

FDM-MFVT: Few-step Sampling Diffusion Model for Mask-Free Virtual Try-On

Jiaxin Liu¹, Xiaoye Liang^{1,2}, Lai Jiang^{1,3}, Mai Xu^{1,3}, and Jun Liu¹*

¹ School of Electronic Information Engineering, Beihang University, Beijing, China

² Zhongguancun Academy, Beijing, China

³ State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China

liujiaxinkw@buaa.edu.cn, liujun2019@buaa.edu.cn



Fig. 1: We propose FDM-MFVT, a mask-free few-step virtual try-on framework that achieves high-fidelity virtual try-on with fewer inference steps across multiple scenarios and requires no reference conditions.

Abstract. Image-based Virtual Try-On (IVTON) has greatly advanced through diffusion models, yet existing methods require many sampling steps and depend on masks with costly auxiliary networks. In addition, the absence of large-scale mask-free paired datasets further limits the development of mask-free IVTON. We propose FDM-MFVT, a few-step

* Corresponding author

diffusion model for mask-free IVTON, integrating an Outfit-aware Noise Optimization Module (OANO) and an Instruction-driven Try-on Module (IDT) to enhance efficiency and flexibility. The OANO module initializes the alignment space with noise using the input image and only needs 6 steps to generate a higher-fidelity try-on image compared to 30 steps. The IDT module uses virtual try-on prompts and efficient adaptation to generate high-quality results from garment and person images alone. We further introduce MFVT, a 30,000-pair mask-free IVTON dataset. Experiments show that FDM-MFVT achieves superior quantitative and qualitative results with fewer inference steps than mask-based and mask-free baseline methods.

Keywords: Diffusion model · Image synthesis · Virtual Try-On

1 Introduction

The objective of image-based virtual try-on (IVTON) is to synthesize a realistic image of an individual wearing a target garment, given a person image and a garment image. This task has attracted considerable attention due to its extensive applications in e-commerce, including online retail [15] and virtual fashion design [9, 21, 37]. To produce photorealistic results, deep learning-based methods [11, 17, 23, 26, 40] typically rely on datasets [4, 32] comprising aligned tuples of person images, garment images, agnostic-mask, and other conditional inputs.

Recent advances in diffusion models [2, 35] have demonstrated exceptional performance in image synthesis across a wide range of tasks. Compared with earlier generative adversarial network (GAN)-based methods [16–18, 26, 31], diffusion-based virtual try-on methods [5, 6, 20, 24, 28, 43, 44] achieve substantially higher image fidelity by leveraging the strong generative priors of large-scale diffusion networks. These priors help preserve garment details and improve overall visual quality. However, diffusion models generally require many sampling steps to generate high-fidelity IVTON outputs, which greatly limits their practicality for real-time try-on systems. In addition to high sampling costs, current diffusion-based virtual try-on systems typically depend on garment masks and multiple auxiliary encoders to process conditional inputs such as clothing masks, DensePose [13], and human parsing maps. This multi-condition design significantly increases inference complexity and makes the system highly sensitive to mask quality. Furthermore, progress toward mask-free try-on is hindered by the lack of large-scale, high-quality paired datasets without mask annotations. All in all, existing IVTON methods face three key challenges: (1) High computational overhead caused by the many sampling steps required for high-fidelity diffusion-based synthesis; (2) Heavy reliance on masks and auxiliary encoders, leading to complex inference pipelines and high sensitivity to preprocessing errors; (3) Limited availability of mask-free paired datasets, restricting the development of fully mask-free try-on frameworks.

To address these challenges, we propose FDM-MFVT, a Few-Step Sampling Diffusion Model for Mask-Free Virtual Try-On. We first observe that the ran-

dom noise distribution in diffusion inference plays a critical role in both synthesis quality and the required number of sampling steps. Certain noise patterns enable high-quality generation with substantially fewer steps, highlighting the importance of effective initialization. Based on this insight, we design an Outfit-aware Noise Optimization Module (OANO) that optimizes the random noise distribution conditioned on the input images. This allows the model to start from a more favorable noise state and produce high-quality results with significantly fewer diffusion steps. Second, to streamline inference and completely remove the dependence on masks or additional conditional inputs, we introduce the Instruction-driven Try-on Module (IDT). This module incorporates simulated contextual cues and a lightweight adaptation mechanism, enabling the model to infer all necessary information directly from person and garment images. This design maintains high synthesis quality under few-step sampling while substantially reducing architectural complexity. Finally, to support mask-free IVTON research, we construct MFVT, a large-scale, high-quality mask-free dataset. Extensive experiments show that FDM-MFVT achieves state-of-the-art performance across multiple metrics while significantly improving inference efficiency. Compared with existing mask-free and mask-based VTON methods, our method exhibits superior robustness and real-time capability, as shown in Fig. 1. In summary, our contributions are threefold:

- We propose a mask-free IVTON framework that eliminates dependency on auxiliary conditions and operates effectively with only a few inference steps, enabling practical real-time virtual try-on.
- We design a novel initialization and conditioning strategy that improves diffusion efficiency and garment-person alignment by integrating outfit-aware noise learning with instruction-driven try-on for robust, high-quality synthesis.
- We construct a new large-scale mask-free dataset (MFVT) that fills a critical gap in training mask-free IVTON systems and provides a solid foundation for future research.

2 Related Work

2.1 Image-based Virtual Try-On

Image-based Virtual Try-On aims to synthesize a realistic image of a person wearing a target garment. Early works such as VITON [17] introduces a coarse-to-fine GAN framework that transfers multi garments onto human images while preserving pose and garments structure, and VITON-HD [4] further enhances realism and resolution through a multi-scale generator. Subsequent methods, including ClothFlow [16], modeled garment deformation as dense flow fields, while OOTD [43] provided a large-scale benchmark dataset and methods that rely on human parsing maps and skeleton information. More recently, PICTURE [33] enables photorealistic try-on from unconstrained garment inputs, demonstrating strong generalization to real-world clothing. StableVITON [24] leveraged stable

diffusion model to improve visual fidelity using multiple conditional inputs such as masks and label maps. However, these methods depend on multi-conditional guidance, which significantly increases inference complexity and limits scalability. To reduce reliance on explicit conditions, a few studies have explored mask-free virtual try-on. For example, OmniTry [10] proposed a unified mask-free framework that generalizes virtual try-on to clothing and diverse wearable items. However, current mask-free methods often suffer from blurry garment boundaries, loss of fine texture, and limited robustness across diverse poses. Thus, achieving high-quality mask-free virtual try-on remains an open and challenging research problem.

2.2 Few-Step Inference in Diffusion

Diffusion models achieve impressive generative quality but typically require hundreds of denoising steps, which limits their efficiency. To address this, recent works have explored few-step inference strategies, which can be broadly categorized into three directions, e.g. distillation-based [2, 12, 27, 36, 38], sampler-reformulation-based [39] and noise-optimization-based [47] methods. Early acceleration methods focused on distilling multi-step diffusion trajectories into compact student models. Progressive distillation [36] compresses the sampling process into fewer iterations, while consistency models [38] directly learn mappings that enable one-step or few-step generation. These methods demonstrate that knowledge transfer from teacher to student can significantly reduce inference cost while maintaining fidelity. Another line of work reformulates the reverse process to allow deterministic or non-Markovian sampling. DDIM [39] introduced implicit models that reduce the number of steps by redefining the reverse diffusion dynamics, enabling efficient few-step sampling with controllable interpolation. Such mapping-based formulations highlight that the reverse process can be approximated without sacrificing sample diversity. Beyond distillation and mappings, recent studies investigate how noise scheduling and initialization affect efficiency. The Golden Noise method [47] proposes optimized noise distributions that accelerate convergence and improve sample quality in few-step regimes. This perspective emphasizes that careful design of noise priors can complement architectural or training-based acceleration.

3 Method

3.1 Overview

In this section, we introduce the FDM-MFVT framework, designed to produce high-quality mask-free virtual try-on results with minimal inference steps. As illustrated in Fig. 2, our framework consists of an OANO module and an IDT module. Within the OANO module, given the garment image I_g and the person image I_p , the original noise x_T is refined into optimized noise x'_T . This optimization leverages outfit-specific features to provide a more informative initialization, thereby improving the fidelity and realism of subsequent synthesis

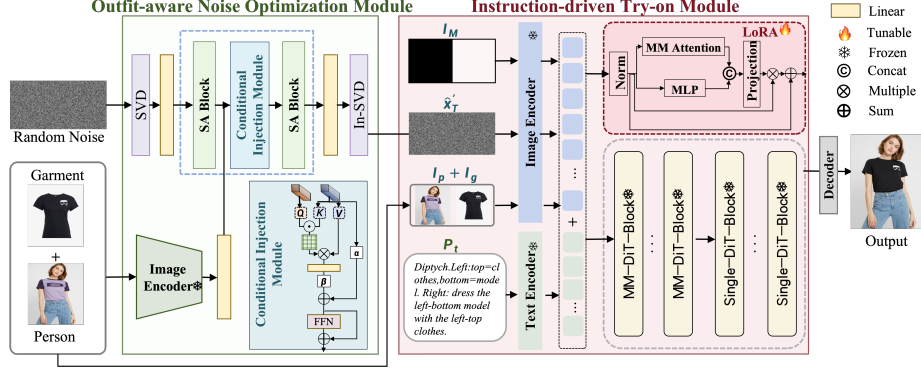


Fig. 2: Illustration of the FDM-MFVT framework. The left part shows the Outfit-aware Noise Optimization (OANO) module, which refines noise initialization. The right part shows the Instruction-driven Try-on (IDT) module, which integrates visual and textual features for mask-free synthesis.

while enabling high-quality outputs with minimal steps. Building on the optimized noise x'_T , the IDT module integrates both visual and textual instructions to guide the try-on process. Specifically, multi-modal encoders extract features from the garment image I_g , person image I_p and mask I_M , as well as textual prompts P_t , which describes the desired composition. These features are fused through attention-based mechanisms and low-rank adaptation layers, enabling precise alignment between clothing and body structure. The resulting representation is progressively refined through MM-DiT and Single-DiT blocks before being decoded into the final try-on image.

3.2 OANO Module

The OANO module is designed to refine noise initialization for mask-free virtual try-on. Instead of relying solely on random noise, OANO module constructs noise pairs (x_T, x'_T) by combining garment and person features with image inversion [1, 3, 29, 41]. These noise pairs provide more informative starting points, capturing structural and semantic features of the input. By training on directly generated noise pairs, OANO module produces optimized noise $\hat{x}'_T$ that enhances the realism and consistency of the final synthesis.

Noise Pair Construction Prior studies [30, 34] suggest that not all noises are equally effective, and the choice of initialization strongly influences generation quality. In our framework, noise pairs are directly constructed rather than learned. Specifically, given an input image, we perform inversion to obtain a target noise representation x'_T that encodes richer semantic features aligned with the garment and person context. This target noise serves as a more informative initialization compared to purely random noise x_T . We then filter the generated noise pairs by evaluating reconstruction quality: only those pairs in which x'_T

yields higher fidelity than x_T are retained. Formally, the construction process is expressed as:

$$\mathcal{D}_{\text{noise}} = \{(x_T, x'_T) \mid Q(x'_T) > Q(x_T)\}, \quad (1)$$

where $Q(\cdot)$ denotes the quality evaluation function. This ensures that the dataset emphasizes informative, outfit-aware noise samples.

Model Architecture Analysis [47] indicates that the singular vectors of x_T and x'_T exhibit strong similarity, which can be exploited for efficient noise optimization. Building on this property, the OANO module takes garment and person images (I_g, I_p) together with random noise x_T as inputs, and outputs optimized noise $\hat{x}'_T$. Specifically, we combine I_g and I_p to obtain I_{gp} , encode its features, and perform singular value decomposition (SVD) on x_T :

$$I_f = L(E(I_{gp})), \quad (2)$$

$$x_T = U \times \Sigma \times V^T, \quad (3)$$

$$x_{T_f} = L(\phi(U, \Sigma, V^T)). \quad (4)$$

where $E(\cdot)$ denotes the image encoder, $L(\cdot)$ a linear layer, and $\phi(\cdot)$ a ternary function integrating SVD components. The Conditional Injection Module then perturbs x_{T_f} using I_f through learnable attention-based operations:

$$\mathbf{V}_T = SA(x_{T_f})\mathbf{W}^V, \quad (5)$$

$$\mathbf{K}_T = SA(x_{T_f})\mathbf{W}^K, \quad (6)$$

$$\mathbf{Q}_R = I_f\mathbf{W}^Q, \quad (7)$$

$$\mathbf{Z}_T = \text{softmax}\left(\frac{\mathbf{Q}_R\mathbf{K}_T^T}{\sqrt{D_K}}\right) \cdot \mathbf{V}_T, \quad (8)$$

$$x'_{T_f} = \alpha Z_T W^E + \beta x_{T_f} + FFN(\alpha Z_T W^E + \beta x_{T_f}), \quad (9)$$

where $SA(\cdot)$ denotes self-attention, W^E is the output weight matrix, and α, β are learnable parameters. Finally, the optimized noise is obtained as:

$$\hat{x}'_T = U \times L(SA(x'_{T_f})) \times V^T. \quad (10)$$

Training The OANO module is trained on directly generated noise pairs $\mathcal{D}_{\text{noise}} = \{(x_T, x'_T, I_g, I_p)\}$, where x_T is random noise and x'_T is the target noise obtained via inversion and quality-based filtering. For each training instance, the module predicts $\hat{x}'_T$, supervised to approximate x'_T using mean squared error (MSE):

$$\mathcal{L}_{\text{OANO}} = \text{MSE}(x'_T, \hat{x}'_T) = \frac{1}{D} \|x'_T - \hat{x}'_T\|_2^2, \quad (11)$$

where D is the dimensionality of the noise tensor. This objective ensures that $\hat{x}'_T$ closely approximates x'_T , providing a robust initialization for downstream synthesis.

3.3 IDT Module

Model Architecture The IDT module is built upon a transformer-based denoising diffusion framework, leveraging the contextual reasoning ability of large-scale Diffusion Transformers (DiTs) [22]. Unlike mask-based methods, the IDT module enables mask-free virtual try-on by embedding both visual and textual modalities into a unified contextual prompt. Specifically, the garment image I_g , the person image I_p , and the mask I_M are first encoded using a variational autoencoder (VAE) encoder:

$$z_i = \text{VAE}(\text{Concat}(I_g, I_p) \odot I_M), \quad (12)$$

where z_i denotes the latent representations of the garment, person, and mask images. The textual instruction prompt P_t is encoded using a pretrained T5 encoder:

$$z_t = \text{T5}(P_t), \quad (13)$$

where z_t represents the textual embedding of the try-on instruction. All latent tokens are then concatenated to form the contextual input sequence:

$$Z = \text{Concat}(z_i, z_t). \quad (14)$$

This unified token sequence Z is subsequently fed into a frozen Inpainting Diffusion Transformer block D_I , together with the noisy latent X , to guide the try-on process:

$$I_t = D_I(Z, X). \quad (15)$$

LoRA Architecture To adapt the frozen DiT blocks for instruction-driven try-on, we employ Low-Rank Adaptation (LoRA). The encoded token sequence Z is first normalized:

$$Z_n = \text{Norm}(Z). \quad (16)$$

The normalized tokens are then processed by two parallel branches: multimodal attention (MM-Attn) and a multilayer perceptron (MLP):

$$Z_{attn} = \text{MM-Attn}(Z_n), \quad Z_{mlp} = \text{MLP}(Z_n). \quad (17)$$

The outputs of these branches are concatenated and projected through a linear layer:

$$Z_{proj} = \text{Linear}(\text{Concat}(Z_{attn}, Z_{mlp})). \quad (18)$$

Finally, the projected features are combined with the normalized input via element-wise multiplication and residual addition:

$$Z_{out} = Z_{proj} \odot Z_n + Z_n, \quad (19)$$

where $\odot$ denotes element-wise multiplication. The resulting Z_{out} serves as the LoRA-adapted output, enabling fine-tuning of the otherwise frozen DiT blocks. **Training** The training of IDT module with LoRA fine-tuning follows the standard denoising diffusion objective. Given a clean target latent x_0 and a noisy

latent x_t at timestep t , the model predicts the added noise $\hat{\varepsilon}_\theta$. The loss function is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{x_0, t, \varepsilon} \left[\|\varepsilon - \hat{\varepsilon}_\theta(x_t, t, Z_{\text{out}})\|^2 \right], \quad (20)$$

where ε is the sampled Gaussian noise, and $\hat{\varepsilon}_\theta$ is the noise predicted by the LoRA-adapted DiT block conditioned on Z_{out} . In practice, this objective ensures that the LoRA parameters learn to align garment, person, and mask features with textual instructions, while keeping the majority of the DiT parameters frozen. This design achieves efficient fine-tuning with minimal additional parameters, enabling instruction-driven virtual try-on in a zero-shot setting.

4 Experiments

4.1 Experimental Setting

Datasets To facilitate mask-free virtual try-on, we construct a large-scale and high-quality dataset, termed MFVT. The dataset provides garment–person–target triplets without explicit segmentation masks, overcoming the limitations of existing mask-based datasets. As shown in Fig. 3, MFVT is built by leveraging a pretrained mask-based virtual try-on model to generate person images conditioned on randomly sampled garment–person pairs from existing open-sourced mask-based datasets [4, 32]. For each sampled pair, the model infers a synthesized person image wearing the target garment. The generated image, together with the original person and garment images, forms the triplets in MFVT. In total, MFVT contains 30,000 pairs across diverse clothing categories: 10,671 upper-body, 9,761 lower-body, and 9,568 dresses. All the images in MFVT are standardized to a resolution of 1024×768 , and divided into training and test sets in a 7:3 ratio. Beyond MFVT, we further conduct experiments on two widely used open-sourced mask-based datasets: VTON-HD [4], DressCode [32], and StreetVTON [8]. VTON-HD is a high-resolution virtual try-on dataset that provides person–garment pairs with segmentation masks and emphasizes realistic garment transfer with fine-grained details. DressCode is a large-scale dataset focusing on fashion garments across upper-body, lower-body and dresses, offering diverse garment styles and person poses and widely adopted in prior virtual try-on research as a standard benchmark. StreetVTON support in-the-wild virtual try-on applications. In conclusion, we utilize MFVT for mask-free virtual try-on experiments, VTON-HD together with DressCode as representative mask-based benchmarks, and have demonstrated the scenario generalization of our method on StreetVTON.

Implementation Details We adopt Flux1.Fill as the backbone architecture for training. The IDT module is trained on two 80 GB NVIDIA A800 GPUs with a batch size of 16 for 15,000 steps. The OANO module is trained on a single 80 GB NVIDIA A800 GPU with a batch size of 64 for 150 epochs. We evaluate our model in two settings: mask-free and mask-based. In the mask-free setting, since only one open-source model is available, we evaluate exclusively on Any2AnyTryon [14], which we retrain using MFVT to ensure a fair comparison.

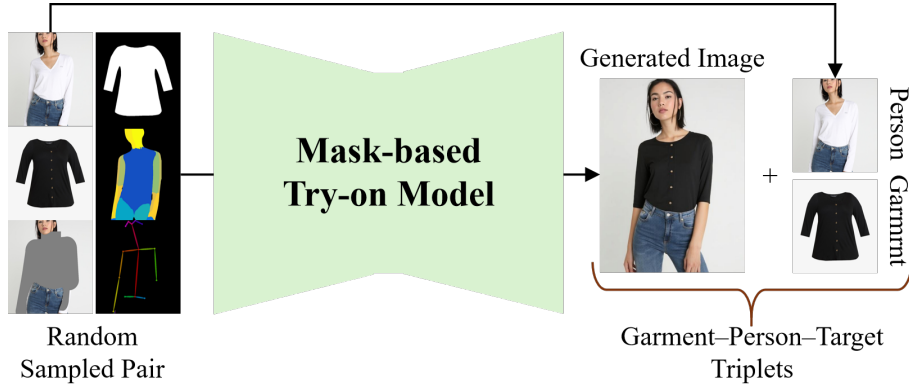


Fig. 3: Construction pipeline of the MFVT dataset.

In the mask-based setting, we compare our method against eight state-of-the-art methods: StableVTON [24], OOTDiffusion [43], CATVTON [6], TPD [43], IDM-VITON [5], ITA-MDT [20], OmniVTON [45] and FastFit [7]. We report both quantitative and qualitative results. Specifically, Quantitative evaluation includes objective metrics such as LPIPS [46], SSIM [42], FID [19], and KID [25], while qualitative evaluation is conducted through visual comparisons, highlighting perceptual realism, garment fidelity, and try-on consistency across diverse poses and identities.

4.2 Quantitative Results

Here, we present a comprehensive comparison of the quantitative results of our model against both mask-free and mask-based baselines.

Mask-free In the mask-free setting, we evaluate our framework against Any2AnyTryon on the MFVT dataset, which includes upper-body, lower-body, and dresses categories. The evaluation employs four metrics: LPIPS, SSIM, FID, and KID. As reported in Tab. 1, under the same number of inference steps, our method consistently surpasses Any2AnyTryon across all metrics. For example, at 6 inference steps, our method reduces FID by 2.906, 3.136 and 2.030 for upper-body, lower-body and dresses, respectively. The results highlight the enhanced fidelity and perceptual similarity of the synthesized images, validating the effectiveness of our framework. Remarkably, our method achieves performance comparable to the 30-step results of Any2AnyTryon even with only 6 inference steps. For instance, in the dress category, Any2AnyTryon requires 30 steps to reach 0.829 SSIM and 8.527 FID, while our FDM-MFVT achieves 0.872 SSIM and 8.431 FID with only 6 steps. This highlights both the efficiency and practicality of our framework in mask-free scenarios.

Mask-based In the mask-based setting, we evaluate our method against eight state-of-the-art virtual try-on methods: StableVTON [24], OOTDiffusion [43], CATVTON [6], TPD [43], IDM-VITON [5], ITA-MDT [20], OmniVTON [45]

Table 1: Quantitative evaluation of our method against Any2AnyTryon on the MFVT dataset, including upper-body, lower-body, and dresses categories. Inf.Step refers to the inference steps.

Method	Dataset	Inf.Step=6				Inf.Step=30			
		LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
Any2AnyTryon	Upper-body	0.324	0.769	11.247	1.421	0.147	0.854	8.471	0.725
	Lower-body	0.475	0.774	11.452	1.213	0.102	0.867	8.527	1.054
	Dresses	0.348	0.771	10.461	1.411	0.164	0.829	8.527	0.947
FDM-MFVT (Ours)	Upper-body	0.087	0.887	8.341	0.725	0.085	0.893	8.341	0.712
	Lower-body	0.083	0.879	8.316	0.968	0.082	0.880	8.219	0.945
	Dresses	0.102	0.872	8.431	0.901	0.091	0.883	8.286	0.872

Table 2: Quantitative comparison of our method and eight baselines on the unpaired test sets of VTON-HD [4] and DressCode [32] Upper-body datasets. Inf.Step refers to the inference steps. The best and second best results are in **bold** and underlined.

Method	VTON-HD				DressCode Upper			
	Inf.Step=6		Inf.Step=30		Inf.Step=6		Inf.Step=30	
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
StableVTON	13.427	3.427	9.717	1.450	22.847	6.547	13.666	3.857
OOTDDiffusion	16.362	5.871	12.268	3.214	37.106	7.839	29.383	4.302
CATVTON	13.949	3.946	10.263	2.370	23.185	3.147	12.720	1.910
TPD	16.235	6.692	13.124	3.527	25.073	<u>2.978</u>	12.957	1.874
IDM-VTON	12.539	<u>2.849</u>	9.261	<u>1.253</u>	21.849	3.041	11.879	<u>1.813</u>
ITA-MDT	<u>11.054</u>	3.014	8.889	1.397	<u>20.437</u>	3.256	11.928	1.871
OmniVTON	13.296	3.781	9.732	1.193	22.348	3.239	12.324	1.869
FastFit	12.316	3.258	<u>8.731</u>	1.264	21.316	3.124	<u>11.542</u>	1.824
FDM-MFVT (Ours)	8.687	1.201	8.659	1.010	11.231	1.401	10.963	1.394

and FastFit [7]. We report both quantitative and qualitative results. Specifically, Quantitative evaluation includes objective metrics such as LPIPS [46], SSIM [42], FID [19], and KID [25]. Experiments are conducted on the unpaired test sets of VTON-HD [4] and DressCode [32] Upper-body, where the input garment differs from the target garment. The evaluation employs two metrics: FID, and KID. All models are tested under identical inference conditions, with comparisons made at both 6 and 30 inference steps to assess efficiency. As shown in Tab. 2, our method consistently outperforms all baselines across both FID and KID. Remarkably, despite not relying on explicit masks, our framework achieves superior performance compared to mask-based methods, demonstrating strong robustness and generalization ability. For example, at 6 inference steps, our method surpasses the second-best baseline (ITA-MDT) by 2.367 FID on VTON-HD [4] and by 9.206 FID on DressCode [32] Upper-body. Moreover, our 6-step outputs already match or exceed the 30-step results of competing methods. For instance, on DressCode Upper-body, our 6-step FID of 11.231 is better than the 30-step FID of FastFit (11.542). Together, the above results confirm that our method advances the state of the art in both effectiveness, efficiency and robustness for virtual try-on tasks, achieving higher fidelity and diversity while requiring fewer inference steps.



Fig. 4: Qualitative results of our method on the StreetVTON dataset.

4.3 Qualitative Results

Here, we present a comprehensive comparison of qualitative results between our method and both mask-free and mask-based baselines, focusing on garment fidelity, texture realism, and overall visual coherence. Furthermore, we validated the scenario generalization ability of our method in the wild dataset as shown in Fig. 4. Fig. 4 demonstrates our method’s ability to maintain garment fit and accessories using the StreetV-TON dataset. Our method achieves more accurate clothing rendering while retaining original image content for a more realistic result. Although our method was not trained in the wild dataset, Fig. 4 shows the high fidelity of the try-on image generated by our method in the wild dataset, demonstrating the generalization ability of our method.

Mask-Free We conduct qualitative experiments on the MFVT dataset to compare our method with Any2AnyTryon across three garment categories: upper-body, lower-body, and dresses. As shown in Fig. 5, each row corresponds to one category, with two representative cases provided per row. The first row showcases upper-body garments such as tank tops and long-sleeve shirts, the second row presents lower-body garments including pants and skirts, and the third row displays dresses with varied styles and textures. Our method consistently produces more realistic and coherent try-on results, accurately preserving garment-specific attributes such as texture, length, and material appearance. For instance, in the second row, our output preserves the color contrast and structural details of the orange top, whereas Any2AnyTryon introduces undesired black-pants artifacts originating from the initial model’s garment. The enhanced garment-body alignment, texture realism, and structural consistency achieved by our method highlight its effectiveness in mask-free virtual try-on scenarios, especially under challenging conditions involving complex patterns and full-body poses.

Mask-based To assess performance in mask-based scenarios, we conduct comparisons between our method and six leading baselines: StableVTON [24], OOT-



Fig. 5: Qualitative comparison between our method and Any2AnyTryon on the MFVT dataset.

Diffusion [43], CATVTON [6], TPD [43], IDM-VITON [5], ITA-MDT [20], OmniVTON [45] and FastFit [7]. All evaluations are performed on the unpaired test sets of the VTON-HD [4] and DressCode [32] datasets. As illustrated in Fig. 6, our method consistently generates more photo-realistic try-on results with sharper textures and cleaner garment boundaries. For instance, in the second row, the black lace camisole preserves its intricate detailing and layered structure in our output, whereas the baseline results exhibit distortions around the wrist, partially retaining elements of the original garment and even hallucinating additional hair, largely due to poor mask quality. Overall, these comparisons highlight the robustness of our framework in mask-based scenarios, achieving superior visual fidelity and garment consistency over existing baselines.

Table 3: Quantitative results of FDM-MFVT ablation.

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
w/o OANO	0.121	0.864	8.417	1.019
w/o IDT	0.135	0.859	8.378	1.142
FDM-MFVT	0.091	0.879	8.363	0.865

4.4 Ablation Study

Ablation Study To evaluate the contribution of each component in our FDM-MFVT framework, we conduct an ablation study by selectively removing the OANO module and the IDT module. As shown in Table 3, removing OANO

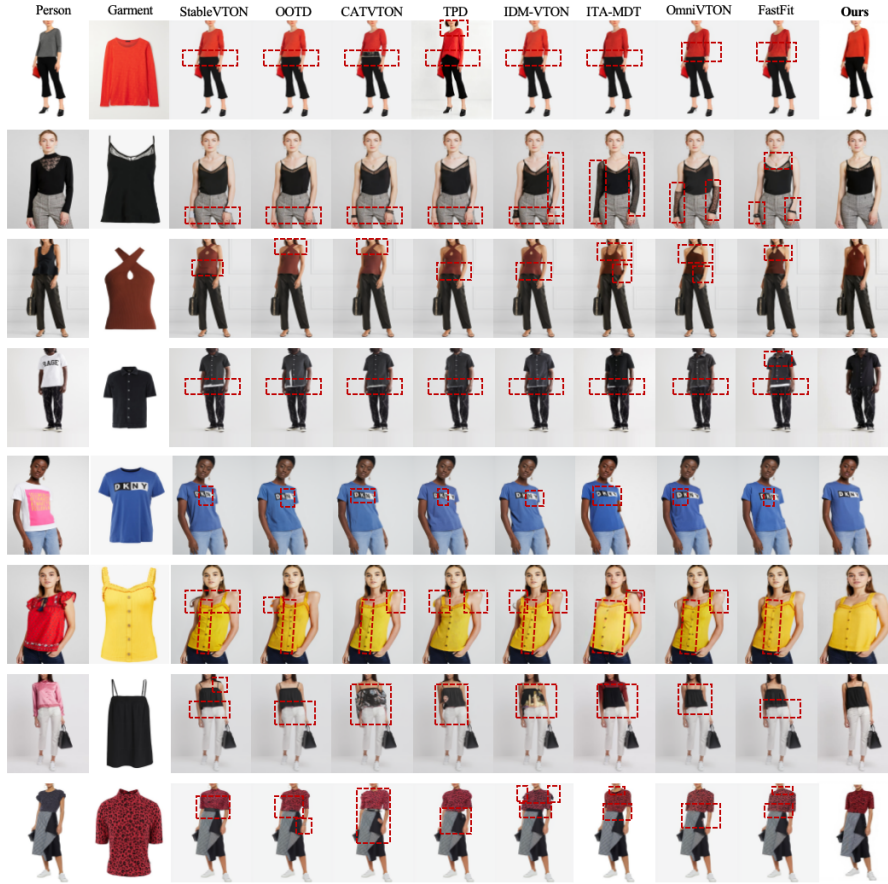


Fig. 6: Qualitative comparison between our method and eight baselines on the unpaired test sets of VTON-HD [4] and DressCode [32] Upper-body datasets.

leads to a noticeable degradation in perceptual quality, with LPIPS increasing from 0.091 to 0.121 and SSIM dropping from 0.879 to 0.864. This confirms the importance of noise refinement in enhancing structural consistency and visual realism. Similarly, excluding IDT module results in further performance decline, with 0.135 and 0.859 decline in LPIPS and SSIM, indicating that instruction-guided alignment plays a crucial role in generating coherent garment-body compositions. The full model achieves the best results across all metrics, including the lowest FID of 8.363 and KID of 0.865, demonstrating the effectiveness of integrating both modules. Qualitative comparisons in Fig. 7 further support these results. Models without OANO or IDT module exhibit noticeable artifacts, misaligned garments, and reduced texture fidelity, whereas our complete framework produces visually consistent and realistic try-on results.



Fig. 7: Qualitative results of FDM-MFVT ablation.

Efficiency Analysis The OANO module significantly reduced the number of inference steps and obtained higher fidelity virtual try-on images in just 0.28s. In addition, the OANO module has a small number of parameters and is easy to train. These factors highlight the lightweight, efficient, and broad application potential of OANO module.

5 Conclusion

In this work, we presented FDM-MFVT, a novel few-step diffusion framework for mask-free image-based virtual try-on. By integrating an OANO module and an IDT module, our method significantly reduced sampling cost while maintaining high-quality synthesis. The OANO module enabled efficient convergence by aligning the initial noise with the input image, achieving comparable results to 30-step baselines with only 6 inference steps. The IDT module further enhanced flexibility by leveraging garment-person inputs without relying on segmentation masks or auxiliary networks. To support mask-free training and evaluation, we introduced MFVT, a large-scale paired dataset comprising 30,000 high-resolution garment-person pairs across diverse categories. Extensive experiments on MFVT, VTON-HD [4], DressCode [32] and StreetVTON [8] demonstrated that FDM-MFVT consistently outperformed both mask-free and mask-based baselines in terms of fidelity, diversity, and efficiency. Ablation studies confirmed the effectiveness of each module, and qualitative comparisons highlighted superior texture preservation and garment-body alignment. Together, our contributions advanced the state of the art in virtual try-on by addressing key limitations of existing diffusion-based methods, offering a scalable, robust, and real-time solution for mask-free try-on synthesis.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 62401027 and 62231002.

References

1. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7877–7888 (2025)
2. Berthelot, D., Autef, A., Lin, J., Yap, D.A., Zhai, S., Hu, S., Zheng, D., Talbott, W., Gu, E.: Tract: Denoising diffusion models with transitive closure time-distillation. arXiv preprint arXiv:2303.04248 (2023)
3. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 22560–22570 (2023)
4. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14131–14140 (2021)
5. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
6. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Jiang, D., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models. arXiv preprint arXiv:2407.15886 (2024)
7. Chong, Z., Lei, Y., Zhang, S., He, Z., Wang, Z., Zhang, X., Dong, X., Wu, Y., Jiang, D., Liang, X.: Fastfit: Accelerating multi-reference virtual try-on via cacheable diffusion models. arXiv preprint arXiv:2508.20586 (2025)
8. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 1414–1423 (2025)
9. De Divitiis, L., Becattini, F., Baecchi, C., Del Bimbo, A.: Disentangling features for fashion recommendation. *ACM Transactions on Multimedia Computing, Communications and Applications* **19**(1s), 1–21 (2023)
10. Feng, Y., Zhang, L., Cao, H., Chen, Y., Feng, X., Cao, J., Wu, Y., Wang, B.: Omnistry: Virtual try-on anything without masks. arXiv preprint arXiv:2508.13632 (2025)
11. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: 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 (2021)
12. Gu, J., Zhai, S., Zhang, Y., Liu, L., Susskind, J.M.: Boot: Data-free distillation of denoising diffusion models with bootstrapping. In: ICML 2023 Workshop on Structured Probabilistic Inference {\&} Generative Modeling. vol. 3 (2023)
13. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7297–7306 (2018)

14. Guo, H., Zeng, B., Song, Y., Zhang, W., Liu, J., Zhang, C.: 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)
15. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: Proceedings of the IEEE international conference on computer vision. pp. 3343–3351 (2015)
16. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10471–10480 (2019)
17. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)
18. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3470–3479 (2022)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
20. Hong, J.W., Ton, T., Pham, T.X., Koo, G., Yoon, S., Yoo, C.D.: Ita-mdt: Image-timestep-adaptive masked diffusion transformer framework for image-based virtual try-on. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28284–28294 (2025)
21. Hsiao, W.L., Grauman, K.: Creating capsule wardrobes from fashion images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7161–7170 (2018)
22. Huang, L., Wang, W., Wu, Z.F., Shi, Y., Dou, H., Liang, C., Feng, Y., Liu, Y., Zhou, J.: In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775* (2024)
23. Issenhuth, T., Mary, J., Calauzenes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: European Conference on Computer Vision. pp. 619–635. Springer (2020)
24. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8176–8185 (2024)
25. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830* (2019)
26. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: European Conference on Computer Vision. pp. 204–219. Springer (2022)
27. Li, Y., Wang, H., Jin, Q., Hu, J., Chemerys, P., Fu, Y., Wang, Y., Tulyakov, S., Ren, J.: Snapfusion: Text-to-image diffusion model on mobile devices within two seconds. *Advances in Neural Information Processing Systems* **36**, 20662–20678 (2023)
28. Liang, X., Qu, Z., Zou, M., Liu, J., Jiang, L., Xu, M., Zhu, Y.: Vtedit-bench: A comprehensive benchmark for multi-reference image editing models in virtual try-on. *arXiv preprint arXiv:2603.11734* (2026)

29. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
30. Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14297–14306 (2023)
31. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In: CVPR workshops. vol. 3, pp. 10–14 (2020)
32. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2231–2235 (2022)
33. Ning, S., Wang, D., Qin, Y., Jin, Z., Wang, B., Han, X.: Picture: Photorealistic virtual try-on from unconstrained designs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6976–6985 (2024)
34. Qi, Z., Bai, L., Xiong, H., Xie, Z.: Not all noises are created equally: Diffusion noise selection and optimization. arXiv preprint arXiv:2407.14041 (2024)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
36. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512 (2022)
37. Sarkar, R., Bodla, N., Vasileva, M.I., Lin, Y.L., Beniwal, A., Lu, A., Medioni, G.: Outfittransformer: Learning outfit representations for fashion recommendation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3601–3609 (2023)
38. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
39. Song, Y., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 9198–9207 (2020)
40. Wang, B., Zheng, H., Liang, X., Shen, X., Shao, L.: Toward photo-realistic virtual try-on by adaptively generating-preserving image content. IEEE Transactions on Image Processing **27**(12), 6283–6294 (2018)
41. Wang, J., Pu, J., Qi, Z., Guo, J., Ma, Y., Huang, N., Chen, Y., Li, X., Shan, Y.: Taming rectified flow for inversion and editing. arXiv preprint arXiv:2411.04746 (2024)
42. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
43. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8996–9004 (2025)
44. Yang, X., Ding, C., Hong, Z., Huang, J., Tao, J., Xu, X.: 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 (2024)
45. Yang, Z., Li, Y., He, S., Li, X., Xu, Y., Dong, J., Du, Y.: Omnivton: Training-free universal virtual try-on. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16702–16711 (2025)

- 46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
- 47. Zhou, Z., Shao, S., Bai, L., Zhang, S., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17688–17697 (2025)