FLDM-VTON: Faithful Latent Diffusion Model for Virtual Try-on

Chenhui Wang[1] , **Tao Chen**[1] , **Zhihao Chen**[1] , **Zhizhong Huang**[2] ,
Taoran Jiang[3] , **Qi Wang**[3] , **Hongming Shan**[1,4*]

[1] Institute of Science and Technology for Brain-inspired Intelligence, MOE Frontiers Center for Brain Science, and Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China

[2] School of Computer Science, Fudan University, Shanghai 200433, China

[3] Suzhou Xiangji Technology Service Co., Ltd., Suzhou 215223, China

[4] Shanghai Center for Brain Science and Brain-inspired Technology, Shanghai 200031, China

chenhuiwang21@m.fudan.edu.cn, hmshan@fudan.edu.cn

Abstract

Despite their impressive generative performance, latent diffusion model-based virtual try-on (VTON) methods lack faithfulness to crucial details of the clothes, such as style, pattern, and text. To alleviate these issues caused by the diffusion stochastic nature and latent supervision, we propose a novel **F**aithful **L**atent **D**iffusion **M**odel for VTON, termed FLDM-VTON. FLDM-VTON improves the conventional latent diffusion process in three major aspects. First, we propose incorporating warped clothes as both the starting point and local condition, supplying the model with faithful clothes priors. Second, we introduce a novel clothes flattening network to constrain generated try-on images, providing clothes-consistent faithful supervision. Third, we devise a clothes-posterior sampling for faithful inference, further enhancing the model performance over conventional clothes-agnostic Gaussian sampling. Extensive experimental results on the benchmark VITON-HD and Dress Code datasets demonstrate that our FLDM-VTON outperforms state-of-the-art baselines and is able to generate photo-realistic try-on images with faithful clothing details.

1 Introduction

Image-based virtual try-on (VTON) aims to transfer a piece of in-shop flat clothes onto one person's body while preserving the details of both the human and the clothes, such as style, pattern, and text. In the past decade, VTON has attracted considerable attention [Wang *et al.*, 2018; Choi *et al.*, 2021; Minar *et al.*, 2020; Lee *et al.*, 2022; Morelli *et al.*, 2023; Gou *et al.*, 2023; Xie *et al.*, 2023], and with the rapid advances of generative artificial intelligence [Vaswani *et al.*, 2017; Ho *et al.*, 2020] it has the great potential to improve users' shopping experience by bridging the gap between users and online shopping.

Prior methods for VTON highly rely on generative adversarial networks (GANs) [Goodfellow *et al.*, 2014; Huang *et al.*, 2023; Wang *et al.*, 2024] to synthesize try-on images. Typically, they first use thin plate spline (TPS)-based [Wang *et al.*, 2018; Li *et al.*, 2021; Fele *et al.*, 2022] or appearance flow-based [Han *et al.*, 2019; He *et al.*, 2022] algorithms to warp the flat clothes to the person's body, and then use GAN to further refine the previously generated try-on images. Nevertheless, due to the mode collapse issue [Bau *et al.*, 2019], GAN-based methods fail to synthesize photo-realistic try-on images and accurately capture intricate clothing details, often leading to flaws on the generated results; see Figure 1.

Recently, the diffusion model has shown remarkable generative capabilities across various tasks, such as image inpainting, image editing, and even segmentation [Ho *et al.*, 2020; Rombach *et al.*, 2022; Song *et al.*, 2020; Chen *et al.*, 2023]. Compared with GAN, the diffusion model offers more stable training and direct likelihood estimation. However, directly applying the diffusion model to high-resolution VTON is infeasible due to limited computational resources. Therefore, current diffusion-based VTON methods [Morelli *et al.*, 2023; Gou *et al.*, 2023] are built upon the latent diffusion model (LDM) [Ramesh *et al.*, 2021] that performs diffusion process in a latent space. Although showing effectiveness in generating realistic try-on images, they often produce *unfaithful* clothing details with respect to the original flat clothes.

We identify the *stochastic nature* and *latent supervision* of LDM as the key limiting factors for the faithfulness. On one hand, the diffusion stochastic nature poses a challenge in preserving clothing details, as indicated by the initial Gaussian noise introduced at the sampling process and the added Gaussian noise at each time-step. On the other hand, the latent supervision falls short in providing image-level supervision for fine clothing details. Thus, generating highly faithful clothing details with respect to the original flat clothes using the diffusion model remains a significant challenge.

To alleviate these issues in LDM, we propose a novel **F**aithful **L**atent **D**iffusion **M**odel for VTON, termed FLDM-VTON. To achieve faithful try-on generation, our FLDM-VTON improves the training of conventional latent diffusion process in two major aspects: **(i)** supplying the model with

*Corresponding author

Figure 1: Comparison between state-of-the-art baselines and our FLDM-VTON on the VITON-HD dataset.

faithful clothes priors by leveraging warped clothes as both the starting point and local condition to mitigate the initial and in-process added stochasticity, respectively, and **(ii)** providing *clothes-consistent faithful supervision* through a novel clothes flattening network to bring additional image-level constraints from the original flat clothes. In addition to the training improvements, our FLDM-VTON also improves the inference process by devising a clothes-posterior sampling, further enhancing the model performance over conventional clothes-agnostic Gaussian sampling.

Contributions. Our contributions are as follows. **(i)** We propose a novel faithful latent diffusion model for VTON to address the unfaithful issue caused by the diffusion stochastic nature and latent supervision. **(ii)** We propose incorporating warped clothes as both the starting point and local condition, supplying the model with faithful clothes priors. **(iii)** We introduce a novel clothes flattening network to constrain the generated try-on images, providing clothes-consistent faithful supervision. **(iv)** We devise a clothes-posterior sampling for faithful inference, further enhancing the model performance over conventional clothes-agnostic Gaussian sampling. **(v)** Extensive experimental results on the VITON-HD and Dress Code datasets demonstrate that our FLDM-VTON outperforms state-of-the-art baselines and is able to generate photo-realistic try-on images with faithful clothing details.

2 Related Work

Image-based virtual try-on. Numerous VTON studies have investigated generating high-resolution photo-like try-on images by transferring a in-shop flat clothes item onto one person's body, enhancing the online shopping experience. Currently, most VTON's works can be divided into two stages: pre-warping and refining. In the pre-warping stage, thin plate spline (TPS)-based [Wang *et al.*, 2018; Li *et al.*, 2021; Fele *et al.*, 2022] or appearance flow-based [Han *et al.*, 2019; Ge *et al.*, 2021; He *et al.*, 2022] algorithms are used to warp the flat clothes to the corresponding positions on the person body. In the refinement stage, convolutional neural network (CNN) or GAN techniques are employed to further refine generated images. Although there are some diffusion-based VTON methods [Morelli *et al.*, 2023; Gou *et al.*, 2023; Zhu *et al.*, 2023], they often generate unfaithful clothing details. In contrast, our FLDM-VTON prioritizes generating realistic try-on images with faithful clothing details.

Diffusion models. Inspired by non-equilibrium statistical physics, diffusion models have recently presented strong capability in image synthesis. To generate high-resolution images with limited computational resources, LDM, *e.g*. DALL-E and Stable Diffusion (SD) series [Ramesh *et al.*, 2021; Rombach *et al.*, 2022; Podell *et al.*, 2023], conducts the diffusion process in the latent space. Moreover, customized or subject-driven image generation has recently become an important research topic, including text-driven [Ruiz *et al.*, 2023; Wei *et al.*, 2024] and image exemplar-based generation [Yang *et al.*, 2023]. DreamBooth [Ruiz *et al.*, 2023] embeds a given subject instance in the output domain of a text-to-image diffusion model by binding the subject to a unique identifier. PAINT-BY-EXAMPLE [Yang *et al.*, 2023] uses CLIP [Radford *et al.*, 2021] to convert the target image as an embedding for guidance. In contrast to their emphasis on preserving general subject information, our FLDM-VTON focuses on preserving more fine-grained clothing details.

3 Methodology

In this section, we present the try-on diffusion model with faithful clothes priors in Sec. 3.1 and clothes-consistent faithful supervision in Sec. 3.2, followed by an overview of the FLDM-VTON and a clothes-posterior sampling for faithful inference in Sec. 3.3.

3.1 Try-on Diffusion with Clothes Priors

Given a person image $\boldsymbol{P} \in \mathbb{R}^{H\times W\times 3}$ and a mask $\boldsymbol{m} \in \{0,1\}^{H\times W}$ indicating the try-on region, one can obtain a clothes-agnostic person image $\boldsymbol{P}^{\mathrm{a}}$ through element-wise multiplication; $\boldsymbol{P}^{\mathrm{a}}$ refers to the person image with the try-on region being masked out. The goal of VTON is to transfer a piece of flat clothes $\boldsymbol{C} \in \mathbb{R}^{H\times W\times 3}$ onto $\boldsymbol{P}^{\mathrm{a}}$, yielding a photo-realistic try-on image $\widehat{\boldsymbol{T}} \in \mathbb{R}^{H\times W\times 3}$ with faithful clothing details.

Current state-of-the-art (SOTA) LDM-based VTON methods [Morelli *et al.*, 2023; Gou *et al.*, 2023] first use TPS or flow-based warping to generate the warped clothes $\boldsymbol{C}^{\mathrm{w}}$ from

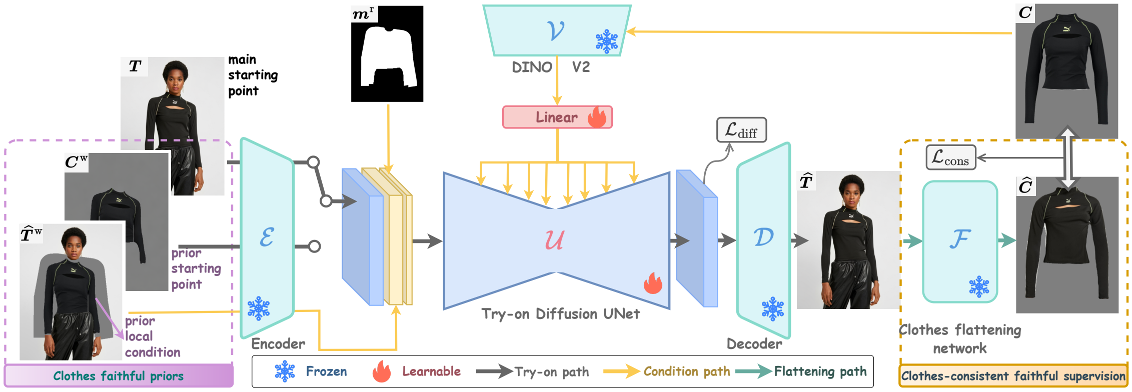

Figure 2: Overview of the proposed FLDM-VTON. Our FLDM-VTON is trained with main and prior denoising input, constrained by DINO-V2, and the proposed clothes flattening network to generate the try-on images $\widehat{\boldsymbol{T}}$.

the flat clothes $\boldsymbol{C}$, and then employ an LDM for realistic refinement. With a pre-trained encoder $\mathcal{E}$ and decoder $\mathcal{D}$, LDM trains a diffusion model in the latent space, involving forward and reverse processes. In the forward process, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ is added at arbitrary time-step t to the resultant latent feature $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{T}) \in \mathbb{R}^{h \times w \times c}$, where $\boldsymbol{T}$ is the ground-truth try-on image. In the reverse process, a diffusion UNet is employed to estimate the added noise $\boldsymbol{\epsilon}$.

Although achieving realistic results, conventional LDM-based VTON methods lack faithfulness to original clothing details. We identify the diffusion stochastic nature as the key limiting factor, which can be reflected in two primary aspects: **(i)** the initial Gaussian noise introduced at the sampling process and **(ii)** the added Gaussian noise at each time-step. To alleviate these, our idea is to supply the model with clothes priors, leveraging warped clothes **(i)** as the starting point to address the initial stochasticity and **(ii)** as the local condition to mitigate the in-process added stochasticity.

Figure 2 presents an overview of our FLDM-VTON. Unlike conventional forward process that takes $\mathcal{E}(\boldsymbol{T})$ as the starting point, ours takes either $\mathcal{E}(\boldsymbol{T})$ or warped clothes feature $\mathcal{E}(\boldsymbol{C}^{\mathrm{w}})$ as the starting point, in which warped clothes feature provides clothes prior at the beginning. To differentiate these two starting points, we refer to $\boldsymbol{z}_0^{\mathrm{m}} = \mathcal{E}(\boldsymbol{T})$ as *main* starting point and $\boldsymbol{z}_0^{\mathrm{p}} = \mathcal{E}(\boldsymbol{C}^{\mathrm{w}})$ as *prior* starting point. Moreover, we leverage the pre-warped try-on image feature $\mathcal{E}(\widehat{\boldsymbol{T}}^{\mathrm{w}})$ as prior local condition for all time-steps, where $\widehat{\boldsymbol{T}}^{\mathrm{w}} = \boldsymbol{C}^{\mathrm{w}} + \boldsymbol{P}^{\mathrm{a}}$.

Next, we detail the concrete forward and reverse processes.

Forward process. We gradually add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$ on the main and prior starting latent features, $\boldsymbol{z}_0^{\mathrm{m}}$ and $\boldsymbol{z}_0^{\mathrm{p}}$, with an arbitrary time-step t, yielding t-th corresponding latent features as follows:

$$\boldsymbol{z}_t^{\mathrm{m}} = \sqrt{\alpha_t}\boldsymbol{z}_0^{\mathrm{m}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{z}_t^{\mathrm{p}} = \sqrt{\alpha_t}\boldsymbol{z}_0^{\mathrm{p}} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad (1)$$

where $\alpha_t := \prod_{s=1}^{t}(1 - \beta_s)$, and β_s is a pre-defined variance schedule [Nichol and Dhariwal, 2021].

Reverse process. We have down-sized mask $\boldsymbol{m}^{\mathrm{r}} \in \{0, 1\}^{h \times w}$ as the denoising condition and pre-warped try-on image feature $\mathcal{E}(\widehat{\boldsymbol{T}}^{\mathrm{w}})$ as the prior local condition. We concatenate the t-th latent feature, prior local condition, and denoising condition along the channel dimension, severing as the input to the diffusion UNet:

$$\boldsymbol{\psi}_t^{\mathrm{m}} = [\boldsymbol{z}_t^{\mathrm{m}}; \mathcal{E}(\widehat{\boldsymbol{T}}^{\mathrm{w}}); \boldsymbol{m}_{\mathrm{r}}], \quad \boldsymbol{\psi}_t^{\mathrm{p}} = [\boldsymbol{z}_t^{\mathrm{p}}; \mathcal{E}(\widehat{\boldsymbol{T}}^{\mathrm{w}}); \boldsymbol{m}_{\mathrm{r}}], \quad (2)$$

where $[\cdot;\cdot]$ denotes concatenation operation.

Given one image pair, we obtain two distinct inputs: the main denoising input $\boldsymbol{\psi}_t^{\mathrm{m}}$ and the prior denoising input $\boldsymbol{\psi}_t^{\mathrm{p}}$. These inputs are individually processed through the same try-on diffusion UNet $\mathcal{U}$ to predict the main starting latent feature $\boldsymbol{z}_0^{\mathrm{m}}$. In addition, we also encode the flat clothes $\boldsymbol{C}$ through DINO-V2 [Oquab *et al.*, 2023], a currently powerful self-supervised visual encoder $\mathcal{V}$, which serves as the global controller being injected into each UNet layer via cross-attention. Therefore, the diffusion training loss function over one single sample and one time-step t is defined as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{diff}} = \frac{1}{2}\big(&\|\mathcal{U}(\boldsymbol{\psi}_t^{\mathrm{m}}, \mathcal{V}(\boldsymbol{C}), t) - \boldsymbol{z}_0^{\mathrm{m}}\|_2^2 \\ &+\|\mathcal{U}(\boldsymbol{\psi}_t^{\mathrm{p}}, \mathcal{V}(\boldsymbol{C}), t) - \boldsymbol{z}_0^{\mathrm{m}}\|_2^2\big),\end{aligned} \quad (3)$$

where $\boldsymbol{\psi}_t^{\mathrm{m}}$ contributes to preserving the photo-realistic quality as established by existing diffusion models and $\boldsymbol{\psi}_t^{\mathrm{p}}$ contributes to enhancing the faithfulness of generated clothes.

3.2 Clothes-consistent Faithful Supervision

Although clothes priors can help enhance the faithfulness from the input, it is still challenging to preserve the fine details such as pattern and text since the training is only supervised by the ground-truth try-on latent feature. To further improve the faithfulness to fine details, we introduce clothes-consistent faithful supervision, drawing inspiration from the fact that the clothes item you try on should be identical to the flat one once you take it off and flatten it out. To this end, we introduce a clothes flattening network $\mathcal{F}$ that can take off clothes from the generated try-on image and flatten it out like the original flat one.

Clothes flattening network. Our clothes flattening network is a two-step method: (i) take-off step and (ii) flatten-out step. The take-off step can be easily done by masking out the generated try-on image with the pre-parsed clothes mask $\boldsymbol{m}^{\mathrm{C}}$. The flatten-out step is an inverse warping process, which is done by training a flattening module to predict

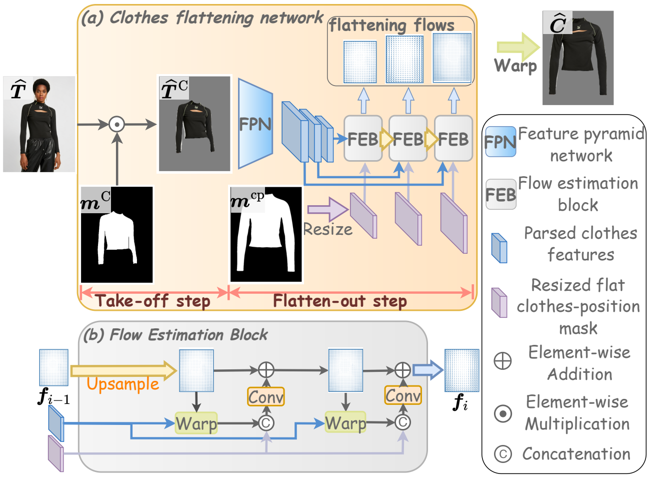

Figure 3: Illustration of the proposed clothes flattening network.

flattening flows. More specifically, our clothes flattening network is designed with a U-shape structure, which utilizes a Feature Pyramid Network (FPN) [Lin *et al.*, 2017] to encode the clothes-parsed feature at multiple scales, and then employs cascaded flow estimation blocks to predict the flattening flows with down-sized flat clothes-position masks. Note that we use five different multi-scale features in our experiments; three of these are illustrated in Figure 3 for simplicity.

Following the SOTA appearance flow training strategy [Ge *et al.*, 2021], we use mixed loss function to train our clothes flattening network, including $\mathcal{L}_1$ loss and perceptual loss $\mathcal{L}_{\text{per}}$ [Johnson *et al.*, 2016] at the image level and the second-order smooth loss $\mathcal{L}_{\text{sec}}$ and the total-variation loss $\mathcal{L}_{\text{TV}}$ at the flow level, which is defined as:

$$\mathcal{L}_{\text{flat}} = \mathcal{L}_1 + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{sec}}\mathcal{L}_{\text{sec}} + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}}, \tag{4}$$

where λ_* are the hyperparameters to adjust the weights among different loss components.

Clothes-consistent supervision. Once the clothes flattening network is trained, we can use the frozen clothes flattening network $\mathcal{F}$ to process the generated try-on images $\widehat{\boldsymbol{T}} = \mathcal{D}(\hat{z}_0^{\text{m}})$, yielding estimated flat clothes $\widehat{\boldsymbol{C}}$. Measuring the difference between estimated flat clothes $\widehat{\boldsymbol{C}}$ and original flat clothes $\boldsymbol{C}$ can further provide clothes consistent faithful supervision for try-on diffusion training, which is defined as:

$$\mathcal{L}_{\text{cons}} = \|\widehat{\boldsymbol{C}} - \boldsymbol{C}\|_1 = \|\mathcal{F}(\widehat{\boldsymbol{T}}) - \boldsymbol{C}\|_1. \tag{5}$$

3.3 Overview of Our FLDM-VTON

Overview. Figure 2 presents the overview of our FLDM-VTON. During the training phase, we handle each sample by deriving its main and prior denoising inputs, *i.e.* ψ_t^{m} and ψ_t^{p}, by concatenating the t-th latent feature with prior local and denoising conditions. And then, our try-on diffusion UNet $\mathcal{U}$ individually processes these two types of inputs guided by the global controller DINO-V2 $\mathcal{V}$, estimating the try-on latent feature with the diffusion loss $\mathcal{L}_{\text{diff}}$ in Eq. (3). Moreover, our FLDM-VTON incorporates a clothes flattening network, providing clothes-consistent faithful supervision with the clothes-consistent loss $\mathcal{L}_{\text{cons}}$ in Eq. (5). The overall training loss for the try-on diffusion UNet is defined as follows:

$$\mathcal{L}_{\text{Try-on}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \tag{6}$$

where λ_{cons} is a trade-off hyperparameter. Note that only the try-on diffusion UNet and the linear layer followed by the DINO V2 are trained.

Faithful inference. During the inference phase, conventional diffusion models initiate inference by a clothes-agnostic noise sampled from a standard Gaussian distribution. However, this introduces significant initial stochasticity, adversely affecting the faithfulness of generated clothing details. To address this issue, we devise a clothes-posterior sampling to further enhance the model performance.

With the introduced clothes prior for model training, our FLDM-VTON can initiate inference from a posterior Gaussian noise that is specifically conditioned by the warped clothes feature $\mathcal{E}(\boldsymbol{C}^{\text{w}})$. Specifically, the clothes-posterior noise is the T-th prior latent feature: $\boldsymbol{z}_T^{\text{p}} = \sqrt{\alpha_T}\mathcal{E}(\boldsymbol{C}^{\text{w}}) + \sqrt{1-\alpha_T}\boldsymbol{\epsilon}$, corresponding to the warped clothes feature $\mathcal{E}(\boldsymbol{C}^{\text{w}})$ after T diffusion forward time-steps using Eq. (1). By doing this, the initial stochasticity of the sampling process is significantly reduced, thereby ensuring the faithfulness of the generated clothing details.

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on two popular high-resolution VTON benchmarks: the VITON-HD dataset [Choi *et al.*, 2021] and Dress Code dataset [Morelli *et al.*, 2022]. Both datasets contain high-resolution paired images of size 512×384: in-shop flat clothes and their corresponding persons wearing the clothes. The VITON-HD dataset contains $13,679$ image pairs for upper-body clothes. The Dress Code dataset includes $15,366$ image pairs for upper-body clothes, $8,951$ image pairs for lower-body clothes, and $29,478$ image pairs for dresses. We follow the official guidelines to divide the data into training and testing sets [Choi *et al.*, 2021; Morelli *et al.*, 2022].

Implementation details. We use SOTA appearance flow-based warping method [Ge *et al.*, 2021; Gou *et al.*, 2023] to generate a warped clothes image aligning the flat clothes with the person. We adopt Adam optimizer to optimize all networks with a mini-batch size of 8 and a learning rate of 2.0×10^{-5} on 4 NVIDIA V100 GPUs. In addition, we employ the encoder and decoder of SD KL-regularized auto-encoder, with a down-sampling factor of $d = 8$ and a latent channel number of $c = 4$, as our encoder $\mathcal{E}$ and decoder $\mathcal{D}$, respectively. We set $T = 1,000$ for latent diffusion training as suggested by SD [Lee *et al.*, 2022], and use the DPM solver [Lu *et al.*, 2022] with 50 sampling steps for inference. Besides, we use FreeU [Si *et al.*, 2023] to reweight the contributions of the backbone and skip connection features with different scaling factors for inference, enhancing the denoising capability of the LDM and reducing low-frequency information.

Figure 4: Qualitative results of different methods and ours on the VITON-HD dataset. Best viewed when zoomed in.

Methods		Paired		Unpaired	
		LPIPS↓	SSIM↑	FID↓	KID↓
CP-VTON	I	0.160	0.831	31.34	2.37
CP-VTON+		0.131	0.847	22.79	1.55
VITON-HD	II	0.116	0.862	12.12	0.32
HR-VITON		0.104	0.878	11.27	0.27
GP-VTON		<u>0.081</u>	<u>0.884</u>	9.19	**0.09**
PAINT-BY-EXAMPLE	III	0.143	0.803	11.94	0.39
LADI-VTON		0.096	0.863	9.47	0.19
DCI-VTON		<u>0.081</u>	0.880	**8.76**	<u>0.11</u>
FLDM-VTON (**ours**)		**0.080**	**0.886**	<u>8.81</u>	0.13

Table 1: Comparison results on the VITON-HD dataset. **I**, **II**, and **III** represent CNN-based, GAN-based, and diffusion-based methods, respectively. The best and second best results are highlighted in **bold** and <u>underlined</u>, respectively.

Evaluation metrics. We evaluate performance under paired and unpaired test settings, where ground-truth try-on images are available for the paired setting while not for the unpaired setting. We adopt Learned Perceptual Image Patch Similarity (LPIPS) [Zhang *et al.*, 2018] and Structural Similarity (SSIM) [Wang *et al.*, 2004] for paired setting while using Fréchet Inception Distance (FID) [Heusel *et al.*, 2017] and Kernel Inception Distance (KID) [Bińkowski *et al.*, 2018] for unpaired setting. Note that we present all qualitative results under the unpaired setting.

4.2 Comparison with SOTA Methods

We comprehensively compare our FLDM-VTON with several SOTA methods, which can be categorized into three groups: CNN-based, GAN-based, and diffusion-based. Specifically, the CNN-based methods are CP-VTON [Wang *et al.*, 2018] and CP-VTON+ [Minar *et al.*, 2020]. For the GAN-based methods, we include VITON-HD [Choi *et al.*, 2021], HR-VITON [Lee *et al.*, 2022], and GP-VTON [Xie *et al.*, 2023]. Moreover, we compare our FLDM-VTON with recently published SOTA diffusion-based methods: PAINT-BY-EXAMPLE [Yang *et al.*, 2023], LADI-VTON [Morelli *et al.*, 2023], and DCI-VTON [Gou *et al.*, 2023]. Due to the extensive size of the Dress Code dataset, we only compare our FLDM-VTON with several SOTA methods, including GP-VTON, PAINT-BY-EXAMPLE, and LADI-VTON.

Quantitative comparison. Tables 1 and 2 present quantitative results on the VITON-HD and Dress Code datasets, respectively, which show that our FLDM-VTON outperforms most competitors across various performance metrics under paired and unpaired settings. We observe that diffusion-based methods perform better in terms of FID and KID, two realistic metrics. However, our FLDM-VTON not only maintains comparable realistic performance but also excels in faithfulness under the paired setting, *i.e.* LPIPS and SSIM.

Qualitative comparison. Figures 1, 4, and 5 present qualitative results, which demonstrate the superior performance of our FLDM-VTON in generating realistic try-on images while preserving faithful clothing details to the original flat clothes. Although CNN-based methods generate try-on images that more closely resemble the original flat clothes, they lack realism and don't achieve a photo-like quality. With adversarial training to enhance the initially warped clothes, GP-VTON stands out with superior performance. However, its qualitative results suggest a relatively simple composition of warped clothes and persons due to GAN's intrinsic

Figure 5: Qualitative results of different methods and ours on the Dress Code dataset. Best viewed when zoomed in.

Methods	Upper				Lower				Dresses			
	Paired		Unpaired		Paired		Unpaired		Paired		Unpaired	
	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
GP-VTON	0.271	0.775	21.58	0.99	0.283	0.759	23.61	1.13	0.221	0.789	17.73	0.70
PAINT-BY-EXAMPLE	0.165	0.858	38.78	2.70	0.197	0.818	23.56	0.99	0.346	0.700	21.51	0.87
LADI-VTON	0.049	0.928	13.26	0.27	**0.051**	0.922	14.80	0.31	0.089	0.868	13.40	0.25
FLDM-VTON (**ours**)	**0.045**	**0.930**	**11.45**	**0.16**	**0.051**	**0.924**	**13.27**	**0.24**	**0.079**	**0.891**	**12.61**	**0.19**

Table 2: Comparison results on the Dress Code dataset. The best results are highlighted in **bold**.

Figure 6: Qualitative results on real-world data.

Figure 7: Results of different source and target clothing types.

model collapse problem. Among diffusion-based methods, PAINT-BY-EXAMPLE and LADI-VTON generate more realistic images. However, the stochastic nature of conventional diffusion models poses a challenge in faithfully preserving crucial clothing details, often leading to substantial distortions. DCI-VTON faces challenges in correctly adapting the clothes' style to the person, though generating relatively faithful images. In contrast, our FLDM-VTON is able to generate more realistic images while effectively handling complex style, pattern, and text.

Application to real-world data. Figure 6 also presents qualitative results on real-world data from the Amazon website, illustrating the robustness of our FLDM-VTON. Our method effectively generates corresponding try-on images for different persons with diverse flat clothes, ensuring photo-realistic quality with faithful clothing details.

More challenging try-on situations. We further examine our FLDM-VTON on challenging try-on situations, such as different types of source and target clothes, by simply *masking out the region related to the source and target clothes*. The challenging try-on results in Figure 7 suggest the strong effectiveness of our FLDM-VTON.

4.3 Ablation Study

Here, we conduct detailed ablation studies to show the effectiveness of different conditions and proposed components for

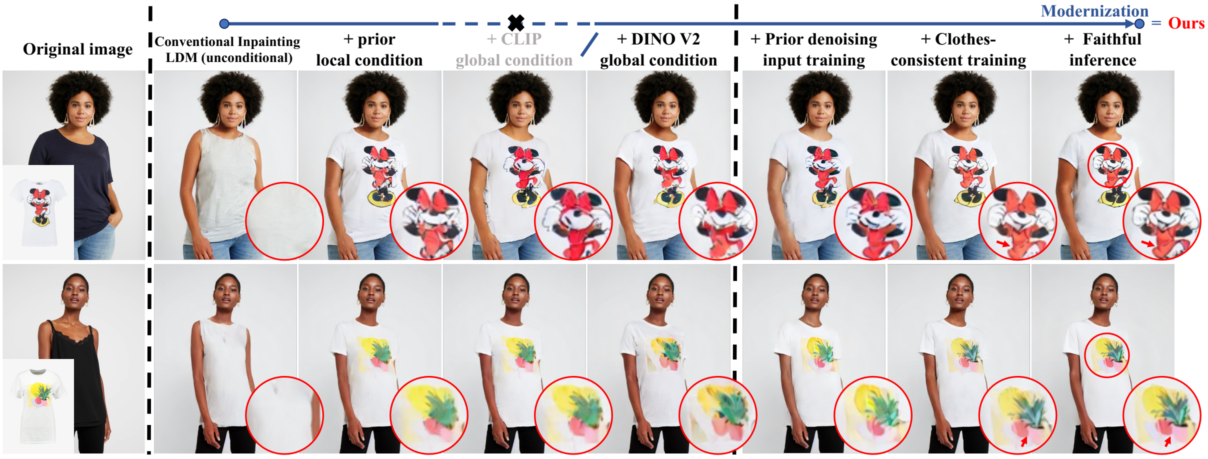

Figure 8: Ablation qualitative results for modernizing our FLDM-VTON on the VITON-HD dataset. Best viewed when zoomed in.

Methods	Paired		Unpaired	
	LPIPS↓	SSIM↑	FID↓	KID↓
Conventional inpainting LDM	0.158	0.788	34.58	3.41
+ Prior local condition (Sec. 3.1)	0.127	0.864	13.43	0.34
+ CLIP global condition	0.121	0.870	13.02	0.32
+ DINO V2 global condition	0.114	0.874	12.24	0.29
+ Prior denoising input training (Sec. 3.1)	0.094	0.876	10.60	0.27
+ Clothes-consistent supervision (Sec. 3.2)	0.082	0.883	8.90	0.15
+ Faithful inference (Sec. 3.3)	**0.080**	**0.886**	**8.81**	**0.13**

Table 3: Ablation quantitative results on the VITON-HD dataset. The best results are highlighted in **bold**.

modernizing our FLDM-VTON. Figure 8 and Table 3 present ablation qualitative and quantitative results, respectively.

Ablation on different conditions. We take the conventional unconditional inpainting LDM as our initial baseline. First, we enhance it with the prior local condition $\widehat{\boldsymbol{T}}^{w}$. We observe that the generated try-on image can capture most coarse-grained clothing details due to the prior information provided by warped clothes. However, the style of complex patterns is significantly distorted. Second, by integrating the CLIP or DINO V2 global conditions, there is a notable improvement in the style of these complex patterns. Notably, DINO V2 outperforms CLIP, which can be attributed to its more discriminative global feature extraction.

Ablation on proposed components. Here, we take the inpainting LDM with the prior local condition and DINO V2 global condition as the initial baseline. We first investigate the impact of the additional training for the prior denoising input. By including this training process, we observe that the style of complex patterns is largely preserved. However, there still remains a challenge in preserving fine clothing details.

Second, we show the effect of clothes-consistent supervision in Figure 9, including the real clothes-parsed image $\boldsymbol{T}^{C}$, estimated flat clothes image $\widehat{\boldsymbol{C}}$, and real flat clothes image $\boldsymbol{C}$. We find that our clothes flattening network effectively transforms the clothes from the try-on state to the flat state. By comparing the estimated and the original flat clothes, we observe that the original flat clothes have more fine clothing details, which further validates the importance of providing clothes-consistent supervision to guide the try-on diffusion process towards higher faithfulness to the original flat clothes. In summary, both the quantitative and qualitative results demonstrate that introducing the clothes flattening network into the training improves the capability of the try-on diffusion UNet to preserve more faithful clothing details.

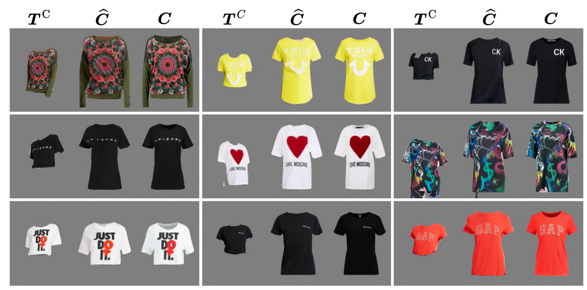

Figure 9: Clothes-consistent results. $\boldsymbol{T}^{C}$, $\widehat{\boldsymbol{C}}$, and $\boldsymbol{C}$ denote the real clothes-parsed image, estimated flat clothes image, and real flat clothes image, respectively. Best viewed when zoomed in.

In addition, we observe that while the faithful inference contributes not as much as other components to the quantitative results in Table 3, it significantly enhances the faithfulness of generated clothing details, as shown in Figure 8, without introducing additional computational burdens.

5 Conclusion

In this paper, we proposed a novel faithful latent diffusion model for virtual try-on. With the introduced faithful clothes priors and clothes-consistent faithful supervision, the proposed FLDM-VTON can significantly alleviate the unfaithful generation issue caused by the diffusion stochastic nature and latent supervision in LDM. In addition, the devised clothes-posterior sampling for faithful inference can further improve the model performance. Extensive experimental results on two popular VTON benchmarks validate the superior performance of our FLDM-VTON—generating photo-realistic try-on images with faithful clothing details.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 62101136).

References

[Bau *et al.*, 2019] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a GAN cannot generate. In *ICCV*, pages 4502–4511, 2019.

[Bińkowski *et al.*, 2018] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *ICLR*, 2018.

[Chen *et al.*, 2023] Tao Chen, Chenhui Wang, and Hongming Shan. BerDiff: Conditional Bernoulli diffusion model for medical image segmentation. In *MICCAI*, 2023.

[Choi *et al.*, 2021] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, pages 14131–14140, 2021.

[Fele *et al.*, 2022] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-VTON: Context-driven image-based virtual try-on network. In *WACV*, pages 3144–3153, 2022.

[Ge *et al.*, 2021] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, pages 8485–8493, 2021.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014.

[Gou *et al.*, 2023] 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 *ACM MM*, pages 7599–7607, 2023.

[Han *et al.*, 2019] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. ClothFlow: A flow-based model for clothed person generation. In *ICCV*, pages 10471–10480, 2019.

[He *et al.*, 2022] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, pages 3470–3479, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. volume 30, 2017.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[Huang *et al.*, 2023] Zhizhong Huang, Siteng Ma, Junping Zhang, and Hongming Shan. Adaptive nonlinear latent transformation for conditional face editing. In *ICCV*, pages 21022–21031, 2023.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[Lee *et al.*, 2022] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, pages 204–219. Springer, 2022.

[Li *et al.*, 2021] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details. In *CVPR*, pages 15546–15555, 2021.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017.

[Lu *et al.*, 2022] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, volume 35, pages 5775–5787, 2022.

[Minar *et al.*, 2020] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. CP-VTON+: Clothing shape and texture preserving image-based virtual try-on. In *CVPRW*, volume 3, pages 10–14, 2020.

[Morelli *et al.*, 2022] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-resolution multi-category virtual try-on. In *CVPR*, pages 2231–2235, 2022.

[Morelli *et al.*, 2023] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LADI-VTON: Latent diffusion textual-inversion enhanced virtual try-on. In *ACM MM*, 2023.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, pages 8162–8171, 2021.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv:2307.01952*, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[Ruiz *et al.*, 2023] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023.

[Si *et al.*, 2023] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion U-Net. *arXiv:2309.11497*, 2023.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004.

[Wang *et al.*, 2018] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, pages 589–604, 2018.

[Wang *et al.*, 2024] Chenhui Wang, Sirong Piao, Zhizhong Huang, Qi Gao, Junping Zhang, Yuxin Li, and Hongming Shan. Joint learning framework of cross-modal synthesis and diagnosis for Alzheimer's disease by mining underlying shared modality information. *Med. Image Anal.*, 91:103032, 2024.

[Wei *et al.*, 2024] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. DreamVideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024.

[Xie *et al.*, 2023] 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 *CVPR*, pages 23550–23559, 2023.

[Yang *et al.*, 2023] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by Example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023.

[Zhang *et al.*, 2018] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *ICCV*, pages 586–595, 2018.

[Zhu *et al.*, 2023] 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 *CVPR*, pages 4606–4615, 2023.