International Journal of Computer Vision (2025) 133:3781–3802
https://doi.org/10.1007/s11263-025-02352-3

TryOn-Adapter: Efficient Fine-Grained Clothing Identity Adaptation for High-Fidelity Virtual Try-On

Jiazheng Xing[1] · **Chao Xu**[1,2] · **Yijie Qian**[1] · **Yang Liu**[2] · **Guang Dai**[3] · **Baigui Sun**[2] · **Yong Liu**[1] · **Jingdong Wang**[4]

Received: 30 March 2024 / Accepted: 6 January 2025 / Published online: 31 January 2025

Abstract

Virtual try-on focuses on adjusting the given clothes to fit a specific person seamlessly while avoiding any distortion of the patterns and textures of the garment. However, the clothing identity uncontrollability and training inefficiency of existing diffusion-based methods, which struggle to maintain the identity even with full parameter training, are significant limitations that hinder the widespread applications. In this work, we propose an effective and efficient framework, termed TryOn-Adapter. Specifically, we first decouple clothing identity into fine-grained factors: style for color and category information, texture for high-frequency details, and structure for smooth spatial adaptive transformation. Our approach utilizes a pre-trained exemplar-based diffusion model as the fundamental network, whose parameters are frozen except for the attention layers. We then customize three lightweight modules (Style Preserving, Texture Highlighting, and Structure Adapting) incorporated with fine-tuning techniques to enable precise and efficient identity control. Meanwhile, we introduce the training-free T-RePaint strategy to further enhance clothing identity preservation while maintaining the realistic try-on effect during the inference. Our experiments demonstrate that our approach achieves state-of-the-art performance on two widely-used benchmarks. Additionally, compared with recent full-tuning diffusion-based methods, we only use about half of their tunable parameters during training. The code will be made publicly available at https://github.com/jiazheng-xing/TryOn-Adapter.

Keywords Virtual Try-On · Large-scale generative models · Diffusion model · Identity preservation

1 Introduction

Virtual try-on (Han et al., 2018; Wang et al., 2018; Choi et al., 2021; Morelli et al., 2022) aims to enable users to naturally try on new category clothes in the target regions by giving an image of the garment and an image of the person while preserving the non-target regions. The core of this task lies in maintaining the pattern and texture of the clothes, termed *clothing identity*, unchanged in various conditions. Considering the scarcity of high-quality paired datasets, the current works usually follow a two-stage design (Choi et al., 2021; Morelli et al., 2022; Lee et al., 2022; Xie et al., 2023; Gou

Communicated by Shengfeng He.

Jiazheng Xing and Chao Xu have contributed equally to this work.

✉ Yong Liu
yongliu@iipc.zju.edu.cn

Jiazheng Xing
jiazhengxing@zju.edu.cn

Chao Xu
xc264362@alibaba-inc.com

Yijie Qian
yijieqian@zju.edu.cn

Yang Liu
ly261666@alibaba-inc.com

Guang Dai
guang.gdai@gmail.com

Baigui Sun
baigui.sbg@alibaba-inc.com

Jingdong Wang
wangjingdong@baidu.com

[1] Laboratory of Advanced Perception on Robotics and Intelligent Learning, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027 Zhejiang, China

[2] Alibaba Group, Hangzhou, China

[3] SGIT AI Lab, State Grid Shaanxi Electric Power Company, Xian, China

[4] Baidu Inc., Beijing, China

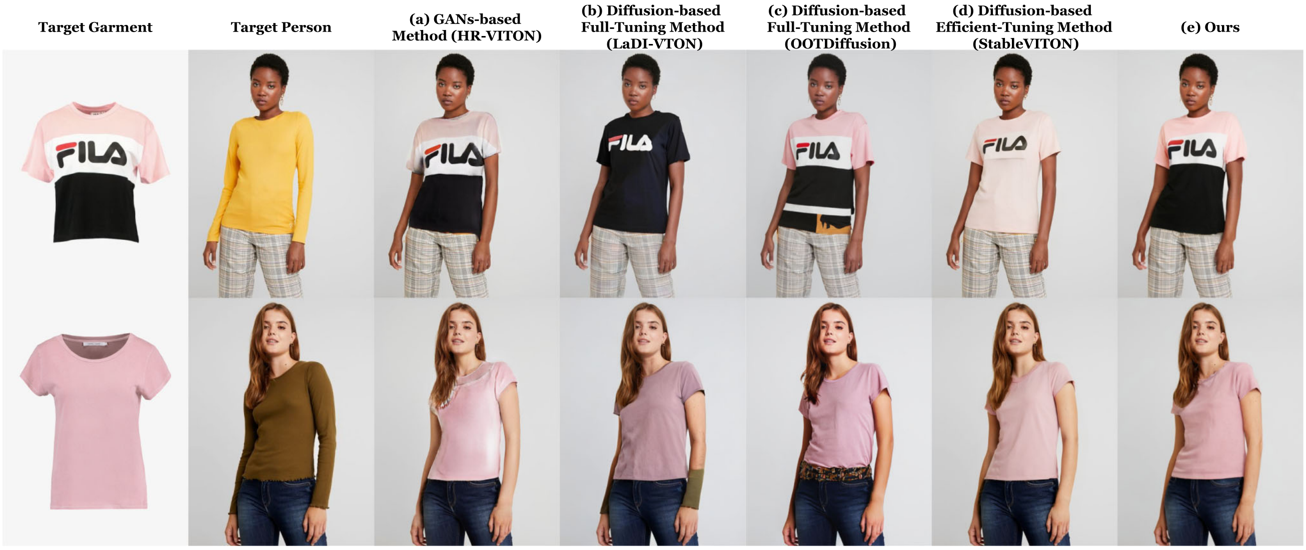
Target Garment
Target Person
(a) GANs-based Method (HR-VITON)
(b) Diffusion-based Full-Tuning Method (LaDI-VTON)
(c) Diffusion-based Full-Tuning Method (OOTDiffusion)
(d) Diffusion-based Efficient-Tuning Method (StableVITON)
(e) Ours

Fig. 1 Performance comparison of four different methods on VITON-HD dataset at 512 × 384 resolution, including our TryOn-Adapter, GANs-based method HR-VITON (Rombach et al., 2022), Diffusion-based method LaDI-VTON (Morelli et al., 2023), StableVITON (Kim et al., 2023) and OOTDiffusion (Xu et al., 2024). Our method generates high-quality results and exhibits strong clothing identity preservation capability, i.e., consistent color style and logo textures, as well as a smooth transition between long and short sleeves

et al., 2023): target garment deformation and composite generation. The former (Ge et al., 2021b; Han et al., 2019; He et al., 2022) focuses on transferring the original clothing into the desired form based on the posture and body shape of the given person. Despite providing a prior warped template, direct blending produces severe artifacts when encountering occlusion and large shape differences. Therefore, the latter is introduced for further refinement with a powerful generative model. Concretely, most of the previous works (Ge et al., 2021b; Bai et al., 2022; Lee et al., 2022; Morelli et al., 2022) have relied on the GANs (Goodfellow et al., 2014), but they suffer from unstable training (Gulrajani et al., 2017) and mode collapse (Miyato et al., 2018), leading to the detail loss in their generated results, especially for the highly patterned garments, as shown in Fig. 1 column 3. More recently, diffusion models (Song et al., 2020; Ho et al., 2020; Rombach et al., 2022) have attracted widespread attention and permeated into the virtual try-on. Two diffusion-based models have regarded try-on as an inpainting task. DCI-VTON (Gou et al., 2023) is built upon the exemplar-based image generation method, harnessing its ability to preserve irrelevant areas while focusing on fusing the warped garment into the target area. LaDI-VTON (Morelli et al., 2023) further employs the textual-inversion technique to refine the target areas. Another approach, OOTDiffusion (Xu et al., 2024), introduces an outfitting UNet to implicitly transform the given garment. However, they cannot achieve satisfactory results due to insufficient exploration of identity-preserving modules, even tuning all parameters of UNet for adaptive learning. As shown in Fig. 1 column 4 (LaDI-VTON), the color and textures of their generated clothes are completely different from the target clothes (row 1), and the transition from long sleeves to short sleeves exhibits obvious artifacts (row 2). The same color and texture degradation exhibited in column 5 (OOTDiffusion).

Although diffusion-based garment composition generations have progressed, they lack in-depth thinking in two key aspects. (1) *Identity controllability*. Previous methods (Yang et al., 2023; Gou et al., 2023) utilize the class token of CLIP embeddings obtained from the reference garment image. However, this global vectorized feature, when directly integrated into the UNet, fails to retain identity cues. By contrast, this work decouples the garment characteristics into three fine-grained factors to simplify identity preservation, i.e., *style* (color and category information), *texture* (high-frequency details such as patterns, logo, and text), and *structure* (smooth transition when under different pose or body shape, as well as a significant difference between the original and target clothing, such as the aforementioned long and short sleeves issues). (2) *Training efficiency*. Diffusion-based methods usually suffer from low training efficiency, especially in a fully fine-tuned manner. To tackle this problem, Parameter-Efficient Fine-Tuning techniques (PEFT), such as ControlNet (Zhang et al., 2023), T2I-Adapter (Mou et al., 2023), and GLIGEN (Li et al., 2023b), employing a small number of training parameters to control the denoising process. It is worthwhile to consider how to introduce efficient training modules or even training-free mechanisms into the try-on task without sacrificing performance. Notably, concurrent work StableVITON (Kim et al., 2023) circumvents full-tuning, yet their tenuous representation of clothing identity (only a single image) results in continued difficulty in producing satisfactory outputs. As shown in Fig. 1 column 6, they exhibit inconsistent color style and logo textures.

To realize more identity-controllable and training-efficient virtual try-on, we propose a novel paradigm, termed **TryOn-Adapter**, which follows the Paint-by-Example framework and customizes three lightweight components according to the decoupled factors to effectively control hierarchical identity cues. Specifically, we start by freezing all the parameters in the UNet blocks except for the attention layers, which transfer the universal pre-trained model to the specific try-on task with minimum trainable parameters. For style preservation, we utilize both patch and class tokens to learn comprehensive style representation, with the former compensating for the lack of detailed identity in the latter. Furthermore, due to the limitation of CLIP in capturing the complex color style, we further enhance the patch tokens with visual features embedded in the VAE encoder through an adaptive transfer module. To avoid disturbing the feature distribution of the pre-trained model, inspired by GLIGEN (Li et al., 2023b), we insert trainable gated self-attention layers in all layers to inject the updated patch tokens into the frozen backbone. Moreover, to preserve the texture, a post-processed high-frequency feature map is incorporated as a texture refinement guidance to highlight the local details. For another factor involving spatial cues, we take the segmentation map, obtained by a rule-based training-free extraction method, as the structure condition to explicitly rearrange the target areas of the body and clothing to conform to the warped cloth. We follow the T2I-Adapter (Mou et al., 2023) to inject the above two conditions into UNet by two lightweight networks incorporated a well-designed position attention module that helps amplify the spatial cues. During the inference phase, we introduce a time-partially function on the training-free technique RePaint (Lugmayr et al., 2022; Avrahami et al., 2023), termed T-RePaint, to further enhance the clothing identity without compromising the overall image fidelity. Additionally, a learnable latent blending module is integrated within the autoencoder to produce more visually consistent results.

In this way, we preserve the hierarchical identity details of the given garment without full fine-tuning, as illustrated in Fig. 1 column 7.

In summary, we present the following contributions.

- We propose a novel, effective, and efficient framework for virtual try-on, **TryOn-Adapter**, to maintain the identity of the given garment with low consumption.
- We decouple clothing identity into fine-grained factors: style, texture, and structure, represented by the global class token and enhanced patch token embeddings, high-frequency feature map, and segmentation maps, respectively. Each factor incorporates a tailored lightweight module and injection mechanism to achieve precise and efficient identity control. Meanwhile, we introduce a training-free technique, T-RePaint, to further reinforce the clothing identity preservation while maintaining the realistic try-on effect during the inference.
- Extensive experiments on two widely used datasets have shown that our method can achieve outstanding performance with minor trainable parameters.

2 Related Work

2.1 Image-Based Virtual Try-On

To avoid distortion of garment image textures and confusion of the identities as much as possible, the image-level virtual try-on (Yang et al., 2020; Issenhuth et al., 2020; Han et al., 2018; Wang et al., 2018; Choi et al., 2021; Morelli et al., 2022; Chen et al., 2023a; Li et al., 2023b) task is typically divided into two stages: the target garment deformation stage and the composite generation stage. For the first stage, the Thin Plain Spine (TPS) method was commonly employed to deform clothing in previous works (Han et al., 2018; Ge et al., 2021a; Minar et al., 2020; Zheng et al., 2019; Yang et al., 2020), which is limited to offering only basic deformation processing. Furthermore, many flow-based works (Ge et al., 2021b; Han et al., 2019; He et al., 2022; Bai et al., 2022) have been proposed, aiming to build the appearance flow field between clothing and corresponding regions of the human body to better deform the clothing for a more natural fit to the body. In our work, we adopt the flow-based method PF-AFN (Ge et al., 2021b) to accomplish the rough deformation of the clothing in the first stage.

The second stage can be classified into two categories: the GANs-based methods and the Diffusion-based methods. GANs-based methods (Ge et al., 2021b; Bai et al., 2022; Lee et al., 2022; Morelli et al., 2022; Lewis et al., 2021) inherit the weaknesses of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), such as unstable training (Gulrajani et al., 2017) and mode drop in the output distribution (Miyato et al., 2018), leading to the problem of detail loss in their generated results. Specifically, FashionGAN (Zhu et al., 2017) generates the image conditioned on textual descriptions and semantic layouts, TryOnGAN (Lewis et al., 2021) trains a pose-conditioned StyleGAN2 (Karras et al., 2020) on unpaired fashion images, and so on. Unlike the former, Diffusion-based methods (Yang et al., 2023; Gou et al., 2023; Morelli et al., 2023; Baldrati et al., 2023; Li et al., 2023a) with a more stable training procedure can provide superior image generation quality. Specifically, MGD (Baldrati et al., 2023) is the first latent diffusion model defined for humancentric fashion image editing, conditioned by multimodal inputs such as text, body pose, and sketches. DCI-VTON (Gou et al., 2023) treats the virtual try-on task as an inpainting task and adds the warped clothes to the input of the diffusion model as the local condition. LaDI-VTON (Morelli et al., 2023) follows the similar paradigm and further exploits the textual inversion technique for the first time in this task. Another research avenue focuses on exploring how to replace the warping network with a reference UNet. OOTDiffusion (Xu et al., 2024) introduces an outfitting UNet to learn garment details in the initial step, subsequently integrating these details into the denoising UNet through outfitting fusion. OutfitAnyone (Sun et al., 2024) leverages a two-stream conditional diffusion model to adeptly handle garment deformation. However, despite incorporating complex network structures, they lack in-depth thinking on the fine-grained identity, thus rendering the control of clothing details challenging. Besides, all of them require extensive full-parameter training, which leads to the issues of high resource consumption. StableVITON (Kim et al., 2023), a more recent work, partially trains the proposed zero cross-attention blocks and SD encoder, but they still directly use the given garment image to provide clothing cues, which makes it difficult for the network to capture details. By contrast, our method decouples the complex clothing into fine-grained features and tailors them with carefully chosen fine-tuning techniques to significantly enhance the preservation of the given garment without incurring excessive training consumption.

2.2 Diffusion Models

Recently, the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020; Sohl-Dickstein et al., 2015) has emerged as a critical technology in image synthesis, renowned for its ability to generate high-fidelity images from a normal distribution by reversing the noise addition process. In response to the computational complexity and resource requirements of DDPM, the Latent Diffusion Model (Rombach et al., 2022) (LDM) efficiently performs diffusion and denoising in the latent space through its optimized encoder-decoder architecture, streamlining the generation process. Based on the unique advantages shown by the Diffusion model in preserving image details, many works in text2image generation (Gal et al., 2022; Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Wei et al., 2023; Ramesh et al., 2022; Li et al., 2023b), image editing (Mou et al., 2023; Zhang et al., 2023; Saharia et al., 2022; Nichol et al., 2021), and subject-driven generation (Chen et al., 2023b; Bhunia et al., 2023; Yang et al., 2023; Shi et al., 2023; Wang et al., 2022) have recently emerged. The success of these previous works have provided ample inspiration for image-based virtual Try-On.

3 Method

3.1 Architecture Overview

In this paper, we propose a novel **TryOn-Adapter** to preserve the identity of the given garment while requiring relatively minimal training resources. The non-rigid warping preservation in virtual try-on is a challenging task. Previous diffusion-based methods (Morelli et al., 2023; Gou et al., 2023; Baldrati et al., 2023; Kim et al., 2023; Li et al., 2023a) do not decompose clothing identity adequately, resulting in unsatisfactory results, so we tackle this problem by dividing it into three factors, i.e., style, texture, and structure, and each factor is equipped with a special lightweight design: The Style Preserving module (Sect. 3.2) aims to preserve the overall style of the garment. The Texture Highlighting module (Sect. 3.3) focuses on refining high-frequency details. The Structure Adapting module (Sect. 3.3) compensates for unnatural areas caused by clothing changes. The T-RePaint (Sect. 3.4) further reinforces the clothing identity preserving without compromising the overall image fidelity during the inference.

Specifically, as illustrated in Fig. 2, when given an image $I_p \in \mathbb{R}^{3\times H\times W}$ of a person and an image $I_c \in \mathbb{R}^{3\times H'\times W'}$ of a target garment, our method aims to generate an image $\hat{I} \in \mathbb{R}^{3\times H\times W}$, where the person in I_p is depicted wearing the garment from I_c. For input preprocessing, we first remove the original clothing from I_p using a provided garment mask $m \in \{0, 1\}^{1\times H\times W}$, resulting in a clothing-agnostic RGB image $I_a \in \mathbb{R}^{3\times H\times W}$, which retains non-target regions such as head and background. To obtain the warped garment image $I_c^w \in \mathbb{R}^{3\times H'\times W'}$ and its mask $I_m^w \in \{0, 1\}^{1\times H'\times W'}$, we employ the garment warping model which here we choose GP-VTON (Xie et al., 2023) to warp the original garment image I_c into a coarse try-on shape based on the person's pose and other mask information in I_p. For the denoising process, the UNet model takes the pixel-wise addition of coarse warped garment image I_c^w and the clothing-agnostic image I_a, along with the noisy person I_t and mask m, as inputs. Besides, the target garment I_c, high-frequency map I_{HF} extract from I_c^w, and human segmentation map I_{seg} respectively serve as representations of style, texture, and structure, enabling fine-grained identity control. Note that our TryOn-Adapter is trained under a self-reconstruction manner, where I_c is the exact garment worn by I_p, and I_{seg} is also obtained from I_p. During the inference, I_c and the clothing on I_p are different, and I_{seg} is generated by the proposed precise yet user-friendly method. For the latent space reconstruction process, the independent Enhanced Latent Blending Module is inserted into the autoencoder to further maintain consistent visual quality (Sect. 3.4).

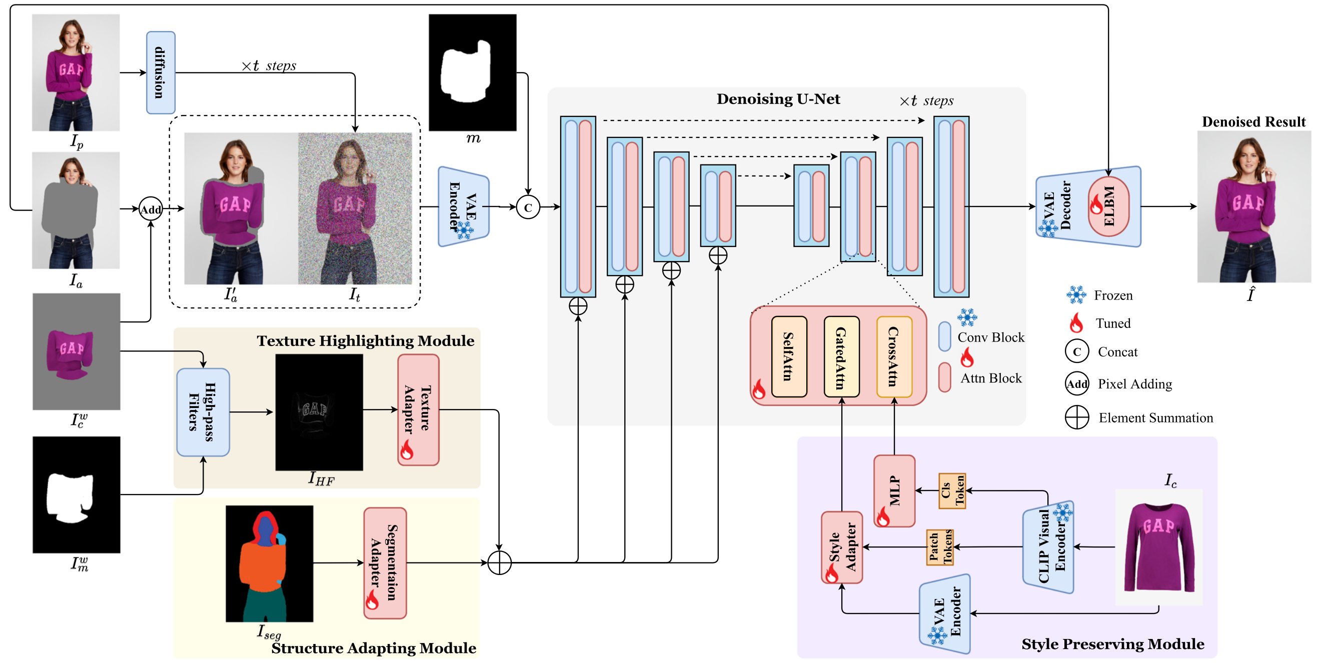

Fig. 2 The overall architecture of our **TryOn-Adapter** is composed of five parts: (1) the pre-trained stable diffusion model with fixed parameters except for attention layers; (2) the Style Preserving module aimed to preserve the overall style of the garment, including color and category information; (3) the Texture Highlighting module focuses on refining the high-frequency details. (4) the Structure Adapting module compensates for unnatural areas caused by clothing changes. (5) the Enhanced Latent Blending Module focuses on consistent visual quality

3.2 Style Preserving Module

For this module, we extract as much style information as possible from the reference image to inject into the denoising U-Net to control the overall style of the generated garment, including color and category information. First, we input garment image I_c through the frozen CLIP visual encoder to get the class token feature $\mathbf{T}_{cls} \in \mathbb{R}^{1\times 1024}$ and patch tokens features $\mathbf{T}_{patch} \in \mathbb{R}^{256\times 1024}$. The former is used as a coarse condition and added several additional fully connected layers to decode this feature given by:

$$\mathbf{h}_{cls} = \mathrm{MLPs}(\mathbf{T}_{cls}), \tag{1}$$

where $\mathbf{h}_{cls} \in \mathbb{R}^{1\times 1024}$. Besides, unlike common image editing tasks, virtual try-on needs to guarantee the identity of the reference image fully, so we further introduce the latter to supplement fined style cues. However, although these two features embody sufficient style cues, they are not sensitive to color information. To enhance the color perception of patch tokens and guide the alignment of CLIP features with the output domain of the diffusion model, we design a style adapter to fuse CLIP patch embeddings $\mathbf{T}_{patch}$ and VAE visual embeddings $\mathbf{F}_{vae} \in \mathbb{R}^{4\times 28\times 28}$, as shown in Fig 3. Formally:

$$\mathbf{F}_{patch} = \mathrm{MHA}(\mathbf{T}_{patch}, \mathbf{F}'_{vae}, \mathbf{F}'_{vae}) + \mathbf{T}_{patch}, \tag{2}$$

$$\mathbf{h}_{patch} = \mathrm{FFN}(\mathbf{F}_{patch}) + \mathbf{F}_{patch}, \tag{3}$$

where the $\mathbf{F}_{vae}$ is obtained by the reference image I_c through the pre-trained Stable Diffusion VAE encoder and $\mathbf{F}'_{vae} \in \mathbb{R}^{784\times 1024}$ is obtained by $\mathbf{F}_{vae}$ through a series of flatten and mapping operations. Moreover, MHA and FFN indicate multi-head attention and feed-forward network. To inject the coarse feature $\mathbf{h}_{cls}$ and fined feature $\mathbf{h}_{patch}$ into UNet, we do not merge them and replace the text tokens in the original stable diffusion model, as it was considered a naive solution that impedes the network from understanding the content in the reference image and the connection to the source image, as mentioned in Paint-by-Example (Yang et al., 2023). Therefore, followed by GLIGEN (Li et al., 2023b), $\mathbf{h}_{cls}$ and $\mathbf{h}_{patch}$ are fed into the diffusion process through cross-attention and gated self-attention, respectively. We denote $\mathbf{v} = [v_1, \ldots, v_M]$ as the visual feature tokens of an image. Therefore, the attention block of our denoising U-net consists of three attention layers, as shown in Fig. 2, which can be written as:

$$\mathbf{v} = \mathbf{v} + \mathrm{SelfAttn}(\mathbf{v}), \tag{4}$$

$$\mathbf{v} = \mathbf{v} + \beta \cdot \tanh(\gamma) \cdot \mathrm{SelfAttn}([\mathbf{v}, \mathbf{h}_{patch}]), \tag{5}$$

$$\mathbf{v} = \mathbf{v} + \mathrm{CrossAttn}(\mathbf{v}, \mathbf{h}_{cls}), \tag{6}$$

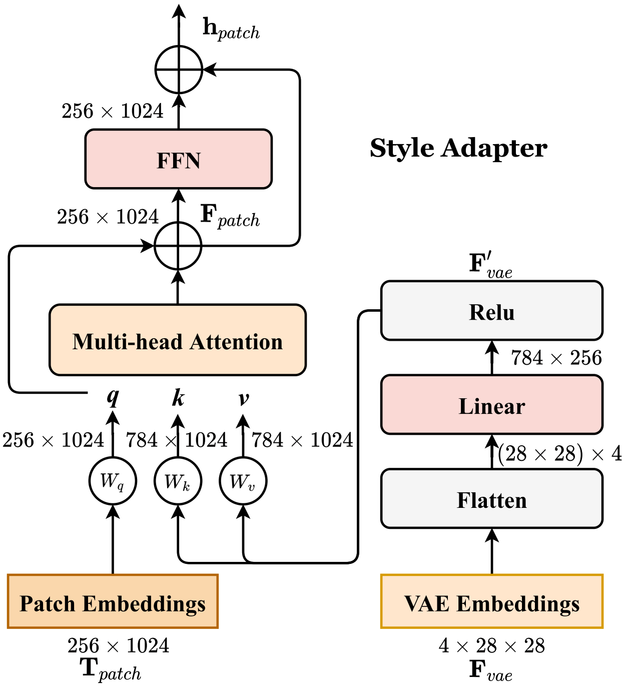

Fig. 3 The architecture of the style adapter

where γ is a learnable scalar initialized as 0, and β is a constant to balance the importance of the adapter layer. Through gated self-attention, we have effectively introduced the guidance of style features to the generation process while avoiding the destruction of pre-trained weights.

3.3 Texture Highlighting Module and Structure Adapting Module

After the Style Preserving module, the discriminative style features have been combined into the UNet. However, they struggle to preserve complex textures, such as patterns and logos, and exhibit obvious artifacts during the clothing transition when the original and target clothing are significantly difference, as well as in some cases involving challenging poses and body shapes. To address these issues, it is crucial to integrate explicit spatial conditions and maintain consistency between these guidance features and the generated image features to achieve perfect preservation. Therefore, inspired by T2I-Adapter (Mou et al., 2023), we introduce two lightweight designs, the Texture Highlighting module and the Structure Adapting module. As shown in Fig. 2, they incorporate the high-frequency texture information for texture refinement and utilize the human body segmentation map for unnatural transition areas correction. Both conditions are encoded into multi-scale features and injected into the intermediate features of the denoising U-Net for precise control.

For texture preservation, we extract the high-frequency map I_{HF} of the warped garment by the sobel operator that highlights the complex texture and patterns of the garment, especially the logo and text. Besides, we observe that the edge information occasionally provides incorrect guidance since I_c^w is just an offline rough result without adaptive refinement. To avoid introducing such ambiguous cues, we erode the edges of the clothes, given by:

$$I_{HF} = 0.5 \times \left(\left|I_c^w \otimes \mathbf{K}_s^x\right| + \left|I_c^w \otimes \mathbf{K}_s^y\right|\right) \odot \left(I_m^w \ominus \mathbf{K}_e\right), \tag{7}$$

where $\mathbf{K}_s^x$, $\mathbf{K}_s^y$, $\mathbf{K}_e$ denote the horizontal, vertical Sobel kernels and erosion kernel. $\otimes, \odot, \ominus$ refer to convolution product, Hadamard product, and erosion operation. The visual illustration for the texture highlighting map generation is as shown in Fig. 4a.

For structure guidance, we utilize the segmentation map I_{seg}, which provides human posture information and explicitly indicates the clothing and body areas, serving as the strong prior information for correcting the discordant areas that appear after the garment change, such as the transition between long and short sleeves. Unlike previous methods (Choi et al., 2021; Li et al., 2023b; Xie et al., 2023; Cui et al., 2023) that use networks to predict the target segmentation map, this work avoids redundant off-the-shell networks, proposing a rule-based training-free segmentation extraction method to achieve precise results yet user-friendly process. The core idea of this design is to combine the existing cloth-agnostic segmentation map I_{seg}^{ca}, warped cloth Mask I_m^w, and the human body densepose (Güler et al., 2018) map I_{dp} to obtain the decomposed segmentation map. Specifically, we first form a preliminary composed image I_{caw} by performing a per-pixel OR ($\vee$) operation to merge the I_{seg}^{ca} with the I_m^w in the binary logical space, i.e., $I_{caw} = I_{seg}^{ca} \vee I_m^w$. Next, we combine I_{caw} with the densepose map I_{dp} to complete the missing arm parts. To remove the overlapping parts between I_{caw} and I_{dp} and the noise in I_{dp} itself, we use the per-pixel AND ($\wedge$) operation and connectivity-based filtering (Filter_l) to obtain the modified human pose map I'_{dp}. This process removes noise and irrelevant details by excluding connected components with pixel counts below the threshold l (here set to 12), given by:

$$\text{Filter}_l(\cdot) = \{p \in I | size(C_p) \geq l\}, \tag{8}$$

$$I'_{dp} = \text{Filter}_l(I_{dp} - (I_{dp} \wedge I_{caw})), \tag{9}$$

where I represents the image to be filtered, p represents a pixel in the image, C_p represents the connected region adjacent to pixel p, and $size(C_p)$ represents the number of pixels in the connected component. Finally, the I'_{dp} is merged with I_{caw} through the per-pixel OR ($\vee$) operation to obtain the recomposed segmentation map I_{seg}, i.e., $I_{seg} = I_{caw} \vee I'_{dp}$.

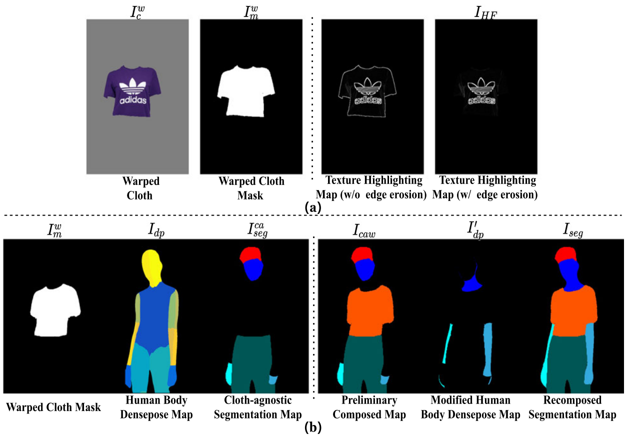

Fig. 4 **a** Visual illustration for the texture highlighting map generation. **b** Visual illustration for the target segmentation map generation

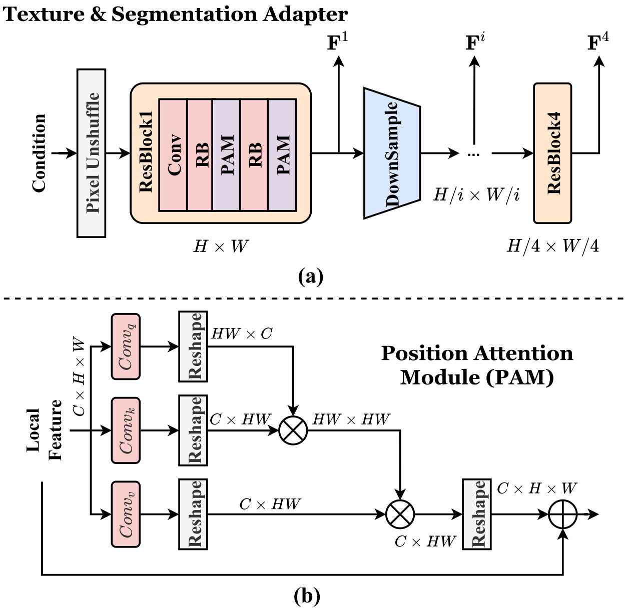

Fig. 5 **a** The architecture of the texture and segmentation adapter. Every ResBlock consists of a convolution layer, two resnet layers, and two position attention modules. **b** The architecture of the position attention module

The visual illustration for the target segmentation map generation is shown in Fig. 4b.

In practice, we follow the network design in T2I-Adapter (Mou et al., 2023), and add Position Attention Modules (PAM) inspired by DANet (Fu et al., 2019) to establish rich contextual relationships on local features in each resblock is shown in Fig. 5a. The architecture design of PAM is depicted in Fig. 5b, which enhances the representation of spatial information for the texture highlighting map and the recomposed segmentation map. Concretely, we introduce a Texture Adapter for the high-frequency map and a Segmentation Adapter for the segmentation map to obtain multi-scale conditional features $\mathbf{F}_{HF} = \{\mathbf{F}_{HF}^1, \mathbf{F}_{HF}^2, \mathbf{F}_{HF}^3, \mathbf{F}_{HF}^4\}$, $\mathbf{F}_{seg} = \{\mathbf{F}_{seg}^1, \mathbf{F}_{seg}^2, \mathbf{F}_{seg}^3, \mathbf{F}_{seg}^4\}$. These multi-scale features are correspond to the intermediate feature $\mathbf{F}_{enc} = \{\mathbf{F}_{enc}^1, \mathbf{F}_{enc}^2, \mathbf{F}_{enc}^3, \mathbf{F}_{enc}^4\}$ in the denoising UNet encoder. Both adapters have the same network structure, as shown in Fig. 5a. Finally, the conditional features $\mathbf{F}_{HF}$, $\mathbf{F}_{seg}$, and $\mathbf{F}_{enc}$ are weighted and added at each scale to update $\mathbf{F}_{enc}$, obtaining $\mathbf{F}'_{enc}$ with:

$$\mathbf{F}'_{enc} = \mathbf{F}_{enc} + \omega \cdot \mathbf{F}_{seg} + (1 - \omega) \cdot \mathbf{F}_{HF}, \tag{10}$$

where $\omega \in (0, 1)$ is a hyperparameter. The intermediate features of UNet are updated by injecting this explicit information, allowing it to focus on complex textural details and relationships of individual spatial parts.

3.4 Diffusion Model for Virtual Try-On

In this work, we implement our method based on a pre-trained diffusion model built upon Stable Diffusion (Rombach et al., 2022), i.e., Paint-by-Example (Yang et al., 2023), and added the identity preserving modules into this model to control the generation. The diffusion model includes two parts: an autoencoder (VAE), which can compress input images into latent space and reconstruct them, and a U-Net to perform denoising in the latent space directly. As shown in Fig. 2, for the first part, we embed the ground-truth image I_p and inpainting image I'_a through the pre-trained encoder of VAE into the latent space, obtaining $\mathbf{z}_0$ and $\mathbf{z}'_a$. The forward process is executed at $\mathbf{z}_0$ at a given timestamp t, with:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \tag{11}$$

where $\mathbf{z}_t$ indicates the noisy feature map at step t, α_t decreases with the timestep t, and $\epsilon \in \mathcal{N}(0, \mathbf{I})$ is the Gaussian noise. For the generative process, we concatenate $\mathbf{z}_t$, $\mathbf{z}'_a$, and the resized mask m as the U-Net's input $\mathbf{z}'_t = [\mathbf{z}_t, \mathbf{z}'_a, m]$. The style features $\mathbf{h}_c = [\mathbf{h}_{cls}, \mathbf{h}_{patch}]$, texture condition $\mathbf{F}_{HF}$, and structure guidance $\mathbf{F}_{seg}$ are also injected into the UNet. Finally, our TryOn-Adapter is optimized via the objective:

$$\mathbb{E}_{\mathbf{z},t,\mathbf{h}_c,\mathbf{F}_{HF},\mathbf{F}_{seg},\epsilon\in\mathcal{N}(0,\mathbf{I})}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{z}'_t, t, \mathbf{h}_c, \mathbf{F}_{HF}, \mathbf{F}_{seg}\right)\right\|_2^2\right], \tag{12}$$

where the θ denotes the all learnable parameters.

To further reinforce the clothing identity preservation, inspired by previous works (Lugmayr et al., 2022; Avrahami et al., 2023; Corneanu et al., 2024), we utilize a training-free technique (i.e., RePaint) in the latent space during the inference. RePaint is aimed at sampling known regions (i.e., unknown mappings) and replacing them at each denoising step in the inference process. Warped target garment images I_c^w contain crucial prior information for preserving

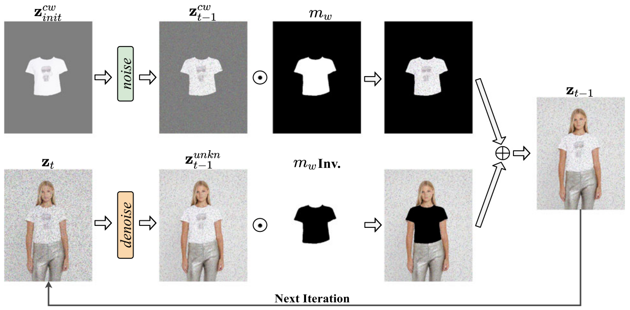

Fig. 6 Overview of our T-RePaint for $T' \leqslant t < T$

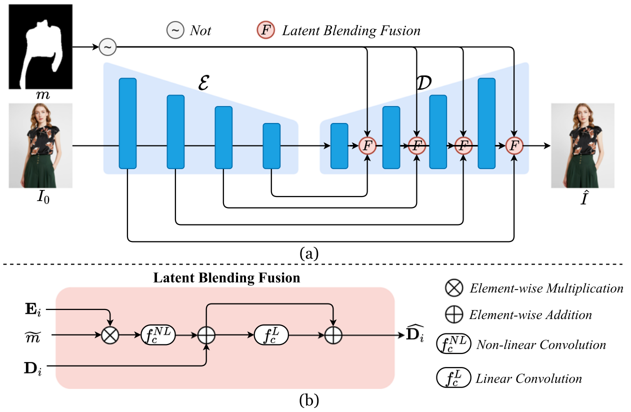

Fig. 7 **a**: Overview of our Enhanced Latent Blending Module. The autoencoder is frozen, and only the Latent Blending Fusion operation is learnable. **b** The architecture of Latent Blending Fusion operation

the identity, so applying RePaint to them further enhances the preservation effect. We observe that applying RePaint at all denoising steps results in noticeable noise at the RePaint edges and lacks realistic try-on effects in the final generated image. To address this problem, we propose a T-RePaint approach, applying RePaint only in the early denoising steps. Specifically, given a range of time steps $[1, T]$, the RePaint process starts from time step T and ends on step T' $(T' < T)$. We feed I_c^w to the VAE encoder to obtain the warped garment feature $\mathbf{z}^{cw}_{init}$, and the warped garment mask I_m^w is resized as m_w. We use $\mathbf{z}_T$ to denote a noise sampled from the Gaussian distribution, and $\mathbf{z}_0$ to denote the final image synthesis. Since the forward process is defined by Markov Chain at Eq. 11, we can sample the warped garment feature at any time step t to obtain the intermediate feature $\mathbf{z}^{cw}_{t-1}$, i.e., $\mathbf{z}^{cw}_{t-1} \sim nosie\left(\mathbf{z}^{cw}_{init}, t\right)$. Meanwhile, we use $\mathbf{c}$ to denote all conditions in the denoising process based on the diffusion model, so the unknown regions' denoising at step t can be defined as $\mathbf{z}^{unkn}_{t-1} \sim denosie\left(\mathbf{z}_t, \mathbf{c}, t\right)$. Thus, we achieve the reverse step with the composition of $\mathbf{z}^{cw}_{t-1}$ and $\mathbf{z}^{unkn}_{t-1}$ controlled by the content keeping mask m_w, given by:

$$\begin{cases} \mathbf{z}_{t-1} = m_w \odot \mathbf{z}^{cw}_{t-1} + (1 - m_w) \odot \mathbf{z}^{unkn}_{t-1} \;\; \left(T' \leqslant t < T\right) \\ \qquad\qquad \mathbf{z}_{t-1} = \mathbf{z}^{unkn}_{t-1} \;\; \left(1 \leqslant t < T'\right) \end{cases}. \tag{13}$$

Our T-RePaint is shown in Fig. 6 for $T' \leqslant t < T$.

As mentioned before, the VAE enables the denoising network to operate in a lower-dimensional latent space, thereby reducing the computational cost in the diffusion network. However, due to data loss deriving from the spatial compression performed by the autoencoder, the latent space might struggle to capture high-frequency details precisely, which can easily lead to distortion of faces or hands in the generated images. For the distortion problem, some previous methods (Gou et al., 2023; Li et al., 2023a) blend the background areas from the person image (e.g., face, hands) with the foreground areas (clothing) from the generated image at the pixel level, but bring about identifiable artifacts and blurred at the same time. By contrast, inspired by recent works (Morelli et al., 2023; Li et al., 2019; Zhu et al., 2023b; Avrahami et al., 2023), we propose the Enhanced Latend Blending Module (ELBM), which utilizes a background mask to directly copy the background region of the encoders' features from different layers and combines them with the corresponding ones of the decoder through some skip connections and learnable parameters. In this way, the VAE Decoder's difficulty in capturing high-frequency information is alleviated by blending enhanced background information into the decoding process. Specifically, we use I_0 to denote the original image and m to denote the background mask. Given the encoder $\mathcal{E}$, the decoder $\mathcal{D}$ and the input I_0, the i-th feature map comes from the encoder and the decoder can be represented as $\mathbf{E}_i$ and $\mathbf{D}_i$, respectively. The enhanced latent blending process is formulated as:

$$\widehat{\mathbf{D}_i} = \mathbf{D}_i + f_c^{NL}\left(\mathbf{E}_i\right) \otimes \widetilde{m}, \tag{14}$$

$$\widehat{\mathbf{D}_i} = \widehat{\mathbf{D}}_i + f_c^{L}\left(\widehat{\mathbf{D}}_i\right), \tag{15}$$

where $\otimes$ is element-wise multiplication, $\widetilde{m} = 1 - m$. f_c^{NL} and f_c^{L} represent learnable non-linear and linear convolution. Unlike LaDI-VTON's (Morelli et al., 2023) EMASC, we further integrate the output $\widehat{\mathbf{D}_i}$ of Eq. 14 with a linear convolution and residual connection, as shown in Eq. 15, to reduce the probability of a disconnected feeling at the foreground-background junction. The training process only employs a frozen autoencoder and trainable convolution layers under the supervision of reconstruction and VGG loss. Our ELBM is illustrated in Fig 7. Through this design, the consistent visual quality of synthesized images has been significantly enhanced.

Table 1 Quantitative comparisons on the VITON-HD dataset (Choi et al., 2021)

Method	Reference	Tunable	512 × 384						1024 × 768			
		Params	LPIPS ↓	SSIM ↑	FID_p ↓	KID_p ↓	FID_u ↓	KID_u ↓	LPIPS ↓	SSIM ↑	FID_u ↓	KID_u ↓
VITON-HD (Choi et al., 2021)	CVPR(21)	–	0.116	0.863	11.01	3.71	12.96	4.09	0.117	0.883	11.79	3.77
PF-AFN* (Ge et al., 2021b)	CVPR(21)	–	0.087	0.886	–	–	9.48	–	0.096	0.898	9.81	–
FS-VTON* (He et al., 2022)	CVPR(22)	–	0.091	0.883	–	–	9.55	–	0.097	0.896	9.67	–
HR-VTON (Lee et al., 2022)	ECCV(22)	–	0.097	0.878	10.88	4.48	13.06	4.72	0.105	0.889	13.91	4.63
SDAFN* (Bai et al., 2022)	ECCV(22)	–	0.092	0.882	–	–	9.40	–	0.108	0.892	9.78	–
GP-VTON* (Xie et al., 2023)	CVPR(23)	–	0.080	0.894	–	-	9.20	–	0.091	0.898	9.23	–
TryOnDiffusion* (Zhu et al., 2023a)	CVPR(23)	–	–	–	–	–	23.35	10.84	–	–	–	–
Paint-by-Example (Yang et al., 2023)	CVPR(23)	923 M	0.143	0.843	9.97	1.72	11.04	2.09	–	–	–	–
MGD* (Baldrati et al., 2023)	ICCV(23)	859 M	–	–	10.60	3.26	12.81	3.86	–	–	–	–
LaDI-VTON (Morelli et al., 2023)	ACMMM(23)	1003 M	0.104	0.872	8.96	1.67	9.93	1.91	–	–	–	–
DCI-VTON (Gou et al., 2023)	ACMMM(23)	923 M	0.072	0.892	5.57	0.57	8.76	0.87	–	–	–	–
WarpDiffusion* (Li et al., 2023a)	Arxiv(23)	>859 M	0.078	0.896	-	-	8.90	-	0.089	0.901	9.19	-
StableVITON (Kim et al., 2023)	CVPR(24)	611 M	0.082	0.865	7.11	1.47	9.76	1.71	–	–	–	–
StableVITON (RePaint) (Kim et al., 2023)	CVPR(24)	611 M	0.077	0.889	6.17	1.06	9.17	1.32	–	–	–	–
OOTDifussion* (Xu et al., 2024)	Arxiv(24)	1719 M	0.071	0.878	-	-	8.81	0.82	–	–	–	-
OOTDifussion (Xu et al., 2024)	Arxiv(24)	1719 M	0.074	0.874	5.98	0.87	8.89	0.91	0.078	0.881	9.23	1.04
TryOn-Adapter	-	**510 M**	0.071	0.894	5.57	0.56	8.63	0.79	0.075	0.901	8.97	0.85
TryOn-Adapter (RePaint)	-	**510 M**	**0.069**	**0.897**	**5.54**	**0.53**	**8.62**	**0.78**	**0.072**	**0.903**	**8.95**	**0.84**

The bold indicates the highest results. *Note* *Denotes results reported in their official papers, which may differ in metric implementation. 'Tunable Params' indicates the trainable parameters in the diffusion model. Without access to the open-source code for WarpDiffusion (Li et al., 2023a), we are unable to determine the exact number of trainable parameters, yet their paper's full fine-tuning method implies it exceeds 859 M

4 Experiments

4.1 Experimental Setup

Datasets. We mainly conduct qualitative and qualitative evaluations of our TryOn-Adapter on VITON-HD (Han et al., 2018), which comprises 13,679 image pairs. Each pair comprises a front-view upper-body woman and an upper garment under the resolution of 1024 × 768. Followed by the previous works (Morelli et al., 2023; Gou et al., 2023; Xie et al., 2023), we split the dataset into 11,674/2032 training/testing pairs. To prove that our method can have excellent results in more diverse scenarios, we further conduct experimental evaluations on Dreesscode (Morelli et al., 2022), which contains 53,792 front-view full-body person and garments pairs from different categories, i.e., upper, lower, and dresses.

Evaluation Metrics To quantitatively evaluate our model, we use various metrics for the similarity and realism assessment. For similarity evaluation, we aim to assess the generated image's coherence compared to the ground truth, which can test the model's capability of ID preservation. This evaluation is mainly validated on paired images, for which we employed two widely used metrics: Structural Similarity (SSIM) for pixel level and Learned Perceptual Image Patch Similarity (LPIPS) for feature level. For realism assessment, the aim is to ensure that the generated images exhibit consistent visual quality and realistic try-on effects. Both paired images and unpaired images should be measured, and we use the Frechet Inception Distance (FID) and Kernel Inception Distance (KID) as our metrics at the feature level.

Implementation Details We build our diffusion model based on Paint-by-Example (Yang et al., 2023), including an autoencoder with latent-space downsampling factor $f = 8$ and a UNet denoiser. We utilize its pre-trained model and freeze all parameters except attention layers. We first train the ELBM module. For the diffusion model, the style preserving module is separately trained with the texture highlighting and structure adapting modules. We generate the images at 512 × 384 and 1024 × 768 resolutions, and the reference image I_c is resized at 224×224. We set $\omega = 0.5$ in Sect. 3.3. For optimizing, we utilize AdamW (Loshchilov and Hutter, 2017) optimizer with the learning rate of 1×10^{-5}, and we trained on 4 NVIDIA Tesla A100 GPUs for 40 epochs. For the inference, we utilize the PLMS (Liu et al., 2022) sampling method, with 100 sampling steps, and we set $T' = 50$ in T-RePaint (see Sect. 3.4).

4.2 Quantitative and Qualitative Evaluations

Quantitative Evaluations As shown in Table 1, we quantitatively compare our method with the previous traditional methods at two resolutions, 512 × 384 and 1024 × 768, on the VITON-HD dataset (Choi et al., 2021), including VITON-HD (Choi et al., 2021), PF-AFN (Ge et al., 2021b), FS-VTON (He et al., 2022), HR-VTON (Lee et al., 2022), SDAFN (Bai et al., 2022), GP-VTON (Xie et al., 2023), and diffusion-based methods including TryOnDiffusion (Zhu et al., 2023a), Paint-by-Example (Yang et al., 2023), MGD (Baldrati et al., 2023), LaDI-VTON (Morelli et al., 2023), DCI-VTON (Gou et al., 2023), WarpDiffusion (Li et al., 2023a), StableVITON (Kim et al., 2023), OOTDiffusion[1] (Xu et al., 2024). For resolution 512 × 384, GP-VTON (Xie et al., 2023) has achieved the best performance among traditional methods, showing excellent structural similarity (SSIM) results. However, its performance in authenticity is not as good as the diffusion-based methods. In full-tuning diffusion-based methods, due to the specified adaptation-based architecture for fine-grained identity factors, our method not only reduces the trainable parameters to nearly half compared to other methods but also achieves state-of-the-art performance across all metrics. Besides, our method has seen a significant performance improvement compared to our baseline Paint-by-Example (Yang et al., 2023), thanks to the three identity-preserving modules we designed. Additionally, in the unpaired setting, which is closer to real-world application scenarios, our KID and FID scores show significant advantages compared to other outperforming methods, such as DCI-VTON (Gou et al., 2023). Compared with the method StableVITON (Kim et al., 2023), which also employs the RePaint technique and efficient training, our method exhibits more excellent performance compared to StableVITON (Kim et al., 2023) (rows 14, 18). Besides, the table results show that StableVITON's ability to preserve identity heavily relies on RePaint (rows 13, 14) even though it has more trainable parameters than ours. This also demonstrates that a single image is insufficient to fully

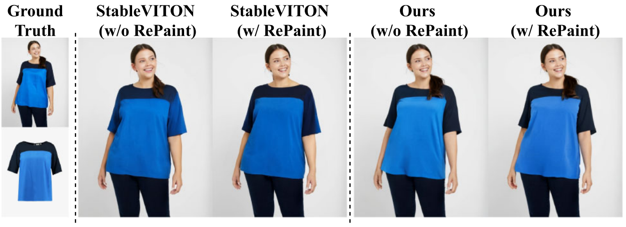

Fig. 8 Qualitative evaluation on the VITON-HD dataset (Choi et al., 2021) with StableVITON (Kim et al., 2023) and our **TryOn-Adapter** at 512 × 384 resolution to compare the impact of RePaint on each method. The results verify that StableVITON heavily relies on RePaint to preserve identity

[1] The results of OOTDiffusion are reproduced using the official code (https://github.com/levihsu/OOTDiffusion.git), but we find that the generated outcomes vary with changes in the seed. For practicality and fairness in comparison, we run each case only once with the random seed in the following quantitative and qualitative experiments.

Table 2 Quantitative results on the Dresscode dataset (Morelli et al., 2022)

Method	Upper		Lower		Dresses		All					
	$\text{FID}_u \downarrow$	$\text{KID}_u \downarrow$	$\text{FID}_u \downarrow$	$\text{KID}_u \downarrow$	$\text{FID}_u \downarrow$	$\text{KID}_u \downarrow$	LPIPS ↓	SSIM ↑	$\text{FID}_p \downarrow$	$\text{KID}_p \downarrow$	$\text{FID}_u \downarrow$	$\text{KID}_u \downarrow$
PF-AFN* (Ge et al., 2021b)	14.32	–	18.32	–	13.59	–	–	–	–	–	–	–
FS-VTON* (He et al., 2022)	13.16	–	17.99	–	13.87	–	–	–	–	–	–	–
HR-VTON* (Lee et al., 2022)	16.86	–	22.81	–	16.12	–	0.086	0.901	–	–	–	–
CP-VTON (Wang et al., 2018)	48.31	35.25	51.29	38.48	25.94	15.81	0.186	0.842	28.44	21.96	31.19	25.17
PSAD (Morelli et al., 2022)	17.51	7.15	19.68	8.90	17.07	6.66	0.058	0.918	8.01	4.90	10.61	6.17
SDAFN* (Bai et al., 2022)	12.61	–	16.05	–	11.80	–	0.063	0.916	–	–	–	–
GP-VTON* (Xie et al., 2023)	11.98	–	16.07	–	12.26	–	0.050	0.925	–	–	–	–
LaDI-VTON (Morelli et al., 2023)	13.26	2.67	14.80	3.13	13.40	2.50	0.064	0.906	4.14	1.21	6.48	2.20
MGD* (Baldrati et al., 2023)	–	–	–	–	–	–	–	–	5.74	2.11	7.33	2.82
WarpDiffusion* (Li et al., 2023a)	–	–	–	–	–	–	0.088	0.895	–	–	8.61	–
TryOn-Adapter	11.58	1.63	14.10	3.09	11.58	1.66	0.049	0.926	3.48	0.93	6.15	1.17
TryOn-Adapter (RePaint)	**11.55**	1.61	**14.08**	3.07	**11.56**	1.64	**0.045**	**0.929**	**3.44**	0.91	**6.13**	1.15

The bold indicates the highest results. Note: * denotes results reported in their official papers, which may differ in metric implementation

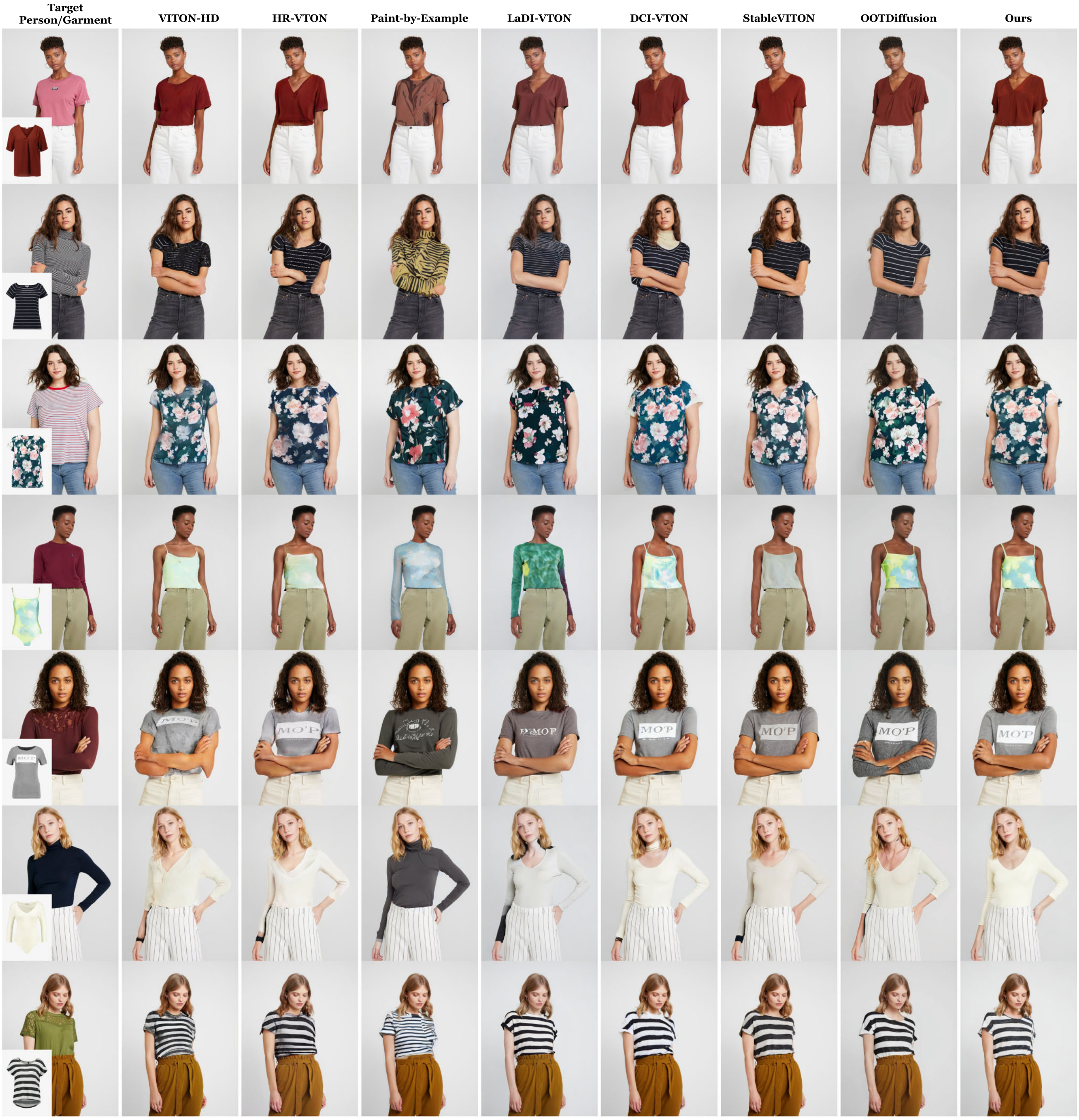

Fig. 9 Qualitative comparison on the VITON-HD dataset (Choi et al., 2021) with VITON-HD (Choi et al., 2021), HR-VTON (Lee et al., 2022), Paint-by-Example (Yang et al., 2023), LaDI-VTON (Morelli et al., 2023), DCI-VTON (Gou et al., 2023), StableVITON (Kim et al., 2023), OOTDiffusion (Xu et al., 2024), and our **TryOn-Adapter** at 512 × 384 resolution

capture the complexity of clothing identity. Conversely, our TryOn-Adapter itself has a strong ability to preserve garment identity (row 17), thanks to our decoupling of the identity preservation problem. Since our T-RePaint can bring some performance improvement and incurs no additional cost, we incorporate it into our approach (row 18). We also conduct qualitative comparisons to confirm this phenomenon, as shown in Fig. 8. Compared to the current work OOTDiffusion (Xu et al., 2024), where both the reported results in its paper (row 15) and our reproduced results (row 16) are provided, our TryOn-Adapter uses fewer than a third of the trainable parameters (510 M *v.s.* 1719 M), yet achieves better

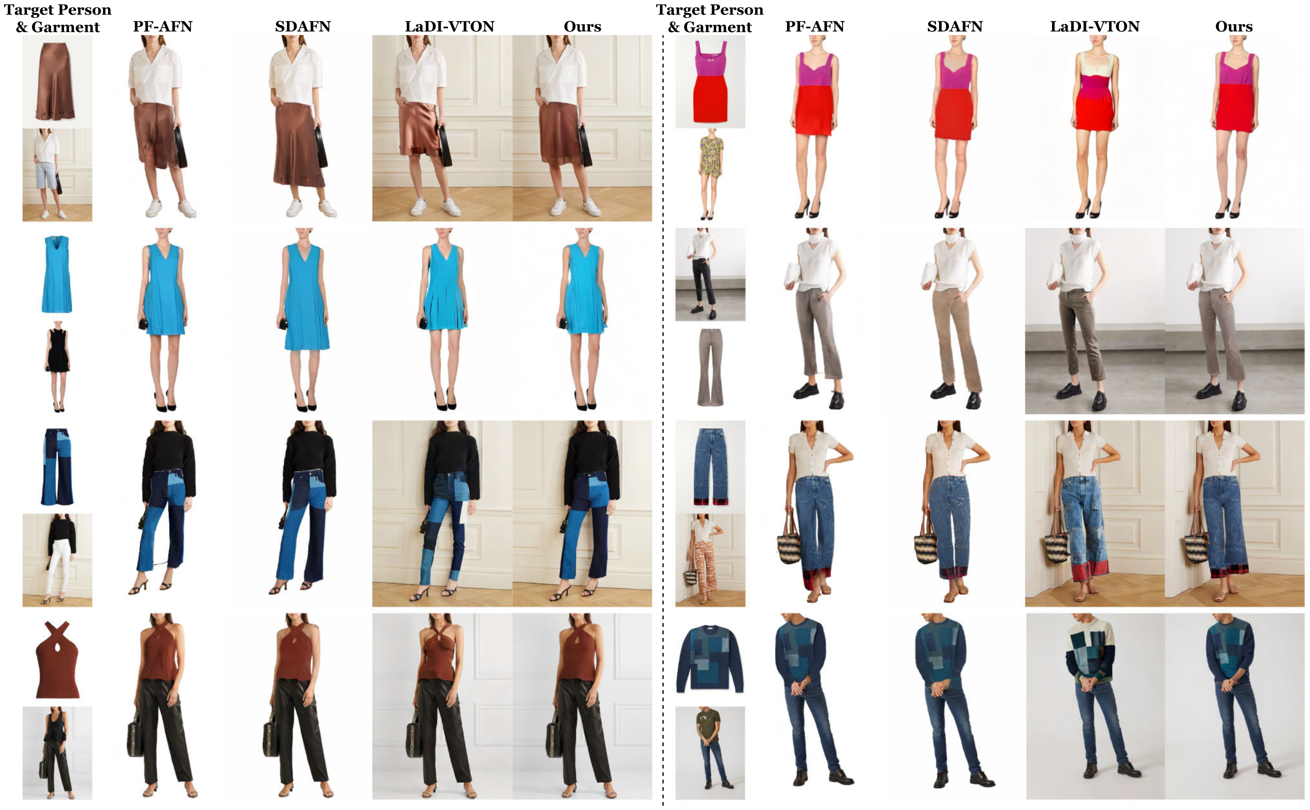

Fig. 10 Qualitative comparison on the Dresscode dataset (Morelli et al., 2022) with PF-AFN (Ge et al., 2021b), SDAFN (Bai et al., 2022), LaDI-VTON (Morelli et al., 2023), and our **TryOn-Adapter** at 512 × 384 resolution

performance in all metrics. This proves the necessity of our fine-grained exploration and analysis of identity preservation in clothing. For resolution 1024 × 768, we only include methods that provide either 1024 × 768 results in their paper or a 1024 × 768 checkpoint in their official code. Since most of the previous work (Ge et al., 2021b; He et al., 2022; Lee et al., 2022; Wang et al., 2018; Morelli et al., 2022; Bai et al., 2022; Li et al., 2023a) does not provide results for $\mathbf{FID_p}$ and $\mathbf{KID_p}$ at 1024 × 768 resolution, we have followed suit. As shown in Table 1, it can be seen that our TryOn-Adapter achieves state-of-the-art performance even at high resolutions, which is consistent with the conclusion drawn in 512 × 384 part.

To further quantitatively evaluate our TryOn-Adapter, we compare our method on the Dresscode dataset (Morelli et al., 2022) with the previous traditional methods, including PF-AFN (Ge et al., 2021b), FS-VTON (He et al., 2022), HR-VTON (Lee et al., 2022), SDAFN (Bai et al., 2022), CP-VTON (Wang et al., 2018), PSAD (Morelli et al., 2022), SDAFN (Bai et al., 2022), GP-VTON (Xie et al., 2023), and diffusion-based methods including MGD (Baldrati et al., 2023), LaDI-VTON (Morelli et al., 2023), and WarpDiffusion (Li et al., 2023a). As shown in Table 2, our TryOn-Adapter's performance has reached the most excellent results among all metrics under various settings.

Qualitative Evaluations Figure 9 shows the qualitative comparison of the results produced by different methods in the unpaired setting on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution. As depicted in the figure, although traditional methods like VITON-HD (Choi et al., 2021) and HR-VTON (Lee et al., 2022) (as in columns 2 and 3) can preserve the identity of the target garment, the resulting garments exhibit some distortion when worn on a person, appearing unnatural. As for diffusion-based methods, the target garment can be worn naturally on a person, but it cannot guarantee the identity of the clothing. Paint-by-Example (Yang et al., 2023) (as in column 4) and LaDI-VTON (Morelli et al., 2023) (as in column 5) cannot guarantee the style of the target garment, especially the color information. DCI-VTON (Gou et al., 2023) compared to the previous two, has made great progress in style-preserving but has not effectively addressed the problem of long and short sleeves (as in column 6, rows 2 and 6), and the patterns and textures of the garments are not clear enough (as in column 6, rows 3 and 4). Meanwhile, StableVITON (Kim et al., 2023) follows an efficient training strategy with ControlNet (Zhang et al., 2023), but it does not decouple the clothing identity preservation issue. This results in noticeable color discrepancies (as in column 7, rows 4, 5, and 6) and a lack of fidelity in fine texture details (as in col-

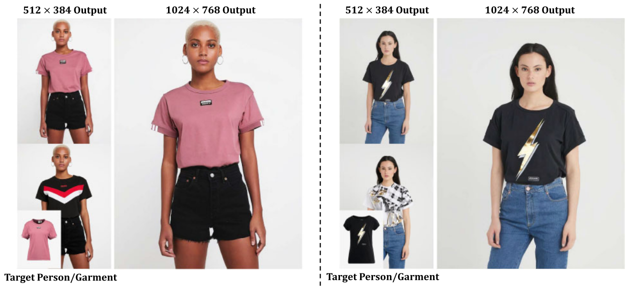

Fig. 11 Qualitative results of our TryOn-Adapter on the VITON-HD dataset (Choi et al., 2021) at the resolutions of 1024 × 768 and 512 × 384. Please zoom in for more details

umn 7, rows 4 and 5) compared to the target clothing in its output. OOTDiffusion (Xu et al., 2024) introduces numerous training parameters, but the exploration of fine-grained clothing identity preservation is still insufficient, resulting in some unsatisfactory generated results. Specifically, its generated clothing exhibits color deviation (as in column 8, rows 2, 3, 4, 5, and 7) and texture deviation (as in column 8, rows 2 and 3) compared to the original clothing. It encounters issues with confusion between long and short sleeves (as in column 8, row 5) and abnormal body structure (as in column 8, row 2). Compared to the above diffusion-based methods, our method benefits from the well-designed three adapter modules, effectively addressing the shortcomings. Consequently, our method can ensure a commendable preservation of garment identity (as in column 9, rows 1, 2, 4, and 7), featuring enhanced color fidelity (as in column 9), sharper illustration of intricate textures (as in column 9, rows 2, 3, 4, and 5), and better management of long/short sleeve transformations while naturally worn (as in column 9, rows 2, 4, and 6).

For further qualitative evaluations, we report in Fig. 10 sample images generated by our model and by the competitors using officially released weights on the Dresscode dataset (Morelli et al., 2022) at 512 × 384 resolution. Compared to traditional methods such as PF-AFN (Ge et al., 2021b) and SDAFN (Bai et al., 2022), our method's try-on results will have a more realistic try-on effect without the unnatural signs of pasting from the warped garment onto the target person. Compared to the diffusion-based method LaDI-VTON (Morelli et al., 2023), our method has a distinct advantage in preserving the garment's identity, including elements like the style and texture details of the clothing.

Moreover, echoing the 1024 × 768 resolution in quantitative experiments, we provide the try-on results with higher resolution of our TryOn-Adapter in Fig. 11. Our 1024 × 768 resolution outputs are equally impressive as that at 512 × 384 resolution, demonstrating its ability to consistently attain superior performance across the resolutions.

User Study of Virtual Try-On We further evaluate our method against different methods, including VITON-HD (Choi et al., 2021), HR-VTON (Lee et al., 2022), Paint-by-Example (PbE) (Yang et al., 2023), LaDI-VTON (Morelli et al., 2023), DCI-VTON (Gou et al., 2023), GP-VTON (Xie et al., 2023), StableVITON (Kim et al., 2023), and OOTDiffusion (Ruiz et al., 2023) through a user study on different virtual try-on generation results in the VITON-HD dataset. We randomly select 300 unpaired sets from the test dataset, each containing a target garment image and a target person image. We survey 28 non-experts for this study, asking them to choose an image with the most satisfactory performance among the generated results of our model and baselines according to the following two questions: (1) Which image is the most photo-realistic? (2) Which image preserves the details of the target clothing the most? As shown in Fig. 12, our approach received over 45% support for both questions. The results demonstrate that our method can generate naturally realistic images while effectively preserving target garment details during the virtual try-on process.

4.3 Ablation Study and Further Analysis

Effectiveness of Individual Adaptation Components To demonstrate the effectiveness of our proposed adaptation, we conduct ablation experiments on the VITON-HD (Choi et al., 2021) dataset. To more intuitively verify the effectiveness of each adaptation module, all results do not employ T-RePaint. Meanwhile, all tests use ELBM to prevent inaccuracies from reconstruction affecting result comparisons. We choose two baselines for comparison. One freezes all training parameters and uses Paint-by-Example's (Yang et al., 2023) pre-trained model for inference, while the other is based on the former but only trains the attention layers and the CLIP class token's linear mapping layer related to the cross attention. As shown in Table 3, we gradually incorporate our designed adaptations, and the model's performance strengthens step by step. As shown in Fig. 13, the visual comparison of our generated results for each stage will be more intuitive. The frozen baseline is a semi-finished result (column 2), where the garment is detached from the body. For the second baseline (column 3), which is only fine-tuned on the attention layers, the generated clothing style diverges from the target clothing, and the boundary between the limbs and the garment is unclear. After adding the style adaptation (column 4), the clothing can naturally be worn on the person, and the clothing style has been significantly improved, but the details and textures of the clothing are not clear enough, and the shadow exists in the neck area. After combining texture adaptation (column 5), the representation of the clothing's detailed texture has been enhanced, but the high-frequency map lacks the ability to determine whether the shadow on the neck area is skin or

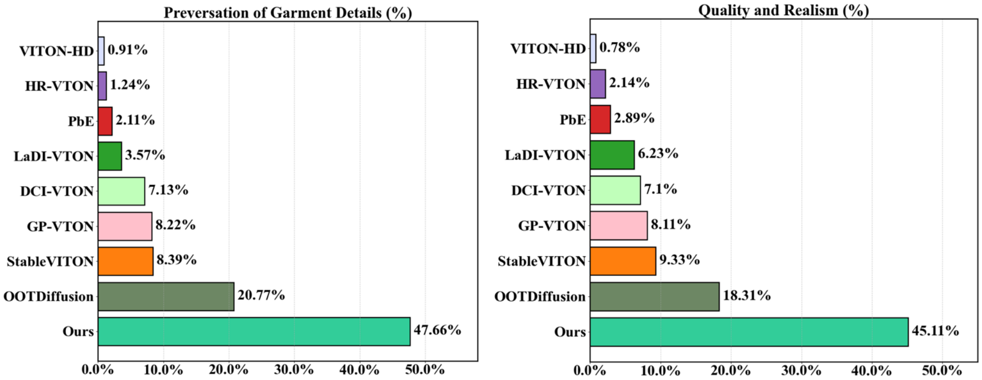

Fig. 12 User study results on VITON-HD dataset at 512 × 384 resolution. We compare our method with VITON-HD (Choi et al., 2021), HR-VTON (Lee et al., 2022), Paint-by-Example (PbE) (Yang et al., 2023), LaDI-VTON (Morelli et al., 2023), DCI-VTON (Gou et al., 2023), GP-VTON (Xie et al., 2023), StableVITON (Kim et al., 2023), and OOTDiffusion (Xu et al., 2024)

Table 3 Effectiveness of our Adapter components on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	Params	Tunable Params	LPIPS↓	SSIM↑	FID_u ↓	KID_u ↓
Frozen	859 M	0 M	0.227	0.791	23.48	14.67
Frozen + fine-tuned attention layers	859 M	84 M	0.119	0.849	11.00	2.29
+ style adaptation	1048 M	273 M	0.079	0.887	8.89	0.94
+ texture adaptation	1129 M	354 M	0.074	0.892	8.73	0.82
+ segmentation adaptation	1212 M	435 M	0.071	0.894	8.63	0.79

“Params" and “Tunable Params" indicate the total and trainable parameters in the diffusion model, respectively

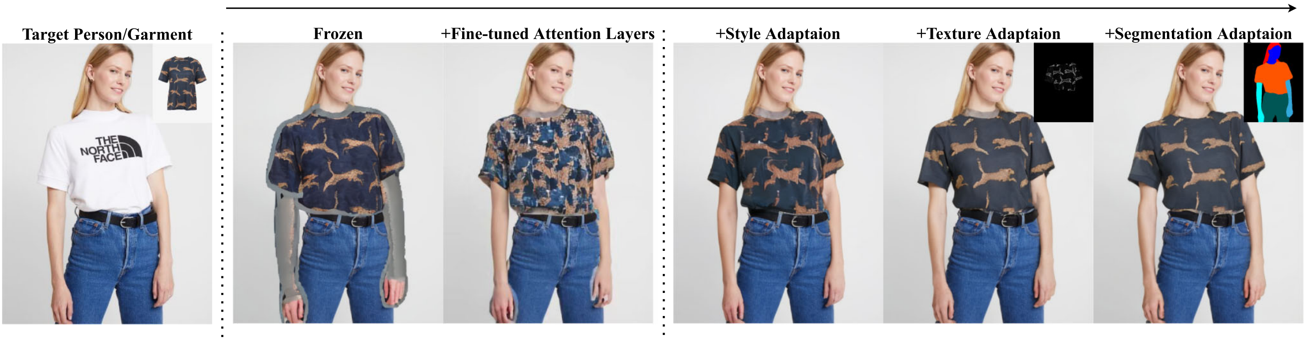

Fig. 13 Visual effectiveness of individual adaptation components in our TryOn-Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

a collar. After introducing segmentation adaptation (column 6), the neck shadow issue was successfully resolved.

Analysis on Style Adapter To analyze the impact of our style adapter designed for patch token in the Style Preserving module, we conduct experimental evaluations on the VITON-HD dataset (Choi et al., 2021). As shown in Table 4, with the addition of the style adapter, all quantitative metrics have been improved. For a clear visual representation comparison on qualitative evaluation, we maintain consistency with previous ablation studies here by using ELBM and not T-RePaint. As shown in Fig. 14, it can be seen that after adding the style adapter, there has been a noticeable improvement in the color difference between the generated garment (column 4) and the target garment (column 2) compared to the generated results without the style adapter (column 3). Meanwhile, the logos and textures on the garment also became clearer after integrating this module. The above results demonstrate the significance of integrating VAE embeddings, while also proving the effectiveness of our style adapter.

Qualitative Evaluation of Texture Highlighting Module and Structure Adapting Module To verify the robustness of the Texture Highlighting Module and Structure Adapting Mod-

Table 4 Quantitative analysis of the Style Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	$LPIPS \downarrow$	$SSIM \uparrow$	$FID_u \downarrow$	$KID_u \downarrow$
w/o style adapter	0.073	0.892	8.69	0.81
Ours	**0.069**	**0.897**	**8.62**	**0.78**

The bold indicates the highest results in each ablation study

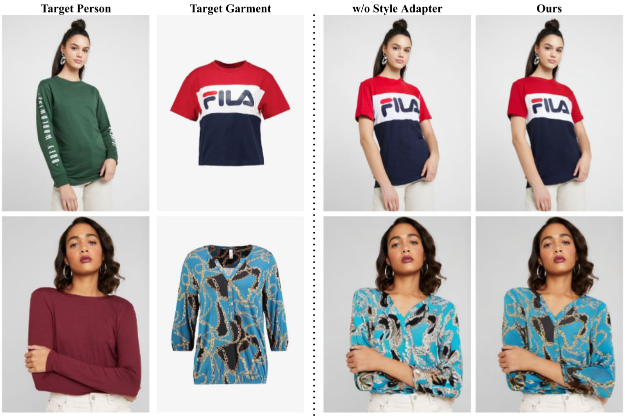

Fig. 14 Visual effectiveness of the Style Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

ule, we supply more convincing visual results as shown in Fig. 15. Here, we also use ELBM and do not use T-RePaint. The example on the left in this figure proves that the Texture Highlighting Module can effectively enhance the texture of the target garment, especially the details of cartoon patterns. And, the example on the right demonstrates that the Structure Adapting Module is capable of addressing the problem of long and short sleeves well, bringing about a realistic try-on effect.

Qualitative Comparison Between the Diffusion Model and GANs To analyze the performance of the Diffusion Model and GAN in virtual try-on, we compare our TryOn-Adapter with the latest GANs-based method GP-VTON (Xie et al., 2023), where both use the same warped garment and the latter employs the GANs as the generative model as shown in Fig. 16. The try-on results generated by GP-VTON (Xie

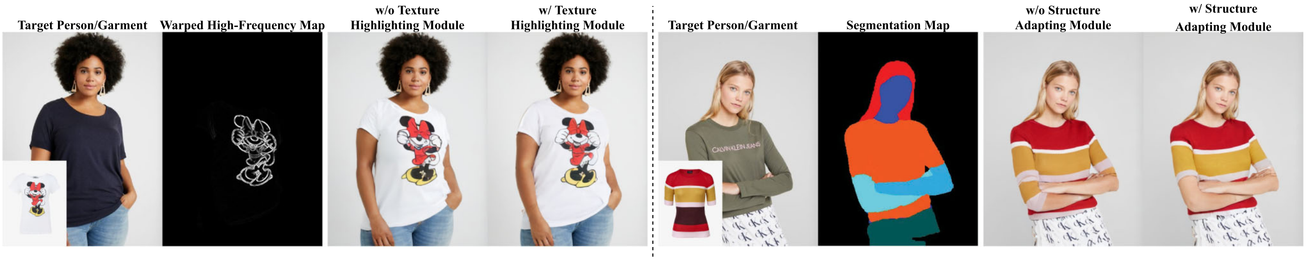
Target Person/Garment
Warped High-Frequency Map
w/o Texture Highlighting Module
w/ Texture Highlighting Module
Target Person/Garment
Segmentation Map
w/o Structure Adapting Module
w/ Structure Adapting Module

Fig. 15 Qualitative evaluation of Texture Highlighting Module and Structure Adapting Module in our TryOn-Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

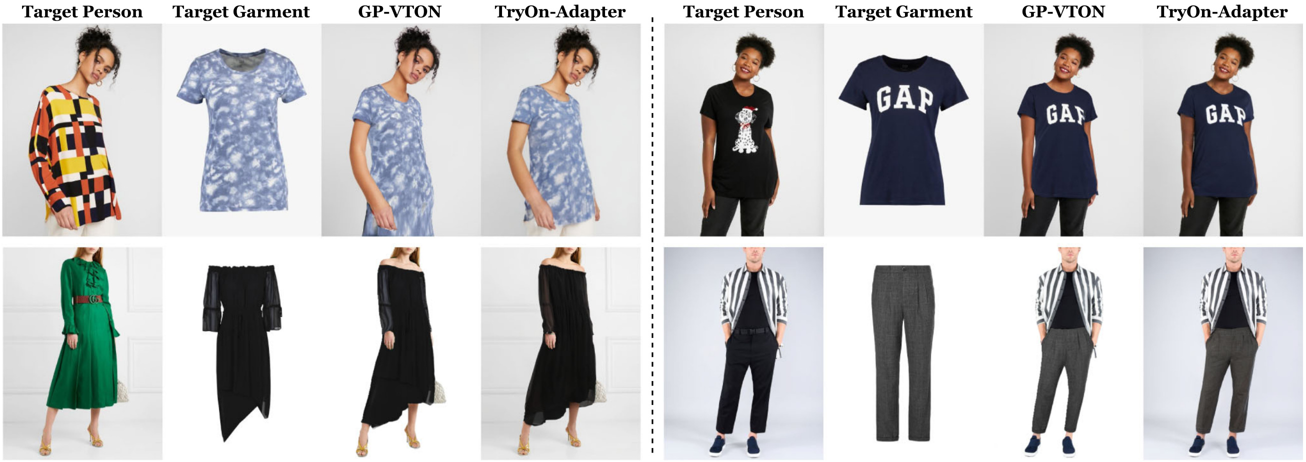

Fig. 16 Qualitative Comparison between the Diffusion Model and GAN (GP-VTON (Xie et al., 2023) vs. TryOn-Adapter) on the VITON-HD (Choi et al., 2021) and Dresscode (Morelli et al., 2022) datasets at 512 × 384 resolution

Table 5 Quantitative analysis of PAM in Texture and Segmentation Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	$LPIPS \downarrow$	$SSIM \uparrow$	$FID_u \downarrow$	$KID_u \downarrow$
w/o PAM	0.071	0.895	8.65	0.80
Ours	**0.069**	**0.897**	**8.62**	**0.78**

The bold indicates the highest results in each ablation study

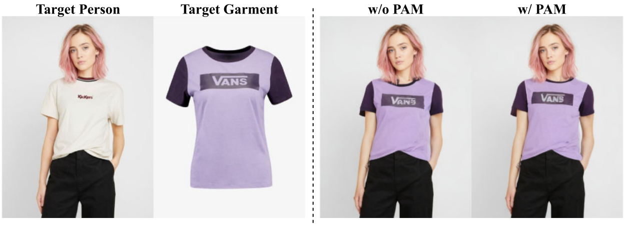

Fig. 17 Quantitative evaluation of PAM in Texture and Segmentation Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

et al., 2023) (based on GANs) are prone to distortions and deformations, as seen in the waist area of the first row's left image and the logo area of the right image. Furthermore, the outputs from GP-VTON lack a realistic sense of actual try-on, resembling a warped garment pasted onto the target person, as in the second row, especially with the arm in the sleeve on the left image, which is almost completely forgotten. Besides, GP-VTON exhibits noticeable jaggies around the clothing when zoomed in. Therefore, diffusion models possess more powerful generative capabilities than GANs.

Analysis on PAM in Texture and Segmentation Adapter To analyze the impact of the position attention module (PAM) in Texture and Segmentation Adapter, we conduct experiments on the VITON-HD dataset (Choi et al., 2021). For the quantitative evaluation, we can see all quantitative metrics are improved after adding PAM, as shown in Table 5. For the qualitative evaluation, we don't use T-RePaint for a direct visual comparison. As shown in Fig. 17, the logos in the generated images are more evident after integrating PAM, benefiting from the enhanced local spatial representation by PAM, which allows the Adapters to interpret the high-frequency information in the images better.

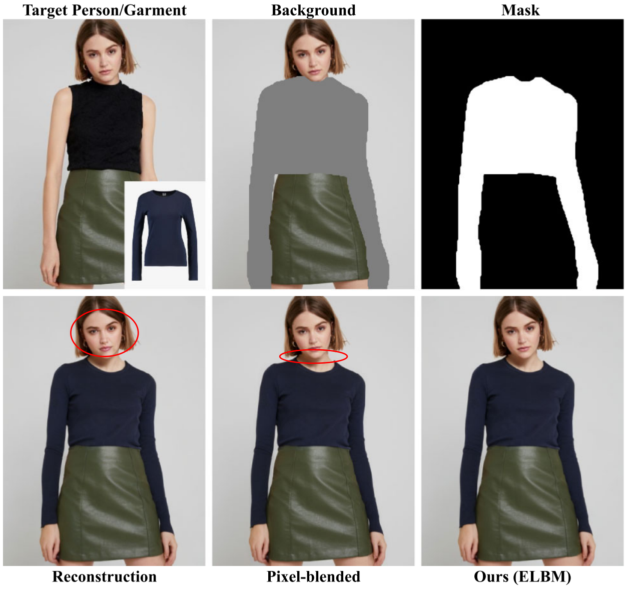

Fig. 18 Qualitative evaluation of Enhanced Latent Blending Module (ELBM) on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution. The reconstruction result is from the VAE of Stable Diffusion, and the pixel-blended result is from the combination of the original target person image and the reconstruction result image at the pixel level

Table 6 Quantitative analysis of PAM in Texture and Segmentation Adapter on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Task	ELMB	f_c^{NL}	f_c^L	$LPIPS \downarrow$	$SSIM \uparrow$
Reconstruction	w/o	×	×	0.024	0.937
Reconstruction	w/	✓	×	0.021	0.954
Reconstruction	w/	✓	✓	**0.020**	**0.956**
Try-On (paired)	w/o	×	×	0.076	0.867
Try-On (paired)	w/	✓	×	0.071	0.895
Try-On (paired)	w/	✓	✓	**0.069**	**0.897**

The bold indicates the highest results in each ablation study

Analysis on Enhanced Latent Blending Module (ELBM) To analyze the impact of the Enhanced Latent Blending Module, we conduct qualitative and quantitative evaluations on the VITON-HD dataset (Choi et al., 2021). For the qualitative evaluation, we compare three approaches at the final image synthesis stage on virtual try-on: reconstruction, pixel-blended, and our ELBM. Given the original person image I_0 and background mask m, the reconstruction I_{re} result is from the VAE of Stable Diffusion, and the pixel-blended result I_{pb} is from the combination of the original target person image and the reconstruction result image at the pixel level, i.e., $I_{pb} = I_0 \otimes (1 - m) + I_{re} \otimes m$. As shown in Fig. 18, the reconstruction result (column 1, row 2) exhibits some degree of distortion and deformation in the human face, whereas the pixel-blended result (column 2, row 2) preserves the critical facial features well but introduces noise and shadows at the junction of the neck due to the rough combination of I_0 and I_{re}. Our ELBM (column 3, row 2) effectively addresses the above problems, preserving high-frequency background information, such as the face and hands, while avoiding introducing any noise that may result from image combination. Please zoom in for more details.

For the quantitative evaluation, we conduct experiments on two tasks, including image reconstruction and paired virtual try-on, and we ablate the impacts of our two convolutions f_c^{NL} and f_c^L in latent blending fusion of ELBM. As shown

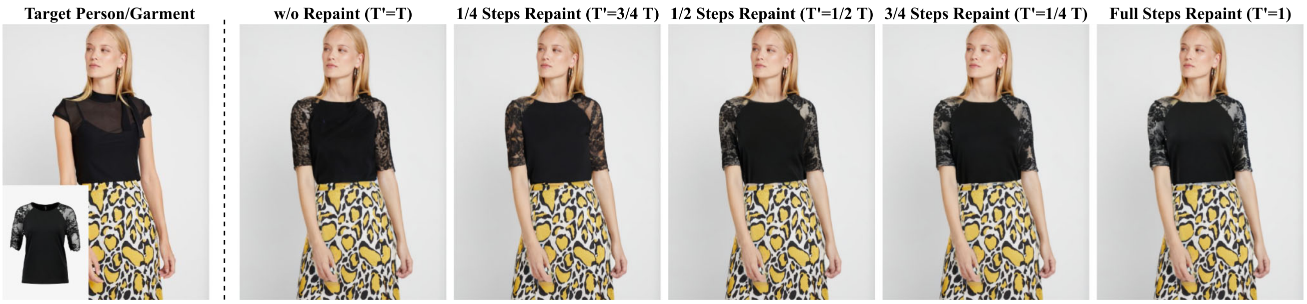
Target Person/Garment
w/o Repaint (T'=T)
1/4 Steps Repaint (T'=3/4 T)
1/2 Steps Repaint (T'=1/2 T)
3/4 Steps Repaint (T'=1/4 T)
Full Steps Repaint (T'=1)

Fig. 19 Qualitative evaluation comparing the impact of varying numbers of RePaint steps on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Table 7 Quantitative evaluation comparing the impact of varying numbers of RePaint steps on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	$SSIM$ ↑	$LPIPS$ ↓	FID_u ↓	KID_u ↓
w/o RePaint	0.894	0.071	8.63	0.79
1/4 Steps RePaint	0.896	0.070	**8.62**	**0.78**
1/2 Steps RePaint	**0.897**	**0.069**	**8.62**	**0.78**
3/4 Steps RePaint	0.895	0.071	8.67	0.82
Full Steps RePaint	0.896	0.074	8.99	1.09

The bold indicates the highest results in each ablation study

in Tab 6, the ELBM we propose not only improves the reconstruction capabilities of the Stable Diffusion autoencoder in the reconstruction task but also elevates the overall performance of the final virtual try-on pipeline, resulting in superior evaluation metrics. At the same time, the second and fifth rows in Table 6 represent LaDI-VTON (Morelli et al., 2023), while the third and sixth rows represent our proposed ELBM. Our ELBM exhibits better performance, which demonstrates the effectiveness of deep fusion in Eq. 15.

Analysis on T-RePaint To evaluate the impact of varying numbers of RePaint steps, we conduct quantitative and qualitative experiments on the VITON-HD dataset (Choi et al., 2021). For qualitative evaluation, utilizing RePaint for half of the denoising steps ($T' = 1/2\ T$) during the inference achieves a balance between preserving the identity of the garment and realizing a realistic try-on effect, thereby attaining the best generative outcomes, as illustrated in Fig. 19. Meanwhile, a larger T' yields a more realistic try-on effect but poorer texture ID preservation (see columns 2 and 3), and vice versa. Especially with T' set to 1, the generated image's garment depends heavily on the warped garment, which ensures ID preservation but can lead to distortions if the warped garment is distorted. Additionally, employing RePaint in full steps severely undermines the realism of the generated images. This is illustrated in Fig. 19 last column, where there is a noticeable disconnection at the intersection of skirts and tops, and the shoulders are barely discernible. For the qualitative evaluation, results across various metrics also indicate that setting $T' = 1/2\ T$, i.e., using RePaint for half of the steps during the inference, yields the best performance, as shown in Table 7.

Table 8 Quantitative comparison between the different methods of generating segmentation maps on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	$MIoU_{cloth}$ ↑	$MIoU_{all}$ ↑
VITON-HD (Choi et al., 2021)	0.8997	0.9598
Ours	**0.9662**	**0.9762**

MIoU$_{cloth}$ represents the result computed only within the clothing area, and **MIoU**$_{all}$ represents the result computed for the entire body excluding the neck

The bold indicates the highest results in each ablation study

Comparison Between the Different Methods of Generating Segmentation Maps To demonstrate the effectiveness of our designed training-free segmentation map generation method, we qualitatively and quantitatively compared our generated results with the results produced by VITON-HD (Choi et al., 2021) using a trainable segmentation generator network. As for qualitative comparison, as the Training Set of the VITON-HD (Choi et al., 2021) dataset provides the Ground Truth (GT) for the segmentation map, we calculated the **MIoU** (Long et al., 2015) for the generated results of different methods against the GT. The calculation of **MIoU** includes **MIoU**$_{cloth}$ and **MIoU**$_{all}$, where the former computes only in the clothing area, while the latter computes in the entire body area. Since the segmentation generation network of VITON-HD (Choi et al., 2021) does not generate the neck area, the calculation of **MIoU**$_{all}$ excludes the neck area. As shown in Table 8, our method outperforms VITON-HD in both metrics, demonstrating the effectiveness of our approach. For the quantitative comparison, we conduct experiments on both the Training Set and Testing set

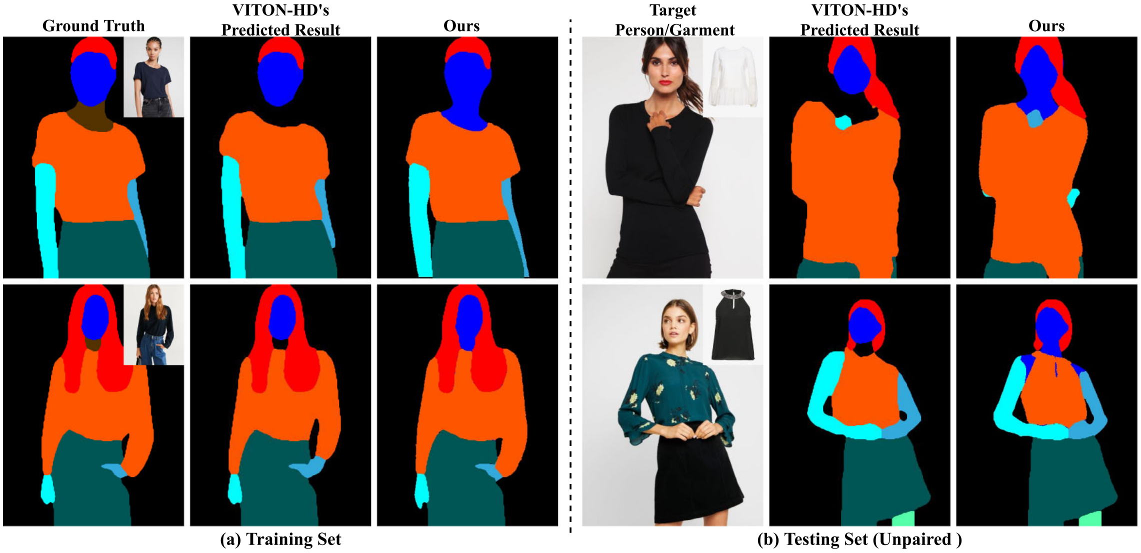
Ground Truth
VITON-HD's Predicted Result
Ours
Target Person/Garment
VITON-HD's Predicted Result
Ours
(a) Training Set
(b) Testing Set (Unpaired)

Fig. 20 Qualitative Comparison between the different methods of generating segmentation maps on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution. VITON-HD's predicted results come from a segmentation generator network, while our results are generated by a training-free method

Table 9 Quantitative comparison between Full Fine-tuning and Parameter Efficient Fine-tuning (PEFT) on the VITON-HD dataset (Choi et al., 2021) at 512 × 384 resolution

Method	LPIPS ↓	SSIM ↑	FID_u ↓	KID_u ↓	Params (Tunable)	Time (1 epoch)
Full Fine-tuning	**0.068**	**0.897**	8.63	0.79	1285 M	1.38 h
PEFT (Ours)	0.069	**0.897**	**8.62**	**0.78**	**510 M**	**0.83 h**

The bold indicates the highest results in each ablation study

(unpaired). As shown in Fig. 20, the results generated by both methods are very close to the Ground Truth on the Training Set. However, on the Testing Set, our method shows significantly better results. For example, in the first row, the result generated by VITON-HD is missing a hand, and in the second row, the generated cloth style is incorrect. Overall, our segmentation map generation method is very user-friendly, demonstrating good performance and requiring no network training parameters.

Comparison Between Full Fine-Tuning and Parameter Efficient Fine-Tuning (PEFT) Our method is built upon Paint-by-Example (Yang et al., 2023) and its pre-trained weights, thus inheriting the ability to manipulate specific areas while keeping others unchanged. Consequently, we only need to fine-tune the attention layers and the designed adapters that receive the critical identity cues to adapt to the try-on task. As shown in Table 9, full fine-tuning only offers limited performance boosting on LPIPS and SSIM but leads to significant computational costs. We can also infer that the decoupled clothing identity, in conjunction with the injection modules we designed, has reduced the training difficulty and requirements of preserving the given garment. Therefore, considering the balance between performance and consumption, PEFT emerges as the preferred option.

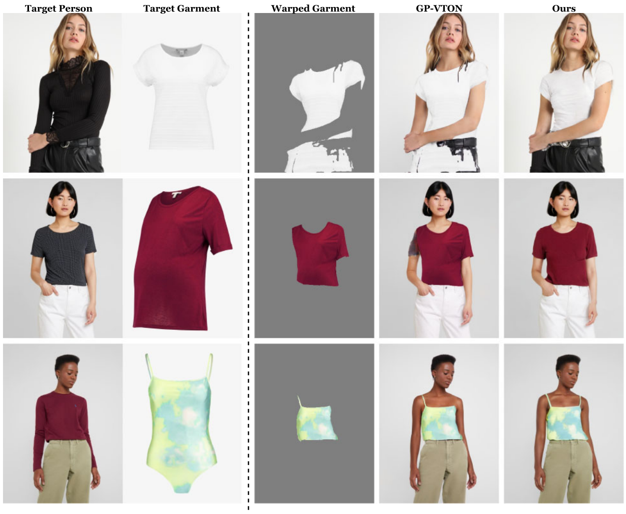

Fig. 21 Performance comparison of GP-VTON (Xie et al., 2023) and our TryOn-Adapter in handling poor warped garments on the VITON-HD dataset (Choi et al., 2021) at the resolutions of 512 × 384

Discussion of Meeting Poor Warped Garments To assess the impact of poor warped garments on our method, we provide the detailed qualitative analysis in Fig. 21. First, based on our observations of the warped garments produced by GP-VTON (Xie et al., 2023), we find that the warped garments fit

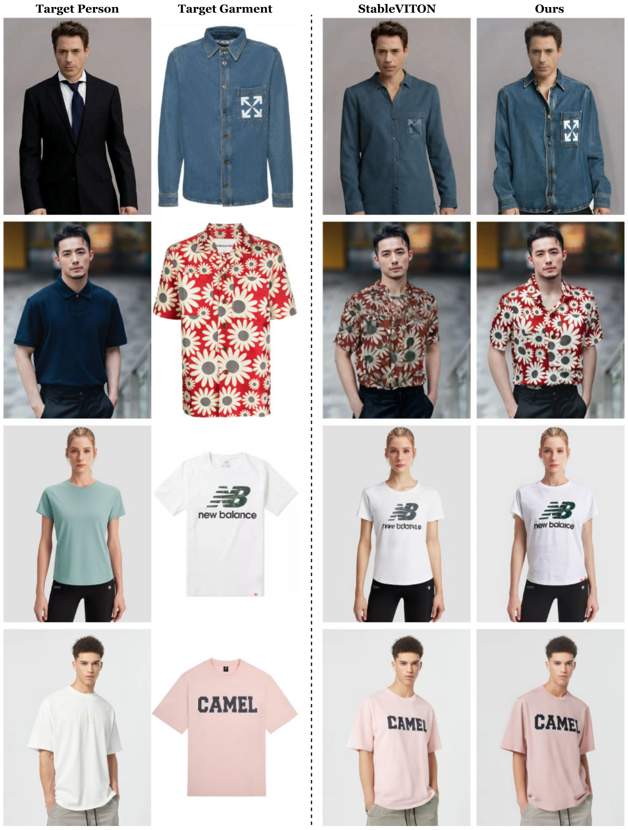

Fig. 22 Qualitative evaluation of our TryOn-Adapter and StableVITON (Kim et al., 2023) on in-the-wild data at 512 × 384 resolution

the target mask area almost accurately, and the texture of the garments does not show noticeable degradation. The problem is that the garment may become excessively stretched (as in column 3, row 1) or incomplete (as in column 3, rows 2 and 3) after passing through the warping network. To be precise, the warping network merges the black clothes and pants of the target person in the first case (as in row 1), resulting in excessive stretching of the warped garment. In the second case (as in row 2), the warping network struggles with the invisible areas of the target garment, which is captured from a non-frontal angle, leading to missing parts in the left sleeve of the warped garment. Additionally, the warping network tends to lose localized clothing parts, as demonstrated in the third case (as in row 3). GP-VTON utilizing the same warped garments as ours directly maps these poor warping results to the final generation (as shown in column 4). Fortunately, our generative model effectively addresses these issues (as shown in column 5). Concretely, to tackle excessive stretching of the garment, our mask m will confine the generation range. For incomplete warped garments, such as those missing sleeves or straps, our designed style adapter will provide compensation.

Qualitative Evaluation on In-the-Wild Data To demonstrate the versatility of our TryOn-Adapter, we conduct experiments using an in-the-wild dataset at 512 × 384 resolution and compare our model's performance with that of StableVITON (Kim et al., 2023). We leverage network weights trained on the VITON-HD (Choi et al., 2021) dataset to perform inference on these wild data. The results are presented in Fig. 22. The wild data are web-crawled: the models in the first[2] and second[3] rows are sourced from RED, while those in the third[4] and fourth[5] rows are obtained from Taobao. The target garments are collected from both Taobao and eBay. The results clearly show that our TryOn-Adapter outperforms StableVITON (Kim et al., 2023) on the in-the-wild dataset. Specifically, our approach more effectively maintains the color and style of the target garments (as in rows 1, 2, 4), preserves texture details (as in row 2) and logos (as in rows 1, 3, 4) of clothing, and ensures a more natural fit on the target person (as in row 2). These experiment results highlight the robustness of our model.

5 Conclusion

Virtual try-on has gained widespread attention due to significantly enhancing the online shopping experience for users. We revisit two critical aspects of diffusion-based virtual try-on technology: identity controllability, and training efficiency. We propose an effective and efficient framework, termed TryOn-Adapter, to tackle these three issues. We first decouple clothing identity into fine-grained factors: style, texture, and structure. Then, each factor incorporates a customized lightweight module and fine-tuning mechanism to achieve precise and efficient identity control. Meanwhile, we introduce a training-free technique, T-RePaint, to further reinforce the clothing identity preservation without compromising the overall image fidelity during the inference. In the final image try-on synthesis stage, we design an enhanced latent blending module for image reconstruction in latent space, enabling the consistent visual quality of the generated image. Extensive experiments on two widely used datasets have shown that our method can achieve outstanding performance with minor trainable parameters.

Limitations Although we satisfactorily resolve the issues of efficiently preserving the identity of the given garment and maintaining consistent visual quality for final try-on synthesis. However, like most previous works, our method is still

[2] https://www.xiaohongshu.com/explore/63256fd90000000011016565.

[3] https://www.xiaohongshu.com/explore/6620e7bc000000000d031fef.

[4] https://m.tb.cn/h.gnbGB3Cx9XZkJpU.

[5] https://m.tb.cn/h.gmEvtXdLMlOqFOp.

a certain distance away from achieving widespread practical application due to the limitation of the datasets. At the same time, to avoid the extra data preprocessing, we will focus more on reference-net-based approaches and aim to propose innovative methods that balance computational cost with performance, thereby advancing the virtual try-on field. Furthermore, there is a lack of targeted quantitative evaluation metrics for virtual try-on tasks. We plan to develop a more granular evaluation from overall style, local texture, and structure for virtual try-on assessment, but progress is slow due to data scarcity.

Data Availability We claim to release the dataset and code upon acceptance. The datasets generated and analyzed during the current study will be available in our open-source repository.

References

Avrahami, O., Fried, O., & Lischinski, D. (2023). Blended latent diffusion. *ACM Transactions on Graphics (TOG), 42*(4), 1–11.

Bai, S., Zhou, H., Li, Z., Zhou, C., & Yang, H. (2022). Single stage virtual try-on via deformable attention flows. In *European conference on computer vision* Berlin: Springer (pp. 409–425).

Baldrati, A., Morelli, D., Cartella, G., Cornia, M., Bertini, M., & Cucchiara, R. (2023). Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. arXiv preprint arXiv:2304.02051

Bhunia, A. K., Khan, S., Cholakkal, H., Anwer, R. M., Laaksonen, J., Shah, M., & Khan, F. S. (2023). Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5968–5976).

Chen, C. Y., Chen, Y. C., Shuai, H. H., & Cheng, W. H. (2023a). Size does matter: Size-aware virtual try-on via clothing-oriented transformation try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7513–7522).

Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., & Zhao, H. (2023b). Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481

Choi, S., Park, S., Lee, M., & Choo, J. (2021). 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).

Corneanu, C., Gadde, R., & Martinez, A. M. (2024). Latentpaint: Image in painting in latent space with diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4334–4343).

Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., & Lazebnik, S. (2023). Street tryon: Learning in-the-wild virtual try-on from unpaired person images. arXiv preprint arXiv:2311.16094

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems, 34*, 8780–8794.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019) Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3146–3154).

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618

Ge, C., Song, Y., Ge, Y., Yang, H., Liu, W., & Luo, P. (2021a). 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).

Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., & Luo, P. (2021b). 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).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27).

Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., & Zhang, L. (2023). Taming the power of diffusion models for high-quality virtual try-on with appearance flow. arXiv preprint arXiv:2308.06101

Güler, R. A., Neverova, N., & Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7297–7306).

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in neural information processing systems* (Vol. 30).

Han, X., Wu, Z., Wu, Z., Yu, R., & Davis, L. S. (2018). Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7543–7552).

Han, X., Hu, X., Huang, W., & Scott, M. R. (2019). Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10471–10480).

He, S., Song, Y. Z., & Xiang, T. (2022). 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).

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems, 33*, 6840–6851.

Issenhuth, T., Mary, J., & Calauzenes, C. (2020). Do not mask what you do not need to mask: A parser-free virtual try-on. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part, X.X. 16*. Berlin: Springer (pp. 619–635).

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020) Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).

Kim, J., Gu, G., Park, M., Park, S., & Choo, J. (2023). Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. arXiv preprint arXiv:2312.01725

Lee, S., Gu, G., Park, S., Choi, S., & Choo, J. (2022). High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European conference on computer vision*. Berlin: Springer (pp. 204–219).

Lewis, K. M., Varadharajan, S., & Kemelmacher-Shlizerman, I. (2021). Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG), 40*(4), 1–10.

Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2019). Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457

Li, X., Kampffmeyer, M., Dong, X., Xie, Z., Zhu, F., Dong, H., & Liang, X. (2023a). Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on. arXiv preprint arXiv:2312.03667

Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., & Lee, Y. J. (2023b). Gligen: Open-set grounded text-to-image generation.

In *Proceedings of the, I.EEE/CVF conference on computer vision and pattern recognition* (pp. 22511–22521).

Li, Z., Wei, P., Yin, X., Ma, Z., & Kot, A. C. (2023c). Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22788–22797).

Liu, L., Ren, Y., Lin, Z., & Zhao, Z. (2022). Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11461–11471).

Minar, M. R., Tuan, T. T., Ahn, H., Rosin, P., & Lai, Y. K. (2020). Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. *CVPR Workshops, 3*, 10–14.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957

Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., & Cucchiara, R. (2022). 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).

Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., & Cucchiara, R. (2023). Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., & Qie, X. (2023). T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., & Chen, M. (2021). Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents vol. 1(2), 3. arXiv preprint arXiv:2204.06125

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K. (2023). Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22500–22510).

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Mahdavi, S. S., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems, 35*, 36479–36494.

Shi, J., Xiong, W., Lin, Z., & Jung, H. J. (2023). Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning, PMLR* (pp. 2256–2265).

Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502

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

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., & Yang, M. (2018). Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 589–604).

Wang, Q., Liu, L., Hua, M., He, Q., Zhu, P., Cao, B., & Hu, Q. (2022). Hs-diffusion: Learning a semantic-guided diffusion model for head swapping. arXiv preprint arXiv:2212.06458

Wei Y, Zhang, Y., Ji, Z., Bai, J., Zhang, L., & Zuo, W. (2023). Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint arXiv:2302.13848

Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., & Liang, X. (2023). 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).

Xu, Y., Gu, T., Chen, W., & Chen, C. (2024). Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. arXiv preprint arXiv:2403.01779

Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., & Wen, F. (2023). Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18381–18391).

Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P. (2020). Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7850–7859).

Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3836–3847).

Zheng, N., Song, X., Chen, Z., Hu, L., Cao, D., & Nie, L. (2019). Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 266–274).

Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., & Kemelmacher-Shlizerman, I. (2023a). Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4606–4615).

Zhu, S., Urtasun, R., Fidler, S., Lin, D., & Change Loy, C. (2017). Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE international conference on computer vision* (pp. 1680–1688).

Zhu, Z., Feng, X., Chen, D., Bao, J., Wang, L., Chen, Y., Yuan, L., & Hua, G. (2023b). Designing a better asymmetric vqgan for stablediffusion. arXiv preprint arXiv:2306.04632

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.