Cascaded Diffusion Models for Virtual Try-On: Improving Control and Resolution

Guangyuan Li[1*], Yongkang Wang[1*], Junsheng Luan[1], Lei Zhao [1†], Wei Xing [1†], Huaizhong Lin [1†], Binkai Ou[2],

[1]College of Computer Science and Technology, Zhejiang University
[2] Innovation Research & Development, BoardWare Information System Limited
{cslgy, wykang, cszhl, wxing, linhz}@zju.edu.cn

Abstract

Previous virtual try-on methods have employed ControlNet architecture in exemplar-based inpainting diffusion models to guide the generation of try-on images, preserving the garment's features and enhancing the realism of the generated images. While these methods have maintained the identity of the garment and improved the naturalness of the generated images, they still face the following limitations: (1) For garments with complex features, such as intricate text, patterns, and uncommon styles, they struggle to retain these detailed features in the generated try-on images. (2) They are limited to generating try-on images at a maximum resolution of 1K, which may not meet the demands of real-world scenarios, where higher resolutions might be required. To address the aforementioned issues, in this paper, we propose a Cascaded Diffusion Model for virtual try-on to enhance both image controllability and resolution. We call it CDM-VTON. Specifically, we design two diffusion models: the Multi-Conditioned Diffusion Model (MC-DM) and the Super-Resolution Diffusion Model (SR-DM). The former generates low-resolution try-on images while preserving the garment's complex features, and the latter enhances the resolution of these images. Additionally, we incorporate a multi-control integration module in the MC-DM, which injects multiple control conditions into a frozen denoising U-Net to ensure that the generated try-on images retain complex garment features. Our experimental results demonstrate that our method outperforms previous approaches in preserving garment details and generating authentic virtual try-on images, both qualitatively and quantitatively.

Introduction

Image-based virtual try-on (VTON) is an important computer vision task that aims to generate seamless photos of a person wearing target garments by inputting images of the person and the target garments. Due to its convenience and potential to provide personalized shopping experiences for e-commerce users, there is significant interest in generating realistic virtual try-on images. The key challenges VTON faces are how to naturally match the garments to the human body in various poses and gestures while maintaining the patterns and textures of the garments without distortion (Han et al. 2018; Wang et al. 2018).

Figure 1: We compare our method with diffusion-based virtual try-on methods at 2K resolution. We use the DiffBIR method to perform super-resolution reconstruction on the generated results of the comparison methods. Specifically, the generated try-on images have an initial resolution of 512×384, and after super-resolution reconstruction, the resolution of the images is increased to 2048×1536. As can be seen, compared to other methods, our approach generates 2K try-on images with superior detail retention and realism, particularly in the depiction of garment textures, where it maximally preserves their intricate details. **Best viewed when zoomed in.**

Previous image-based VTON methods (Choi et al. 2021; Ge et al. 2021; He, Song, and Xiang 2022; Lee et al. 2022; Wang et al. 2018; Xie et al. 2023) mainly relied on Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). These methods initially deform the garment image to align with the given person image, and then integrate the warped garment image with the person image in a generator for synthesis. However, GAN-based VTON methods face the following issues: (1) The deformation methods (Duchon 1977; Jaderberg et al. 2015; Li, Huang, and Loy 2019) they use

*These authors contributed equally.
†Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

cannot handle challenging poses. (2) The images they generate often lack a degree of realism and may fail to produce finer details.

Recently, diffusion models (DMs) have demonstrated outstanding performance in various visual tasks (Li et al. 2022; Kawar et al. 2022; Ruiz et al. 2023; Saharia et al. 2022). Compared to GANs, DMs demonstrate superior capability in generating images with fine-grained realism. When DMs are applied to VTON tasks (Baldrati et al. 2023; Gou et al. 2023; Yang et al. 2023), a key challenge is ensuring the controllability of the generated results, particularly in maintaining the complex textures and patterns of the target garment. To control the generated results, some methods (Kim et al. 2024; Zeng et al. 2024; Choi et al. 2024) employ the ControlNet (Zhang, Rao, and Agrawala 2023) architecture to provide more garment-agnostic person representations as control conditions. While these methods can generate try-on results that preserve garment identity, they struggle to effectively retain complex features in the generated images for garments with intricate features, such as detailed text, patterns, or uncommon styles. In addition, they can only generate try-on images up to a 1K resolution, whereas higher resolutions, such as 2K, may be required for practical applications.

To cope with the aforementioned issues, We propose a Cascaded Diffusion Model for Virtual Try-On (CDM-VTON), which significantly enhances the controllability of generated results while also generating high-resolution (HR) try-on images. Specifically, our proposed model consists of the Multi-Conditioned Diffusion Model (MC-DM) and the Super-Resolution Diffusion Model (SR-DM). The MC-DM generates low-resolution (LR) try-on results with complex garment features, and the SR-DM upsamples these generated LR images to HR images. To enhance the controllability of the generated try-on results, we design a Multi-control Integration Module (MIM) within the MC-DM, which injects multiple control conditions into the frozen denoising U-Net. To improve the resolution of the generated images, we fine-tune a pre-trained diffusion model using the ControlNet architecture, obtaining the SR-DM specifically employing for VTON tasks. Inspired by (Zeng et al. 2024), we further integrate DINO-V2 (Oquab et al. 2023) into ControlNet to extract more detailed feature information, thereby enhancing the control over content generation in try-on images. The main contributions are summarized as follows:

(1) We propose CDM-VTON, a novel virtual try-on model, which can improve both the controllability and resolution of the generated try-on images.

(2) We design the Multi-Conditioned Diffusion Model, which utilizes the multi-control integration module to inject multiple control conditions into the frozen denoising U-Net. This model ensures that the generated try-on images contain the complex features of the garments.

(3) We introduce the Super-Resolution Diffusion Model specifically for VTON tasks to improve the resolution of the generated try-on images, catering to practical applications.

(4) Experimental results indicate that our method outperforms previous approaches in both preserving garment details and generating realistic virtual try-on images.

Related Works

Image-based Virtual Try-On

Image-based virtual try-on aims to integrate a given garment image onto a target person, generating realistic try-on images. Research approaches for this task mainly fall into two categories: GAN-based methods (Choi et al. 2021; Ge et al. 2021; Dong et al. 2022; He, Song, and Xiang 2022; Lee et al. 2022; Wang et al. 2018; Xie et al. 2023) and DM-based methods (Baldrati et al. 2023; Gou et al. 2023; Yang et al. 2023). GAN-based models typically first warp the target garment and then synthesize the try-on image conditioned on the warped garment and the person's image. However, for complex or atypical human poses, GAN-based methods struggle to generate satisfactory try-on images. Currently, several DM-based methods leverage the powerful generative capabilities of diffusion models to address the aforementioned issues. For instance, PBE (Yang et al. 2023) employs a robust diffusion model that can semantically alter image content based on example images. LaDI-VTON (Morelli et al. 2023) utilizes textual inversion within the diffusion model to map the visual features of garments into the CLIP embedding space. DCI-VTON (Gou et al. 2023) uses the warped garment as a conditional input to the diffusion model to better preserve the garment's features. However, the above DM-based methods fail to effectively control the texture and pattern of the generated garments, leading to distorted try-on images.

Adding Conditional Control

To achieve fine-grained control in image synthesis, some methods incorporate various conditional controls into text-to-image (T2I) diffusion models (Zhang et al. 2024). For instance, ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2024) propose fine-tuning additional modules that encode spatial information, such as edges, depth, and human poses, to control the diffusion model and generate the desired images. IP-Adapter (Ye et al. 2023) introduces conditional T2I diffusion models to control image generation with both textual and visual prompts. Some studies have applied the aforementioned techniques to VTON tasks to achieve controllable generation of try-on images. For example, StableVITON (Kim et al. 2024) and CAT-DM (Zeng et al. 2024) employ the ControlNet architecture to learn the semantic correspondence between garments and human bodies within the latent space of pre-trained diffusion models in an end-to-end manner. IDM-VTON (Choi et al. 2024) utilizes two different modules to encode the semantics of garment images. While these methods ensure the controllability of generated images, they struggle to effectively preserve the complex features of garments. Furthermore, they are limited to generating try-on results with a maximum resolution of 1K, restricting their applicability in real scenarios.

Diffusion-based Super-Resolution

Currently, diffusion-based super-resolution (SR) methods (Lin et al. 2024; Wang et al. 2024; Sun et al. 2024; Rao et al.

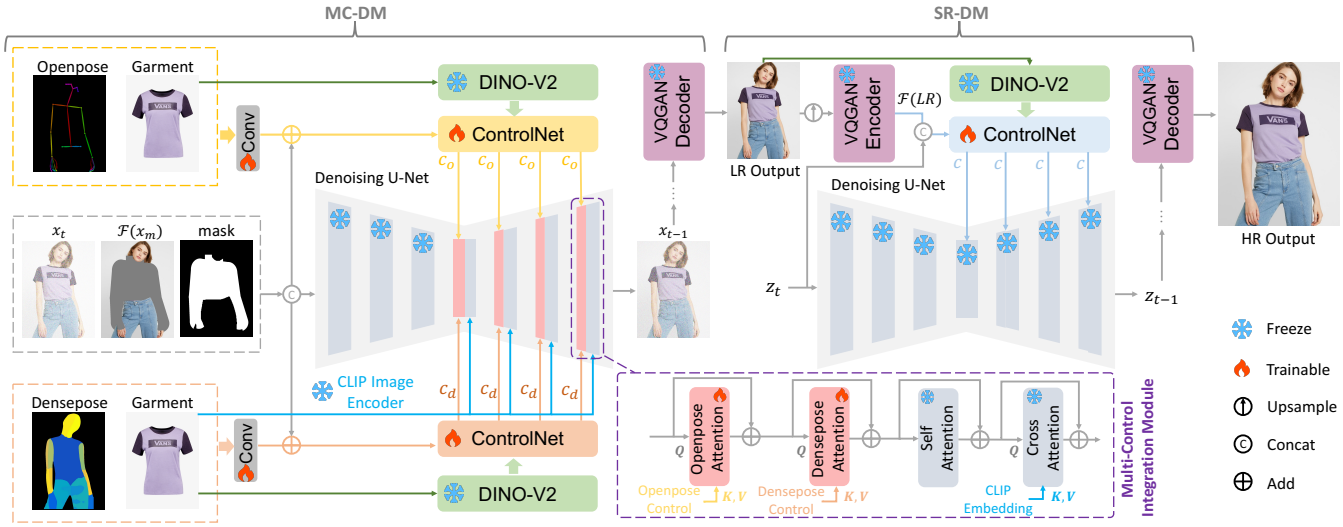

Figure 2: The overall framework of CDM-VTON consists of two latent diffusion models: MC-DM and SR-DM. MC-DM primarily comprises a frozen pre-trained diffusion model and two trainable ControlNets. SR-DM mainly consists of a frozen pre-trained diffusion model and a trainable ControlNet. We replace CLIP with DINO-V2 as the feature extractor for image conditions in ControlNet. Additionally, to better integrate control conditions into the frozen pre-trained diffusion model, we design a Multi-Control Integration Module, which includes trainable Openpose Attention, trainable Densepose Attention, as well as frozen Self-Attention and frozen Cross-Attention.

2025; Li et al. 2024) primarily focus on leveraging diffusion models (DMs) to enhance the quality of the final SR image. For example, StableSR (Wang et al. 2024) is a pioneering work in this field, utilizing the prior knowledge of pre-trained DMs to improve the fidelity of SR images. DiffBIR (Lin et al. 2024) combines traditional image restoration models with pre-trained text-to-image DMs, mitigating the adverse effects of image degradation on the reconstruction process. DWTrans (Li et al. 2023) first employs the pre-trained DMs to generate reference images, then uses these generated reference images to guide the reconstruction of the low-resolution (LR) image. CoSeR (Sun et al. 2024) can extract cognitive embeddings from LR images and enhance reconstruction results using implicit diffusion priors. Inspired by the aforementioned methods, we utilize the ControlNet architecture to fine-tune pre-trained DMs to develop a SR model specifically for the VTON task, aimed at enhancing the resolution of generated try-on images.

Methodology

Overall Architecture

To preserve the complex features of garment and enhance the resolution of generated try-on images, we propose the Cascaded Diffusion Model (CDM-VTON). It consists of MC-DM, a novel multi-conditioned diffusion model designed to improve the controllability of garment content in virtual try-on, and SR-DM, a super-resolution conditional diffusion model aimed at increasing the resolution of try-on images. The overall architecture of CDM-VTON is shown in Fig. 2. In MC-DM, the frozen denoising U-Net takes the noisy x_t, the latent of the masked-out person image $\mathcal{F}(x_m)$, and the mask as inputs. Besides the given x_t, $\mathcal{F}(x_m)$, and the mask, ControlNet generates two sets of control vectors, openpose c_o and densepose c_d, by incorporating additional control conditions (garment and openpose/densepose). We design a Multi-Control Integration Module (MIM) to better integrate these control vectors into the frozen denoising U-Net. After t iterations and decoding, MC-DM outputs low-resolution (LR) try-on images with complex garment features. In SR-DM, the frozen denoising U-Net takes the noisy z_t as input, and the trainable ControlNet takes the latent of the LR try-on image as input. The goal of training the ControlNet is to preserve the fidelity of the input image. After t iterations and decoding, SR-DM outputs the high-resolution (HR) try-on image with high realism. The following sections will provide a detailed explanation of MC-DM and SR-DM.

Multi-Conditioned Diffusion Model

MC-DM employs the ControlNet architecture based on the pre-trained Stable Diffusion (SD) model (Rombach et al. 2022), retaining the generative capabilities of the SD while incorporating additional control conditions. To better control and preserve the complex features of the garments in the generated try-on images, we employ two ControlNets. The first takes openpose and garment as inputs and outputs the control variable c_o. The second takes densepose and garment as inputs and outputs the control variable c_d. Intuitively, the control variables c_o and c_d can provide different and effective conditional information to the frozen U-Net. To more effectively integrate these control variables, we design the Multi-Control Integration Module (MIM), which consists of trainable attention modules and frozen attention modules. The specific designs will be elaborated below.

Stable Diffusion Model. The Stable Diffusion model is a large-scale diffusion model trained on the LAION-5B (Schuhmann et al. 2022) dataset, based on the latent diffusion model. Using a frozen encoder E to convert the input image x into latent features $z_0 = E(x)$, we can define a forward diffusion process in the latent space:

$$q\left(\mathbf{z}_t \mid \mathbf{z}_0\right) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \tag{1}$$

where $t \in [1, ..., T]$, T is the number of steps, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \Pi_{s=1}^{t}\alpha_s$, β_t is variance schedule. The objective function used during the training of the Stable Diffusion model is as follows:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\mathbf{y},\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, t, \tau_\theta(\mathbf{y})\right)\right\|_2^2\right], \tag{2}$$

where $\epsilon_\theta(\cdot)$ represents the denoising network, $\tau_\theta(\cdot)$ is the CLIP (Radford et al. 2021) text/image encoder to condition the text/image prompt $\mathbf{y}$.

ControlNet Architecture. Although the SD model is a large-scale diffusion model trained on LAION-5B, it is not directly applicable to virtual try-on tasks. Therefore, we introduce the ControlNet architecture to control the SD model, ensuring that the generated try-on image maintain pixel consistency with the target garment. As shown in Fig. 2, we employ two trainable ControlNets to provide different control variables, c_o and c_d, to the frozen denoising U-Net. Given a set of conditions, including noise x_t, the latent of the masked image $\mathcal{F}(x_m)$, the mask m, the time step t, the garment image g, and additional control conditions (such as openpose o or densepose d), ControlNet first generates the corresponding control variables c_o or c_d. These vectors are then integrated into the frozen denoising U-Net by the Multi-Control Integration Module (MIM) to guide the try-on image generation process. We lock all parameters of the SD model and copy the parameters of the SD Encoder blocks and SD Middle block into ControlNet. During training, we specifically perform gradient updates on the parameters of ControlNet and the openpose attention and densepose attention in MIM, Similar to Eq. 2:

$$\mathbb{E}_{x,\mathcal{F}(x_m),m,g,d,o,t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, \mathcal{F}(x_m), m, g, d, o, t\right)\right\|_2^2\right]. \tag{3}$$

Multi-Control Integration Module. Current virtual try-on methods using the ControlNet architecture simply integrate control variables into a frozen denoising U-Net. This fusion approach fails to fully utilize the control information provided by ControlNet, resulting in generated try-on images that fail to preserve the complex features of the garment. To address this issue, we design a Multi-Control Injection Module (MIM) that integrates control variables c_o and c_d into the denoising U-Net using an attention mechanism. As shown in Fig. 2, the MIM module enhances the original attention modules in the denoising U-Net by introducing trainable openpose attention and densepose attention, while keeping the self-attention and cross-attention components in a frozen state. The benefit of this structure is that it enhances all attention modules applied in the middle and decoder of the frozen denoising U-Net. Specifically, for openpose attention, the query Q is derived from the latent features of the denoising U-Net, while K and V are sourced from the openpose control c_o. Similarly, for densepose attention, the keys K and values V are derived from the densepose control c_d. For the frozen cross-attention, we utilize the CLIP Embedding as K and V.

Super-Resolution Diffusion Model

Current virtual try-on methods can only generate try-on images with a maximum resolution of 1K, whereas practical application scenarios may require higher resolution images. The most straightforward approach is to upsample the generated try-on images using a super-resolution (SR) model. However, directly applying existing SR methods (*e.g.*, StableSR (Wang et al. 2024), DiffBIR (Lin et al. 2024)) may result in distorted upsampled images, as they lack prior knowledge about the person and the garments. Therefore, we introduce a super-resolution diffusion model (SR-DM) specifically designed for virtual try-on, as shown in Fig. 2. Inspired by IRControlNet (Lin et al. 2024), SR-DM also employs the ControlNet architecture to control image generation. The difference is that to enable pixel-level controllability in SR-DM, we utilize DINO-V2 (Oquab et al. 2023) as the feature extractor for the LR images in ControlNet. Compared to CLIP (Radford et al. 2021), DINO-V2 encodes images not only as global tokens but also as patch tokens, which helps retain information from the LR images and provides detailed representations (Zeng et al. 2024). We integrate the LR image features extracted by DINO-V2 into ControlNet using fully connected layers and cross-attention.

During training, only the parameters of ControlNet will be updated. We aim to minimize the following latent diffusion objective:

$$\mathbb{E}_{z_t,c,t,\epsilon,\mathcal{F}(LR)}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c, t, \mathcal{F}(LR)\right)\right\|_2^2\right], \tag{4}$$

where $\mathcal{F}(LR)$ refers to the latent features extracted by the VQGAN Encoder. During inference, the generated LR try-on image is used as a conditional input to the model. After t iterations, the model generates the HR try-on image with high fidelity that preserves the garment's detailed features.

Experiments

Experiments Setting

Datasets. We employ two publicly available datasets, DressCode (Morelli et al. 2022) and VITON-HD (Choi et al. 2021), to evaluate the virtual try-on task. Both datasets consist of paired images of garments and their corresponding human models wearing the garments. The DressCode dataset is categorized into three classes: upper body, lower body, and dresses. The testing experiments are conducted under two settings: paired and unpaired. In the paired setting, the input garment image and the garment worn by the human model are the same item. Conversely, the human model tries on different garment in the unpaired setting.

Figure 3: Qualitative comparison of our proposed MC-DM with other baseline methods. **Left**: Comparison results on VITON-HD (Choi et al. 2021). **Right**: Comparison results on DressCode (Morelli et al. 2022). **Best viewed when zoomed in.**

Figure 4: Qualitative comparison of our proposed method with other baseline methods at 2K resolution using the VITON-HD (Choi et al. 2021) and DressCode (Morelli et al. 2022). For baseline methods, we employ DiffBIR (Lin et al. 2024) for SR reconstruction. **Best viewed when zoomed in**.

Baselines. To demonstrate the superiority of our proposed method, we compare our method with a GAN-based virtual try-on method: HR-VITON (Lee et al. 2022), a diffusion-based inpainting method: Paint-by-Example (PBE) (Yang et al. 2023), and five diffusion-based virtual try-on methods: LaDI-VTON (Morelli et al. 2023), DCI-VTON (Gou et al. 2023), StableVITON (Kim et al. 2024), CAT-DM (Zeng et al. 2024), and IDM-VTON (Choi et al. 2024). For a fair comparison, all methods are evaluated on images with a resolution of 512×384. For assessing the results of super-resolution, we use DiffBIR (Lin et al. 2024) to perform 2× or 4× upsampling on the try-on images generated by the aforementioned comparison methods.

Evaluation Metrics. We conduct quantitative evaluations under both unpaired and paired settings. Specifically, in the unpaired setting, we employ the Fréchet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018) to evaluate the realism of the generated results. In the paired setting, where ground truth is available, we use the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) and Structural Similarity Index Measure (SSIM) (Wang et al. 2004) to assess the quality of the generated images.

Implementation Details. In the experiment, we train MC-DM and SR-DM separately. Specifically, MC-DM is conducted using two NVIDIA A6000 (48GB) GPUs with image resolutions of 512×384. We use the AdamW optimizer with a learning rate set to $2e$-5. SR-DM is trained on two NVIDIA A100 (80GB) GPUs, employing the AdamW optimizer with a learning rate of $5e$-5. In MC-DM, we use Paint-by-Example (Yang et al. 2023) as the frozen pre-trained diffusion model. In SR-DM, we use IRControlNet (Lin et al. 2024) as the frozen pre-trained diffusion model.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
HR-VITON	0.878	0.1045	11.265	2.73
PBE	0.802	0.1428	11.939	3.85
LaDI-VTON	0.864	0.0964	9.480	1.99
DCI-VTON	0.880	0.0804	8.754	1.10
StableVITON	<u>0.888</u>	<u>0.0732</u>	8.233	**0.49**
CAT-DM	0.877	0.0803	8.933	1.37
IDM-VTON	0.872	0.1021	<u>6.292</u>	1.02
Ours	**0.896**	**0.0636**	**6.062**	<u>0.58</u>

Table 1: Quantitative results on VITON-HD (Choi et al. 2021). **Bold** and <u>underline</u> denote the best and the second best result, respectively.

Qualitative Results

We conduct a qualitative analysis under the unpaired setting. Fig. 3 shows our method's capability in generating

Method	DressCode-Upper				DressCode-Lower				DressCode-Dresses			
	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
PBE	0.872	0.1209	20.32	7.01	0.804	0.2108	24.95	7.36	0.761	0.2516	31.25	19.09
LaDI-VTON	0.915	0.0649	17.40	5.92	**0.910**	**0.0596**	17.90	5.45	0.854	**0.1076**	16.13	4.76
CAT-DM	0.927	0.0507	12.62	1.89	0.902	0.0621	14.83	2.82	0.863	0.1091	14.30	3.36
IDM-VTON	0.921	0.0625	8.62	1.53	0.897	0.0813	12.37	2.09	0.856	0.1219	11.36	2.57
Ours	**0.933**	**0.0494**	**8.48**	**1.10**	**0.910**	0.0761	**11.84**	**1.52**	**0.864**	0.1090	**10.92**	**2.19**

Table 2: Quantitative results on DressCode (Morelli et al. 2022). **Bold** and underline denote the best and the second best result, respectively.

Figure 5: Qualitative results of different variant models on the VITON-HD (Choi et al. 2021). **Best viewed when zoomed in.**

Figure 6: Qualitative results of MC-DM employing different super-resolution models on the VITON-HD (Choi et al. 2021). **Best viewed when zoomed in.**

Figure 7: Qualitative comparisons on in-the-wild image. We show generated virtual try-on images using our proposed method compared with other methods. Our method outperforms other baseline methods in generating authentic images and preserving fine-grained details of garments. **Best viewed when zoomed in.**

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
StableVITON*	**0.852**	0.0927	6.085	0.42
CAT-DM*	0.830	0.1008	6.321	0.75
IDM-VTON*	0.834	0.0984	5.034	0.61
MC-DM*	0.847	**0.0912**	**4.859**	**0.41**
StableVITON+	**0.854**	0.0871	5.650	0.39
CAT-DM+	0.838	0.0959	5.910	0.64
IDM-VTON+	0.840	0.0914	4.632	0.49
MC-DM+	0.851	**0.0857**	**4.386**	**0.37**

Table 3: Quantitative results for the VITON-HD (Choi et al. 2021) using DiffBIR (*) (Lin et al. 2024) and SR-DM (+). **Bold** and underline denote the best and the second best result, respectively.

controllable try-on images. Compared to baseline methods, our proposed MC-DM generates realistic images and effectively preserves the original text, texture, and style of the garments. Specifically, for the VITON-HD dataset, our MC-DM method retains the text and patterns on the garments more effectively than other diffusion-based virtual try-on methods. For the DressCode dataset, our method better preserves garment styles, such as cuffs and pant legs, in the generated try-on images. This indicates that our designed MC-DM effectively controls the generation of try-on images and preserves complex detailed features about the garments. To evaluate the performance of super-resolution, we employ DiffBIR (Lin et al. 2024) to perform 4× upsampling

Variant	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
Bare Model	0.802	0.1428	11.939	3.85
w/ c_o	0.849	0.0931	8.887	1.46
w/ c_d	0.851	0.0925	8.864	1.44
w/ c_o & *w/* c_d	<u>0.878</u>	<u>0.0804</u>	<u>7.533</u>	<u>1.02</u>
Full Model	**0.896**	**0.0636**	**6.062**	**0.58**

Table 4: Quantitative results of different variant models on the VITON-HD (Choi et al. 2021). **Bold** and <u>underline</u> denote the best and the second best result, respectively.

Method	SSIM ↑	LPIPS ↓	FID ↓	KID ↓
MC-DM+StableSR	0.845	<u>0.0908</u>	4.915	0.46
MC-DM+DiffBIR	<u>0.847</u>	0.0912	4.859	<u>0.41</u>
MC-DM+SUPIR	0.833	0.0945	<u>4.712</u>	0.43
MC-DM+SR-DM	**0.851**	**0.0857**	**4.386**	**0.37**

Table 5: Quantitative results of MC-DM employing different super-resolution models on the VITON-HD (Choi et al. 2021). **Bold** and <u>underline</u> denote the best and the second best result, respectively.

(512×384 to 2048×1536) on the try-on images generated by baseline methods to achieve a 2K resolution, as shown in Fig. 4. As can be seen, the 2K resolution try-on images generated by our method effectively restore the original detailed information of the garments. This is attributed to (1) MC-DM can generate try-on images that preserve the original garment details, and (2) our fine-tuning of the super-resolution model specifically for the virtual try-on task.

Quantitative Results

Quantitative analysis is conducted under both the paired setting (evaluated using SSIM and LPIPS) and the unpaired setting (evaluated using FID and KID). As shown in Tab. 1 and Tab. 2, our proposed method outperforms other methods on most metrics, demonstrating its effectiveness in generating high-quality and realistic try-on images. Tab. 3 reports the quantitative metrics of various methods after reconstruction using different SR techniques. Here, we perform a 2× upsampling (from 512×384 to 1024×768) to compute SSIM and LPIPS metrics. As shown, our MC-DM exhibits excellent performance across different SR methods, indicating that the try-on images generated by MC-DM are highly controllable. Furthermore, compared to the results reconstructed using DiffBIR, the results reconstructed with our SR-DM show further improvement, demonstrating that SR-DM can further enhance the quality and realism of the try-on images.

Ablation Study

In this section, we conduct ablation studies on the various modules of the proposed method using the VITON-HD (Choi et al. 2021) dataset.

Effect of MC-DM. To validate the role of the multi-control in MC-DM, we design five variant models: (1) without any control, *i.e.*, all ControlNet branches and the MIM module are removed, named as Bare Model; (2) using only the openpose ControlNet branch without the MIM module, termed as *w/* c_o; (3) using only the densepose ControlNet branch without the MIM module, named as *w/* c_d; (4) using both the openpose and densepose ControlNet branches without the MIM module, named as *w/* c_o & *w/* c_d; (5) the Full Model, which includes both the openpose and densepose ControlNet branches along with the MIM module. The results are shown in Fig. 5 and Tab. 4. As can be seen, the Bare Model fails to effectively preserve the garment patterns. Although *w/* c_o and *w/* c_d can control the generation of try-on images based on the garments, they cannot capture the fine details of the garments. Compared to *w/* c_o & *w/* c_d, the Full Model, which includes the MIM module, generates try-on images that effectively retain the details and style of the garments. This indicates that the MIM module can successfully integrate control variables into the frozen network.

Effect of SR-DM. To validate the effectiveness of SR-DM, we replace SR-DM with another blind SR model, StableSR (Wang et al. 2024), DiffBIR (Lin et al. 2024), and SUPIR (Yu et al. 2024), to perform 4× upsampling on the results generated by MC-DM. The results are shown in Fig. 6 and Tab. 5. We observe that DiffBIR fails to effectively reconstruct the text and texture of the garments because it lacks prior knowledge related to garments or human models. This results in distortion of the garments in the generated try-on images, as indicated by the red arrows in Fig. 6. In contrast, SR-DM is fine-tuned using a large number of garment and human model images and employs DINO-V2 as the feature extractor for ControlNet. Consequently, SR-DM can effectively reconstruct the detailed information of the garments. Previous study (Zeng et al. 2024) has demonstrated that using DINO-V2 as the extractor in ControlNet outperforms alternatives like CLIP (Radford et al. 2021), IP-Adapter (Ye et al. 2023), and SeeCoder (Xu et al. 2024).

Results on In-the-Wild Image

We evaluate our method on the challenging in-the-wild images, comparing it with other diffusion-based VTON methods, StableVITON and IDM-VTON. Fig. 7 shows the results of customizing a pair of garment and human model images. As can be seen, compared to other methods, our method not only preserves the garment patterns but also accurately generates complex details, such as text and designs on the garments, which demonstrates that our method is well-suited for application in real-world scenarios.

Conclusion

In this paper, we propose a Cascaded Diffusion Model for virtual try-on, which consists of the Multi-Conditioned Diffusion Model (MC-DM) and the Super-Resolution Diffusion Model (SR-DM). The MC-DM utilizes a multi-control integration module to generate try-on images that preserve the garment's identity. The SR-DM enhances the resolution of the try-on images to meet the requirements of practical applications. Extensive experiments across various datasets demonstrate that our method outperforms previous approaches in both preserving garment details and generating high-resolution try-on images.

Acknowledgments

This work was supported in part by the Zhejiang Province Program (2024C01110, 2022C01222, 2023C03199, 2023C03201), the National Program of China (62172365, 2021YFF0900604, 19ZDA197), the Macau project on key technology research and display system development for new personalized controllable dressing dynamic display, the Ningbo Science and Technology Plan Project (2022Z167, 2023Z137), and the MOE Frontier Science Center for Brain Science & Brain-Machine Integration (Zhejiang University).

References

Baldrati, A.; Morelli, D.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23393–23402.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*.

Choi, S.; Park, S.; Lee, M.; and 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*, 14131–14140.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*.

Dong, J.; Chen, X.; Zhang, M.; Yang, X.; Chen, S.; Li, X.; and Wang, X. 2022. Partially Relevant Video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 246–257.

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

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

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

Gou, J.; Sun, S.; Zhang, J.; Si, J.; Qian, C.; and Zhang, L. 2023. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7599–7607.

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

He, S.; Song, Y.-Z.; and 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*, 3470–3479.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35: 23593–23606.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8176–8185.

Lee, S.; Gu, G.; Park, S.; Choi, S.; and Choo, J. 2022. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, 204–219. Springer.

Li, G.; Rao, C.; Mo, J.; Zhang, Z.; Xing, W.; and Zhao, L. 2024. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11365–11374.

Li, G.; Xing, W.; Zhao, L.; Lan, Z.; Sun, J.; Zhang, Z.; Zhang, Q.; Lin, H.; and Lin, Z. 2023. Self-reference image super-resolution via pre-trained diffusion large model and window adjustable transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7981–7992.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Li, Y.; Huang, C.; and Loy, C. C. 2019. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3693–3702.

Lin, X.; He, J.; Chen, Z.; Lyu, Z.; Dai, B.; Yu, F.; Ouyang, W.; Qiao, Y.; and Dong, C. 2024. DiffBIR: Towards Blind Image Restoration with Generative Diffusion Prior. arXiv:2308.15070.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8580–8589.

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and 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*, 2231–2235.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out

more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, C.; Li, G.; Lan, Z.; Sun, J.; Luan, J.; Xing, W.; Zhao, L.; Lin, H.; Dong, J.; and Zhang, D. 2025. Rethinking Video Deblurring with Wavelet-Aware Dynamic Transformer and Diffusion Model. In *European Conference on Computer Vision*, 421–437. Springer.

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

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and 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*, 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.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sun, H.; Li, W.; Liu, J.; Chen, H.; Pei, R.; Zou, X.; Yan, Y.; and Yang, Y. 2024. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25868–25878.

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

Wang, J.; Yue, Z.; Zhou, S.; Chan, K. C.; and Loy, C. C. 2024. Exploiting Diffusion Prior for Real-World Image Super-Resolution. *International Journal of Computer Vision*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; and 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*, 23550–23559.

Xu, X.; Guo, J.; Wang, Z.; Huang, G.; Essa, I.; and Shi, H. 2024. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8682–8692.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and 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*, 18381–18391.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25669–25680.

Zeng, J.; Song, D.; Nie, W.; Tian, H.; Wang, T.; and Liu, A.-A. 2024. CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8372–8382.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Zhang, Q.; Xing, W.; Li, G.; Zhao, L.; Sun, J.; Lan, Z.; Luan, J.; Huang, Y.; and Lin, H. 2024. ArtBank: Artistic Style Transfer with Pre-trained Diffusion Model and Implicit Style Prompt Bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7396–7404.