{A}_g)$ ▷ template-based edit instruction
Stage 2: Diffusion-Based Garment Editing
$I^{\text{edit}}_{\text{garment}} \leftarrow$ FLUX.2 Klein$(I_{\text{garment}}, T_{\text{inst}})$ ▷ apply appearance/structure edit
Stage 3: Try-On Image Generation
$I^{\text{edit}}_{\text{person}} \leftarrow$ FitDiT$(I_{\text{person}}, I^{\text{edit}}_{\text{garment}})$ ▷ render edited garment on person
Stage 4: LLM-Guided Quality Verification
$s \leftarrow$ GPT-5Judge$(I, I^{\text{edit}}, T_{\text{inst}})$ ▷ adherence & quality scoring
$f_{\text{verifier}} \leftarrow$ Finetune(InternVL-3.5, $\{(I, I^{\text{edit}}, T_{\text{inst}}, s)\}_{\sim 5\text{k}}$)
$\mathcal{S}^* \leftarrow \{(I_{\text{garment}}, I_{\text{person}}, I^{\text{edit}}_{\text{garment}}, I^{\text{edit}}_{\text{person}}, T_{\text{inst}}) \mid f_{\text{verifier}}(I, I^{\text{edit}}, T_{\text{inst}}) > t\}$
return $\mathcal{S}^*$ // ∼146k verified samples, 3 categories, 7 edit types

broader fashion editing domain, recent efforts focus on either person-to-person [23] or garment-to-garment [64] transformations. T-FIT [23] guides the model using concept vectors derived from CLIP text embeddings and fine-tunes it via LoRA [22] to preserve non-instructional visual attributes. EditGarment [64] constructs synthetic pairs of original and target garments with corresponding textual captions and trains an InstructPix2Pix [7] model for garment-level editing. Despite these advances, no existing approach can jointly perform VTON or VTOFF and instruction-driven editing within a unified framework. Moreover, the lack of a reliable benchmark with ground-truth supervision limits consistent evaluation of edit quality and semantic fidelity. Our proposed Dress-ED bridges these gaps by unifying VTON, VTOFF, and instruction-based editing for dedicated model training and evaluation.

3 Dress Editing Dataset (Dress-ED)

We introduce the **Dress Editing Dataset** (**Dress-ED**), a large-scale benchmark unifying instruction-driven fashion editing for both VTON and VTOFF. Dress-ED extends Dress Code [39] by adding natural-language editing instructions and corresponding edited images, aligning textual intent with visual garment semantics. Each sample supports *appearance* edits (*e.g.*, color, pattern, material) and *structural* edits (*e.g.*, sleeve length, neckline, details), enabling fine-grained and controllable garment manipulation. This unified formulation provides a comprehensive benchmark for evaluating and training instruction-guided fashion editing models.

3.1 Automated Dataset Generation Pipeline

We construct Dress-ED through a fully automated multimodal pipeline, summarized in Alg. 1 and illustrated in Fig. 2, which converts static garment–person pairs into instruction-driven editing samples. Built upon the Dress Code dataset [39], which contains paired images $(I_{\text{person}}, I_{\text{garment}})$ across three representative clothing types (*i.e.*, upper-body, lower-body, and dresses), the pipeline integrates

multimodal understanding, diffusion-based generation, and LLM-guided verification to produce coherent instruction-grounded samples. Each finalized instance in Dress-ED is formulated as:

$$\mathcal{S} = \left(I_{\text{garment}},\, I_{\text{person}},\, I^{\text{edit}}_{\text{garment}},\, I^{\text{edit}}_{\text{person}},\, T_{\text{inst}}\right), \tag{1}$$

where $(I_{\text{person}}, I_{\text{garment}})$ are the original inputs, $I^{\text{edit}}_{\text{garment}}$ and $I^{\text{edit}}_{\text{person}}$ are their edited counterparts, and T_{inst} denotes the natural-language instruction.

Stage 1: Structured Attribute Extraction and Instruction Generation. Given the in-shop garment I_{garment} and the corresponding on-model image I_{person}, we employ Qwen3-VL [3] to extract structured garment semantics:

$$A_g = f_{\text{MLLM}}(I_{\text{garment}}, I_{\text{person}}), \tag{2}$$

where A_g is a structured JSON with both descriptive text and explicit attributes (*e.g.*, color, pattern, material, neckline, sleeve length). Natural-language edit instructions are then synthesized via rule-based templates:

$$T_{\text{inst}} = f_{\text{temp}}(A_g), \tag{3}$$

covering two main categories, *i.e.*, *appearance edits* (color, pattern, material) and *structural edits* (adding, removing, reshaping parts). For example, a "blue cotton shirt with short sleeves" may yield "change the sleeve length to long" or "make the shirt yellow", forming $(I_{\text{garment}}, T_{\text{inst}})$ pairs.

Stage 2: Edited Garment Generation. Given the $(I_{\text{garment}}, T_{\text{inst}})$ pair, we employ the diffusion-based editor FLUX.2 Klein [29] to synthesize edited in-shop garments:

$$I^{\text{edit}}_{\text{garment}} = f_{\text{diff}}(I_{\text{garment}}, T_{\text{inst}}), \tag{4}$$

yielding approximately 300k edited samples across categories, which are subsequently refined through automated quality verification to ensure realism and instruction adherence.

Stage 3: Edited Try-On Generation. We take I_{person} and the edited garment $I^{\text{edit}}_{\text{garment}}$, and employ the state-of-the-art VTON model FitDiT [25] to generate its try-on result:

$$I^{\text{edit}}_{\text{person}} = f_{\text{VTON}}(I_{\text{person}}, I^{\text{edit}}_{\text{garment}}), \tag{5}$$

where the person is realistically rendered wearing the modified garment. The generated try-on images are subsequently validated through the same automated quality verification pipeline to ensure semantic consistency and visual fidelity.

Stage 4: Automated Quality Verification. To ensure semantic correctness and visual fidelity, we adopt a three-step verification pipeline. First, each triplet $(I, I^{\text{edit}}, T_{\text{inst}})$ is evaluated by GPT-5 [40], which outputs an edit score $s \in [0, 100]$ assessing instruction adherence, content preservation, and realism. Second, a subset of $\sim$5k annotated samples is used to fine-tune InternVL-3.5 [53], distilling GPT-5's scoring capability into a scalable multimodal evaluator:

$$f_{\text{verifier}} \leftarrow \text{Finetune}(\text{InternVL-3.5}, \{(I^{\text{edit}}, T_{\text{inst}}, s)\}), \tag{6}$$

Table 1: Comparison of fashion and general-purpose editing datasets. While existing datasets lack instruction-guided garment-person correspondences, **Dress-ED** introduces 146k verified samples enabling, for the first time, instruction-driven VTON and VTOFF.

Dataset	# Samples	# Editing Types	Automatic Generated	MLLM Filtering	Supported Tasks
General-Purpose Editing Datasets					
MagicBrush [66]	10,388	7	✗	✗	Visual Attributes Editing
RefEdit [44]	20,000	5	✓	✗	Visual Attributes Editing
UltraEdit [69]	4,108,262	9	✓	✗	Visual Attributes, Style, Background Editing
HQ-Edit [24]	450,000	7	✓	✓	Visual Attributes and Style Editing
AnyEdit [65]	2,500,000	25	✓	✓	Visual Attributes, Camera Movements, Scene Editing
ImgEdit [63]	1,200,000	11	✓	✓	Visual Attributes and Garment Editing
Fashion-Oriented Datasets					
DeepFashion2 [18]	873,234	-	-	-	Cloth Retrieval, Segmentation, Detection, Pose Estimation
Street TryOn [11]	14,453	-	-	-	Person-to-Person VTON
VITON-HD [9]	13,679	-	-	-	VTON, VTOFF
Dress Code [39]	53,792	-	-	-	VTON, VTOFF
EditGarment [64]	20,596	6	✓	✓	Instruction-Driven Garment Editing
Dress-ED (Ours)	146,460	7	✓	✓	Instruction-Driven VTON and VTOFF, Person Editing, Garment Editing, VTON, VTOFF

where I^{edit} denotes either $I^{\text{edit}}_{\text{garment}}$ or $I^{\text{edit}}_{\text{person}}$ depending on the verification target.

Finally, f_{InternVL} is applied to all edited garments and try-on results, and samples with predicted scores below a threshold t are filtered out:

$$\mathcal{S}^* = \{(I_{\text{garment}}, I_{\text{person}}, I^{\text{edit}}_{\text{garment}}, I^{\text{edit}}_{\text{person}}, T_{\text{inst}}) \mid \quad f_{\text{verifier}}(I^{\text{edit}}, T_{\text{inst}}) > t\},$$

where t is set to 80 after manual verification, retaining only samples for which both $I^{\text{edit}}_{\text{garment}}$ and $I^{\text{edit}}_{\text{person}}$ obtain a score above this threshold.

This process yields ∼146k verified high-quality samples across all garment categories, forming the final Dress-ED.

3.2 Dataset Statistics

The final Dress-ED comprises 146,460 visual quadruplets $(I_{\text{garment}}, I_{\text{person}}, I^{\text{edit}}_{\text{garment}}, I^{\text{edit}}_{\text{person}})$ paired with natural-language instructions T_{inst}, covering three representative garment categories: dresses (80,865 samples), upper-body (45,567 samples), and lower-body (20,028 samples), encompassing 49,664 distinct garment identities and 6,073 unique editing instructions. We follow the same train/test protocol as Dress Code [39], splitting

Table 2: Sample distribution by editing type.

Edit Type	Count	Proportion (%)
Add Detail	39,776	27
Change Pattern	39,559	27
Change Color	29,983	20
Modify Structure	15,670	11
Change Material	14,091	10
Remove Element	5,305	4
Fine-Grained	2,076	1
Total	146,460	100.0

the data into 132,201 training and 14,259 test samples with no overlap between partitions. All images are standardized to a resolution of 768×1024.

Editing Distribution. We define seven editing types, grouped into two macro-categories: **appearance edits**, comprising *Change Color*, *Change Pattern*, *Change Material*, and *Fine-Grained*, which modify the visual properties of the garment; and **structural edits**, comprising *Add Detail*, *Remove Element*, and *Modify Structure*, which alter its geometry or composition by adding, removing, or reshaping garment parts. Overall, the dataset exhibits substantial diversity across garment categories and edit types, establishing a comprehensive benchmark for instruction-driven VTON and VTOFF research. Table 2 provides an overview of the editing instruction statistics while a visual taxonomy of the edits is presented in the supplementary material.

3.3 Dataset Comparison

Existing fashion datasets [9, 11, 18, 31, 39] primarily target either VTON or VTOFF, but none include edited counterparts or support instruction-driven transformations. For instance, VITON-HD [9] and Dress Code [39] contain real garment–person pairs but lack text-guided edits. At the same time, general-purpose editing datasets [7,24,44,62,66,69] are built upon data triplets including the editing instruction, but lack domain-specific pairs of garment-person data that suit VTON and VTOFF. EditGarment [64] collects a dataset of garment-only editing pairs, while ImgEdit [63] contains garment data as a subsection of their dataset, lacking structured garment-person association. In contrast, our dataset introduces a novel visual quadruplet structure $(I_{\text{garment}}, I_{\text{person}}, I^{\text{edit}}_{\text{garment}}, I^{\text{edit}}_{\text{person}})$ paired with a natural-language instruction T_{inst}, where each edited garment and corresponding try-on image are coherently generated and verified. This formulation enables a range of instruction-driven fashion editing tasks, including edited VTON, edited VTOFF, person editing, and garment editing, while remaining compatible with standard VTON and VTOFF setups. A detailed comparison between Dress-ED and existing fashion and general-purpose datasets is presented in Table 1.

3.4 Benchmark Tasks and Evaluation

While Dress-ED can support a variety of instruction-driven fashion editing tasks, in this work we focus on two core benchmarks that reflect its primary objectives, *i.e.*, *instruction-driven VTON* and *instruction-driven VTOFF*. Each benchmark targets a specific aspect of instruction-based fashion editing, either the correct rendering of edited garments on people or the accurate generation of these garments in a catalog style. To evaluate both visual realism and semantic edit correctness, we introduce a dual evaluation protocol, detailed below.

Instruction-Driven Virtual Try-On. This benchmark evaluates the ability of a model to generate a realistic edited person image $I^{\text{edit}}_{\text{person}}$ conditioned on an edit instruction T_{inst} and garment context. Following standard VTON literature [9,39],

two complementary configurations are considered. In the *paired setting*, the model receives $(I_{\text{person}}, I_{\text{garment}}, T_{\text{inst}})$ and must generate $I^{\text{edit}}_{\text{person}}$ where the same garment worn by the person is modified according to T_{inst}. This configuration measures localized editing ability while preserving identity, pose, and background. In the *unpaired setting*, the model is provided with $(I_{\text{person}}, I'_{\text{garment}}, T_{\text{inst}})$, where I'_{garment} refers to a different garment image. The goal is to synthesize the person wearing the garment I'_{garment} edited as described by T_{inst}, thereby assessing practical generalization to real-world virtual shopping scenarios.

Instruction-Driven Virtual Try-Off. This benchmark focuses on garment-level editing in the catalog domain. Given a person image I_{person} and a textual instruction T_{inst}, the model must generate an edited catalog garment image $I^{\text{edit}}_{\text{garment}}$. The instruction specifies both the garment identity and the intended modification (*e.g.*, "make the skirt red"), and the target supervision is the corresponding edited in-shop image provided in Dress-ED. This task isolates the capacity to semantically manipulate garment appearance and structure independent of person-specific context.

On Using Real Images as Ground-truth Data. For rigorous quantitative evaluation, the test phase employs an inverse-editing protocol. The model is tasked with reconstructing the original, unedited person image I_{person} from its edited counterpart $I^{\text{edit}}_{\text{person}}$ using an inverse instruction T^{r}_{inst}. This guarantees that reference-based metrics are calculated strictly against real-world ground truth, accurately measuring localized editing ability while preserving identity, pose, and background without synthetic bias. We adopt this evaluation setting for all of our tasks (VTON paired, unpaired, and VTOFF).

Evaluation Metrics. We evaluate all benchmarks using two complementary metric families that jointly quantify visual fidelity and instruction adherence.

- **Visual Fidelity.** We employ standard perceptual and distributional metrics, including FID [42], KID [6], SSIM [54], LPIPS [67], and DISTS [13], computed independently for both VTON and VTOFF outputs.
- **Edit Correctness.** To evaluate semantic alignment with the editing instruction, we compute DINO-I [41] similarity between the generated image and the ground-truth image to measure content-level consistency.

Benchmark Protocol. All evaluations are conducted on the held-out test split (14,259 samples), with no overlap of garment identities across splits. For VTON, both paired and unpaired settings are reported separately. This unified protocol provides a standardized basis for evaluating instruction-driven fashion editing across VTON and VTOFF paradigms.

4 Proposed Method Towards Editing

We propose a unified multimodal diffusion architecture for instruction-driven editing, termed as **Dress-EM** (**Dress Editing Model**) and applicable to both VTON and VTOFF tasks. Our framework unifies a pretrained multimodal large

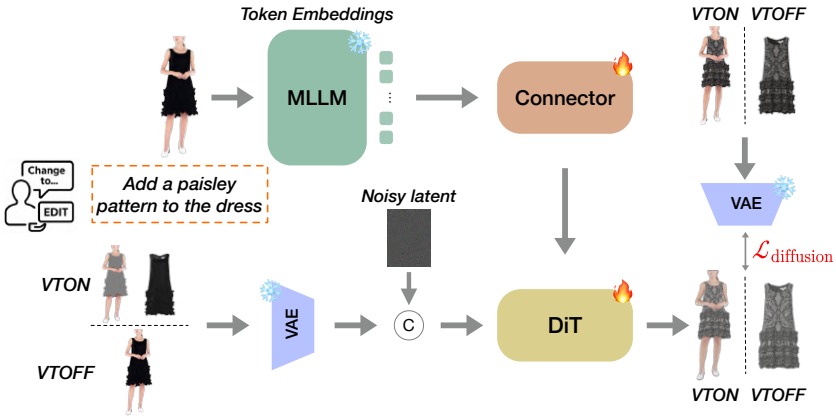

Fig. 3: Overview of the proposed Dress-EM architecture for instruction-driven fashion editing.

language model (MLLM), a lightweight connector, and a diffusion transformer (DiT) into a cohesive architecture, as illustrated in Fig. 3. The only distinction between the two tasks lies in the inputs provided to the VAE encoder: for VTOFF, the input is the original person image, while for VTON, the encoder receives both the masked person and garment features, depending on the paired or unpaired setting.

Architecture Overview. Given a visual reference I, the image of the original person, and a text instruction t, the MLLM (*i.e.*, InternVL-3.5 [53]) jointly encodes the multimodal context through a single forward pass. This produces a sequence of fused token embeddings $\boldsymbol{E}_{\text{MLLM}} = \{e_1, \ldots, e_N\}$ capturing fine-grained cross-modal relationships between the garment, the person, and the desired modification. To isolate semantically relevant tokens, we retain only those aligned with the instruction and the image, discarding system prefix and irrelevant embeddings.

Connector. The connector is a lightweight Transformer component that bridges the MLLM feature space and the DiT conditioning space. Specifically, it compresses $\boldsymbol{E}_{\text{MLLM}}$ into a multimodal latent representation $\boldsymbol{z}_c = f_{conn}(\boldsymbol{E}_{\text{MLLM}})$ and projects the mean pooled token embedding through a linear layer to produce a global guidance vector $\boldsymbol{g} \in \mathbb{R}^{2048}$ and a Token Refinement Pathway $z_c \in \mathbb{R}^{77\times4096}$.

Global Guidance Pathway (g). This pathway extracts a high-level semantic summary of the multimodal context produced by the MLLM. The connector aggregates the MLLM tokens via masked mean pooling to obtain a single context representation, which is then normalized and projected through a linear layer to produce a 2048-dimensional global guidance vector g.

Token Refinement Pathway (z_c). This pathway preserves fine-grained multimodal information by refining the full sequence of MLLM tokens. The tokens are processed by a lightweight stack of Transformer blocks that resolve cross-modal interactions and produce a refined multimodal embedding $z_c \in \mathbb{R}^{77\times4096}$.

Diffusion Transformer. Following the DiT paradigm [45], the edited image is generated in the latent space of a pretrained VAE. The DiT receives two complementary conditioning sources: *(i)* the multimodal embedding $\boldsymbol{z}_c$ and a global guidance vector $\boldsymbol{g}$ produced by the MLLM and connector, which encode the semantic action of the edit, and *(ii)* the VAE latent features $\boldsymbol{z}_{\text{VAE}}$, which capture the spatial and structural details of the visual reference. For the VTON task, the $\boldsymbol{z}_{\text{VAE}}$ is the concatenation along the width dimension of the latents of the masked person and the latents of the garment. Instead, for the VTOFF task, it is the latent of the garment. These signals allow the DiT to understand what to modify and where to apply it jointly. Formally, the denoising process is defined as

$$\hat{\boldsymbol{x}}_{t-1} = \text{DiT}(\boldsymbol{x}_t, \boldsymbol{z}_{\text{VAE}}, \boldsymbol{z}_c, \boldsymbol{g}, t), \tag{7}$$

where the DiT predicts the noise residual conditioned on both semantic and visual features. After denoising, the recovered latent $\boldsymbol{x}_0$ is passed through the VAE decoder to obtain the final edited image I^{edit}.

Optimization Objective. The model is optimized via the standard diffusion loss $\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t, t}\left[\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|^2\right]$.

5 Experiments

We conduct extensive experiments on Dress-ED to validate our proposed approach and to establish benchmark results for instruction-driven fashion editing. All evaluations follow the benchmark protocols introduced in Sec. 3.4. Results are reported for instruction-driven VTON in both paired and unpaired configurations, as well as instruction-driven VTOFF. We use the real images as the ground truth to ensure a proper evaluation that is not influenced by the quality upper bounds of the ground-truth synthetic images. Therefore, given a synthetic image $I^{\text{edit}}_{\text{person}}$, we use the instruction T^r_{inst} to edit the image in order to predict I_{person}.

5.1 Implementation and Training Details

All models are trained on the proposed dataset using a resolution of 768×1024. We employ AdamW [37] as optimizer, using a cosine scheduler with a learning rate of 1×10^{-4} and a batch size of 4 for 50k steps. We use DeepSpeed [46] and train on 16 NVIDIA A100 GPUs. As diffusion backbone, we start from Stable Diffusion 3 medium [14], while we employ InternVL-3.5-8B [53] as MLLM. During training, the VAE and the MLLM are kept frozen, while the connector and the DiT are trained end-to-end. During training, we utilize both original and edited images as ground-truth targets. This mixed supervision strategy forces the model to maintain robust alignment with real-world textures while simultaneously learning complex, instruction-driven structural and appearance modifications.

5.2 Baselines and Competitors

We compare our method against both task-specific and general-purpose diffusion-based approaches relevant to VTON, VTOFF, and instruction-guided editing.

Table 3: Quantitative results on instruction-driven VTON (paired, unpaired) and VTOFF on Dress-ED. We report FID, KID, SSIM, LPIPS, DISTS, and DINO-I. ↓ indicates lower is better, ↑ indicates higher is better.

	Paired						Unpaired		
Method	SSIM ↑	LPIPS ↓	DISTS ↓	FID ↓	KID ↓	DINO-I ↑	FID ↓	KID ↓	DINO-I ↑
Edited VTON									
FLUX.2 Klein [29]	0.9291	0.0743	0.0853	4.32	1.71	0.9302	6.29	1.80	**0.6549**
Qwen-Image-Edit [56]	0.9173	0.0963	0.1135	8.51	3.42	0.8802	10.19	3.24	0.6453
Any2AnyTryon [20]	**0.9429**	0.0702	0.0795	5.34	1.98	0.9074	6.36	1.94	0.4996
CatVTON [10]	0.9314	0.0801	0.1022	4.57	1.72	0.9144	6.40	1.91	0.6476
Dress-EM (Ours)	0.9417	**0.0628**	**0.0733**	**3.79**	**1.13**	**0.9383**	**5.82**	**1.44**	0.6476
Edited VTOFF									
FLUX.2 Klein [29]	0.8694	0.2800	0.2579	17.45	7.48	0.7468	-	-	-
Qwen-Image-Edit [56]	0.8463	0.2970	0.2792	24.09	9.56	0.7192	-	-	-
Any2AnyTryon [20]	0.8757	0.2878	0.2474	19.03	6.26	0.6864	-	-	-
CatVTON [10]	**0.8800**	0.2415	0.2058	6.78	1.45	0.7918	-	-	-
Dress-EM (Ours)	**0.8800**	**0.2060**	**0.1892**	**5.06**	**0.73**	**0.8071**	-	-	-

For fairness, all task-specific models are retrained on Dress-ED using either their official source codes, when available, or faithful re-implementations following the original architectures and training protocols. For all comparisons, input resolution and evaluation splits are kept identical across methods.

VTON and VTOFF Models. We include recent state-of-the-art approaches for standard VTON and VTOFF, namely CatVTON [10] and Any2AnyTryon [20]. Since CatVTON does not natively support text input, we extend its architecture by incorporating instruction embeddings following standard conditioning practices in diffusion-based editing. This adaptation enables fair comparison under instruction-driven settings. To ensure consistency, CatVTON is retrained with Stable Diffusion 3 (medium variant), matching the pretrained backbone used in our Dress-EM model.

General-purpose Editing Models. We further evaluate large-scale multimodal diffusion models in a zero-shot setting to examine their capability for instruction-driven garment manipulation. Specifically, we include FLUX.2 Klein [29] and Qwen-Image-Edit [56], both capable of text-guided editing across generic domains. These models serve as strong foundation-level baselines for assessing semantic alignment and edit fidelity in complex fashion edits. A critical challenge is to extend these models to multi-image scenarios such as VTON. We address this scenario by concatenating image tokens at the input of the DiT [45] for both models, while preserving their spatial relationships through positional encoding.

5.3 Editing Results

We report quantitative and qualitative comparisons between Dress-EM and competing methods across all benchmark settings, using the metrics defined in Sec. 3.4.

Quantitative Results. Table 3 summarizes quantitative results across all editing tasks and settings. Overall, Dress-EM consistently outperforms all other methods,

Table 4: Per-edit-type quantitative results for instruction-driven VTON and VTOFF on Dress-ED. We report DISTS (↓) and DINO-I (↑) for the categories Change Color, Change Pattern, Change Material, Fine-Grained, Add Detail, Remove Element, and Modify Structure.

Method	Color		Pattern		Material		Fine-Gr.		Add		Remove		Struct	
	DISTS	DINO	DISTS	DINO	DISTS	DINO	DISTS	DINO	DISTS	DINO	DISTS	DINO	DISTS	DINO
Edited VTON														
FLUX.2 Klein [29]	0.081	0.939	0.100	0.908	0.085	0.939	0.081	0.936	0.075	0.941	0.084	0.935	0.078	0.943
Qwen-Image-Ed [56]	0.106	0.896	0.171	0.820	0.093	0.893	0.089	0.922	0.086	0.894	0.082	0.926	0.080	0.941
Any2AnyTryon [20]	0.070	0.937	0.099	0.865	0.072	0.923	0.076	0.917	0.074	0.906	**0.068**	0.940	0.074	0.929
CatVTON [10]	0.096	0.929	0.117	0.889	0.099	0.922	0.098	0.912	0.096	0.919	0.098	0.931	0.097	0.921
Dress-EM (Ours)	**0.066**	**0.949**	**0.087**	**0.920**	**0.070**	**0.942**	**0.072**	**0.938**	**0.067**	**0.944**	0.070	**0.947**	**0.069**	**0.946**
Edited VTOFF														
FLUX.2 Klein [29]	0.252	0.757	0.299	0.707	0.245	0.755	0.227	0.790	0.237	0.754	0.255	0.728	0.232	0.816
Qwen-Image-Ed [56]	0.267	0.757	0.379	0.650	0.238	0.757	0.233	0.761	0.227	0.725	0.260	0.729	0.221	0.766
Any2AnyTryon [20]	0.226	0.765	0.267	0.699	0.220	0.769	0.228	0.775	0.238	0.707	0.232	0.745	0.227	0.762
CatVTON [10]	0.210	0.807	0.236	0.747	0.197	0.817	0.197	0.802	0.181	0.807	0.204	0.804	0.187	0.808
Dress-EM (Ours)	**0.180**	**0.828**	**0.219**	**0.771**	**0.183**	**0.827**	**0.181**	**0.821**	**0.174**	**0.811**	**0.187**	**0.828**	**0.172**	**0.819**

demonstrating strong generalization across both paired and unpaired scenarios. In the paired edited VTON setting, Dress-EM achieves the best results across nearly all metrics, including the lowest FID, KID, LPIPS, and DISTS values. These results indicate superior image realism and faithful preservation of garment structure and identity compared to CatVTON [10] and Any2AnyTryon [20]. The strong DINO-I further validates the ability of the model to generate precise and semantically consistent edits. In the more challenging unpaired edited VTON configuration, specialized models demonstrate robust performance, maintaining a competitive edge even when handling garments unseen during the original try-on process. Dress-EM maintains high DINO-I value despite the increased complexity, confirming the robustness of its MLLM-guided conditioning and connector design.

For the edited VTOFF task, which reconstructs the edited in-shop garment from the dressed person, Dress-EM confirms the best overall performance across nearly all metrics, including the lowest FID, KID, and DISTS, and the highest DINO-I. These results demonstrate its strong capability to recover fine garment details and produce semantically faithful edits.

Per-Edit-Type Analysis. Table 4 reports a fine-grained breakdown across the seven edit categories defined in Dress-ED, evaluated using DISTS and DINO-I metrics. Dress-EM consistently achieves the lowest DISTS values (indicating higher perceptual similarity) and the highest DINO-I scores across nearly all edit types for both VTON and VTOFF. Notably, the largest improvements occur in appearance-driven edits such as color, pattern, and material, where multimodal conditioning effectively captures localized texture and tone changes. Structural modifications, including sleeve length and shape adjustments, show consistent quantitative gains, indicating that our model effectively preserves spatial coherence while performing localized structural edits.

Qualitative Results. A visual comparison is presented in Fig. 4, which showcases representative examples across both VTON (unpaired) and VTOFF settings. As it can be seen, Dress-EM produces high-quality and photorealistic results

Fig. 4: Qualitative results for edited VTON unpaired setting (first two rows) and edited VTOFF (last two rows), showing realistic and instruction-consistent edits.

that accurately follow the editing instructions while preserving person identity, pose, and garment details, further confirming the effectiveness of the proposed approach.

6 Conclusion

We introduced Dress-ED, the first large-scale dataset and benchmark for instruction-driven fashion editing that unifies VTON, VTOFF, and text-guided garment modification within a single framework. Built through a fully automated multimodal pipeline integrating MLLM-based garment understanding, diffusion-based synthesis, and LLM-guided verification, Dress-ED comprises over 146k verified, semantically consistent samples across diverse edit types. Beyond dataset construction, it establishes unified benchmarks and a multimodal diffusion baseline, enabling systematic evaluation of controllable and interpretable fashion generation. We hope this work fosters future research on multimodal garment understanding and interactive editing.

Acknowledgments. This work was supported by EU Horizon projects ELIAS (No. 101120237) and ELLIOT (No. 101214398), and by the FIS project GUIDANCE (No. FIS2023-03251). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources.

References

1. AI, F.: FASHN Human Parser: SegFormer for Fashion Human Parsing. https://huggingface.co/fashn-ai/fashn-human-parser (2024) 19
2. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable Flow: Vital Layers for Training-Free Image Editing. In: CVPR (2025) 4
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025) 3, 6, 19
4. Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single Stage Virtual Try-On Via Deformable Attention Flows. In: ECCV (2022) 3
5. Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing. In: ICCV (2023) 2, 3
6. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. arXiv preprint arXiv:1801.01401 (2018) 9
7. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: CVPR (2023) 2, 4, 5, 8
8. Chen, C.Y., Chen, Y.C., Shuai, H.H., Cheng, W.H.: Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network. In: ICCV (2023) 3
9. Choi, S., Park, S., Lee, M., Choo, J.: VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In: CVPR (2021) 2, 7, 8
10. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: CatVTON: Concatenation Is All You Need for Virtual Try-On with Diffusion Models. In: ICLR (2025) 2, 3, 12, 13, 23
11. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street TryOn: Learning In-the-Wild Virtual Try-On from Unpaired Person Images. In: WACV (2024) 7, 8
12. Cui, A., McKee, D., Lazebnik, S.: Dressing in Order: Recurrent Person Image Generation for Pose Transfer, Virtual Try-On and Outfit Editing. In: ICCV (2021) 3
13. Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image Quality Assessment: Unifying Structure and Texture Similarity. IEEE Trans. PAMI **44**(5), 2567–2581 (2020) 9
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICML (2024) 11
15. Fele, B., Lampe, A., Peer, P., Struc, V.: C-VTON: Context-Driven Image-Based Virtual Try-On Network. In: WACV (2022) 3
16. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding Instruction-based Image Editing via Multimodal Large Language Models. In: ICLR (2024) 4
17. Garibi, D., Patashnik, O., Voynov, A., Averbuch-Elor, H., Cohen-Or, D.: ReNoise: Real Image Inversion Through Iterative Noising. In: ECCV (2024) 4

18. Ge, Y., Zhang, R., Wang, X., Tang, X., Luo, P.: DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In: CVPR (2019) 7, 8
19. Girella, F., Talon, D., Liu, Z., Ruan, Z., Wang, Y., Cristani, M.: LOTS of Fashion! Multi-Conditioning for Image Generation via Sketch-Text Pairing. In: ICCV (2025) 3
20. Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., Liu, J.: Any2AnyTryon: Leveraging Adaptive Position Embeddings for Versatile Virtual Clothing Tasks. In: ICCV (2025) 4, 12, 13, 23
21. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: VITON: An Image-Based Virtual Try-On Network. In: CVPR (2018) 2, 3
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022) 5
23. Huang, S., Li, H., Zheng, C., Ge, M., Gao, W., Wang, L., Liu, L.: Text-Driven Fashion Image Editing with Compositional Concept Learning and Counterfactual Abduction. In: CVPR (2025) 2, 5
24. Hui, M., Yang, S., Zhao, B., Shi, Y., Wang, H., Wang, P., Zhou, Y., Xie, C.: HQ-Edit: A High-Quality Dataset for Instruction-based Image Editing. In: ICLR (2025) 2, 7, 8
25. Jiang, B., Hu, X., Luo, D., He, Q., Xu, C., Peng, J., Zhang, J., Wang, C., Wu, Y., Fu, Y.: FitDiT: Advancing the Authentic Garment Details for High-fidelity Virtual Try-on. In: CVPR (2025) 2, 3, 6, 19, 26
26. Jiao, G., Huang, B., Wang, K.C., Liao, R.: UniEdit-Flow: Unleashing Inversion and Editing in the Era of Flow Models. arXiv preprint arXiv:2504.13109 (2025) 4
27. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: ECCV (2024) 19
28. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In: CVPR (2024) 2, 3
29. Labs, B.F.: FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2 (2025) 3, 4, 6, 12, 13, 23, 26
30. Lee, S., Kwak, J.g.: Voost: A Unified and Scalable Diffusion Transformer for Bidirectional Virtual Try-On and Try-Off. In: SIGGRAPH Asia (2025) 3
31. Lepage, S., Mary, J., Picard, D.: LRVS-Fashion: Extending Visual Search with Referring Instructions. arXiv preprint arXiv:2306.02928 (2023) 8
32. Li, Q., Qiu, S., Han, J., Xu, X., Seyfioglu, M.S., Koo, K.K., Bouyarmane, K.: DiT-VTON: Diffusion Transformer Framework for Unified Multi-Category Virtual Try-On and Virtual Try-All with Integrated Image Editing. In: CVPR Workshops (2025) 2, 3
33. Li, Y., Zhou, H., Shang, W., Lin, R., Chen, X., Ni, B.: Anyfit: Controllable virtual try-on for any combination of attire across any scenario. In: NeurIPS (2024) 2
34. Liu, J., He, Z., Wang, G., Li, G., Lin, L.: One Model For All: Partial Diffusion for Unified Try-On and Try-Off in Any Pose. arXiv preprint arXiv:2508.04559 (2025) 3
35. Liu, S., Han, Y., Xing, P., Yin, F., Wang, R., Cheng, W., Liao, J., Wang, Y., Fu, H., Han, C., Li, G., Peng, Y., Sun, Q., Wu, J., Cai, Y., Ge, Z., Ming, R., Xia, L., Zeng, X., Zhu, Y., Jiao, B., Zhang, X., Yu, G., Jiang, D.: Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761 (2025) 4
36. Lobba, D., Sanguigni, F., Ren, B., Cornia, M., Cucchiara, R., Sebe, N.: Inverse Virtual Try-On: Generating Multi-Category Product-Style Images from Clothed Individuals. In: ICLR (2026) 2, 3

37. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: ICLR (2019) 11
38. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In: ACM Multimedia (2023) 3
39. Morelli, D., Matteo, F., Marcella, C., Federico, L., Fabio, C., Rita, C.: Dress Code: High-Resolution Multi-Category Virtual Try-On. In: ECCV (2022) 2, 4, 5, 7, 8, 19, 21
40. OpenAI: Introducing GPT-5 (2025) 3, 4, 6, 24
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193 (2023) 9
42. Parmar, G., Zhang, R., Zhu, J.Y.: On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In: CVPR (2022) 9
43. Patel, M., Wen, S., Metaxas, D.N., Yang, Y.: Steering Rectified Flow Models in the Vector Field for Controlled Image Generation. In: ICCV (2025) 4
44. Pathiraja, B., Patel, M., Singh, S., Yang, Y., Baral, C.: Refedit: A benchmark and method for improving instruction-based image editing model on referring expressions. In: ICCV (2025) 7, 8
45. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: ICCV (2023) 3, 11, 12
46. Rajbhandari, S., Rasley, J., Ruwase, O., He, Y.: ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. In: SC (2021) 11, 19
47. Ren, B., Tang, H., Meng, F., Runwei, D., Torr, P.H., Sebe, N.: Cloth Interactive Transformer for Virtual Try-On. ACM TOMM **20**(4), 1–20 (2023) 3
48. Rout, L., Chen, Y., Ruiz, N., Caramanis, C., Shakkottai, S., Chu, W.S.: Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In: ICLR (2025) 4
49. Sheynin, S., Polyak, A., Singer, U., Kirstain, Y., Zohar, A., Ashual, O., Parikh, D., Taigman, Y.: Emu edit: Precise image editing via recognition and generation tasks. In: CVPR (2025) 4
50. Velioglu, R., Bevandic, P., Chan, R., Hammer, B.: TryOffDiff: Virtual-Try-Off via High-Fidelity Garment Reconstruction using Diffusion Models. In: BMVC (2024) 2, 3
51. Velioglu, R., Bevandic, P., Chan, R., Hammer, B.: MGT: Extending Virtual Try-Off to Multi-Garment Scenarios. In: ICCV Workshops (2025) 3
52. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward Characteristic-Preserving Image-based Virtual Try-On Network. In: ECCV (2018) 3
53. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint arXiv:2508.18265 (2025) 3, 6, 10, 11, 25
54. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Processing **13**(4), 600–612 (2004) 9
55. Wei, C., Xiong, Z., Ren, W., Du, X., Zhang, G., Chen, W.: OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision. In: ICLR (2025) 4
56. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-Image Technical Report. arXiv preprint arXiv:2508.02324 (2025) 4, 12, 13, 23

57. Xarchakos, I., Koukopoulos, T.: TryOffAnyone: Tiled Cloth Generation from a Dressed Person. arXiv preprint arXiv:2412.08573 (2024) 3
58. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: GP-VTON: Towards General Purpose Virtual Try-on via Collaborative Local-Flow Global-Parsing Learning. In: CVPR (2023) 3
59. Xie, Z., Li, H., Ding, H., Li, M., Cao, Y.: HieraFashDiff: Hierarchical Fashion Design with Multi-stage Diffusion Models. In: AAAI (2024) 4
60. Yan, K., Gao, T., Zhang, H., Xie, C.: Linking garment with person via semantically associated landmarks for virtual try-on. In: CVPR (2023) 3
61. Yang, L., Zeng, B., Liu, J., Li, H., Xu, M., Zhang, W., Yan, S.: EditWorld: Simulating World Dynamics for Instruction-Following Image Editing. In: ACM Multimedia (2024) 4
62. Yang, S., Hui, M., Zhao, B., Zhou, Y., Ruiz, N., Xie, C.: Complex-Edit: CoT-Like Instruction Generation for Complexity-Controllable Image Editing Benchmark. arXiv preprint arXiv:2504.13143 (2025) 8
63. Ye, Y., He, X., Li, Z., Lin, B., Yuan, S., Yan, Z., Hou, B., Yuan, L.: Imgedit: A unified image editing dataset and benchmark. In: NeurIPS (2025) 7, 8
64. Yin, D., Guo, J., Lu, H., Wu, F., Lu, D.: EditGarment: An Instruction-Based Garment Editing Dataset Constructed with Automated MLLM Synthesis and Semantic-Aware Evaluation. In: ACM Multimedia (2025) 2, 5, 7, 8
65. Yu, Q., Chow, W., Yue, Z., Pan, K., Wu, Y., Wan, X., Li, J., Tang, S., Zhang, H., Zhuang, Y.: Anyedit: Mastering unified high-quality image editing for any idea. In: CVPR (2025) 4, 7
66. Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. In: NeurIPS (2023) 2, 4, 7, 8
67. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018) 9
68. Zhang, S., Yang, X., Feng, Y., Qin, C., Chen, C.C., Yu, N., Chen, Z., Wang, H., Savarese, S., Ermon, S., et al.: HIVE: Harnessing Human Feedback for Instructional Visual Editing. In: CVPR (2024) 4
69. Zhao, H., Ma, X.S., Chen, L., Si, S., Wu, R., An, K., Yu, P., Zhang, M., Li, Q., Chang, B.: Ultraedit: Instruction-based fine-grained image editing at scale. In: NeurIPS (2024) 2, 7, 8
70. Zhou, Z., Liu, S., Han, X., Liu, H., Ng, K.W., Xie, T., Cong, Y., Li, H., Xu, M., Pérez-Rúa, J.M., Patel, A., Xiang, T., Shi, M., He, S.: Learning Flow Fields in Attention for Controllable Person Image Generation. In: CVPR (2025) 3
71. Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&m vto: Multi-garment virtual try-on and editing. In: CVPR (2024) 2, 3, 4
72. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: TryOnDiffusion: A Tale of Two UNets. In: CVPR (2023) 2, 3

A Additional Implementation Details

Training Configuration. During training, we employ DeepSpeed [46] for memory optimization and use `bfloat16` precision to ensure training stability and efficiency. As mentioned in the main paper, the weights of the VAE and the MLLM are kept frozen. Only the connector module and the DiT parameters are optimized end-to-end.

B Additional Dataset Details

B.1 Mask Preprocessing

FitDiT uses bounding boxes as input masks. These bounding boxes are extracted from the label maps of a clothed person extracted using a parser. A visual example is provided in Fig. 5 The label maps provided in Dress Code [39] have significant inconsistencies in the hand area, partially masking hands too. More powerful segmentation networks such as Sapiens [27] are better at segmenting body parts but struggle to isolate garments from each other, resulting in artifacts, for example upper-body and lower-body items labeled as one. For this reason, we employ a new segmentation model [1], specifically trained on garment data and capable of isolating body parts. We then extract the bounding box mask from the new segmentation maps and apply FitDiT [25] to virtually try-on the edited ground-truth garment to the person, while preserving hands.

Fig. 5: Sample of label maps, masks, in-shop garment and FitDiT result.

B.2 Rule-based Templates

The dataset generation pipeline employs Qwen3-VL [3] to extract garment attributes, which are converted into paired editing instructions via rule-based templates (T_{inst}). For each valid edit, we generate both a forward and a reverse instruction (to undo or invert the transformation), with category-aware constraints for upper-body, lower-body, and dresses. Our final dataset includes seven distinct edit types.

Change Color. This edit is generated only for solid garments. A target color is sampled from a predefined palette, excluding colors already present in the source description: "*Change the color of the garment to* `[Target Color]`".

Change Pattern. A new pattern is sampled from a predefined pattern set (excluding the source pattern). The instruction depends on the source pattern: if the source is solid, the edit is additive: "*Add a* `[Target Pattern]` *pattern to the garment*"; otherwise, it is a replacement: "*Replace the* `[Source Pattern]` *pattern with a* `[Target Pattern]` *pattern*".

Change Material. This edit is also restricted to solid garments. The target material is sampled from a material pool excluding the source, while preserving garment color in the prompt: "*Make the garment* `[Garment Color] [Target Material]`".

Modify Structure. A structural attribute (*e.g.*, sleeves or neckline) is selected from category-specific valid attributes, and a different target value is sampled: "*Change the* `[Attribute]` *to* `[Target Value]`".

Add Detail. A new distinctive feature is sampled from category-specific addable features, excluding features already present on the garment. Generic form: "*Add a* `[Feature]` *to the garment*", or, when location is defined: "*Add a* `[Feature]` *at the* `[Location]` *of the garment*".

Remove Element. A removable feature is selected from *verified* distinctive features that satisfy category and location-specific constraints: "*Remove the* `[Feature]` *from the garment*".

Fine-Grained. It consists of coloring small details or feature replacement. For *Color Detail*, a color-editable feature is selected from *verified* distinctive features under category/location constraints, and recolored to a new color: "*Change the color of the* `[Feature]` *to* `[Target Color]`". Instead, for *Replace Detail*, a feature replacement follows predefined category-specific
source→target pairs (*e.g.*, bow→buttons, belt→drawstring), used only when the required source feature (and location, if needed) exists: "*Replace the* `[Source Feature]` *with* `[Target Feature]`".

B.3 Attribute Selection

To ensure diversity while preserving plausibility, template slots are populated from predefined attribute banks with category-aware constraints:

Colors (global and detail-level): *red*, *blue*, *green*, *yellow*, *black*, *white*, *purple*, *orange*, *pink*, *brown*, *gray*, *navy*, *beige*, *lavender*, *maroon*, *teal*.

Patterns: *polka dot*, *striped*, *floral*, *geometric*, *checkered*, *paisley*, *camouflage*, *leopard print*, *gingham*, *houndstooth*, *abstract*.

Materials: *denim*, *leather*, *silk*, *wool*, *cotton*, *linen*.

Structural Attributes: *long sleeves*, *short sleeves*, *sleeveless*, *cap sleeves*, *puffy sleeves*, *three-quarter sleeves*, *v-neck*, *crew neck*, *scoop neck*, *square neck*, *high neck*, *collar*.

Addable Details (category-specific)
- *Upper-body: bow* (neck), *pockets*, *chest pocket*, *zipper*, *buttons*.
- *Lower-body: belt*, *drawstring*.

Fig. 6: Samples from Dress-ED across the different edit types, including the original person-garment pair, the corresponding edited outputs, and the textual instruction.

– *Dresses: belt*, *fitted belt*, *pockets*, *drawstring*, *bow* (neck/waist).

Removable/Colorable/Replaceable Details: sampled from verified distinctive features under category- and location-specific rules (*e.g.*, *logo*, *text*, *zipper*, *drawstring*, *pocket*, *embroidery*, *fringe*, *belt*, *bow*, *buttons*, *sash*, *stripes*, *feather trim*), including predefined replacement pairs (*e.g.*, *belt* ↔ *bow*, *belt* ↔ *drawstring*, *bow* ↔ *buttons*).

B.4 Dataset Examples and Statistics

To provide further details on Dress-ED, we report the visual taxonomy of our data in Fig. 7. As mentioned in the main paper, the dataset follows the standard split of Dress Code [39], ensuring no overlap of garment identities between training and testing sets. Table 5 reports the distribution of samples across garment categories. Additional examples from Dress-ED are provided in Fig. 6, showing the original

Category	Count	(%)
Dresses	80,865	55.2
Upper-body	45,567	31.1
Lower-body	20,028	13.7
Total	146,460	100.0

Table 5: Sample distribution by garment category.

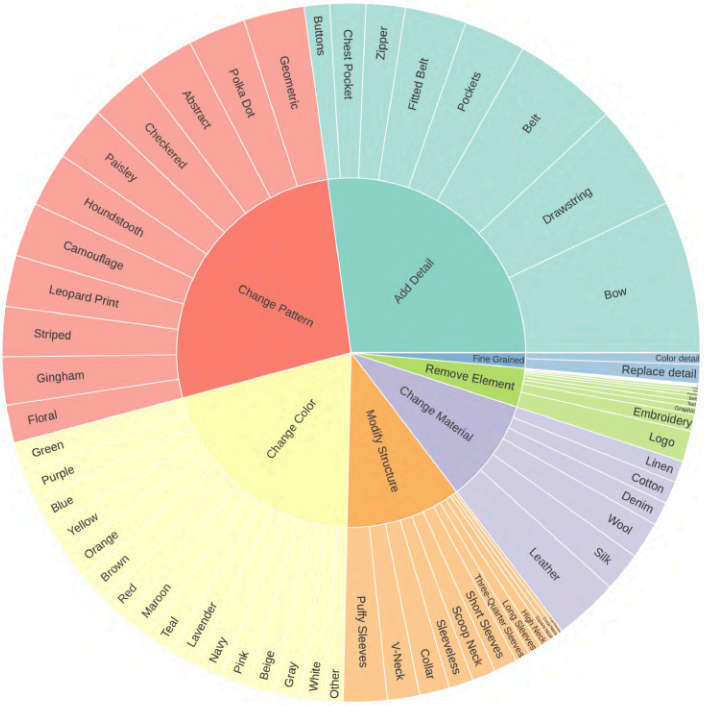

Fig. 7: Editing instruction taxonomy showing the distribution of edit types in Dress-ED, covering both appearance-level changes (*e.g.*, color, pattern, material) and structural modifications (*e.g.*, sleeve length, neckline, shape).

person-garment pair, the corresponding instruction, and the resulting edited outputs.

C Additional Quantitative Results

Forward vs. Inverse Editing. To address artifact inversion and validate the generalization of our approach, we compare forward editing (real → synthetic) and inverse editing (synthetic → real, as in the main protocol). Results are shown in Table 6, considering both edited VTON and VTOFF on the paired test set of Dress-ED. As it can be seen, this comparison shows almost identical results across directions, indicating that the model learns semantic editing rather than simply reversing artifacts. This is further supported by our mixed supervision strategy, which includes real images as targets, preventing collapse to the distribution of synthetic images.

Comparison with Commercial Models. To assess the performance of commercial models, we compare Dress-EM with GPT-Image-1.5 on a random 100-sample subset of the test set (paired edited VTON). Results, reported in Table 7, show

Table 6: Forward vs. reverse editing on Dress-ED (paired).

	Edited VTON				Edited VTOFF			
	SSIM ↑	LPIPS ↓	DISTS ↓	DINO-I ↑	SSIM ↑	LPIPS ↓	DISTS ↓	DINO-I ↑
Forward	0.9317	0.0627	0.0724	0.9493	0.8454	0.2084	0.1871	0.8212
Inverse	0.9417	0.0628	0.0733	0.9383	0.8800	0.2060	0.1892	0.8071

Table 7: Comparison with commercial models on edited VTON.

	SSIM ↑	LPIPS ↓	DISTS ↓	DINO-I ↑
Dress-EM (Ours)	**0.9437**	**0.0622**	**0.0721**	**0.9415**
GPT-Image-1.5	0.9323	0.0838	0.0842	0.9406

that, despite their generality, frontier image editing models underperform compared to a domain-specific model trained on Dress-ED, highlighting the relevance of the proposed benchmark.

D Additional Qualitative Results

In Fig. 8, we provide additional qualitative comparisons for instruction-driven VTON and VTOFF across the range of edit types included in Dress-ED. We compare our Dress-EM model with task-specific methods retrained on Dress-ED (*i.e.*, CatVTON [10] and Any2AnyTryon [20]) as well as general-purpose editing models evaluated in a zero-shot setting (Qwen-Image-Edit [56] and FLUX.2 Klein [29]). Zero-shot models typically struggle to preserve garment identity, maintain viewpoint consistency, or follow the instruction faithfully. Retrained VTON/VTOFF baselines provide more coherent edits, but still exhibit artifacts or incomplete transformations. In contrast, Dress-EM produces visually realistic and instruction-consistent results across appearance and structural edits, confirming the trends observed in the quantitative and qualitative analyses reported in the main paper.

E Additional Discussion and Clarifications

This section provides further explanations on several components of our pipeline. We outline the motivations behind key design choices, describe their practical implications, and discuss limitations relevant to interpreting the dataset and evaluation results.

E.1 On the Use of Synthetic Edited Images

The edited garments and try-on results in Dress-ED are produced through a controlled multimodal generation pipeline. This choice is deliberate: instruction-driven fashion editing requires tightly aligned pairs of garment, person, instruction, and edited outputs, which are extremely difficult to obtain consistently through

Fig. 8: Qualitative results for edited VTON (left, unpaired setting) and VTOFF (right), showing realistic and instruction-consistent edits.

manual annotation. Generative editing ensures that the modification directly corresponds to the instruction and that the garment identity is preserved before and after editing, which is essential for evaluating semantic edit correctness.

Rather than replacing in-the-wild data, the synthetic construction complements it. Dress-ED provides a controlled and reproducible benchmark for measuring whether models follow the intended transformation, while remaining compatible with future real-world data collection or domain adaptation. In this sense, the dataset serves as a structured foundation on top of which broader and more diverse fashion-editing scenarios can be explored.

E.2 LLM-as-Judge and Evaluation Consistency

GPT-5 [40] is employed solely as a *non-generative evaluator*: it does not participate in any image synthesis, nor does it influence VTON/VTOFF models. Its role is restricted to assessing whether an edited result follows the instruction while preserving garment identity and visual plausibility. Using a strong evaluator at this stage helps establish clean supervision signals for filtering, especially for fine-grained structural edits that are difficult to evaluate automatically.

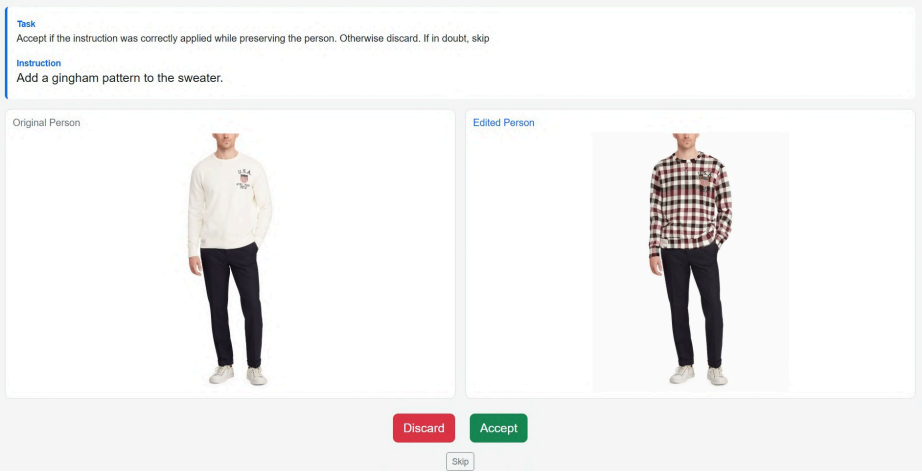

Fig. 9: User study interface.

To enable reproducible large-scale verification, we distill GPT-5's judgments into a smaller open-source MLLM (InternVL-3.5 [53]). Although this introduces some dependence between data construction and evaluation, our final benchmark reports model-agnostic evaluation metrics (*i.e.*, including DINO and standard perceptual scores) which remain independent of both GPT-5 and InternVL. This design ensures that LLM-based judging improves data reliability without dominating or biasing downstream evaluation.

User Study on MLLM Filtering Scores. To evaluate the coherence of our fine-tuned MLLM, we conducted a user study using the interface illustrated in Fig. 9. Participants were tasked with assessing the edit correctness of triplets drawn from our **unfiltered** data, $(T_{\text{inst}}, I_{\text{person}}, I^{\text{edit}}_{\text{person}})$, by choosing to either keep or discard each sample. We consider "good" an MLLM score > 80 and "bad" otherwise, as reported in the main paper.

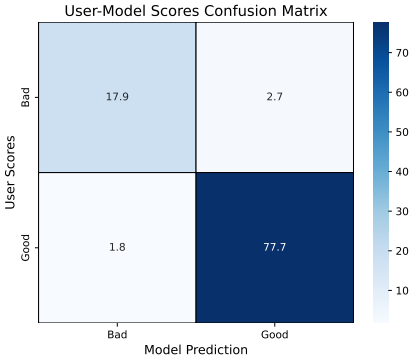

Fig. 10: Confusion matrix of our user study.

Based on 221 collected samples, we computed the confusion matrix between human annotations and MLLM predictions (Fig. 10). The results reveal a strong correlation between the fine-tuned MLLM and user preferences across both high- and low-quality samples. Notably, 77.7% of the images were classified as "good" by both humans and InternVL, while 17.9% were mutually identified as "bad", demonstrating the model's ability to identify noise in the raw data. In total, the MLLM achieved a 95.6% accuracy rate in replicating human judgment, validating its efficacy as a robust filtering tool.

E.3 Template-Based Instruction Generation

Our editing instructions are generated using rule-based templates grounded in the structured garment attributes extracted by the MLLM. While template-driven synthesis limits the linguistic variety compared to free-form human annotations and may introduce more regular phrasing, this structure is intentional. Instruction-driven fashion editing benefits from inputs that are precise, unambiguous, and tightly coupled to garment semantics; templates ensure that each instruction directly reflects a well-defined attribute change rather than a loosely worded or ambiguous request. More importantly, the template space spans a broad range of appearance and structural edits, and can be extended with minimal effort to support additional modification types or linguistic variations.

E.4 Limitation and Future Work

Our work establishes a strong baseline for instruction-driven fashion editing; however, certain design choices open exciting avenues for future exploration.

Data Curation and Domain Scalability. To overcome the prohibitive cost and scarcity of manually annotated editing pairs, Dress-ED prioritizes scalability through an automated synthesis pipeline. While this design choice enables the training of data-hungry diffusion models, the resulting distribution is bounded by the generative capabilities of the underlying teacher models (FLUX.2 Klein [29] and FitDiT [25]). We mitigate potential domain artifacts through our rigorous MLLM-based verification pipeline (filtering samples with a score lower than 80), ensuring high visual fidelity. Future works could further bridge the gap between synthetic and in-the-wild distributions by incorporating small-scale, real-world human feedback loops or few-shot adaptation on physical garment pairs.

Category Expansion. Our framework focuses on three primary garment categories (dresses, upper-body, and lower-body), which constitute the vast majority of e-commerce catalogs. While the proposed architecture is category-agnostic, extending the instruction taxonomy to handle highly occluded accessories (*e.g.*, scarves, jewelry) or complex multi-layer styling (*e.g.*, jackets over hoodies) remains a distinct challenge. We view this as a natural next step for the field, building upon the foundational editing capabilities established in this work.