Layering Virtual Try-On

Chun Feng, Bowei Chen, Mengyi Shan, and Ira Kemelmacher-Shlizerman

University of Washington, Seattle WA 98195, USA
{chunfeng, boweiche, shanmy, kemelmi}@cs.washington.edu

Fig. 1: We propose a Layering Virtual Try-On (LVTON) method that allows for swapping, adding, and removing clothing layers on any given person image. Given a person image, a garment image, and a textual instruction (e.g., swap, add, or remove), it generates a try-on image reflecting the desired garment edit while preserving inner layers when necessary. We showcase three sequential editing examples that exhibit natural layering, accurate garments, consistent facial identity, and realistic shading within complex backgrounds.

Abstract. In the real world, fashion is about layering: adding a jacket over a shirt, or a sequence of adding and removing layers, rather than just a single-layer swap. This fundamental real-world task remains a challenge in existing Virtual Try-On (VTON) methods, which excel at single-layer replacement but are not designed to layer or de-layer an existing outfit. This paper proposes Layering Virtual Try-On (LVTON), a layering benchmark and method that preserves an existing outfit while enabling sequential layering. We find that current VTON paradigms are fundamentally ill-equipped for LVTON, as their reliance on cloth-agnostic representations and single-item datasets discards essential layering context. Our key insight is that the LVTON challenge must be disentangled into two distinct competencies: (1) General VTON Priors (e.g., deformation,

identity preservation) and (2) Specific Layering Knowledge (e.g., layering order and occlusion reasoning). First, our model obtains general VTON priors by being trained on data produced by an automatic data generation pipeline that synthesizes samples from fashion videos via segmentation and inpainting. Second, the model is fine-tuned on a small, dedicated LVTON dataset to learn the layering logic. Our method achieves state-of-the-art results on our LVTON benchmark and demonstrates superior generalizability on traditional VTON benchmarks, setting new state-of-the-art results when fine-tuned and exhibiting zero-shot capabilities.

Keywords: Image Diffusion Models · Image Edit · Virtual Try-On

Fig. 2: Illustration of the Layering VTON (LVTON) challenge and the motivation for our two-stage paradigm. (a) LVTON task aims to composite a new garment (e.g., a sweater vest) over an existing outfit while preserving inner layers. (b) Traditional VTON paradigms mainly rely on a cloth-agnostic representation by masking out the original garment, which discards the inner layer. (c) The direct training dilemma: a model trained only on layering data (w/o Stage 1) learns correct composition logic (preserving the shirt) but fails to retain garment identity (e.g., jacket buttons and sleeves). Our two-stage approach (w/ Stage 1) leverages general VTON priors to resolve this issue.

1 Introduction

Virtual Try-On (VTON) has emerged as a transformative technology in computer vision, with the potential to revolutionize e-commerce, digital fashion, and augmented reality. In its canonical form, the task involves rendering a target garment onto an image of a person to synthesize a photorealistic result. Driven by advances in diffusion models and generative adversarial networks, state-of-the-art (SOTA) methods [7–10,12,16,18,19,25,27,28,32,37,40,45,46,51,52] have achieved remarkable fidelity in single-garment replacement.

Despite impressive progress, existing approaches remain constrained to simplified single-garment scenarios, limiting their applicability in real-world dressing

contexts. Human apparel is inherently layered; individuals wear multiple garments that occlude and interact with one another: a shirt collar protruding from a sweater, or an inner t-shirt visible beneath an open jacket. The capability to compose a new garment onto an existing outfit rather than merely replacing it, represents a critical yet overlooked frontier for VTON's practical deployment. In this work, we formally define this task as **Layering Virtual Try-On (LVTON)**. Unlike conventional VTON, which replaces the existing garment, LVTON aims to composite a new garment on top of the person's existing outfit, while preserving the visibility of the inner layers, as illustrated in Fig. 2 (a).

Existing VTON paradigms are fundamentally ill-equipped to address LVTON. We identify two root causes. First, a methodological limitation prevalent in many dominant approaches: a cloth-agnostic person representation is needed [7–9, 12, 18, 25, 27, 28, 32, 37, 40, 45, 46, 52]. This process, which masks out the original garment to simplify the single-replacement task, irreversibly discards the essential contextual information — the very outfit upon which a new layer must be added. Second, a more pervasive systemic data bias: popular benchmarks, such as VITON-HD [8] and DressCode [33], consist exclusively of single-item try-on examples. Models trained on this data have never observed the complex interplay of layered garments and thus cannot learn the required composition logic. As shown in Fig. 2 (b), previous methods consequently fail on the LVTON task, often erasing the inner layer or producing severe artifacts.

The failure of these mask-based approaches necessitates a paradigm shift towards mask-free editing. Large-scale pre-trained diffusion models [44], which excel at high-fidelity, text-guided image editing without relying on explicit masks, thus emerge as a natural starting point for LVTON. A straightforward solution would be to fine-tune such a model on a dedicated LVTON dataset. Yet, this approach faces two critical hurdles inherent to real-world data collection. First is extreme data scarcity: collecting large-scale, high-quality layering pairs with diverse garments and subjects is prohibitively expensive and practically infeasible. Therefore, any viable **LVTON solution must be highly data-efficient**. Second is the inevitable pose variance. Unlike standard VTON datasets with perfectly aligned pairs, it is practically impossible for a subject to maintain an identical pose while donning an additional layer. This introduces spatial misalignment, significantly compounding the task difficulty compared to standard fixed-pose VTON. Our preliminary experiments reveal that fine-tuning on such limited and pose-variant data is a critical flaw. The model is forced to learn complex layering logic, spatial alignment, and fundamental VTON capabilities (e.g., texture fidelity, garment deformation) simultaneously from an insufficient number of examples. As illustrated in Fig. 2 (c), this leads to a loss of fidelity, where the model fails to maintain the garment's identity (e.g., buttons) and struggles with realistic deformation, even as it learns the correct layering logic (e.g., preserving the inner shirt collar).

This leads to our key insight: the challenges of LVTON are two-fold and must be disentangled. We posit that LVTON requires (1) general VTON priors (i.e., robust garment deformation, placement, and identity preservation) and

(2) specific layering knowledge (i.e., occlusion reasoning and composition logic). The failure of direct training stems from forcing a model to learn both complex knowledge simultaneously from scarce LVTON data. We argue that general VTON priors and specific layering knowledge are distinct yet complementary. By disentangling them, we can exploit abundant single-layer data to learn robust priors, while using only limited layering data for specialization. From this, we propose a two-stage disentangled training paradigm.

Stage 1: Learning general VTON priors. The goal of this stage is to equip the model with robust VTON priors without relying on scarce LVTON data. However, standard VTON datasets are ill-suited for this purpose due to the dual challenges identified above: (i) their reliance on cloth-agnostic masking discards the context vital for layering, and (ii) their aligned poses fail to provide the spatial deformation supervision needed to handle the pose variance inherent in LVTON. If the model does not learn to handle pose shifts in this stage, it cannot effectively support the layering task in stage 2. To address this, we introduce an automatic data generation pipeline that leverages ubiquitous fashion videos. This approach produces training pairs that are, by design, both mask-free and pose-variant. By learning from video frames with natural body movements, the model is forced to master complex garment deformation and identity preservation, thereby establishing a powerful prior robust to the spatial misalignment encountered in real-world layering.

Stage 2: Learning the specific layering knowledge. With the general VTON priors from stage 1, we now focus on learning the specific layering knowledge by fine-tuning on our curated LVTON dataset. We acknowledge that building a comprehensive, unbiased layering dataset at scale remains a profound challenge. However, our disentangled approach turns this limitation into a demonstration of data efficiency. Because the model is already an expert in general VTON from stage 1, it does not need to relearn fundamental garment deformation or texture fidelity. Instead, it can dedicate its capacity to quickly learning the nuanced compositional logic of layering from this scarce data.

Our extensive experiments validate the effectiveness of our disentangled paradigm. On our newly collected LVTON benchmark, our method surpasses existing VTON baselines, producing photorealistic LVTON results where prior art fails. Furthermore, general VTON priors learned in stage 1 prove highly transferable; it exhibits strong zero-shot performance on standard VTON benchmarks. With standard fine-tuning, our model achieves state-of-the-art results on these traditional tasks, demonstrating the robustness and generality of our approach. Finally, our model achieves sequential outfit editing, as shown in Fig. 1.

Our contributions are summarized as follows:

- We define and address Layering Virtual Try-On, a novel and critical task for real-world VTON application.
- We propose a highly data-efficient, disentangled training paradigm that decomposes the LVTON task, enabling the model to learn complex layering logic from a strictly limited dataset and to generalize to in-the-wild scenarios.

- We introduce an automatic and scalable data generation pipeline from videos, enabling the model to learn general VTON priors without requiring real-world paired data.
- We experimentally validate our paradigm, achieving SOTA on our LVTON benchmark and demonstrating competitive performance on traditional VTON benchmarks.

2 Related Works

Virtual Try-On systems. Virtual Try-On (VTON) has seen significant progress, with the majority of existing methods [7–9,12,18,25,27,28,32,37,40,45,46, 51,52] operating under a cloth-agnostic setting. Early GAN-based works [8,28,45] typically employ a two-stage process involving garment warping and try-on generation, both of which heavily rely on precise masks to define the synthesis region. More recent diffusion-based approaches [7, 12, 18, 27, 32, 37, 40, 46, 51, 52] merge these stages into an end-to-end process, but still require the cloth-agnostic representation as a conditional input. As discussed in Sec. 1, this fundamental reliance on masks precludes the modeling of layering try-on, as all information about the person's existing outfit is erased.

Acknowledging this limitation, a few recent works have proposed mask-free VTON [10, 16, 19, 51]. These methods synthesize the try-on result directly on the original, fully-clothed person image, for example, by utilizing different conditioning strategies [10, 19, 51] or leveraging powerful backbones [16]. However, while this removes the mask dependency, it remains insufficient for true layering capabilities. The core issue persists: these mask-free methods are still trained on traditional VTON datasets [8, 33] or custom-collected benchmarks [16] that lack the necessary compositional examples of a person sequentially adding garments. These benchmarks predominantly contain single-item replacement pairs rather than the sequential "before-and-after" examples needed for layering. Therefore, while our method operates in the mask-free setting, our primary contribution is not the mask-free paradigm itself. Instead, we address the critical data bottleneck that prevents existing mask-free methods from achieving realistic layering.

Generative data synthesis for training. Deep generative models have increasingly been employed to synthesize paired training data, circumventing the high cost of manual collection [5, 13, 15, 20, 22, 26, 31, 34, 38, 42, 47]. A common strategy involves leveraging inpainting models to alter garments on a single image, thereby creating a pseudo-pair [16, 17, 19, 24, 29, 30, 36, 39, 41, 49]. However, a critical limitation of these approaches is that the synthesized input and target are pose-invariant. While acceptable for standard VTON, this lack of pose variation is fundamentally insufficient for LVTON. As established in Sec. 1, a model trained solely on pose-invariant pairs learns trivial pixel-to-pixel correspondence rather than the complex spatial deformation required to align a garment to a shifted pose. To address this, our stage 1 introduces a cross-frame data generation paradigm that mines pose-variant pairs from unstructured videos. By pairing an inpainted frame with a temporally distant frame from the same video, we construct training

samples where the subject's pose differs naturally. This forces the model to learn robust garment deformation and alignment priors, which are indispensable for handling the spatial misalignment in LVTON. Crucially, by mastering complex spatial and deformation priors, we reduce the burden on subsequent training phases. This directly enables the high sample efficiency of our stage 2, allowing the model to learn specific layering logic from a strictly limited dataset.

3 Methods

Our method implements the two-stage disentangled training paradigm, designed to address the data and methodological challenges of VTON for layered apparel. We first detail the two core stages of our paradigm: (1) an automatic data generation pipeline to learn the general VTON priors (Sec. 3.1), and (2) a specialized fine-tuning strategy using an augmented, real-world dataset to learn the specific layering knowledge (Sec. 3.2). We then describe our model architecture and model training (Sec. 3.3).

Task Definition. Given a person image $\mathcal{I}_M$, a garment image $\mathcal{I}_G$, a target pose $\mathcal{I}_P$, and a text description $\mathcal{T}$, our goal is to generate a realistic image $\mathcal{I}_T$ that reflects the desired garment and pose edit while preserving the person's identity and the exact visual integrity of non-target (inner) clothing layers. Standard VTON aims to swap an entire garment. In contrast, the text prompt $\mathcal{T}$ specifies the intended interaction (e.g., "add a dark blue denim blouse," "remove a gray sweater," or "swap X for Y"), requiring the model to modify only a specified clothing region while preserving the non-target (inner) regions from $\mathcal{I}_M$.

3.1 Stage 1: Learning the General VTON Priors

As discussed in Sec. 1, the core challenge of LVTON lies in the incompatibility of standard mask-based VTON methods, which discard essential contextual cues. To overcome this incompatibility, we first learn general VTON priors in a mask-free manner. This stage models garment deformation, shading, and spatial placement, providing a strong foundation for context-aware editing.

Our approach is built on a key observation: abundant online videos feature people wearing the same outfits in diverse poses. We utilize this property to generate a scalable, challenging synthetic dataset. The process involves four steps (illustrated in Fig. 3, (a)): **(1)** Data Collection: we collect video sequences of subjects in consistent outfits but with varying poses. **(2)** Frame Extraction: we extract frames and process them using [1,35,48] to obtain tuples $\{(F_i, M_i, G_i, P_i, \mathcal{T}_i)\}_{i=1}^{N}$, where F_i is the image, M_i the garment mask, G_i the segmented garment, P_i the pose, and $\mathcal{T}_i$ is the corresponding text description of the garment G_i. **(3)** Synthesis: we employ an inpainting model [3] to synthesize a new version of each frame, F_i', featuring a novel outfit guided by a Large Language Model (LLM) generated prompt. **(4)** Training Pair Construction: we construct the final training pairs by sampling distinct frames F_i and F_j from the same sequence ($i \neq j$). Each training

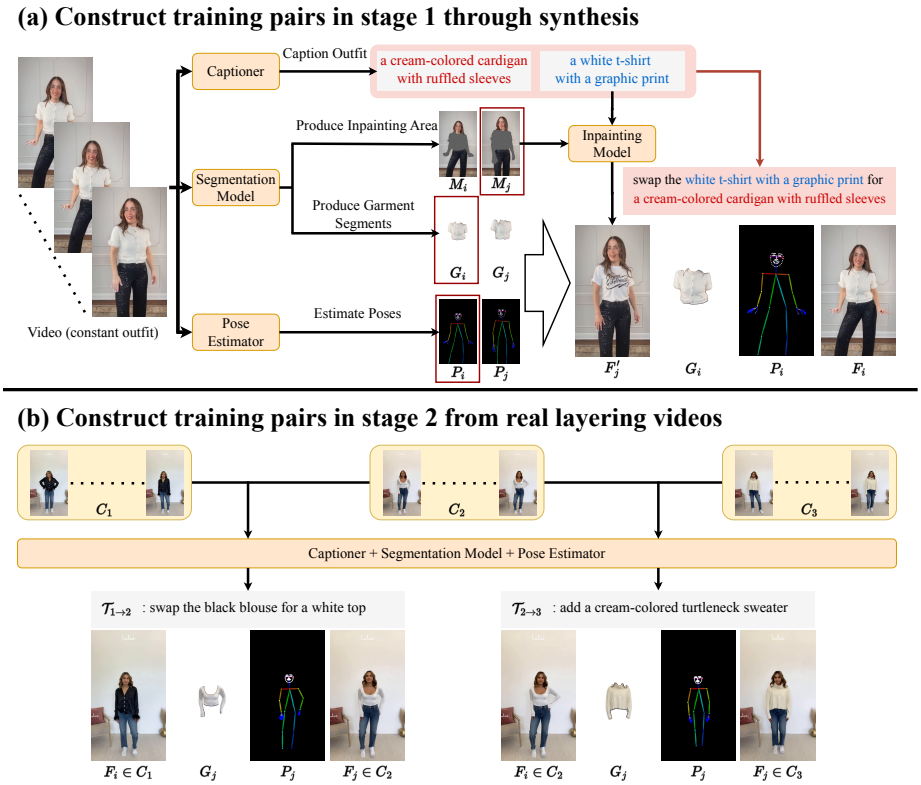

Fig. 3: Our two-stage pipeline. (a) The first stage builds general VTON priors by synthesizing mask-free, pose-mismatched training pairs from videos. This forces the model to learn robust spatial deformation, ensuring accurate alignment even during standard fixed-pose inference. (b) The second stage uses a small, real-world layering dataset to teach the model specific layering knowledge (e.g., "add").

instance is constructed as $(\mathcal{I}_M, \mathcal{I}_G, \mathcal{I}_P, \mathcal{T}, \mathcal{I}_T) := (F'_i, G_j, P_j, \mathcal{T}_j, F_j)$. More details can be found in Appendix B.1.

This video-based construction is pivotal for disentangling the learning process. Crucially, the objective of stage 1 is **NOT** to approximate LVTON, but rather to isolate the fundamental mechanics of mask-free editing. First and foremost, the task forces the model to reason about garment placement directly from context without masks. More critically, by enforcing $i \neq j$, our pipeline guarantees that the input pose (P_i in F'_i) and target pose (P_j) are spatially misaligned. As emphasized in Sec. 1, real-world layering inevitably introduces pose variance. If the model were trained on pose-invariant pairs (where $i = j$), it would collapse into learning trivial pixel alignment. By solving the harder task of "deform-and-place" in this stage, the model acquires robust spatial priors that serve as the indispensable structural foundation for the delicate layering composition in stage 2. Finally, sourcing from "in-the-wild" videos exposes the model to diverse lighting and backgrounds, preventing overfitting to sterile studio environments.

Fig. 4: Visualizations of fine-tuning data. **Top**: A real-world "add" sequence collected from video. **Bottom**: The corresponding "remove" sequence generated by temporal reversal augmentation.

3.2 Stage 2: Fine-tuning for Layering Knowledge

While stage 1 equips the model with strong general VTON priors, it lacks supervision over explicit compositional operations (e.g., "add"). We therefore fine-tune the model on a real-world dataset designed to teach such compositional reasoning. As noted in Sec. 1, building a layering dataset at scale is infeasible. Our disentangled paradigm turns this limitation into a demonstration of data efficiency: based on general VTON priors learned in stage 1, the model can dedicate its entire learning capacity to mastering composition logic. We propose two contributions: a data collection pipeline and a novel augmentation technique to expand the dataset effectively. More details can be found in Appendix B.2.

Data collection. We collect videos of individuals sequentially adding or swapping garments. We partition video frames into temporal clusters $\{C_1, C_2, \cdots, C_K\}$ based on stable outfit configurations. We employ a Vision-Language Model [1] to automatically generate textual descriptions $\mathcal{T}_{m\to m+1}$ for each transition (e.g., "add a gray sweater"). We then sample a source frame $F_i \in C_m$ and a target frame $F_j \in C_{m+1}$ from adjacent clusters to form the training tuple as $(\mathcal{I}_M, \mathcal{I}_G, \mathcal{I}_P, \mathcal{T}, \mathcal{I}_T) := (F_i, G_j, P_j, \mathcal{T}_{m\to m+1}, F_j)$.

Temporal reversal augmentation. To mitigate the scarcity of our fine-tuning data, we introduce a temporal reversal augmentation technique. This stems from the observation that a "dressing" sequence, when reversed, becomes a valid "undressing" sequence. For each forward training instance $(F_i, G_j, P_j, \mathcal{T}_{m\to m+1}, F_j)$, we create a corresponding reverse instance $(F_j, G_i, P_i, \mathcal{T}_{m+1\to m}, F_i)$, where the reverse text prompt $\mathcal{T}_{m+1\to m}$ is generated by applying inversion rules (e.g., "add `<garment>`" becomes "remove `<garment>`") (Fig. 4). This effectively doubles our real-world data and enables learning both garment addition and removal.

3.3 Model Architecture and Training

Model architecture. We build on the pre-trained Qwen-Image-Edit [44], a model with strong high-fidelity image editing capability. As shown in Fig. 5, it comprises three main components: (1) a Variational AutoEncoder (VAE) as the

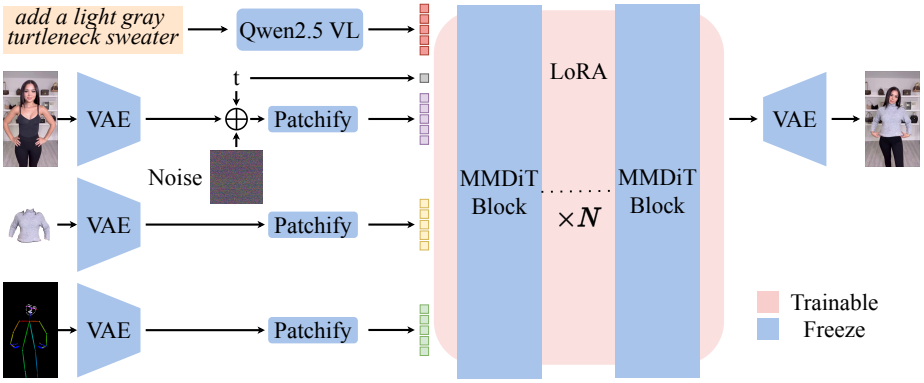

Fig. 5: An overview of model architecture. We apply Low-Rank Adaptation (LoRA) to fine-tune Qwen-Image-Edit. The model conditions on four inputs: (1) a text prompt (e.g., "add a light gray turtleneck sweater"), processed by the Qwen2.5-VL; (2) a noised person image latent, created by adding noise at timestep t; (3) a garment image latent; and (4) a pose image latent.

image tokenizer, (2) Qwen2.5-VL [1] as the multimodal condition encoder, and (3) a Multimodal Diffusion Transformer (MMDiT) [14] as the diffusion backbone. Its key innovation is a dual-stream design, where Qwen2.5-VL encodes high-level semantics from text and reference images, while the VAE provides low-level visual latents. The MMDiT fuses both streams, achieving a balance between semantic consistency and visual fidelity.

Model training. We train the model using LoRA [23] via our two-stage training paradigm. We first train it on our synthetic dataset (stage 1) and then fine-tune it on the real-world layering dataset (stage 2). The model architecture (Sec. 3.3), training objective (standard denoising loss), and all optimization details are kept identical across stage 1 and stage 2. More details can be found in Appendix A.

Table 1: Quantitative comparisons on our LVTON dataset. Our method performs best across all metrics. Note: absolute FID scores are inflated due to our small test set (532 images). Best results are in **bold**. ↑ indicates higher is better, ↓ indicates lower is better.

Model	SSIM ↑	LPIPS ↓	FID ↓	KID$_{\times 10^3}$ ↓
OmniTry [16]	0.810	0.176	54.585	3.165
Any2AnyTryon [19]	0.787	0.229	89.567	18.645
Nano Banana [11]	0.798	0.195	49.179	2.362
HunyuanImage-3.0-Instruct [6]	0.767	0.246	88.070	15.561
Ours	**0.843**	**0.127**	**48.957**	**1.463**

4 Experiments

In this section, Sec. 4.1 and Sec. 4.2 evaluate the proposed method on Layering VTON and traditional VTON tasks, respectively; Sec. 4.3 provides user study and layering-specific evaluation; Sec. 4.4 shows in-the-wild results; and Sec. 4.5 presents ablation studies to validate our key design choices.

Fig. 6: Qualitative results of LVTON. Our method successfully composites the target garment while preserving the inner layer. Baselines often erase inner details or introduce artifacts. Regions corresponding to the errors made by the baselines are highlighted with red boxes in their respective predictions for comparison (e.g., Nano Banana fails to preserve the lower portion of the black dress).

4.1 Layering Virtual Try-On

Datasets and evaluation metrics. Our training paradigm employs two distinct stages to address the scarcity of real-world layering data. For stage 1 (general VTON priors), we construct a synthetic dataset of 29,151 training pairs from 362 videos (Sec. 3.1). For stage 2 (layering knowledge), we curate a real-world LVTON dataset of 5,768 training pairs and 532 test pairs from 60 videos, expanded via temporal reversal augmentation (Sec. 3.2). We evaluate using SSIM [43], LPIPS [50], FID [21], and KID [4] (note that absolute FID values are inflated due to the small test set). More details can be found in Appendix C.

Baselines and results. We conduct a comparison with mask-free methods, OmniTry [16], and Any2AnyTryon [19], as well as three industrial models: HunyuanImage-3.0-Instruct [6], Nano Banana [11], and Google Tryon (a variant of [2]), on our LVTON dataset. To ensure a fair comparison, we adapt these baselines by providing layering instructions (e.g., "add a shirt") as text prompts. For Nano Banana, evaluation is conducted via its public API. The Google TryOn comparison is limited to qualitative results because it lacks a public API. HunyuanImage-3.0-Instruct is evaluated via its official open-source implementation. Quantitative and qualitative results are shown in Tab. 1 and Fig. 6, respectively. Our model outperforms all baselines across metrics, setting a new SOTA for LVTON.

Table 2: Quantitative comparison with SOTA methods on the VITON-HD [8] and DressCode [33] benchmarks. Ours is our complete two-stage model, fine-tuned on the benchmarks. Stage 1-only is our stage 1 model, evaluated in a zero-shot setting (Sec. 4.5). Best results are in **bold**.

Model	VITON-HD				DressCode			
	SSIM ↑	LPIPS ↓	FID ↓	KID$_{\times 10^3}$ ↓	SSIM ↑	LPIPS ↓	FID ↓	KID$_{\times 10^3}$ ↓
Methods Using Cloth-Agnostic Representation								
VITON-HD [8]	0.862	0.117	12.117	3.230	-	-	-	-
HR-VITON [28]	0.878	0.105	11.270	2.730	0.936	0.065	13.820	2.710
GP-VTON [45]	0.894	0.079	9.197	0.880	0.948	0.268	11.980	8.170
LADI-VTON [32]	0.876	0.091	9.410	1.600	0.915	0.063	13.260	2.670
DCI-VTON [18]	0.880	0.081	8.754	1.100	0.937	**0.031**	10.820	1.890
StableVITON [27]	0.852	0.084	8.698	0.880	0.911	0.050	11.266	0.720
OOTDiffusion [46]	0.878	0.071	8.810	0.820	0.898	0.073	3.950	0.720
IDM-VTON [9]	0.870	0.102	6.290	1.201	0.920	0.062	8.640	2.904
FitDit [25]	0.899	0.066	4.731	**0.190**	0.926	0.043	2.638	0.499
Mask-Free Methods								
CATVTON [10]	0.870	0.057	5.425	0.411	0.892	0.046	3.992	0.818
OmniTry [16]	0.698	0.368	57.871	0.450	0.842	0.188	25.364	0.410
Any2AnyTryon [19]	0.839	0.088	6.934	0.740	0.889	0.138	15.719	0.502
MuGa-VTON [12]	0.898	0.073	7.240	0.520	0.937	0.046	8.930	0.940
Stage 1-only	0.887	0.086	5.235	1.550	0.898	0.101	4.710	0.890
Ours	**0.928**	**0.055**	**3.235**	0.200	**0.948**	0.042	**2.598**	**0.315**

Fig. 7: Qualitative comparison on the VITON-HD [8] and DressCode [33] benchmarks. Ours represents our complete two-stage model. Stage 1-only represents our stage 1 model evaluated zero-shot (analysis in Sec. 4.5). Red boxes highlight artifacts and failure cases in the baseline results. Ours consistently avoids these issues, demonstrating superior fidelity and generalization.

4.2 Traditional Virtual Try-On

Datasets and evaluation metrics. We also evaluate our model on traditional VTON benchmarks (VITON-HD [8], DressCode [33]) using the same metrics.

Baselines and results. We benchmark our full two-stage model, Ours, against SOTA methods. As our model requires pose and text inputs, we use the same VLM and pose estimator [1, 48] to generate them for the benchmarks. As shown

in Tab. 2, Ours outperforms all existing methods in SSIM and FID, setting a new state of the art on these benchmarks. The corresponding qualitative results in Fig. 7 also showcase our model's superior fidelity and generalization.

Table 3: User study results (5-point Likert scale). Our method outperforms baselines across both general generation quality (i.e., Realism, Instr.) and LVTON-specific layering metrics (i.e., Occlusion, Inner-Vis., Overall-Lay.).

Method	Realism	Instr.	Occlusion	Inner-Vis.	Overall-Lay.
OmniTry [16]	3.370	3.521	3.603	3.125	2.945
Any2Any [19]	3.194	3.069	3.639	2.859	2.761
Nano Banana [11]	4.278	2.361	3.486	3.394	2.833
Ours	**4.597**	**4.694**	**4.750**	**4.535**	**4.569**

4.3 Layering-Specific Evaluation and User Study

Because standard metrics (e.g., SSIM, LPIPS) do not explicitly capture LVTON-specific challenges like occlusion and inner-layer preservation, we conducted a 30-participant user study. Using a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree), users evaluated our method against baselines across five criteria: (i) ***Visual Realism***: the generated image looks visually realistic and natural; (ii) ***Instruction Following***: the image correctly aligns with the given text instruction; (iii) ***Occlusion Correctness***: the occlusion relationships between garments are accurate (e.g., outer garments appropriately cover inner garments); (iv) ***Inner-Layer Visibility***: relevant inner-layer details (e.g., collars, sleeves, hems) are properly preserved; and (v) ***Overall Layering Quality***: the garment layering appears coherent and physically plausible.

As shown in Tab. 3, our approach consistently outperforms baselines. Most notably, it achieves significant improvements in the layering-specific metrics (Occlusion Correctness, Inner-Layer Visibility, and Overall Layering Quality), demonstrating superior capability in handling complex multi-layer interactions. To complement our human evaluation, we introduce a quantitative metric, Masked Preservation Score (MPS), that measures the structural and perceptual preservation of non-target regions. As detailed in Appendix E, our method improves upon the SOTA on MPS, increasing SSIM by 0.005 and reducing LPIPS by 0.006.

4.4 In-the-Wild Generalization

Our approach demonstrates strong generalization to in-the-wild images sourced directly from the internet (Fig. 8). We evaluate the model on three types of layering edits — adding an outer layer, swapping an existing garment, and replacing an inner layer — across a wide range of unconstrained scenarios (e.g., outdoor scenes, varied poses, different genders, and diverse garment types).

Despite the complexity of these in-the-wild images, the model maintains consistent subject identity and realistic shading. Furthermore, it accurately renders the target garments while flawlessly preserving the structural integrity and visibility of the non-target layers. These results confirm that our training paradigm is highly **data-efficient**: the model effectively learns specific layering logic from the limited stage 2 dataset and successfully generalizes to arbitrary, real-world dressing contexts without loss of visual fidelity.

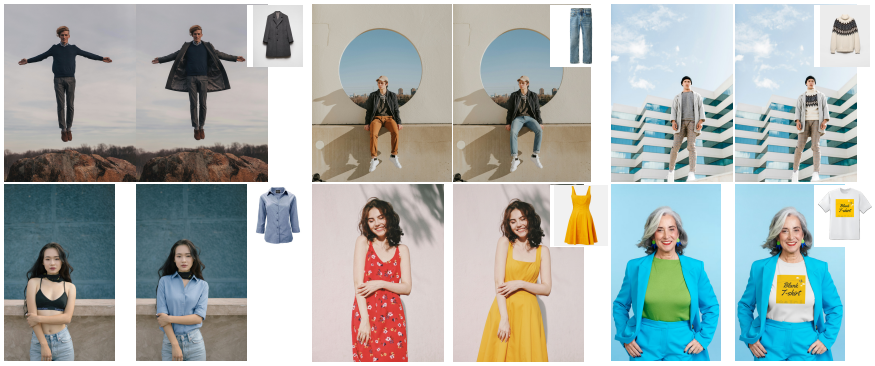

Fig. 8: In-the-wild layering editing. Despite being fine-tuned on a strictly limited stage 2 dataset, our model successfully generalizes to complex, out-of-distribution scenarios sourced from the internet. It accurately executes specific compositional instructions ("add," "swap," and inner-layer swaps) while naturally preserving the visual integrity of the non-target layers and overall high image fidelity.

While our method successfully handles various scenarios, it remains bounded by stage 2 data distribution. Improvements may require scaling the data collection pipeline to cover broader layering distribution. More details can be found in Appendix G.

Table 4: Ablation study on training components, evaluated on our layering dataset. This validates the necessity of two training stages and the temporal reversal augmentation. Ours achieves the best results, confirming our hypothesis. Best results are in **bold**.

Model	SSIM ↑	LPIPS ↓	FID ↓	KID$_{\times 10^3}$ ↓
Stage 2-only	0.817	0.192	58.561	13.242
Stage 1-only	0.821	0.183	56.663	12.286
Ours w/o Aug	0.832	0.131	49.353	3.003
Ours	**0.843**	**0.127**	**48.957**	**1.463**

4.5 Ablation Study

Analysis of disentangled training stages. We evaluate three model variants on our LVTON dataset: (1) **Stage 2-only**: trained from scratch only on our real-world layering data. This simulates "direct training" approach. (2) **Stage 1-only**: our stage 1 model, which only learns "swap" logic. (3) **Ours w/o Aug**: our pipeline without the temporal reversal augmentation in stage 2.

As shown in Tab. 4, the results strongly corroborate our hypothesis. First, Stage 2-only achieves the lowest quantitative scores, confirming that forcing the model to learn complex LVTON logic solely from scarce data is ineffective. Conversely, Stage 1-only exhibits a different failure mode. Since it is trained exclusively on "swap" (garment replacement) scenarios, it struggles to generalize to the "add" (layering) task. Finally, the performance drop in Ours w/o Aug validates the effectiveness of our temporal reversal augmentation.

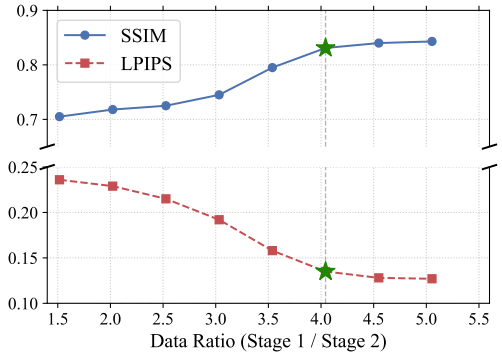

Fig. 9: Ablation study on stage 1 data scale using the LVTON benchmark. The curves illustrate SSIM and LPIPS as a function of the data ratio between stage 1 and stage 2. Increasing the ratio yields consistent improvements, validating the efficacy of stage 1 priors. We adopt the saturation point of 4.0× (green star) as the optimal configuration.

Validation of general VTON priors. To confirm that stage 1 provides robust, generalizable priors, we evaluate Stage 1-only on VITON-HD and DressCode in a zero-shot setting. Crucially, our synthetic dataset was strictly deduplicated against these benchmarks to prevent data leakage and memorization. As Tab. 2 shows, Stage 1-only achieves highly competitive results without seeing target domain data, even surpassing fully trained methods like IDM-VTON. This confirms that stage 1 builds a strong foundation for both general and layering VTON tasks.

Impact of stage 1 data scale. We investigate the relationship between the scale of the stage 1 synthetic data and the final LVTON performance. We train several models using synthetic data subsets of varying ratios relative to the fixed-size stage 2 data, and then fine-tune all models on the identical LVTON dataset. The results, plotted in Fig. 9, reveal a clear and significant trend: performance (SSIM/LPIPS) improves substantially as the volume of stage 1 data increases. Notably, these performance gains begin to plateau around a 4.0× data ratio (Stage 1 / Stage 2), suggesting a point of saturation. This study strongly validates our hypothesis: the stage 1 training is crucial for building robust and generalizable VTON priors, enabling the model to achieve high fidelity when fine-tuning on the limited LVTON data.

5 Limitations

Our method has three limitations. (1) Stage 1 synthetic data quality is bound by upstream segmentation [35] and inpainting [3] models. Any artifacts or errors may be learned, degrading our VTON priors. (2) While the proposed method demonstrates strong generalization to in-the-wild layering scenarios, the limited scale and diversity of the stage 2 dataset restrict generalization to atypical or complex outfit interactions. (3) Our temporal reversal augmentation imperfectly simulates garment removal, missing real-world physical dynamics like inner-layer fabric pulling or hair disruption. Future data collection of authentic garment removal sequences can address this gap.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., Chan, K., Chen, Y., Dieleman, S., Du, Y., et al.: Imagen 3. arXiv preprint arXiv:2408.07009 (2024)
3. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., et al.: Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv e-prints pp. arXiv–2506 (2025)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
5. Bonifacio, L., Abonizio, H., Fadaee, M., Nogueira, R.: Inpars: Unsupervised dataset generation for information retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2387–2392 (2022)
6. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
7. Chen, M., Chen, X., Zhai, Z., Ju, C., Hong, X., Lan, J., Xiao, S.: Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In: European Conference on Computer Vision. pp. 124–142. Springer (2024)
8. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14131–14140 (2021)
9. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
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. arXiv preprint arXiv:2407.15886 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Deria, A., Mahapatra, D., Bozorgtabar, B., Chakraborty, M., Chakraborty, S., Roy, S.: Muga-vton: Multi-garment virtual try-on via diffusion transformers with prompt customization. arXiv preprint arXiv:2508.08488 (2025)
13. Eigenschink, P., Reutterer, T., Vamosi, S., Vamosi, R., Sun, C., Kalcher, K.: Deep generative models for synthetic data: A survey. IEEE Access **11**, 47304–47320 (2023)
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: Forty-first international conference on machine learning (2024)
15. Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., Tian, Y.: Scaling laws of synthetic images for model training... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7382–7392 (2024)

16. Feng, Y., Zhang, L., Cao, H., Chen, Y., Feng, X., Cao, J., Wu, Y., Wang, B.: Omnitry: Virtual try-on anything without masks. arXiv preprint arXiv:2508.13632 (2025)
17. Fu, Y., Chen, C., Qiao, Y., Yu, Y.: Dreamda: Generative data augmentation with diffusion models. arXiv preprint arXiv:2403.12803 (2024)
18. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7599–7607 (2023)
19. Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., Liu, J.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. arXiv preprint arXiv:2501.15891 (2025)
20. Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: Synthclip: Are we ready for a fully synthetic clip training? arXiv preprint arXiv:2402.01832 (2024)
21. 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. Advances in neural information processing systems **30** (2017)
22. Honovich, O., Scialom, T., Levy, O., Schick, T.: Unnatural instructions: Tuning language models with (almost) no human labor. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14409–14428 (2023)
23. 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. ICLR **1**(2), 3 (2022)
24. Huang, Y., Zhang, P., Liu, R., Liang, J.: Can generated images serve as a viable modality for text-centric multimodal learning? arXiv preprint arXiv:2506.17623 (2025)
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. arXiv preprint arXiv:2411.10499 (2024)
26. Josifoski, M., Sakota, M., Peyrard, M., West, R.: Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. arXiv preprint arXiv:2303.04132 (2023)
27. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8176–8185 (2024)
28. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision. pp. 204–219. Springer (2022)
29. Li, N., Shih, K.J., Plummer, B.A.: Enhancing virtual try-on with synthetic pairs and error-aware noise scheduling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21238–21247 (2025)
30. Li, R., Jin, X., et al.: Real: Realism evaluation of text-to-image generation models for effective data augmentation. arXiv preprint arXiv:2502.10663 (2025)
31. Liu, R., Xu, G., Jia, C., Ma, W., Wang, L., Vosoughi, S.: Data boost: Text data augmentation through reinforcement learning guided conditional generation. arXiv preprint arXiv:2012.02952 (2020)
32. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladivton: Latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the 31st ACM international conference on multimedia. pp. 8580–8589 (2023)

33. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2231–2235 (2022)
34. Ng, N., Cho, K., Ghassemi, M.: Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. arXiv preprint arXiv:2009.10195 (2020)
35. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
36. Shivashankar, C., Miller, S.: Semantic data augmentation with generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 863–873 (2023)
37. Sun, K., Cao, J., Wang, Q., Tian, L., Zhang, X., Zhuo, L., Zhang, B., Bo, L., Zhou, W., Zhang, W., et al.: Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. arXiv preprint arXiv:2407.16224 (2024)
38. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems **36**, 48382–48402 (2023)
39. Valvano, G., Agostino, A., De Magistris, G., Graziano, A., Veneri, G.: Controllable image synthesis of industrial data using stable diffusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5354–5363 (2024)
40. Velioglu, R., Bevandic, P., Chan, R., Hammer, B.: Enhancing person-to-person virtual try-on with multi-garment virtual try-off. arXiv preprint arXiv:2504.13078 (2025)
41. Wang, S., Feng, Y., Lan, T., Yu, N., Bai, Y., Xu, R., Wang, H., Xiong, C., Savarese, S.: Text2data: Low-resource data generation with textual control. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 21252–21260 (2025)
42. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D., Hajishirzi, H.: Self-instruct: Aligning language models with self-generated instructions. In: Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers). pp. 13484–13508 (2023)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
44. 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)
45. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23550–23559 (2023)
46. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8996–9004 (2025)
47. Yang, Y., Malaviya, C., Fernandez, J., Swayamdipta, S., Bras, R.L., Wang, J.P., Bhagavatula, C., Choi, Y., Downey, D.: Generative data augmentation for commonsense reasoning. arXiv preprint arXiv:2004.11546 (2020)
48. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4210–4220 (2023)

49. Yin, Y., Kaddour, J., Zhang, X., Nie, Y., Liu, Z., Kong, L., Liu, Q.: Ttida: Controllable generative data augmentation via text-to-text and text-to-image models. arXiv preprint arXiv:2304.08821 (2023)
50. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
51. Zhang, X., Lin, E., Li, X., Luo, Y., Kampffmeyer, M., Dong, X., Liang, X.: Mmtryon: Multi-modal multi-reference control for high-quality fashion generation. arXiv preprint arXiv:2405.00448 (2024)
52. Zhu, L., Li, Y., Liu, N., Peng, H., Yang, D., Kemelmacher-Shlizerman, I.: M&m vto: Multi-garment virtual try-on and editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1346–1356 (2024)