DS-VTON: AN ENHANCED DUAL-SCALE COARSE-TO-FINE FRAMEWORK FOR VIRTUAL TRY-ON

Anonymous authors
Paper under double-blind review

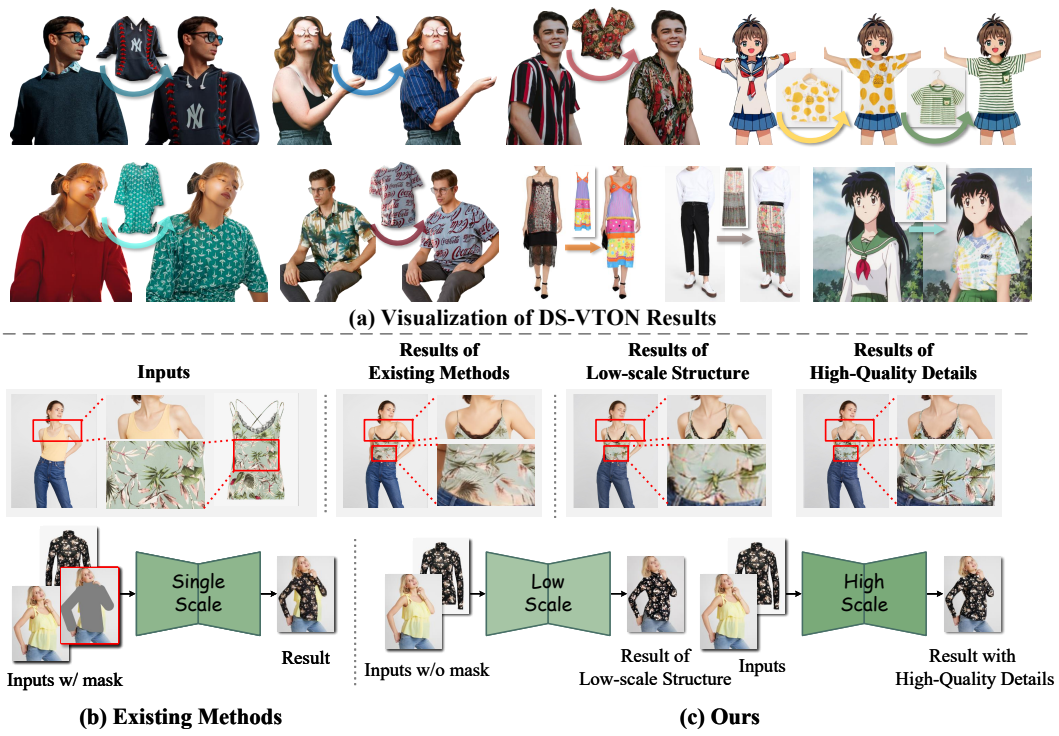

Figure 1: (a) DS-VTON results across diverse scenarios. (b) Existing methods (Kim et al., 2024; Xu et al., 2025; Choi et al., 2024; Zhou et al., 2025) adopt a single-scale pipeline with masked inputs, limiting their ability to capture full-body semantics and garment structure. (c) In contrast, DS-VTON adopts an enhanced dual-scale coarse-to-fine framework combined with a mask-free strategy.

ABSTRACT

Despite recent progress, most existing virtual try-on methods still struggle to simultaneously address two core challenges: accurately aligning the garment image with the target human body, and preserving fine-grained garment textures and patterns. These two requirements map directly onto a coarse-to-fine generation paradigm, where the coarse stage handles structural alignment and the fine stage recovers rich garment details. Motivated by this observation, we propose DS-VTON, an enhanced dual-scale coarse-to-fine framework that tackles the try-on problem more effectively. DS-VTON consists of two stages: the first stage generates a low-resolution try-on result to capture the semantic correspondence between garment and body, where reduced detail facilitates robust structural alignment. In the second stage, a blend-refine diffusion process reconstructs high-resolution outputs by refining the residual between scales through noise–image blending, emphasizing texture fidelity and effectively correcting fine-detail errors from the low-resolution stage. In addition, our method adopts a fully mask-free generation strategy, eliminating reliance on human parsing maps or segmentation masks. Extensive experiments show that DS-VTON not only achieves state-of-the-art performance but consis-

tently and significantly surpasses prior methods in both structural alignment and texture fidelity across multiple standard virtual try-on benchmarks.

1 INTRODUCTION

Given a garment image and a person image, the goal of virtual try-on is to synthesize a photorealistic image of the person wearing the specified garment (Han et al., 2018). As a key enabling technology for online fashion and e-commerce, virtual try-on has attracted increasing attention in recent years (Choi et al., 2021; Ge et al., 2021b; Gou et al., 2023; Lee et al., 2022; Zhang et al., 2024; Morelli et al., 2022; 2023; Choi et al., 2024; Chong et al., 2025; Zhou et al., 2025).

This task involves two fundamental challenges: (1) accurately fitting the garment onto the human body, and (2) preserving fine-grained garment textures. Existing methods fall into two main categories: Generative Adversarial Networks (GANs) (Goodfellow et al., 2020) and Diffusion Models (Ho et al., 2020; Rombach et al., 2022). Early GAN-based approaches (Choi et al., 2021; Ge et al., 2021a; Lee et al., 2022; Xie et al., 2023) typically follow a two-stage pipeline: a warping module first aligns the garment with the target pose, followed by a generation module to synthesize the final image. While warping helps preserve garment appearance, the subsequent generation stage often leads to detail loss due to imperfect feature fusion. Recent diffusion-based methods (Kim et al., 2024; Zhu et al., 2023; Xu et al., 2025; Choi et al., 2024; Zhou et al., 2025) have become popular. The denoising process naturally progresses from coarse to fine: early steps capture global structure, while later steps refine texture details (Balaji et al., 2022; Choi et al., 2022). This progressive generation order aligns well with the requirements of virtual try-on, where both alignment and texture fidelity are essential. **However, relying on a single-stage diffusion process remains inherently limited**. In practice, existing approaches still struggle to ensure accurate garment-body alignment and high-fidelity detail reconstruction. Without explicitly disentangling structure from detail, the unified framework often produces compromised visual quality.

To overcome the limitations of conventional single-stage denoising, we introduce a dual-scale coarse-to-fine framework that explicitly separates global structure alignment from fine-grained texture restoration. **Low-resolution stage:** In the first stage, the model generates a coarse try-on result by suppressing high-frequency content, emphasizing structural alignment through the low-resolution representation. **High-resolution stage:** In the second stage, a **blend-refine diffusion process explicitly** transforms the low-resolution output into high resolution, restoring fine textures and correcting fine-detail errors from the first stage.

Beside, traditional virtual try-on methods rely on human parsing masks for spatial guidance, our approach adopts a fully **mask-free strategy** that eliminates this dependency by leveraging the strong semantic priors embedded in pretrained diffusion models. Our main contributions are as follows:

- We propose a novel dual-scale, mask-free framework that enhances the coarse-to-fine process and is particularly well-suited for the try-on task.
- We introduce a **blend-refine diffusion process** that **explicitly** bridges two complex distributions, enabling controllable transition from coarse alignment to fine-detail restoration.
- Extensive experiments on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) demonstrate state-of-the-art performance, validating the effectiveness of our method both qualitatively and quantitatively.

2 RELATED WORKS

GAN-based virtual try-on. Earlier methods (Choi et al., 2021; Ge et al., 2021a; Lee et al., 2022; Xie et al., 2023), which are based on Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), typically decompose the virtual try-on task into two stages: (1) warping the garment to align with the human body shape, and (2) integrating the warped garment with the human image to generate the final result. For instance, ACGPN (Yang et al., 2020) employs a warping module based on Thin-Plate Spline (TPS) (Duchon, 1977) to deform the garment. PFAFN (Ge et al., 2021b) proposes a parser-free method that guides the garment warping process using learned appearance

flows. VITON-HD (Choi et al., 2021) introduces a specialized normalization layer and generator design to better handle garment-body misalignment during synthesis.

However, a key limitation of GAN-based approaches is their constrained capacity in capturing both precise spatial alignment and fine-grained garment details. As a result, these methods often rely heavily on the warping stage to encode garment appearance early in the pipeline. Yet, the subsequent integration stage frequently introduces artifacts or detail loss, degrading the overall realism of the try-on result.

Diffusion-based virtual try-on. With the rapid advancement of diffusion models (Ho et al., 2020; Rombach et al., 2022), powerful virtual try-on methods have emerged (Kim et al., 2024; Zhu et al., 2023; Choi et al., 2024; Zhou et al., 2025; Sun et al., 2024). Early methods like DCI-VTON (Gou et al., 2023) use a two-stage pipeline: warping the garment to match the body, then blending it with the person image using a diffusion model. DT-VTON (Zhang et al., 2024) splits the task into structural alignment and texture replacement. More recent methods (Morelli et al., 2023; Kim et al., 2024; Zhu et al., 2023; Choi et al., 2024) employ a single diffusion process for direct try-on synthesis, with improved conditioning strategies. For instance, LaDI-VTON (Morelli et al., 2023) uses textual inversion to encode garment identity, while IDM-VTON (Choi et al., 2024) introduces GarmentNet for structural and appearance guidance. Leffa (Zhou et al., 2025) proposes a Leffa loss to guide attention weights during the final 500 denoising steps for better texture recovery. Complementary to these, FitDiT (Jiang et al., 2024) adopts a DiT-based architecture (Peebles & Xie, 2023) with an aggressive rectangular mask to address alignment issues in mask-based pipelines.

Despite progress, diffusion-based methods still face challenges, such as garment fragmentation due to imprecise segmentation and poor rendering of fine patterns like flowers or text, which remain key areas for improvement.

3 METHODS

Notations. Given a person image $\mathbf{x}_p \in \mathbb{R}^{H\times W\times 3}$ and a garment image $\mathbf{x}_g \in \mathbb{R}^{H\times W\times 3}$, the virtual try-on task generates a realistic output $\mathbf{x}_r \in \mathbb{R}^{H\times W\times 3}$, where the person wears the given garment. In the low-resolution stage, the inputs $\mathbf{x}_p$ and $\mathbf{x}_g$ are downsampled to $\tilde{\mathbf{x}}_p, \tilde{\mathbf{x}}_g \in \mathbb{R}^{h\times w\times 3}$, with $h = H/\sigma$, $w = W/\sigma$, and σ denoting the downsampling ratio. These are used to generate a low-resolution result $\tilde{\mathbf{x}}_r \in \mathbb{R}^{h\times w\times 3}$. For simplicity, we assume that $\tilde{\mathbf{x}}_r$ is upsampled to the original resolution in the high-resolution stage, and do not introduce a separate symbol.

Mask-free strategy. Previous virtual try-on methods (Kim et al., 2024; Xu et al., 2025; Choi et al., 2024; Zhou et al., 2025; Jiang et al., 2024) typically use external human parsers to generate body segmentation masks for garment localization. However, in diffusion-based models like Stable Diffusion (Rombach et al., 2022), which already encode strong human-structure priors, this is unnecessary. We therefore adopt a fully mask-free design: the model directly consumes the garment and person images during training and inference, without parser-based pseudo-masks or segmentation guidance.

3.1 OVERVIEW

Our method is built upon Stable Diffusion (Rombach et al., 2022). The backbone architecture follows the dual U-Net framework (Zhang, 2023; Hu, 2024; Xu et al., 2024), which has also demonstrated strong performance in virtual try-on tasks (Choi et al., 2024; Xu et al., 2025). Further architectural details are provided in Subsection 3.2. As shown in Figure 2, both the low-resolution and high-resolution stages share this network architecture.

In the low-resolution stage, images inherently emphasize structural information because fine-grained textures are suppressed. This stage therefore serves two main purposes. First, it enables the model to capture the human body shape and garment category, ensuring accurate structural alignment between the person and the clothing. Second, it provides coarse but semantically reliable garment structure, such as the placement of stripes or patterns, offering guidance for fine-detail reconstruction in the high-resolution stage.

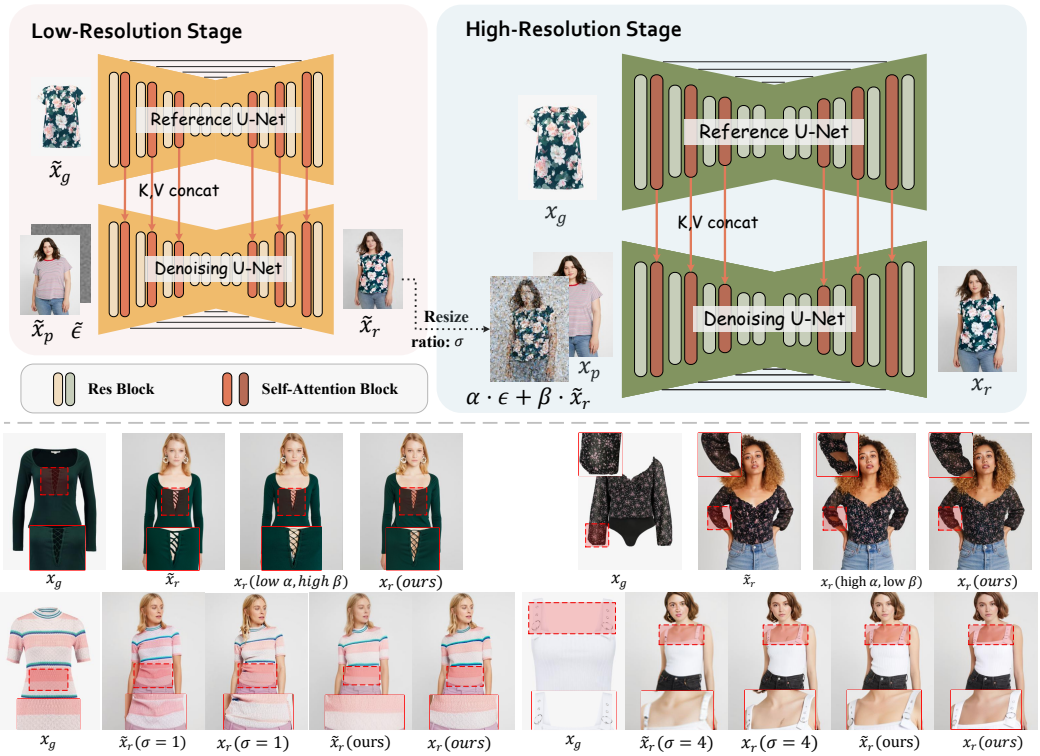

Figure 2: Upper panel: Two-scale generation pipeline. A low-resolution stage produces a coarse try-on result, then refined by a high-resolution stage; both stages share the same network architecture (see Section 3). Lower panel: Results with different settings; ours uses $\sigma = 2$ and $\alpha = \beta = \frac{1}{2}$(see Subsections 3.3 and 3.4). With proper two-stage settings, the second stage leverages the reliable coarse structure from the first stage to correct fine-detail errors and generate high-quality try-on results.

In the high-resolution stage, the key challenge is how to effectively utilize the low-resolution result $\tilde{\mathbf{x}}_r$. Here, we introduce our blend-refine diffusion process, which defines the noisy latent input as

$$\mathbf{x}_T = \alpha \cdot \boldsymbol{\epsilon} + \beta \cdot \tilde{\mathbf{x}}_r, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and α, β are balancing coefficients. The model is trained to gradually denoise this latent, transitioning $\mathbf{x}_T$ to $\mathbf{x}_r$ by converting the noise component $\alpha \cdot \boldsymbol{\epsilon}$ into the residual term $\mathbf{x}_r - \beta \cdot \tilde{\mathbf{x}}_r$, such that the final result satisfies $\mathbf{x}_0 = \mathbf{x}_r$. We detail the full formulation, as well as the design choices for α and β, in Subsection 3.4.

3.2 Network Architecture

We adopt a dual U-Net architecture, where the reference U-Net encodes garment features and integrates them into the main denoising U-Net via self-attention layers, a structure proven effective for garment fidelity and visual quality (Choi et al., 2024). Following the approach of (Chong et al., 2025), we remove all cross-attention layers, relying solely on self-attention, which improves performance and efficiency, as shown by our ablation studies (see Appendix A.2.1). To enhance computational efficiency, we execute the reference U-Net once per sample, using it solely as a conditioning module, as in (Li et al., 2024).

We initialize our U-Net weights using those from Stable Diffusion 1.5 (Rombach et al., 2022). While SDXL (Podell et al., 2025) offers greater generative power, our goal is a lightweight yet effective framework, so we retain SD1.5 as the backbone. Preliminary experiments with transformer-based architectures like SD3 and SD3.5 are discussed in Appendix A.3. In conclusion, SD1.5 strikes the optimal balance between model simplicity and try-on performance, forming the foundation of our pipeline.

3.3 LOW-RESOLUTION STAGE

In this stage, we first downsample the garment image $\mathbf{x}_g$ and the person image $\mathbf{x}_p$ by the downsampling ratio σ. The downsampled garment image $\tilde{\mathbf{x}}_g$ is encoded by a VAE (Kingma, 2013) and then fed into the reference U-Net. The downsampled person image $\tilde{\mathbf{x}}_p$ is also VAE-encoded and concatenated along the feature channels with a Gaussian noise tensor of the same shape. This combined latent is passed to the denoising U-Net to generate a low-resolution result $\tilde{\mathbf{x}}_r \in \mathbb{R}^{h\times w\times 3}$. Both training and inference follow the standard diffusion process used in Stable Diffusion (Rombach et al., 2022; Ho et al., 2020; Song et al., 2020).

The only hyperparameter in this stage is the downsampling ratio σ, with $\sigma \in \{1, 2, 4\}$. As illustrated in Figure 2, when $\sigma = 1$, the low-resolution stage operates at the same resolution as the high-resolution stage, **violating the purpose of leveraging lower resolution for improved structural modeling and introducing artifacts due to difficulty**. $\sigma = 2$ corresponds to downsampling from 768×1024 to 384×512, and so on. Our observations show that both $\sigma = 2$ and $\sigma = 4$ enhance human-body structural understanding, **but, as also evident in Figure 2,** $\sigma = 4$ **sacrifices structural detail**, making $\sigma = 2$ the most reliable choice for accurate garment reconstruction. Quantitative and qualitative comparisons across different σ values are provided in Subsection 4.3, and $\sigma = 2$ is adopted for all experiments in this work.

3.4 HIGH-RESOLUTION STAGE

In this stage, $\mathbf{x}_g$ and $\mathbf{x}_p$ are used in the same way as in the low-resolution stage: $\mathbf{x}_g$ is passed to the reference U-Net, while $\mathbf{x}_p$ is encoded and concatenated with the latent, which is then input to the denoising U-Net. The main difference from the low-resolution stage lies in the initialization of the latent and the denoising process.

3.4.1 REFORMULATING THE DENOISING PROCESS WITH BLEND-REFINE

DDPM. Denoising Diffusion Probabilistic Mode (DDPM) (Ho et al., 2020) aims to approximate the true data distribution by leveraging the diffusion probabilistic model framework (Sohl-Dickstein et al., 2015), which defines a Markov chain of length T that gradually transforms pure Gaussian noise into a sample from the data distribution. Compared with earlier diffusion models, DDPM incorporates ideas from score matching (Song & Ermon, 2019), simplifying the objective by training the model to predict only the noise component $\boldsymbol{\epsilon}$, which approximates the score function (*i.e.*, the gradient of the log-density). The forward and reverse diffusion processes are defined as:

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t}\,\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon} \\ &= \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad (2)\end{aligned}$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}. \qquad (3)$$

Equation 2 and Equation 3 define the forward and reverse diffusion processes, respectively. In these equations, $\boldsymbol{\epsilon}, \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ denotes the noise predicted by the model. The parameters α_t and $\bar{\alpha}_t$ are predefined noise scheduling terms. In this framework, $\mathbf{x}_0 \sim p_{data}(\mathbf{x}_0)$ denotes a sample from the true data distribution, while $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. All formulations above hold for $t = 1, 2, \ldots, T$.

Blend-refine diffusion reformulation. Inspired by (Yue et al., 2023), we reformulate the forward and reverse processes to **explicitly** incorporate the low-resolution result $\tilde{\mathbf{x}}_r$ as a structural prior for high-resolution generation. **Rather than constructing a Markov chain only between Gaussian noise and $p_{\text{data}}(\mathbf{x}_0)$, we aim to build a transition path from $\tilde{\mathbf{x}}_r$ to the high-resolution result $\mathbf{x}_r$.** In the original formulation of (Yue et al., 2023), $\mathbf{x}_T = \boldsymbol{\epsilon} + \tilde{\mathbf{x}}_r$ with $\mathbf{x}_0 = \mathbf{x}_r$, which offers limited flexibility to control the relative influence of $\tilde{\mathbf{x}}_r$ and $\boldsymbol{\epsilon}$. To overcome this limitation, we introduce a simple extension with two coefficients for more flexible initialization, as defined in Equation 1, while keeping $\mathbf{x}_0 = \mathbf{x}_r$. **This straightforward modification yields significant improvements, as shown in Subsection 4.3.** Under this formulation, the blend-refine forward and reverse diffusion processes

become:

$$\begin{aligned}\mathbf{x}_t &= \sqrt{\alpha_t}\,\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\,(\alpha\cdot\boldsymbol{\epsilon} + \beta\cdot\tilde{\mathbf{x}}_r) \\ &= \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,(\alpha\cdot\boldsymbol{\epsilon} + \beta\cdot\tilde{\mathbf{x}}_r)\,, \end{aligned} \tag{4}$$

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}. \tag{5}$$

Equation 4 and Equation 5 define the forward and reverse processes, respectively. The model is trained to predict $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t) \approx \alpha\cdot\boldsymbol{\epsilon} + \beta\cdot\tilde{\mathbf{x}}_r$, such that the noise component in $\mathbf{x}_t$ is replaced by a composition of $\boldsymbol{\epsilon}$ and the low-resolution result $\tilde{\mathbf{x}}_r$. Except for this reformulated initialization and noise target, the rest of the denoising process remains consistent with the original DDPM. All formulations above hold for $t = 1, 2, \ldots, T$.

3.4.2 Controlling noise-structure balance

In the high-resolution stage, generation starts from the latent $\mathbf{x}_T = \alpha\cdot\boldsymbol{\epsilon} + \beta\cdot\tilde{\mathbf{x}}_r$, a combination of stochastic noise and a structural prior. Here, α controls randomness and β controls structural guidance. **Intuitively, β should reflect the similarity between the two-scale distributions, while α should correspond to the size of the probability space associated with the residual term $\mathbf{x}_r - \beta\cdot\tilde{\mathbf{x}}_r$.** As illustrated in Figure 2, an excessively large β forces overreliance on the low-resolution input, suppressing fine-detail recovery, whereas a large α introduces noise that can disrupt structure. Empirically, setting $\alpha = \beta = 0.5$ provides a balanced trade-off between fidelity and flexibility. Further discussion is provided in Subsection 4.3.

4 Experiments

4.1 Experimental setup

Datasets. In this paper, we conduct experiments on the VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) datasets. All ablation studies are carried out on VITON-HD. While VITON-HD contains only upper-body garments, DressCode includes three garment categories: upper-body, lower-body, and dresses. Both datasets consist of paired images, each containing a person image and a corresponding garment image. However, as our method is mask-free, the original paired data alone is insufficient for training. To address this, we use IDM-VTON (Choi et al., 2024) to synthesize additional training data. Further details are provided in Appendix A.1.1.

Implementation details. For network initialization, both the reference U-Net and the denoising U-Net are initialized with pretrained weights from Stable Diffusion 1.5 (Rombach et al., 2022). As detailed in Subsections 3.3 and 3.4, we set the downsampling ratio $\sigma = 2$ and use $\alpha = \beta = 0.5$ to initialize $\mathbf{x}_T$ in the high-resolution stage. Consequently, the low-resolution stage operates at a resolution of 384×512, while the high-resolution stage produces outputs at 768×1024. During inference, the two stages are executed sequentially with 20 sampling steps each, using the DDIM sampler (Song et al., 2020). Additional training details are provided in Appendix A.1.2.

Baselines. We compare our method with several recent state-of-the-art approaches, including CatVTON (Chong et al., 2025), IDM-VTON (Choi et al., 2024), Leffa (Zhou et al., 2025), OOT-Diffusion (Xu et al., 2025), and FitDiT (Jiang et al., 2024), using their official model weights and inference code. For fairness, we standardize the number of inference steps to 30 across all methods. All methods, except FitDiT, are trained separately on each dataset and evaluated accordingly. FitDiT provides only a single set of pretrained weights, jointly trained on VITON-HD, DressCode, and CVDD (Jiang et al., 2024), and is included in our evaluation for completeness.

Evaluation metrics. Previous virtual try-on methods typically evaluate performance under both paired and unpaired settings. The paired setting involves reconstructing the original person image with the same garment, while the unpaired setting replaces it with a different one (Choi et al., 2021). As most prior approaches rely on masking the garment region, they support both settings. In contrast, our method is mask-free and is therefore evaluated only under the unpaired setting, which better reflects real-world scenarios. We adopt Fréchet Inception Distance (FID) (Parmar et al., 2022) and

Figure 3: Qualitative comparison on the VITON-HD dataset. DS-VTON(LR) denotes the low-resolution result, and DS-VTON(HR) represents the final high-resolution result.

Kernel Inception Distance (KID) (Bińkowski et al., 2018) as quantitative metrics. We also conduct a user study to assess perceptual quality: for VITON-HD, we randomly sample 100 results per method; for DressCode, we sample 33, 33, and 34 results from the dresses, lower-body, and upper-body categories, respectively. Participants are asked to select the result they think performs better in the try-on task. All evaluations are conducted at a resolution of 768×1024.

4.2 Quantitative and Qualitative Results

Qualitative comparison. Figure 3 presents a qualitative comparison between DS-VTON and recent baseline methods on the VITON-HD (Choi et al., 2021) dataset. Row numbers referenced below correspond to Figure 3. We evaluate each method in terms of two key aspects: structural alignment and detail preservation. In terms of structural alignment, most mask-based methods fail to accurately capture body pose and garment shape, as illustrated in Row 1. FitDiT (Jiang et al., 2024), which uses a larger rectangular mask, performs relatively better but introduces noticeable artifacts: it fails to reconstruct the hands and exhibits visible artifacts at the junction of the upper and lower garments in Row 1. OOTDiffusion (Xu et al., 2025) alters the original skin tone in Row 2. CatVTON (Chong et al., 2025), IDM-VTON (Choi et al., 2024), and Leffa (Zhou et al., 2025) also show varying degrees of misalignment. In contrast, DS-VTON consistently achieves accurate alignment across a wide range of poses and garment types. Regarding detail preservation, CatVTON and IDM-VTON fail to retain fine-grained textures on complex garments (Row 3), with IDM-VTON generating oversimplified or missing patterns. While Leffa, OOTDiffusion, and FitDiT better preserve textures, they each show limitations: FitDiT achieves texture preservation at the cost of distorting the person's actual body shape and dressing structure (Rows 2 and 4), OOTDiffusion introduces artifact patterns (Row 5),

Table 1: Quantitative comparisons on the VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) datasets. FitDiT (Jiang et al., 2024), trained jointly on VITON-HD, DressCode, and CVDD (Jiang et al., 2024), is included for completeness. In contrast, all other methods are trained individually on each dataset.

Dataset	VITON-HD			DressCode		
Method	FID ↓	KID ↓	User Study ↑	FID ↓	KID ↓	User Study ↑
OOTDiffusion (Xu et al., 2025)	<u>9.02</u>	<u>0.63</u>	4.1	7.10	2.28	7.2
IDM-VTON (Choi et al., 2024)	9.10	1.06	11.6	5.51	1.42	9.1
CatVTON (Chong et al., 2025)	9.40	1.27	3.4	5.24	1.21	5.2
Leffa (Zhou et al., 2025)	9.38	0.92	4.7	6.17	1.90	7.5
FitDiT (Jiang et al., 2024)	9.33	0.89	<u>19.7</u>	<u>4.47</u>	<u>0.41</u>	<u>34.3</u>
DS-VTON (ours)	**8.24**	**0.31**	**56.5**	**4.21**	**0.34**	**36.7**

Table 2: Ablation study on dual-scale design and downsampling ratio σ.

Version	FID ↓	KID ↓
$\sigma = 1$	8.97	1.01
$\sigma = 1, \alpha = \beta = \frac{1}{2}$	8.77	0.61
$\sigma = 4, \alpha = \beta = \frac{1}{2}$	8.41	0.57
$\boldsymbol{\sigma = 2, \alpha = \beta = \frac{1}{2}}$	**8.24**	**0.31**

Table 3: Ablation study on coefficients α, β under fixed $\sigma = 2$.

Version	FID ↓	KID ↓
$\boldsymbol{\sigma = 2, \alpha = \beta = \frac{1}{2}}$	**8.24**	**0.31**
$\sigma = 2, \alpha = \frac{2}{3}, \beta = \frac{1}{3}$	8.46	0.55
$\sigma = 2, \alpha = \frac{1}{3}, \beta = \frac{2}{3}$	8.26	0.35
$\sigma = 2, \alpha = \beta = 1$	8.75	0.94

and Leffa exhibits reduced pattern clarity in complex regions (Row 3). Furthermore, both FitDiT and Leffa suffer from tonal inconsistencies—FitDiT produces noticeably brighter garments (Row 6), while Leffa tends to generate darker outputs. In contrast, DS-VTON preserves fine textures while maintaining tonal fidelity throughout.

Quantitative comparison. We conduct experiments on both the VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) datasets. As shown in Table 1, DS-VTON achieves substantial improvements across both benchmarks. CatVTON (Chong et al., 2025) generates images at 384×512, which we upsample to 768×1024 for comparison. To confirm that this does not bias results, we also evaluate it at native resolution, obtaining FID and KID scores of 9.36 and 1.19, respectively—indicating minimal degradation due to upsampling. Unlike prior methods (Chong et al., 2025; Zhou et al., 2025; Choi et al., 2024) that rely on explicit mask generation and inpainting, DS-VTON is entirely mask-free. This enables accurate, high-quality outputs that are robust and reproducible without dependence on mask quality.

4.3 Ablation Study

Here we present several ablation studies to validate the rationale of our design. **In addition, we evaluate an alternative refinement method used in SDXL; see Subsection A.3 for details.**

Ablation on dual-scale design. To validate the effectiveness of our dual-scale design, we train four variants on the VITON-HD dataset. As shown in Table 2, Row 1 ($\sigma = 1$) corresponds to training the model directly at high resolution (768×1024) without a refinement stage, aligning with the $\sigma = 1$ (LR) column in Figure 4. **Applying the mask-free strategy at this resolution leads to poor structural results, highlighting the need for coarse-level guidance.** Row 2 ($\sigma = 1, \alpha = \beta = \frac{1}{2}$), shown in the $\sigma = 1$ (HR) column, adds a refinement stage. While some structural errors are alleviated, relying solely on the second stage to recover both structure and detail still results in failures, due to the lack of reliable low-resolution guidance. This also violates our design principle of decoupling structure modeling (stage one) from detail refinement (stage two). Rows 3 and 4 evaluate the full dual-scale pipeline with $\sigma = 4$ and $\sigma = 2$, respectively. When garments are structurally simple, both settings perform reasonably well. However, as shown in Row 1 of Figure 4, $\sigma = 4$ **introduces visible information loss in complex cases.** These qualitative results align with the quantitative trends: among the four variants, the single stage ($\sigma = 1$) and the dual stage with $\sigma = 1$, $\sigma = 2$, or $\sigma = 4$, the

Figure 4: Visualized results under varying downsampling ratios σ.

Figure 5: Visualized results under different $\mathbf{x}_T$ initialization settings ($\mathbf{x}_T = \alpha \cdot \boldsymbol{\epsilon} + \beta \cdot \tilde{\mathbf{x}}_r$).

single stage model performs worst, $\sigma = 4$ shows moderate improvement, and $\sigma = 2$ achieves the best overall performance.

Ablation on different initializations of $\mathbf{x}_T$. When fixing the downsampling ratio to $\sigma = 2$, the coefficients α and β determine how strongly the low-resolution output influences the high-resolution stage. As shown in Figure 5, setting α too high, e.g., $\alpha = 1, \beta = 1$ or $\alpha = \frac{2}{3}, \beta = \frac{1}{3}$, causes structural distortions in the final result even though the low-resolution output already provides accurate guidance. In both cases the red stripe on the garment is distorted, and in the $\alpha = \frac{2}{3}, \beta = \frac{1}{3}$ configuration the "GANT" text also becomes warped. **We attribute this to excessive randomness in the high-resolution stage, which then attempts to re-establish structure that should have been fully resolved in the low-resolution stage.** Among the remaining settings, $\alpha = \frac{1}{3}, \beta = \frac{2}{3}$ and $\alpha = \frac{1}{2}, \beta = \frac{1}{2}$ produce generally good results. However, the $\alpha = \frac{1}{3}, \beta = \frac{2}{3}$ setting still exhibits minor distortion in the red stripe, indicating insufficient restoration of fine details. **Here, the low-resolution result exerts too strong an influence: while structurally accurate, it lacks fine details, and the high β prevents the high-resolution stage from effectively correcting those errors.** Based on these observations, we adopt $\alpha = \beta = \frac{1}{2}$ as our default setting. The quantitative results in Table 3 support this choice.

5 Conclusions

We propose DS-VTON, a dual-scale framework that employs blend-refine denoising to bridge low- and high-resolution generation. This design enables a more effective coarse-to-fine process and, combined with a mask-free strategy, achieves significant improvements in both visual quality and robustness over existing methods. Furthermore, the paradigm is inherently scalable and generalizable, with clear potential for extension to higher resolutions and to broader generation tasks beyond virtual try-on, such as personalized image synthesis.

REFERENCES

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11472–11481, 2022.

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 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.

Youngjin Choi, Seunghyun Kwak, Kyungjune Lee, Hyojin Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision (ECCV)*, pp. 206–235, Cham, 2024. Springer Nature Switzerland.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive Theory of Functions of Several Variables: Proceedings of a Conference held at Oberwolfach April 25–May 1, 1976*, pp. 85–100. Springer Berlin Heidelberg, 1977.

Patrick Esser, Shubham Kulal, Andreas Blattmann, Reza Entezari, Jonas Müller, Harsh Saini, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16928–16937, 2021a.

Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8485–8493, 2021b.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 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.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7543–7552, 2018.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8153–8163, 2024.

Bin Jiang, Xiaoxiao Hu, Dongdong Luo, Qian He, Chen Xu, Jing Peng, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*, 2024.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 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.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pp. 204–219. Springer, 2022.

Yexin Li, Haoyu Zhou, Weichen Shang, Runyu Lin, Xinyu Chen, and Bingbing Ni. Anyfit: Controllable virtual try-on for any combination of attire across any scenario. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017.

Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. 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.

Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8580–8589, 2023.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11410–11420, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Daniel Podell, Zana English, Kenneth Lacey, Andreas Blattmann, Tobias Dockhorn, Jonas Müller, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

K. Sun, J. Cao, Q. Wang, L. Tian, X. Zhang, L. Zhuo, and D. Gao. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*, 2024.

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. 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.

Yuhao Xu, Tao Gu, Weifeng Chen, and Aoxue Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8996–9004, 2025.

Zhen Xu, Jing Zhang, Jun Hao Liew, Hongdong Yan, Jianwen Liu, Chunyan Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Hongwen Yang, Rongyao Zhang, Xiaonan Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7850–7859, 2020.

Hailin Ye, Jing Zhang, Shichao Liu, Xiangyu Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Zongyu Yue, Jingyun Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 13294–13307, 2023.

L. Zhang. Reference-only controlnet. `https://github.com/Mikubill/sd-webui-controlnet/discussions/1236`, 2023. Accessed: 2023-04.

Shufang Zhang, Minxue Ni, Shuai Chen, Lei Wang, Wenxin Ding, and Yuhong Liu. A two-stage personalized virtual try-on framework with shape control and texture guidance. *IEEE Transactions on Multimedia*, 26:10225–10236, 2024.

Ziqian Zhou, Shichao Liu, Xiangyu Han, Hao Liu, Kwan-Yee Ng, Ting Xie, and Shimin He. Learning flow fields in attention for controllable person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4606–4615, 2023.

A Appendix

A.1 Experimental details

A.1.1 Additional training data generation and associated challenges

Our method adopts a mask-free paradigm, in contrast to prior approaches (Choi et al., 2024; Xu et al., 2025; Li et al., 2024), which rely on paired person-garment images by masking out the garment region in the person image and reconstructing it using the standalone garment image. In our case, the person image remains unaltered throughout the training process. To train the low-resolution stage, each training sample requires three images: (1) a garment image, (2) a person image wearing that garment, and (3) another person image of the same identity but wearing a different garment. The third image is constructed by randomly sampling another garment of the same category from the dataset and synthesizing a new person image using IDM-VTON (Choi et al., 2024). For the high-resolution stage, we additionally require the low-resolution output corresponding to the original person-garment pair. To obtain this, we use our trained low-resolution model to generate a coarse try-on result by inputting the target garment and the synthesized person image (i.e., with a different garment). This output is then used as the low-resolution input for the high-resolution stage.

In summary, each high-resolution training sample consists of: (1) the target garment image, (2) a person image wearing a different garment, (3) the corresponding low-resolution try-on result for the target garment, and (4) the ground-truth high-resolution image of the person wearing the target garment.

While the above construction enables mask-free supervision, it inevitably introduces certain artifacts. Specifically, in the low-resolution stage, the reference U-Net encodes the garment image, while the

denoising U-Net operates on the concatenation of the noisy latent and the person image—here, a synthesized image of the same identity wearing a different garment.

Since the synthesized person image is generated by a model such as IDM-VTON, it may exhibit variations beyond the garment itself, including changes in hairstyle, background, or the presence of accessories. As a result, the model may inadvertently learn to alter these unrelated regions during training. Ideally, such issues would not occur if fully disentangled and clean ground-truth data were available. Fortunately, we find this side effect to be limited in practice, as most synthesized person images remain visually consistent with the original identity.

A.1.2 More implementation details

All experiments are conducted on 8 NVIDIA A6000 GPUs. For both VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022), the low-resolution stage is trained with a batch size of 8, and the high-resolution stage with a batch size of 2. Both stages are optimized using the AdamW optimizer (Loshchilov et al., 2017) with a learning rate of $1e{-}6$. For VITON-HD, the low-resolution and high-resolution stages are trained for 15,000 and 30,000 steps, respectively (approximately 5 and 24 hours). For DressCode, we jointly train all three garment categories, with the two stages trained for 20,000 and 30,000 steps (approximately 7 and 24 hours). During inference, both stages use 20 DDIM sampling steps. On a single A6000 GPU, the low-resolution stage takes about 1 second per sample, while the high-resolution stage takes about 4 seconds. Although our method includes two stages, the low-resolution stage is much faster than the high-resolution stage, so the total runtime remains acceptable. **Compared with the methods in our earlier comparison, our approach is only slower than CatVTON, comparable to FitDiT, and faster than the others.**

A.1.3 Additional Comparisons with Other Methods

We detail here the data requirements of the methods introduced in Subsection 4.1 to clarify the comparisons. For completeness, we also present paired-setting results on VITON-HD. All competing methods except ours require an agnostic mask; Leffa additionally uses a DensePose map; FitDiT requires human body keypoints; and IDM-VTON depends on both detailed garment descriptions and a DensePose map. Our method, by contrast, is trained only on paired images of the same person wearing different garments.

Because our approach does not mask the original garment region, the input and expected output under the paired setting are identical. Nevertheless, we include two evaluation variants:

- **DS-VTON (direct)**: directly uses the garment and person images to generate the result.
- **DS-VTON (train-way)**: first employs IDM-VTON to create an image of the person wearing another garment, then applies our method to re-dress the original garment. This procedure involves two rounds of unpaired try-on and inevitably introduces additional artifacts, so it does not fully reflect our method's capability.

The quantitative results appear in Table 4.

A.2 More ablation studies

A.2.1 Ablation on removing cross-attention layers in U-Net

Several previous try-on methods (Choi et al., 2024; Xu et al., 2025; Li et al., 2024) based on dual U-Net architectures incorporate additional conditional encoders, such as CLIP (Radford et al., 2021) or IP-Adapter (Ye et al., 2023), to inject garment information via cross-attention. However, these approaches introduce extra complexity, and it remains unclear whether they actually improve performance. To investigate this, we conduct an ablation study on the VITON-HD (Choi et al., 2021) dataset by comparing two versions of our low-resolution stage (384×512): (1) our baseline model, which does not include any cross-attention layers, and (2) a variant built on the baseline by adding cross-attention layers to both the reference U-Net and the denoising U-Net. In the latter, garment features are encoded using CLIP and injected via cross-attention. As shown in Figure 6 and Table 5, adding cross-attention does not help preserve garment details; on the contrary, it introduces noticeable distortions. This is also reflected in the performance metrics.

Table 4: Quantitative comparisons on the VITON-HD dataset under the paired setting.

Method	FID ↓	KID ↓	SSIM ↑	LPIPS ↓
OOTDiffusion	6.47	1.24	0.88	0.08
IDM-VTON	5.84	0.77	0.87	0.06
CatVTON	5.70	0.50	0.88	0.09
Leffa	5.76	0.55	0.89	0.06
FitDiT	7.27	0.73	0.84	0.09
DS-VTON (direct)	**4.75**	0.43	**0.90**	**0.05**
DS-VTON (train-way)	5.23	**0.31**	0.89	0.06

Figure 6: Comparison between our baseline architecture and the variant with cross-attention. In the variant, garment features encoded by CLIP are injected into both the reference U-Net and the denoising U-Net via cross-attention layers.

A.3 Discussions

How about utilizing the DiT architecture for mask-free try-on? We also explored using the DiT (Peebles & Xie, 2023) architecture for mask-free try-on. Specifically, we implemented two variants based on SD3 and SD3.5 (Esser et al., 2024), constructing a dual-DiT structure analogous to the dual U-Net design. The reference DiT encodes the garment image, while the person image is concatenated with the latent tensor along the sequence dimension and passed to the denoising DiT. To integrate the two branches, we follow a structure-aligned fusion strategy: since the reference DiT and denoising DiT share the same architecture, we concatenate the latent features from the corresponding transformer block of the reference DiT into the denoising DiT block, specifically on the key and value (K, V) inputs of the attention layer, before computing self-attention. We also remove the text encoder input entirely, so the joint attention layers in DiT degenerate into pure self-attention. However, both versions failed to converge during training. We speculate that directly applying DiT to the mask-free try-on task may be suboptimal, and therefore did not pursue further investigation.

How about adopting a refinement mechanism similar to that of SDXL? We refer to the refinement method in SDXL (Podell et al., 2025) as the SDXL strategy. During training, a separate model is trained to denoise only the final 200 steps of the high-resolution diffusion process. At inference, a low-resolution result is generated, upsampled, and injected with Gaussian noise at timestep 200, after which the refinement model denoises to produce the final output. Although this method can yield visually plausible results for simple patterns, its overall performance is inferior, as shown in Figure 7. **The reason, we believe, lies in the difference between data distributions involved.** Although our blend-refine diffusion process also begins with adding noise, it differs fundamentally in how it relates the two stages. In SDXL, the refinement model learns to denoise samples drawn solely from the high-resolution distribution, and it lacks an explicit mechanism to relate this with the low-resolution result. In contrast, our blend-refine diffusion bridges the gap between the low-resolution and high-resolution distributions. This connection is key: it allows the model to better capture the transformation between the two distributions involved. Importantly,

Table 5: Ablation study on the effect of cross-attention layers in U-Net.

Version	FID ↓	KID ↓
With cross-attention	9.07	0.94
DS-VTON (low-resolution stage)	**8.88**	**0.72**

Figure 7: Comparison of qualitative results between our baseline and the SDXL variant. Quantitatively, the SDXL variant achieves an FID/KID of **8.98/0.65**, while our DS-VTON achieves **8.24/0.31**.

while the original diffusion process constructs a mapping from a simple, tractable distribution (such as Gaussian noise) to a complex data distribution, the blend-refine diffusion builds a direct bridge between two complex distributions.

Difference between Our Blend-Refine Process and the DCI-VTON Refinement Branch. DCI-VTON also adds noise to the first-stage result and attempts to recover the final output in its second-stage refinement branch. However, it does not explicitly transition the data distribution from the first stage to the second. Its second stage contains two branches: a refinement branch and a reconstruction branch. In the refinement branch, when the added noise is weak, the noisy input is nearly identical to the first-stage result, so the denoised output remains overly similar to the input and struggles to reach the desired final distribution. This is why **DCI-VTON requires two branches** and adopts a perceptual (VGG) loss rather than a pixel-wise loss in the refinement branch. The refinement branch provides only implicit guidance, while the reconstruction branch handles explicit Gaussian noise prediction. By contrast, DS-VTON needs only a single branch. Our blend-refine denoising explicitly transfers the distribution from coarse to fine by altering the noise formulation itself, rather than relying on pure Gaussian noise. This change enables a principled transition between the low- and high-resolution distributions and provides stronger controllability—an essential design choice of our dual-scale framework. To illustrate, suppose we follow DCI-VTON's approach. Let $\tilde{\mathbf{x}}_r$ denote the low-resolution output, $\boldsymbol{\epsilon}$ the noise, and $\mathbf{x}_r$ the high-resolution ground truth. From a data distribution perspective, if we adopt pure Gaussian noise as the only perturbation mechanism, there is a fundamental mismatch between the training and inference processes. During training, at any timestep t, the noisy latent is constructed by adding noise to $\tilde{\mathbf{x}}_r$, i.e., it is solely a function of $\tilde{\mathbf{x}}_r$ and $\boldsymbol{\epsilon}$. However, this setup does not align with inference. In inference, regardless of how the latent is initialized (whether from pure Gaussian noise or some other distribution), the moment a single denoising step is performed, the latent becomes correlated with the target distribution $\mathbf{x}_r$. This stands in stark contrast to training, where the noisy latent at each timestep is always constructed based on $\tilde{\mathbf{x}}_r$. **The key difference is that we modify the noise formulation: rather than using pure Gaussian noise, our blend-refine denoising bridges the low- and high-resolution data distributions, providing strong controllability and serving as a fundamental design choice in our dual-scale framework.**

Broader impacts. The ability of DS-VTON to generate realistic virtual try-on results at multiple resolutions makes it well-suited for practical deployment in e-commerce scenarios, where different resolutions are often required across platforms and devices. At the same time, as with other generative technologies, DS-VTON may raise concerns related to intellectual property and personal privacy. We encourage its responsible and ethical use.

Limitation and Future Work. As discussed in Subsection A.1.1, one key limitation of our method lies in the data generation process. Due to the reliance on synthesized person images, the model may inadvertently learn to alter regions unrelated to the garment (e.g., hair, accessories, or background). While this issue is not severe in most cases, we acknowledge it as a limitation and consider improving data disentanglement and identity preservation an important direction for future work. Another limitation stems from the use of fixed coefficients α and β to initialize the high-resolution refinement stage. Although this static strategy proves effective, it may be overly rigid. In future work, we plan to investigate adaptive or learnable coefficient scheduling mechanisms, which could offer more flexible and content-aware refinement during generation.

A.4 The Use of Large Language Models

The Large Language Model was used solely for refining the text and improving the clarity and readability of the paper. It did not contribute to the research ideation or experimental design.

A.5 More experiment results

In this section, we present additional qualitative comparisons on the DressCode (Morelli et al., 2022) dataset, more results on the VITON-HD (Choi et al., 2021) dataset, and additional in-the-wild examples.

Figure 8: More results on in-the-wild scenarios.

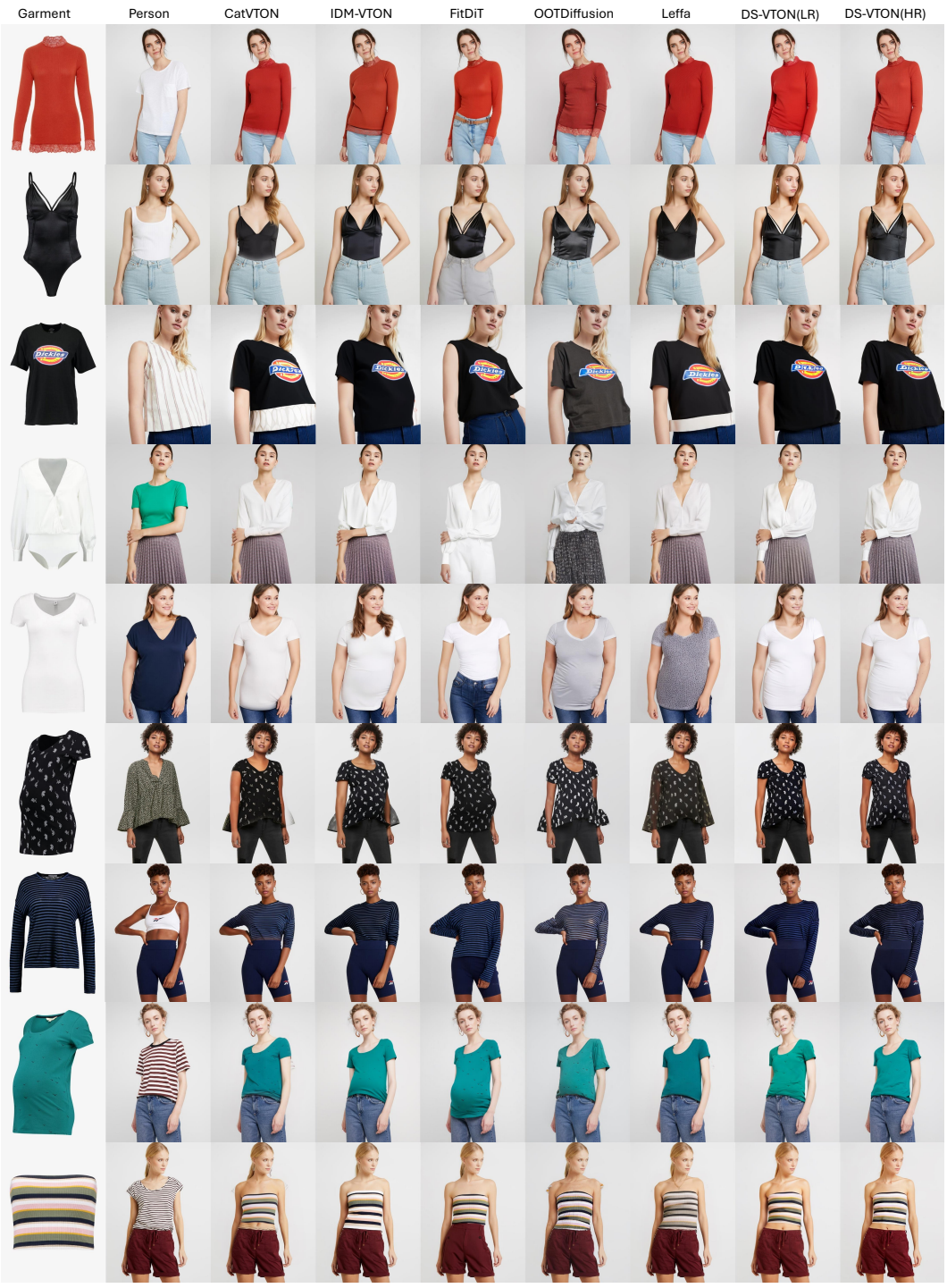

Figure 9: More qualitative comparison on the VITON-HD dataset. DS-VTON (LR) denotes the low-resolution output, and DS-VTON (HR) represents the final high-resolution result.

Figure 10: Qualitative comparison on the DressCode dataset (Dresses category). DS-VTON (LR) denotes the low-resolution output, and DS-VTON (HR) represents the final high-resolution result.

Figure 11: Qualitative comparison on the DressCode dataset (Lower category). DS-VTON (LR) denotes the low-resolution output, and DS-VTON (HR) represents the final high-resolution result.

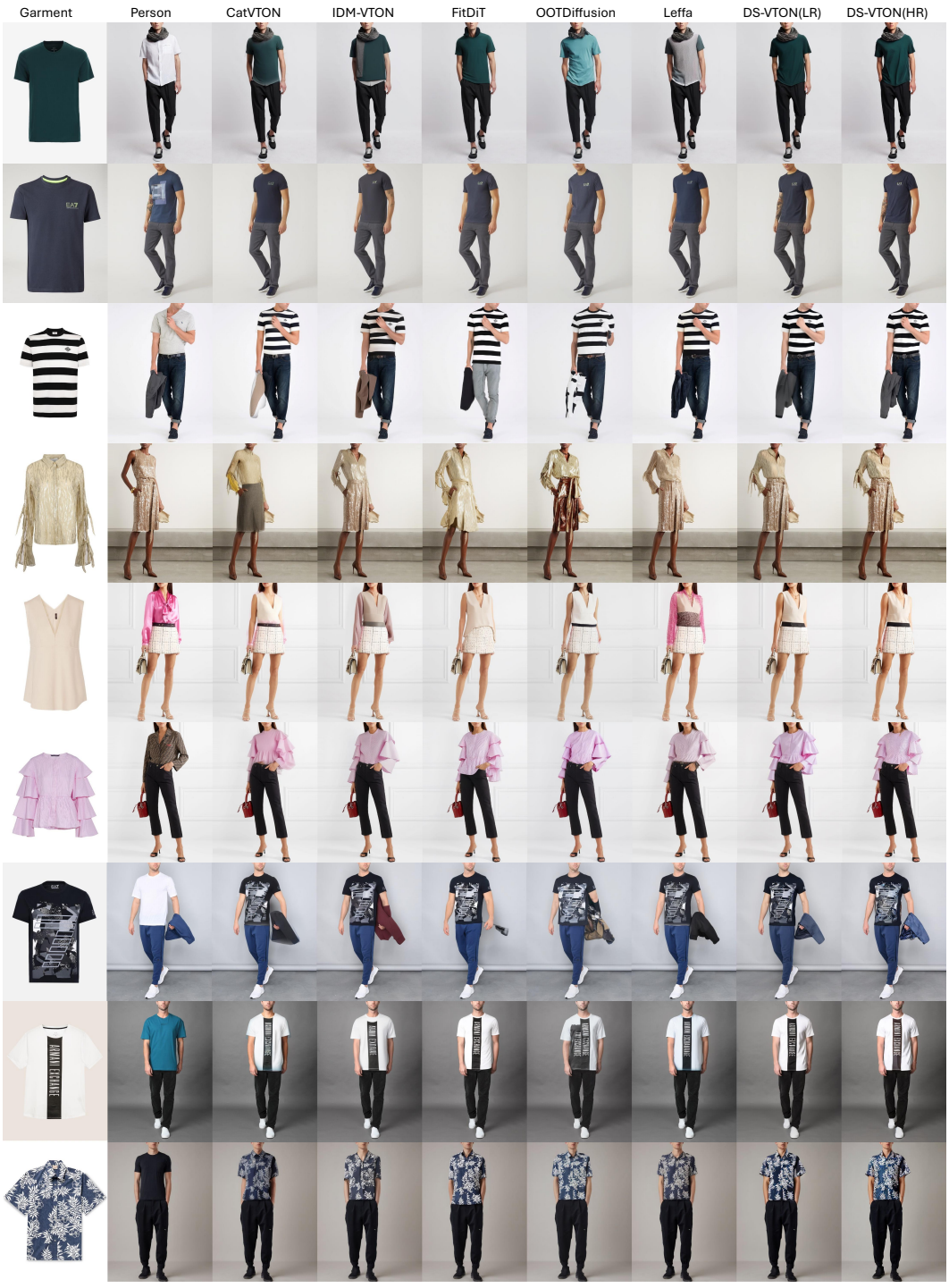

Figure 12: Qualitative comparison on the DressCode dataset (Upper category). DS-VTON (LR) denotes the low-resolution output, and DS-VTON (HR) represents the final high-resolution result.