DIRECTTRYON: ONE-STEP VIRTUAL TRY-ON VIA STRAIGHTENED CONDITIONAL TRANSPORT

Anonymous authors
Paper under double-blind review

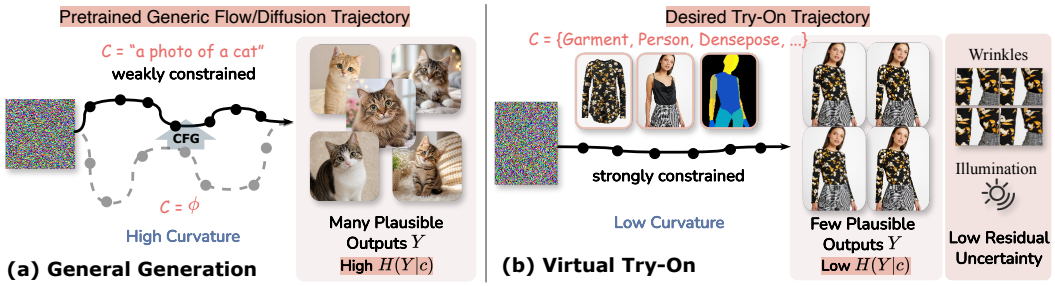

Figure 1: Comparison between general image generation and virtual try-on from the perspective of conditional transport. Virtual try-on is strongly constrained by the person and garment inputs, resulting in fewer plausible outputs and lower residual uncertainty. This suggests a much straighter desired transport path, which is the core intuition of this paper.

ABSTRACT

Diffusion- and flow-matching-based virtual try-on (VTON) methods produce high-quality images but often rely on costly multi-step sampling. Unlike general image generation, VTON is strongly constrained by person and garment images, leaving fewer plausible outputs and lower residual uncertainty. Motivated by the potential for straighter conditional transport in this setting, we propose DirectTryOn, a two-stage framework for one-step VTON. In the first stage, we fine-tune a pretrained teacher model to encourage straighter conditional transport trajectories from noise to the try-on image. This stage uses pure conditional transport, training only on conditional samples and omitting classifier-free guidance at inference. A garment preservation loss encourages retention of garment details across timesteps, while a self-consistency loss aligns velocity predictions across arbitrary timestep pairs. Together with extended fine-tuning, these designs improve one-step generation even before distillation. In the second stage, we distill the straightened teacher into a one-step student model using latent adversarial diffusion distillation to further improve generation quality. Experiments on VITON-HD and DressCode show that DirectTryOn achieves the best performance among the evaluated baselines using a single sampling step, with an approximately 19× end-to-end speedup over the fastest evaluated baseline.

1 INTRODUCTION

Given a garment image and a person image, virtual try-on aims to synthesize a photorealistic image of the person wearing the specified garment (Han et al., 2018). As an important technology for online fashion retail and e-commerce, virtual try-on has attracted growing attention in recent years (Choi et al., 2021; Ge et al., 2021b; Gou et al., 2023; Lee et al., 2022; Morelli et al., 2022; 2023; Choi et al., 2024; Chong et al., 2025; Zhou et al., 2025). Earlier virtual try-on methods (Choi et al., 2021; Ge et al., 2021a; Lee et al., 2022; Xie et al., 2023) were mainly based on generative adversarial networks (GANs) (Goodfellow et al., 2020). Recent advances in diffusion models (Ho et al., 2020; Rombach et al., 2022) and flow matching (Lipman et al., 2023) have shifted the field toward adapting powerful pretrained generative models, such as Stable Diffusion variants (Rombach et al., 2022;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Esser et al., 2024) and Flux (Black Forest Labs, 2024), for virtual try-on (Kim et al., 2024; Zhu et al., 2023; Choi et al., 2024; Zhou et al., 2025). Although these methods achieve promising results by leveraging strong pretrained priors, their reliance on multi-step sampling leads to substantially increased inference cost, which makes them less practical for real-world applications. To improve inference efficiency, existing VTON acceleration methods reduce sampling steps through different strategies. CAT-DM (Zeng et al., 2024) starts diffusion from a coarse try-on result, whereas MC-VTON (Luan et al., 2025) distills a trained model into an eight-step generator. We investigate a complementary but fundamental question: *how can the strong constraints imposed by the person and garment images guide model training for one-step VTON?*

We first revisit this intuition from flow matching. In general, although the sample-wise optimal transport paths are straight, the marginal velocity field can become curved after marginalizing over the data distribution (Lipman et al., 2023). However, in an extreme conditional case where the target is exactly determined by the condition image y, the conditional target distribution collapses to a single point. For each noise sample ϵ, the path is $x_t = ty+(1-t)\epsilon$, with velocity $v_t(x_t \mid y) = y-\epsilon = \frac{y-x_t}{1-t}$. Since conditioning on y leaves no ambiguity over the endpoint, marginalization does not bend the conditional velocity field. The ideal conditional transport therefore remains straight, so one-step and multi-step sampling recover the same output y. This example suggests that when the condition largely determines the target, the conditional transport path can be much straighter than in general image generation, making one-step sampling a natural solution. *VTON is close to this regime: the try-on result is mainly determined by the person image and the garment image, while the remaining uncertainty is mostly limited to local wrinkles, shading, and lighting variations. These plausible outputs usually differ only slightly from one another.*

The practical challenge is that existing VTON models are typically adapted from pretrained generative models, such as Stable Diffusion variants (Rombach et al., 2022; Esser et al., 2024), rather than trained from scratch on task-specific data. This introduces a mismatch: during pretraining, text conditions do not uniquely determine the output, and unconditional training samples are often included, leading to curved transport trajectories. As a result, directly fine-tuned models are not naturally expected to behave consistently under one-step and multi-step sampling.

We propose DirectTryOn, a two-stage framework for one-step VTON. In the first stage, we fine-tune a pretrained model to encourage straighter conditional transport trajectories. We use pure conditional transport to preserve the strong input constraints, training without dropping the garment condition and omitting classifier-free guidance at inference. Since pretrained models tend to emphasize low-frequency information at early timesteps (Park et al., 2025), we introduce a garment preservation loss to retain the fine-grained garment information supplied by these conditions throughout the trajectory. Motivated by the constant velocity in the ideal conditional case, we further introduce a self-consistency loss that aligns velocity predictions across arbitrary timestep pairs to encourage straighter trajectories. Together with extended fine-tuning, these designs yield a teacher with competitive one-step performance even before distillation. In the second stage, we apply Latent Adversarial Diffusion Distillation (LADD) (Sauer et al., 2024) to distill this teacher into a student trained directly for one-step generation, further improving image quality. To support both garment conditioning and preservation, we adopt an MMDiT-style architecture (Esser et al., 2024) with separate garment and noisy-latent streams. The person latent is concatenated with the noisy latent, and the two streams interact through joint self-attention. This design supports garment–person interaction while allowing direct supervision of the garment stream with the garment preservation loss.

Our main contributions are as follows:

1. We formulate one-step VTON as a low-entropy conditional transport problem and analyze why its conditional trajectory can be straighter than that of general image generation.

2. Based on this observation, we propose a two-stage framework that combines pure conditional transport, a garment preservation loss, and a self-consistency loss to train a straightened teacher, followed by LADD-based one-step distillation.

3. Our one-step model achieves state-of-the-art performance on VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022). Timing comparisons show an approximately 19× end-to-end speedup over the fastest baseline.

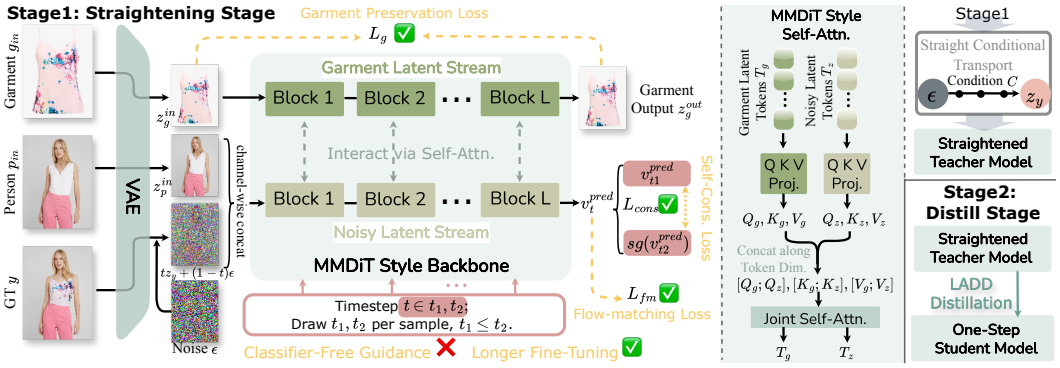

Figure 2: Overview of our framework. Stage 1 trains a teacher model to straighten the conditional transport path for virtual try-on through pure conditional transport, garment supervision, and self consistency loss. Stage 2 distills the straightened teacher into a one-step student model.

2 Related Works

Virtual try-on methods. Earlier GAN-based (Goodfellow et al., 2020) virtual try-on methods (Choi et al., 2021; Ge et al., 2021a; Lee et al., 2022; Xie et al., 2023) typically decompose the task into two stages: (1) warping the garment to align with the target human body shape, and (2) integrating the warped garment with the person image to synthesize the final result. In more recent diffusion-based and flow-matching-based VTON methods, a key question is how to inject garment information into the main network. Earlier methods (Morelli et al., 2023; Gou et al., 2023) typically use an additional image encoder such as CLIP (Radford et al., 2021) and inject garment information through cross-attention. TPD (Yang et al., 2024b) emphasizes the importance of self-attention and shows that directly concatenating the conditional image with the noisy latent along the spatial dimension is more effective than using an additional image encoder with cross-attention for garment information injection. OOTDiffusion (Xu et al., 2025) then adopts a reference U-Net (Zhang, 2023; Hu et al., 2024; Xu et al., 2024) to inject garment information through self-attention layers, while FitDiT (Jiang et al., 2024) can be viewed as a Transformer variant of this design. CatVTON (Chong et al., 2025), by contrast, follows a strategy closer to TPD by concatenating the garment image with the noisy latent along the spatial dimension. JCo-MVTON (Wang et al., 2025) further adopts an MMDiT style (Esser et al., 2024) architecture. Beyond architectural design, existing methods can also be grouped at the training framework level into two categories: mask-based and mask-free. Mask-based methods are constrained by mask accuracy, whereas mask-free methods are constrained by the quality of training data (Sun et al., 2025).

Sampling acceleration. In flow matching (Lipman et al., 2023), the transport path is typically curved after marginalizing over the data distribution, so multi-step sampling is still needed to reach the target distribution. A natural way to reduce the number of sampling steps is therefore to straighten the transport path. This line of work is rooted in Rectified Flow (Liu et al., 2023), which aims to reduce ODE integration error by making trajectories straighter. InstaFlow (Liu et al., 2024) pushes this direction toward one-step and few-step generation by showing that reflow, which reassigns noise-sample pairs and further straightens the probability flow, is crucial for accelerating Stable Diffusion. PeRFlow (Yan et al., 2024) instead performs straightening in a piecewise manner by dividing the full trajectory into multiple time windows. OFM (Kornilov et al., 2024) takes a more theoretical step and aims to recover the straight optimal transport displacement under quadratic cost in a single flow matching training procedure. In the try-on setting, CAT-DM (Zeng et al., 2024) and MC-VTON (Luan et al., 2025) also explore sampling acceleration. CAT-DM initializes diffusion from a GAN generated coarse try-on result instead of Gaussian noise, thereby truncating the denoising trajectory, while MC-VTON adopts LADD (Sauer et al., 2024) to distill a trained virtual try-on model into an 8-step generator.

3 METHODOLOGY

3.1 FRAMEWORK OVERVIEW

Given a person image p_{in} and a garment image g_{in}, our goal is to generate a try-on image y where the person wears the specified garment. We adopt a mask-free setting, where the model directly takes the original person image as input without requiring an explicit garment mask. We perform generation in the latent space. Let z_p^{in}, z_g^{in}, and z_y denote the latents of the person image, garment image, and target image, respectively. Our framework contains two stages. In Stage 1, we train a straightened teacher model by encouraging near straight conditional transport with pure conditional transport, garment supervision, and self consistency loss. In Stage 2, we distill the straightened teacher into a one-step student model for efficient inference. We describe the network architecture in Sec. 3.2, the path straightening objectives in Sec. 3.3, and the distillation stage in Sec. 3.4.

3.2 NETWORK ARCHITECTURE

As shown in Fig. 2, we adopt an MMDiT style architecture and initialize the model weights from SD3 (Esser et al., 2024). We first compare our design with two representative architectures for garment information injection, namely OOTDiffusion (Xu et al., 2025) and CatVTON (Chong et al., 2025).

In OOTDiffusion, the reference U-Net behaves like a large garment encoder: the garment latent is extracted with its own model weights and, in principle, does not need to serve as a query in the subsequent self-attention computation or be updated during denoising. As a result, it only needs to be computed once and can then be reused throughout the entire denoising process. AnyFit (Li et al., 2024) adopts this design. In CatVTON, by contrast, the garment image is concatenated with the noisy latent, so the garment latent can interact with the noisy latent through self-attention and is updated jointly with it, but it does not have its own separate weights. Consequently, the garment related computation has to be repeated at every denoising step. From this perspective, the reference U-Net design is more efficient at inference, whereas the concatenation design is simpler and has a smaller model size.

MMDiT combines the strengths of both designs: different branches can interact through self-attention, while each branch still maintains its own weights for feature processing. Although this design is computationally more expensive, our goal is one-step generation, so we prioritize model capacity and representation ability. Compared with JCo-MVTON (Wang et al., 2025), which also follows an MMDiT style design, we do not introduce a separate branch for the person image. Instead, we concatenate the person image with the noisy latent along the channel dimension, since preserving person identity is relatively simple and does not require an additional stream.

3.3 STAGE1: CONDITIONAL TRANSPORT PATH STRAIGHTENING

We begin with a simple theorem that formalizes the ideal conditional case motivating our method. The proof is given in Appendix A.

Theorem 1. *Under the Optimal Transport conditional path in flow matching, if the conditional input uniquely determines the target, then the conditional transport path is straight, and one-step sampling is equivalent to multi-step sampling.*

Based on this motivation, our view is different from prior works that aim to straighten the marginal transport path itself (Liu et al., 2023; 2024; Yang et al., 2024a; Kornilov et al., 2024). In VTON, the conditional structure already makes the desired transport path close to straight. Therefore, our goal is not to force an inherently curved marginal path to become straight, but to help the model better exploit the strongly constrained conditional structure of try-on generation.

Pure conditional transport. To better understand how classifier-free guidance (Ho & Salimans, 2022) affects final performance, we conduct an experiment on VITON-HD (Choi et al., 2021) based on the architecture described above. Training is performed in the mask-free setting, and evaluation is conducted under the unpaired setting using FID (Parmar et al., 2022). We consider eight configurations obtained by varying three factors: whether unconditional training samples are included,

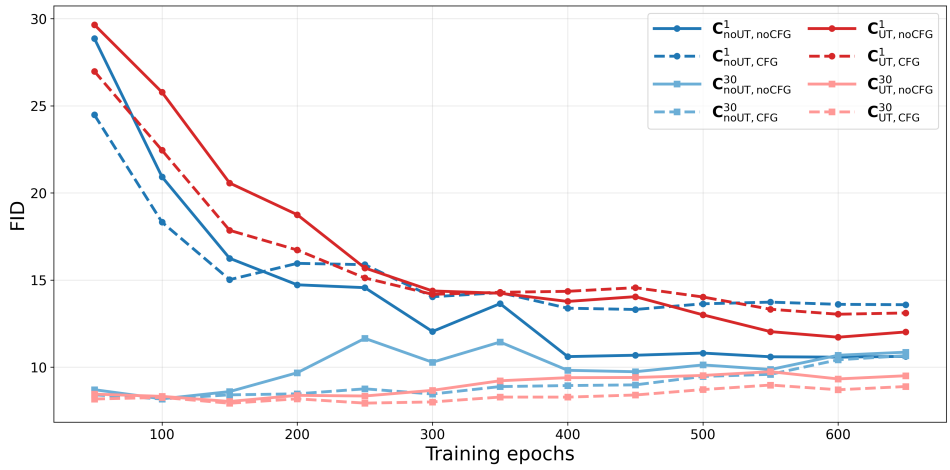

Figure 3: FID curves on VITON-HD under the unpaired evaluation setting. All models are trained in the mask-free setting, and the eight configurations are obtained by varying three factors: whether unconditional training samples are included, whether classifier-free guidance (CFG) is used during inference, and whether inference is performed with 1 or 30 sampling steps. Models are trained for 650 epochs and evaluated every 50 epochs. Detailed discussion is provided in Sec. 3.3.

whether classifier-free guidance (CFG) is used during inference, and whether inference is performed with one step or 30 steps. We denote a configuration by $\mathbf{C}^K_{\mathrm{A,B}}$, where $\mathrm{A} \in \{\mathrm{UT}, \mathrm{noUT}\}$ indicates whether unconditional training samples are included, $\mathrm{B} \in \{\mathrm{CFG}, \mathrm{noCFG}\}$ indicates whether CFG is used during inference, and $K \in \{1, 30\}$ denotes the number of sampling steps used at inference. In the unconditional training setting, we remove the garment condition to form the unconditional branch, and the unconditional ratio is 0.2. When CFG is used during inference, the guidance scale is set to 2. We train the model for 650 epochs in total and evaluate it every 50 epochs, and the results are shown in Fig. 3.

First, comparing $\mathbf{C}^{30}_{\mathrm{noUT,noCFG}}$ with $\mathbf{C}^{1}_{\mathrm{noUT,noCFG}}$, we observe that the best performance under 30-step sampling is reached at around 100 epochs, whereas the performance under one-step sampling continues to improve until around 400 epochs. This suggests that, after 100 epochs, the main improvement is no longer better modeling of the target distribution itself, but a progressively straighter conditional transport path. **This provides evidence that the try-on task naturally admits a much straighter transport path.** Although a fine-tuned pretrained model does not initially possess this property, the mismatch is gradually reduced as fine-tuning continues.

Second, comparing $\mathbf{C}^{1}_{\mathrm{UT,CFG}}$ with $\mathbf{C}^{1}_{\mathrm{UT,noCFG}}$, and $\mathbf{C}^{1}_{\mathrm{noUT,CFG}}$ with $\mathbf{C}^{1}_{\mathrm{noUT,noCFG}}$, we find that **removing CFG during inference consistently improves one-step performance**, regardless of whether unconditional training samples are included. This is consistent with our analysis: the unconditional transport path tends to be more curved, and therefore under one-step sampling it is more likely to deviate from the target distribution, which harms final performance.

Third, comparing $\mathbf{C}^{1}_{\mathrm{UT,noCFG}}$ with $\mathbf{C}^{1}_{\mathrm{noUT,noCFG}}$ shows that even if CFG is not used during inference, including unconditional training samples still leads to worse performance. This indicates that **unconditional training also makes the learned conditional transport path more curved.** Intuitively, although unconditional and conditional training are intended to model two different transport paths, they are optimized within the same model and therefore inevitably interfere with each other.

Based on these observations, we remove unconditional training samples during training and do not use CFG during inference. More importantly, these experiments not only motivate the design of pure conditional transport, but also **support our central claim that the conditional transport path in try-on is naturally much straighter than that in general image generation.**

Garment preservation loss. In flow matching, the marginal velocity at early timesteps tends to emphasize low-frequency information (Park et al., 2025). To examine how this property affects the



Figure 4: Visualization of garment reconstruction during 30-step inference. The leftmost image is the input garment, and the others show the reconstructed garment output from the garment branch at the 1st, 5th, 25th, and 30th sampling steps. Early steps mainly preserve coarse low-frequency structure, while richer high-frequency details emerge at later steps.

try-on setting, we conduct an experiment using the same architecture described in Sec. 3.2, without any additional modifications. The model is trained on VITON-HD for 100 epochs and evaluated with 30-step inference. Since our architecture adopts an MMDiT style design with a separate garment stream, the garment branch also produces a garment latent representation. We denote the garment latent predicted by this branch as z_g^{out}, and the VAE-encoded latent of the input garment as z_g^{in}. For visualization in Fig. 4, we decode z_g^{out} back to the image space. As shown in Fig. 4, at early timesteps the decoded garment output mainly preserves coarse low-frequency structure, while richer high-frequency details emerge at later timesteps. This suggests that pretrained model weights tend to emphasize low-frequency information at early timesteps. When adapted to virtual try-on, the model may therefore underuse or gradually lose fine-grained garment details along the trajectory. Although this issue can be alleviated to some extent with longer fine-tuning, we introduce a more direct training signal by adding a garment preservation loss in the latent space:

$$L_g = \left\| z_g^{out} - z_g^{in} \right\|_2^2 .$$

This objective encourages the garment branch to preserve the input garment representation at every timestep, thereby helping the model maintain garment details across the entire sampling trajectory.

Self consistency loss. To encourage the model to learn that the overall conditional transport path should be close to straight, we introduce an aggressive self consistency loss. Given two arbitrary timesteps t_1 and t_2 with $t_1 \leq t_2$, and recalling that $x_t = ty + (1-t)\epsilon$, let v_t^{pred} denote the velocity vector predicted by the model at timestep t. We define

$$L_{\text{cons}} = \left\| v_{t_1}^{\text{pred}} - \text{stopgrad}\left(v_{t_2}^{\text{pred}} \right) \right\|_2^2 .$$

This loss has two key characteristics. First, we do not impose any constraint on the choice of timesteps. In many previous methods that also aim to enforce consistency, such as Consistency Models (Song et al., 2023), the two timesteps are usually constrained, for example by requiring a fixed interval between them. In our setting, however, we argue that such constraints are unnecessary. Our goal is not to introduce consistency as an additional heuristic property, but to reflect the fact that, in our scenario, the conditional transport path should be nearly straight. Therefore, we adopt the more aggressive choice of using arbitrary timestep pairs without additional restrictions. Second, we apply stop gradient to the prediction at the lower noise level, namely $v_{t_2}^{pred}$. The intuition is that lower noise inputs contain richer target information and thus yield more reliable predictions. We therefore treat the lower noise prediction as the target and force the higher noise prediction to align with it. Without stop gradient, both sides would be updated simultaneously, which in practice degrades performance. We provide ablations on both design choices in Appendix C.

Combining the above modifications, our conditional transport path straightening stage is trained as follows. For each training sample, we randomly choose two timesteps t_1 and t_2, and optimize the model using the original flow matching loss together with the garment preservation loss and the self

consistency loss. The overall training objective is

$$L = \sum_{t\in\{t_1,t_2\}} L_{\text{fm}} + \alpha \sum_{t\in\{t_1,t_2\}} L_g + \beta L_{\text{cons}}.$$

where α and β control the weights of the garment preservation loss and the self consistency loss, respectively. Unless otherwise specified, we set $\alpha = 0.1$ and $\beta = 0.05$. Moreover, the pretrained model is not inherently aligned with the near straight conditional transport required by VTON. We therefore adopt longer fine-tuning in this stage to reduce the mismatch between the pretrained model and the try-on task. More details are provided in Sec. 4.1.

3.4 STAGE2: ONE-STEP DISTILLATION

In this paper, our final goal is to obtain a one-step generator for the try-on task. Although the conditional transport path straightening stage already yields comparable one-step performance (e.g., FID 9.53 under the unpaired setting on VITON-HD), there is still a significant mismatch between training and inference, since the model in this stage is trained with 1000 timesteps but evaluated with one-step sampling. To further reduce this mismatch and improve one-step generation quality, we directly adopt Latent Adversarial Diffusion Distillation (LADD) (Sauer et al., 2024) to distill a one-step model from the model obtained in the straightening stage. Specifically, we use the model trained in the straightening stage as the teacher model and follow the original LADD framework to train a one-step student model. Unless otherwise specified, all distillation settings follow LADD.

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, with all ablation studies performed on VITON-HD. VITON-HD contains only upper-body garments, whereas DressCode includes three garment categories: upper-body, lower-body, and dresses. Both datasets provide paired samples, each consisting of a person image and a corresponding garment image. Such paired data cannot be directly used for mask-free training. Therefore, we first train a mask-based model to generate the required training data. Further details are provided in Appendix B.

Implementation details. For network initialization, the MMDiT style backbone is initialized from Stable Diffusion 3 Medium (Esser et al., 2024). Since both streams in our architecture are image streams rather than text streams, they are both initialized from the image stream of SD3 Medium. We train separate models for VITON-HD and DressCode. In the straightening stage, we train the VITON-HD model for 400 epochs and the DressCode model for 200 epochs. For VITON-HD, the learning rate is linearly warmed up to 1×10^{-4} in the first 5 epochs, kept at 1×10^{-4} for 50 epochs, linearly decayed to 0 over the next 45 epochs, and then fixed at 1×10^{-5} for the remaining 300 epochs. For DressCode, we follow the same schedule pattern, with 5 warmup epochs, 25 epochs at 1×10^{-4}, 20 decay epochs, and 150 epochs at 1×10^{-5}. In the distillation stage, we further train the VITON-HD model for 40 epochs and the DressCode model for 20 epochs, both with a learning rate of 1×10^{-5}. Both stages use the AdamW optimizer (Loshchilov & Hutter, 2019). Further training details, including the LADD implementation, are provided in Appendix B.

Baselines. We compare our method with several recent state-of-the-art approaches, including CatVTON (Chong et al., 2025), Leffa (Zhou et al., 2025), OOTDiffusion (Xu et al., 2025), and Any2anyTryon (Guo et al., 2025), using their official model weights and inference code. Since these methods are not designed for sampling acceleration, we standardize the number of inference steps to 30 for all of them. We would like to compare with methods which also try to explicitly target accelerated sampling, however, CAT-DM (Zeng et al., 2024) only released GC-DM (Zeng et al., 2024) weights and inference code, and MC-VTON has not open-sourced the code, so we do not include them in the comparison baselines.

Evaluation metrics and protocols. Previous virtual try-on methods typically evaluate performance under both paired and unpaired settings. In the paired setting, the model reconstructs the original



Figure 5: Qualitative comparisons on the VITON-HD (Choi et al., 2021) dataset.

Table 1: Quantitative comparisons on the VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022) datasets. Best in **bold** and second best underlined.

Dataset	VITON-HD		DressCode		Time (s) ↓
Method	FID ↓	KID ↓	FID ↓	KID ↓	
OOTDiffusion (Xu et al., 2025)	9.02	0.63	7.10	2.28	9.24
CatVTON (Chong et al., 2025)	9.40	1.27	5.24	1.21	10.98
Leffa (Zhou et al., 2025)	9.38	0.92	6.17	1.90	11.55
Any2anyTryon (Guo et al., 2025)	9.29	1.22	-	-	61.65
DirectTryOn (ours)	**8.59**	**0.56**	**5.08**	**0.95**	**0.48**

person image with the same garment, whereas in the unpaired setting, the garment is replaced with a different one (Choi et al., 2021). In this paper, since we focus on training mask-free models, we report results only under the unpaired setting, which better reflects real-world applications. Following prior works (Chong et al., 2025; Jiang et al., 2024; Zhou et al., 2025), we adopt Fréchet Inception Distance (FID) (Parmar et al., 2022) and Kernel Inception Distance (KID) (Bińkowski et al., 2018) as quantitative metrics for the unpaired setting. We also compare inference speed on a PPU-810E, defined as the end-to-end time from input to output, including preprocessing. All evaluations are conducted at a resolution of 768×1024. For methods that only support 384×512 resolution, including CatVTON and Any2anyTryon, we run inference at 384×512 and then upsample the generated results to 768×1024 for FID and KID evaluation. This protocol does not substantially bias the comparison. For example, evaluating CatVTON directly at 384×512 gives an FID/KID of 9.36/1.19, while evaluating the upsampled results at 768×1024 gives an FID/KID of 9.40/1.27. This resizing protocol is used only for FID/KID evaluation. For inference speed, we evaluate all methods directly at 768×1024 to ensure a fair comparison.

4.2 Qualitative and Quantitative Results

Qualitative comparison. Qualitative results are shown in Fig. 5. Although our method uses only one-step sampling, it still achieves competitive visual quality compared with multi-step baselines. In particular, one-step sampling does not degrade the try-on results, and the garment appearance is well preserved. This observation is consistent with our previous analysis: since virtual try-on is highly constrained by the person and garment conditions, the generation process is nearly deterministic, making one-step generation feasible after conditional path straightening. We further discuss the effect of mask-based and mask-free designs. Both Any2anyTryon and our method are mask-free, while other baselines rely on explicit masks. As shown in the first row, identity-related regions, such as the hair, can sometimes be altered not only by mask-free methods but also by mask-based methods, especially when the mask is inaccurate or overly aggressive. This suggests that using an explicit mask does not necessarily guarantee better preservation of non-garment regions.



Figure 6: Qualitative and quantitative ablation results. From left to right, each variant progressively applies one additional modification to the previous setting: pure conditional transport (PCT), adding the garment preservation loss (GL), adding the self consistency loss (SC), using longer fine-tuning (FT), and finally applying LADD distillation. The numbers below each result denote FID/KID, where lower values indicate better performance.

Quantitative comparison. We report quantitative results in Tab. 1. For Any2anyTryon, since only the weights trained on VITON-HD are publicly available, we compare with it only on the VITON-HD dataset. Our method is mask-free and therefore does not require additional mask preprocessing. Moreover, since we disable CFG during inference and perform only one sampling step, our method takes only 0.48s on a PPU-810E. Even compared with OOTDiffusion, the fastest baseline among all compared methods, which takes 9.24s, our method is still about 19$\times$ faster. More importantly, despite this substantial acceleration, our method still achieves state-of-the-art performance.

4.3 Ablation Study

To validate our proposed modifications, we conduct an ablation study on VITON-HD in the mask-free and unpaired setting. As shown in Fig. 6, the baseline uses unconditional training samples by dropping the garment condition with a ratio of 0.2, and applies CFG at inference with a guidance scale of 2.

For one-step VTON, both designs are harmful: unconditional samples introduce an additional transport path that is not fully specified by the garment-person condition, while CFG extrapolates the sampling direction using the unconditional prediction, which can push the one-step result away from the target conditional distribution. Removing both yields cleaner conditional transport and substantially improves the result. Adding the garment preservation loss and self consistency loss further helps the model learn a straighter conditional transport path. Up to this point, all variants are trained for 100 epochs. Since pretrained weights are not naturally aligned with this near straight conditional structure, extending fine-tuning by 300 epochs further improves performance, yielding the straightened teacher model. Finally, applying LADD further reduces the gap between training and one-step inference, leading to the best overall performance.

5 Conclusion

In this paper, we study one-step virtual try-on from the perspective of conditional transport. We argue that, compared with general image generation, virtual try-on is a more strongly conditioned task, in which the target image is largely constrained by the person image and the garment image. This suggests that its conditional transport path can be much straighter, making one-step generation a natural solution. Based on this observation, we propose a framework consisting of a conditional path straightening stage and a distillation stage. Extensive experiments on VITON-HD and DressCode show that our method achieves strong one-step try-on performance.

We hope this work can also provide useful inspiration for other generation tasks with similarly strong conditional constraints, where exploiting straighter conditional transport paths may lead to more efficient generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the language of the manuscript and organizing the formatting. We have not used generative AI tools for the methodology design, the implementation of experiments, or the generation of experimental results. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the network architecture (Sec. 3.2), the training objectives and loss weights (Sec. 3.3 and Appendix C), the construction of mask-free training data and the LADD distillation implementation (Appendix B), and the evaluation protocol (Sec. 4.1). All experiments use the publicly available VITON-HD and DressCode datasets and the officially released checkpoints of the compared baselines, with all generated results scored by a unified evaluation script.

ETHICS STATEMENT

This work develops a generative image synthesis method for virtual try-on. As with other generative technologies, it may raise concerns related to intellectual property and personal privacy, and we encourage its responsible and ethical use. Further discussion is provided in the Broader impacts section in the appendix.

REFERENCES

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

Black Forest Labs. Flux.1 [dev]. `https://huggingface.co/black-forest-labs/FLUX.1-dev`, 2024. Model card, accessed 2026-04-21.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2021. Online first published May 2019.

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.

Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon 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, 2024.

Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Hanqing Zhao, Xujie Zhang, Dongmei Jiang, 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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633, 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.

Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 770–785, 2018.

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.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018.

Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19085–19096, 2025.

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 and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 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.

Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. *arXiv preprint arXiv:2411.10499*, 2024.

Jeongho Kim, Gyojung 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.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Thirty-eighth Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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.

Yuhan Li, Hao Zhou, Wenxiang Shang, Ran Lin, Xuanhong 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.

Yujie Liang, Xiaobin Hu, Boyuan Jiang, Donghao Luo, Xu Peng, Kai Wu, Chengming Xu, Wenhui Han, Taisong Jin, Chengjie Wang, and Rongrong Ji. Vton-handfit: Virtual try-on for arbitrary hand pose guided by hand priors embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22616–22626, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Junsheng Luan, Guangyuan Li, Lei Zhao, and Wei Xing. Mc-vton: Minimal control virtual try-on diffusion transformer. *arXiv preprint arXiv:2501.03630*, 2025.

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 Workshops*, 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.

Dogyun Park, Taehoon Lee, Minseok Joo, and Hyunwoo J. Kim. Blockwise flow matching: Improving flow matching models for efficient high-quality generation. In *Thirty-ninth Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 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.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

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

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252, 2023.

Xianbing Sun, Yan Hong, Jiahui Zhan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Ds-vton: High-quality virtual try-on via disentangled dual-scale generation. *arXiv preprint arXiv:2506.00908*, 2025.

Aowen Wang, Wei Li, Hao Luo, Mengxing Ao, Chenyu Zhu, Xinyang Li, and Fan Wang. Jco-mvton: Jointly controllable multi-modal diffusion transformer for mask-free virtual try-on. *arXiv preprint arXiv:2508.17614*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jia Wu, Yijing Dai, Tingfeng Cao, Meiling Wu, Tao Luo, Jian Dong Zhang, Guangming Lu, and Xiaoyi Zeng. High-fidelity virtual try-on beyond paired data scarcity via diffusion-based cycle-consistent learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35821–35830, 2026.

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 Arlene 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.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu 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)*, pp. 1481–1490, 2024.

Hanshu Yan, Xingchao Liu, Jiachun Pan, Jun Hao Liew, Qiang Liu, and Jiashi Feng. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *Advances in Neural Information Processing Systems*, 37:78630–78652, 2024.

Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024a.

Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7017–7026, 2024b.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, 2024.

Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8372–8382, 2024.

Lvmin Zhang. [major update] reference-only control. `https://github.com/Mikubill/sd-webui-controlnet/discussions/1236`, 2023. GitHub discussion in the sd-webui-controlnet repository; reference-only ControlNet method proposed by Lvmin Zhang (lllyasviel).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 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.

Zijian Zhou, Shikun Liu, Xiao Han, Haozhe Liu, Kam Woh Ng, Tian Xie, Yuren Cong, Hang Li, Mengmeng Xu, Juan-Manuel Perez-Rua, Aditya Patel, Tao Xiang, Miaojing Shi, and Sen 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)*, pp. 2491–2501, 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 PROOF OF THEOREM 1

Theorem 1. *Under the Optimal Transport conditional path in flow matching, if the conditional input uniquely determines the target, then the conditional transport path is straight, and one-step sampling is equivalent to multi-step sampling.*

Proof. Under the Optimal Transport conditional path in flow matching,

$$X_t = (1-t)X_0 + tX_1,$$

the conditional velocity field can be written as

$$v_t(x \mid c) = \mathbb{E}\left[\frac{X_1 - x}{1-t} \,\middle|\, X_t = x,\ c\right].$$

As the conditional input c uniquely determines the target, i.e.,

$$X_1 = y(c).$$

Therefore,

$$v_t(x \mid c) = \frac{y(c) - x}{1-t}.$$

For one-step sampling, starting from an arbitrary noise ϵ at $t = 0$,

$$x_0 = \epsilon.$$

Then

$$x_1 = x_0 + v_0(x_0 \mid c) = \epsilon + \big(y(c) - \epsilon\big) = y(c).$$

For multi-step sampling, let

$$0 = t_0 < t_1 < \cdots < t_N = 1,$$

and define the update

$$x_{t_{k+1}} = x_{t_k} + (t_{k+1} - t_k)\, v_{t_k}(x_{t_k} \mid c).$$

We prove by induction that

$$x_{t_k} = (1 - t_k)\epsilon + t_k\, y(c), \qquad \forall k.$$

Therefore, all sampled points lie on the same linear path from ϵ to $y(c)$, which implies that the conditional transport path is straight.

For the base case,

$$x_{t_0} = x_0 = \epsilon = (1 - t_0)\epsilon + t_0\, y(c).$$

Assume

$$x_{t_k} = (1 - t_k)\epsilon + t_k\, y(c).$$

Then

$$v_{t_k}(x_{t_k} \mid c) = \frac{y(c) - x_{t_k}}{1 - t_k} = \frac{y(c) - \big((1 - t_k)\epsilon + t_k\, y(c)\big)}{1 - t_k} = y(c) - \epsilon.$$

Therefore,

$$\begin{aligned} x_{t_{k+1}} &= x_{t_k} + (t_{k+1} - t_k)\, v_{t_k}(x_{t_k} \mid c) \\ &= (1 - t_k)\epsilon + t_k\, y(c) + (t_{k+1} - t_k)\big(y(c) - \epsilon\big) \\ &= (1 - t_{k+1})\epsilon + t_{k+1}\, y(c). \end{aligned}$$

Thus the induction holds.

Finally, at $t_N = 1$,

$$x_{t_N} = (1 - t_N)\epsilon + t_N\, y(c) = y(c).$$

Hence multi-step sampling reaches the same final result as one-step sampling. □

B MORE EXPERIMENT DETAILS

Training data. In this paper, we mainly focus on training a mask-free model for one-step generation. Compared with mask-based training, which only requires paired data consisting of a person image and the corresponding garment image, mask-free training requires at least three images: two person images of the same identity wearing different garments, and one garment image corresponding to one of these two person images. Therefore, when using existing datasets such as VITON-HD (Choi et al., 2021) and DressCode (Morelli et al., 2022), which only provide paired data, we need to generate an additional person image wearing a different garment for each pair in order to construct mask-free training data. To this end, we first train a mask-based model and use it to synthesize the required training data. This mask-based model adopts the same network architecture as our mask-free model, with the main difference lying in the inputs. Specifically, the person image is replaced by an agnostic image, where the garment region is masked out using segmentation information (Gong et al., 2018; Cao et al., 2021). In addition, after the patch embedding layer, we add the DensePose image (Güler et al., 2018) of the person to the noisy latent, so that the generated result is encouraged to preserve the original pose. This design helps avoid pose inconsistency caused by inaccurate masks.

LADD implementation. In the distillation stage, we adopt Latent Adversarial Diffusion Distillation (LADD) (Sauer et al., 2024) to compress our straightened teacher model into a one-step student, with one-step generation as the final target. We first use the straightened teacher model offline to generate pseudo ground-truth try-on images. Specifically, for each training sample, we randomly pair a person image with a garment image from the dataset, and use the teacher model to synthesize the corresponding try-on result. The generated result is then treated as the target image for student training. During distillation, the student is initialized from the teacher model and takes the pure noisy target latent concatenated with the person latent as input, while using the garment latent as the try-on condition. It predicts the velocity field, from which we recover the clean latent prediction. The student is trained with a reconstruction loss between its predicted clean latent and the teacher-generated target latent. In addition to this reconstruction objective, we follow LADD and introduce a latent adversarial objective based on frozen teacher features. Instead of using an external image-space discriminator, we re-noise both the teacher-generated target latent and the student-predicted latent to randomly sampled noise levels, and feed them into the frozen teacher transformer. We extract intermediate hidden states from multiple transformer blocks and attach lightweight convolutional discriminator heads to these features. The discriminator is trained to distinguish re-noised teacher targets from re-noised student predictions, while the student is trained to fool the discriminator. Since our model is initialized from Stable Diffusion 3-Medium, which contains 24 transformer blocks, we sample 8 blocks across the transformer depth to provide multi-level adversarial supervision.

Training details. All experiments in this paper are conducted on a cluster of 16 PPU-810E processors, each with 100 GB of memory. For the path-straightening stage, the batch size is set to 8 per processor, resulting in a total batch size of 128. Training takes approximately 6 days for 400 epochs on VITON-HD and 11 days for 200 epochs on DressCode. For the distillation stage, the batch size is set to 3 per processor, resulting in a total batch size of 48. Distillation takes approximately 9 hours for 40 epochs on VITON-HD and 18 hours for 20 epochs on DressCode.

C MORE ABLATION STUDIES

Ablation on the design of the self consistency loss. Here we provide additional ablations on the design of our self consistency loss. Our method contains two special design choices: first, we do not impose any constraint on timestep sampling for the consistency loss; second, we apply `stopgrad` to the lower-noise prediction, rather than simply requiring the predictions at two timesteps to be the same. For this ablation study, we use mask-free training and evaluate under the unpaired setting on VITON-HD. All models are trained for 100 epochs. In all three variants, pure conditional transport and the garment preservation loss are applied. The only difference lies in the design of the self consistency loss. The original version used in our method is denoted as **Ours**. The variant **Fixed Timestep Interval** requires the two sampled timesteps to have a fixed interval of 100, for example 900 and 800. The variant **Without StopGrad** removes the `stopgrad` operator from the original design. As shown in Tab. 2, our original design achieves the best performance.

Table 2: Ablation study on two design choices in the self consistency loss.

Version	FID ↓	KID ↓
Fixed Timestep Interval	14.15	5.52
Without StopGrad	13.26	4.78
Ours	**11.27**	**3.17**

Table 3: Component-wise ablation on VITON-HD under the mask-free, unpaired, and one-step evaluation setting. All variants are trained for 100 epochs in the straightening stage except the FT row, which uses the full 400-epoch fine-tuning. The rows GL + LADD and SC + LADD are trained without PCT.

Configuration	FID ↓	KID ↓
PCT + GL	15.79	6.87
PCT + GL + SC	11.27	3.17
PCT + GL + SC + FT	9.53	1.68
PCT + GL + SC + FT + LADD	**8.59**	**0.56**
GL + LADD	14.22	5.03
SC + LADD	15.15	5.85

Effect of pure conditional transport. Pure conditional transport (PCT) is the core component of our framework. While the pretrained backbone does not inherently satisfy the low-entropy conditional transport formulation, the mismatch is progressively alleviated with longer training: the PCT-only configuration achieves an FID of 20.92 and a KID of 10.85 after 100 epochs, but improves to an FID of 10.60 and a KID of 2.14 under one-step inference after the full 400-epoch schedule. This is consistent with the trend in Fig. 3, where the one-step performance keeps improving long after the 30-step performance saturates. Together with Tab. 3, these results show that PCT establishes the essential conditional transport structure, while GL and SC serve as complementary objectives that improve garment preservation and encourage a straighter transport path.

On the contribution of LADD distillation. The final one-step performance cannot be attributed to LADD alone: our fully trained teacher model already achieves an FID of 8.52 and a KID of 0.54 with only four sampling steps, outperforming all compared baselines under the same evaluation protocol. To further verify the role of the teacher, we apply the same LADD protocol to CatVTON (Chong et al., 2025); the resulting one-step model achieves only 20.19/13.95, far behind our method (8.59/0.56). Together with Tab. 3, these results show that the straightening stage constructs a strong teacher for distillation, and that the final one-step performance relies on the proposed training framework rather than on LADD alone.

Sensitivity to loss weights. We analyze the sensitivity of the final loss to the weighting coefficients α and β. For the garment preservation loss weight analysis, we fix PCT, disable SC, and vary only the GL loss weight. For the self consistency loss weight analysis, we fix PCT and GL, set the GL loss weight to 0.1, and vary only the SC loss weight. All other training and evaluation settings are kept unchanged, and all models are trained for 100 epochs under the mask-free and unpaired setting on VITON-HD.

As shown in Tab. 4, the model is relatively insensitive to the GL loss weight, while the SC loss weight has a stronger influence and its optimal value is relatively small. This trend is consistent with the role of GL and SC as auxiliary objectives rather than standalone optimization targets: setting either weight to zero degrades performance, confirming that both objectives provide useful supervision, while larger weights shift the optimization away from the primary flow-matching objective. A relatively small coefficient therefore does not imply a minor effect; what matters is that the auxiliary objectives remain appropriately balanced with the primary conditional transport objective.

Table 4: Sensitivity of one-step performance to the garment preservation loss weight (left) and the self consistency loss weight (right), on VITON-HD under the unpaired setting.

GL weight	FID ↓	KID ↓
0.0	20.92	10.85
0.1	**15.79**	**6.87**
0.5	16.07	7.03
1.0	17.56	8.29

SC weight	FID ↓	KID ↓
0.0	15.79	6.87
0.05	**11.27**	**3.17**
0.10	13.25	4.71
0.20	14.75	5.46



Figure 7: Results of our distilled one-step student model using five different Gaussian noise initializations. For each initialization, the model performs a single denoising step from the sampled Gaussian noise to the final try-on result. Despite starting from different noise samples, the generated results remain highly similar, with only slight variations in garment wrinkles, demonstrating the robustness of one-step inference to the choice of initial noise.

D Robustness to noise initialization

Here, we provide an additional test of the distilled one-step student model. We randomly sample five different Gaussian noises, while keeping the same garment image and person image as input, and perform one-step inference from each noise. As shown in Fig. 7, the generated results are highly similar, differing only slightly in garment wrinkles. This confirms that our one-step model is robust to different Gaussian noise initializations and does not suffer noticeable quality degradation from different starting noise points.

E Additional Analysis of Classifier-Free Guidance

Recall the CFG ablation in Sec. 3.3. Fig. 3 reveals several additional observations. First, $\mathbf{C}^{30}_{\text{UT,CFG}}$ consistently achieves the best or near-best performance. This is expected: when CFG is enabled at inference and multi-step sampling is used, introducing unconditional training samples helps the model learn a more reliable velocity field and improves the effectiveness of guidance. Second, comparing $\mathbf{C}^{30}_{\text{noUT,noCFG}}$ with $\mathbf{C}^{30}_{\text{UT,noCFG}}$, we observe that unconditional training still improves performance even when CFG is disabled during inference. This suggests that unconditional samples are not only useful for CFG, but also regularize the model during training. In the 30-step setting, this regularization allows the model to better explore the underlying data distribution and produce more robust predictions.

F Comparison with generic acceleration methods

To examine the advantage of our framework over generic diffusion acceleration strategies, we conduct a controlled comparison on VITON-HD using a base model with the same network architecture as ours, conventionally trained with the standard flow-matching objective without any of our proposed components, including PCT, GL, SC, or longer fine-tuning. We evaluate two representative acceleration strategies on this base model: (1) one-step Distribution Matching Distillation (DMD) (Yin et al., 2024), and (2) four-step sampling with the Euler and DDIM (Song et al., 2020) solvers.

For DMD, we perform one-step distillation using the base model as the frozen teacher. Both the generator and the online fake-score network are initialized from the teacher checkpoint, and the generator starts from Gaussian noise and directly predicts the clean latent for one-step inference.

Table 5: Comparison with generic acceleration methods on VITON-HD under the unpaired setting. DMD distills the base model into a one-step generator, while the Euler and DDIM baselines evaluate the same base model with four sampling steps. All strategies use the same base model for a controlled comparison.

Method	FID ↓	KID ↓
DMD (one-step)	20.04	12.08
4-step Euler sampler	10.77	2.52
4-step DDIM	10.16	1.91
DirectTryOn (ours, one-step)	**8.59**	**0.56**

Table 6: Reported and reproduced FID/KID for OOTDiffusion and Leffa on DressCode under the unpaired setting. The original OOTDiffusion paper uses 20 inference steps, while Leffa does not report the number of steps. Our reproductions use the official checkpoints and inference code.

Method / Reported in	Steps	FID ↓	KID ↓
OOTDiffusion (original paper)	20	4.20	0.37
OOTDiffusion (CatVTON)	-	12.57	6.63
OOTDiffusion (VTON-HandFit)	-	6.70	1.86
OOTDiffusion (CCVTON)	-	7.20	2.30
OOTDiffusion (reproduced by us)	**30**	**7.10**	**2.28**
OOTDiffusion (reproduced by us)	**20**	**8.22**	**2.70**
Leffa (original paper)	not reported	4.48	0.62
Leffa (CCVTON)	-	6.61	1.78
Leffa (reproduced by us)	**30**	**6.17**	**1.90**

Timesteps are uniformly sampled from the middle 2%–98% of the 1,000-step SD3 flow-matching schedule with a shift of 3 for distribution matching, and logit-normal timestep sampling is used for fake-score training. The fake-score network is updated at every iteration, while the generator is updated once every five iterations. We train for 40 epochs, corresponding to 19,400 fake-score updates and 3,880 generator updates. Both networks use AdamW with zero weight decay and gradient clipping at 1.0, with linear warm-up for the first 500 fake-score updates and 100 generator updates. Training is performed on 8 GPUs with a per-GPU batch size of 3 and a global batch size of 24. We use FP16 mixed precision, gradient checkpointing, 8-bit AdamW, TF32, and a random seed of 42. The guidance scale is fixed to 2.0 during distillation.

As shown in Tab. 5, our method substantially outperforms both the DMD-based one-step model and the four-step sampling baselines. Note that DDIM performs four sequential denoising steps, allowing progressive refinement, whereas DMD compresses the entire generation process into a single step, which partly explains the gap between them. These results show that directly accelerating a conventionally trained model, either through DMD or few-step sampling, does not achieve the same generation quality as our proposed training and distillation framework.

G Evaluation protocol and metric discrepancies

As described in Sec. 4.1, all baselines are evaluated with their officially released checkpoints and inference code under a unified protocol: 30 inference steps, the same test split, output resolution, and number of generated images, and a single evaluation script (the official script released by CatVTON, which computes FID with `clean-fid`). Our reproduced baseline numbers therefore differ from those reported in the original papers, which use different evaluation scripts and settings. We provide a detailed breakdown for OOTDiffusion (Xu et al., 2025) and Leffa (Zhou et al., 2025) on DressCode in Tab. 6, which also includes the numbers reported by independent evaluations in CatVTON (Chong et al., 2025), VTON-HandFit (Liang et al., 2025), and CCVTON (Wu et al., 2026). CCVTON (Wu et al., 2026) is a recent mask-free VTON method that adopts a cycle-consistent try-off/try-on framework built on a Flux backbone to enable mask-free training beyond paired data.

Using 20 steps for OOTDiffusion yields slightly worse performance than using 30 steps, indicating that the discrepancy cannot be explained by the number of inference steps. We also note that the

Table 7: Quantitative comparison of non-garment region preservation on VITON-HD. SSIM and LPIPS are computed with the clothing region masked to a uniform gray value and with copy-and-paste post-processing disabled for all methods.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	Time (s) ↓
CatVTON	0.947	0.042	9.40	1.27	10.98
OOTDiffusion	0.958	**0.025**	9.02	0.63	9.24
Any2AnyTryon	0.950	0.032	9.29	1.22	61.65
Leffa	0.955	0.029	9.38	0.92	11.55
DirectTryOn (ours)	**0.971**	**0.025**	**8.59**	**0.56**	**0.48**

original OOTDiffusion paper does not explicitly specify whether its reported 4.20/0.37 corresponds to the paired or unpaired setting, while our paper focuses exclusively on the unpaired setting. Meanwhile, our reproduced OOTDiffusion result (7.10/2.28) is very close to that reported by CCVTON (7.20/2.30), and our reproduced Leffa result (6.17/1.90) is also comparable to that reported by CCVTON (6.61/1.78). We therefore attribute the discrepancies primarily to the use of different FID evaluation scripts and settings, and all comparisons in our paper are conducted under the same evaluation protocol to ensure a consistent and fair comparison.

H Preservation of non-garment regions

Under the mask-free setting, non-clothing regions such as hair, face, background, or accessories may be more prone to modification. Since full-image paired reconstruction metrics do not independently measure the preservation of these regions, we provide an additional quantitative comparison. A direct comparison of non-clothing similarity is not fair across methods, because some mask-based approaches explicitly paste the region outside the agnostic mask from the input person image into the final output; for example, CatVTON applies this operation, so its high similarity in these regions may reflect the copy-and-paste procedure rather than the model's intrinsic preservation capability.

To provide a fair comparison, we disable this copy-and-paste operation for the relevant methods. We then replace the clothing region defined by the agnostic mask with a uniform gray value in both the generated and ground-truth images, and compute SSIM (Wang et al., 2004) and LPIPS (Zhang et al., 2018) on the resulting images, thereby removing appearance differences within the clothing region and emphasizing the comparison of non-clothing regions. SSIM and LPIPS are computed on the raw generated outputs before any copy-and-paste post-processing, while FID, KID, and inference time follow the standard evaluation setting reported in the main paper.

As shown in Tab. 7, our method achieves the highest SSIM and the joint-best LPIPS, supporting its ability to preserve non-clothing regions despite its mask-free design.

I Limitations

In this paper, we do not focus on the distinction between mask-based and mask-free settings. Instead, we train our own mask-based model to generate training data for the mask-free setting. Although this strategy provides relatively reliable training data, the generated data is still not perfect. As a result, our method may inherit a common issue in previous methods (Sun et al., 2025): regions unrelated to the target garment, such as hair or accessories, may occasionally be altered. Fortunately, this issue occurs infrequently in practice, and even when it occurs, the artifacts are usually mild.

Typical failure cases of our method fall into two categories. First, when the teacher-generated training data contains artifacts, e.g., caused by inaccurate masks in complex poses, the mask-free model may inherit these artifacts and occasionally alter non-garment regions such as hair, face, or background. Second, for garments with highly complex textures or under severe garment-person occlusion, fine garment details may be imperfectly reconstructed, which is a common difficulty shared by mask-free try-on methods. We leave a more systematic evaluation of failure cases to future work.

J BROADER IMPACTS

The ability of DirectTryOn to generate realistic virtual try-on results efficiently makes it well suited for practical deployment in e-commerce scenarios. At the same time, as with other generative technologies, DirectTryOn may raise concerns related to intellectual property and personal privacy. We encourage its responsible and ethical use.

K ADDITIONAL QUALITATIVE RESULTS ON DRESSCODE

In this section, we present additional qualitative comparison results on the DressCode (Morelli et al., 2022) dataset.



Figure 8: Qualitative comparison on the DressCode dataset.