CAT-DM: Controllable Accelerated Virtual Try-on with Diffusion Model

Jianhao Zeng[1] Dan Song[1*] Weizhi Nie[1] Hongshuo Tian[1]
Tongtong Wang[2] An-An Liu[1*]
[1] Tianjin University [2] Tencent LightSpeed Studio

Figure 1. CAT-DM not only enhances the controllability of the image generation process for virtual try-on but also effectively accelerates the sampling speed of the diffusion models. **Top:** Comparison results with other methods. CAT-DM accurately generates the pattern details on garments and produces images that are sufficiently clear. **Bottom:** CAT-DM requires fewer sampling steps than other diffusion models to generate clear and realistic virtual try-on images. Compared to the default 50 sampling steps of DCI-VTON [8], CAT-DM achieves a 25-fold acceleration.

Abstract

Generative Adversarial Networks (GANs) dominate the research field in image-based virtual try-on, but have not resolved problems such as unnatural deformation of garments and the blurry generation quality. While the generative quality of diffusion models is impressive, achieving controllability poses a significant challenge when applying it to virtual try-on and multiple denoising iterations limit its potential for real-time applications. In this paper, we propose ***C****ontrollable* ***A****ccelerated virtual* ***T****ry-on with* ***D****iffusion* ***M****odel (CAT-DM). To enhance the controllability, a basic diffusion-based virtual try-on network is designed, which utilizes ControlNet to introduce additional control conditions and improves the feature extraction of garment images. In terms of acceleration, CAT-DM initiates a reverse denoising process with an implicit distribution generated by a pre-trained GAN-based model. Compared with previous try-on methods based on diffusion models, CAT-DM not only retains the pattern and texture details of the in-shop garment but also reduces the sampling steps without compromising generation quality. Extensive experiments demonstrate the superiority of CAT-DM against both GAN-based and diffusion-based methods in producing more realistic images and accurately reproducing garment patterns.*

1. Introduction

Image-based virtual try-on has emerged as a prominent and popular research topic within the field of AIGC, particularly focusing on conditional person image generation. By taking a person image and a target garment image as inputs, the objective of this task is to generate a photo of the person seamlessly wearing the desired garment. Requirements are expected in both aspects of person and clothes: 1) the posture and identity such as face and skin should be the same as the person; 2) target garment is naturally warped and seamlessly put on the body without losing characteristics such as

*Corresponding Author: Dan Song (dan.song@tju.edu.cn), An-An Liu (anan0422@gmail.com). The code is available at https://github.com/zengjianhao/CAT-DM.

†The research is supported in part by the National Natural Science Foundation of China (U21B2024, 62232337, 61902277).

pattern and texture.

Existing image-based virtual try-on methods [4, 7, 9–11, 17, 22, 38, 44, 50] primarily rely on Generative Adversarial Networks (GANs) [40]. GAN-based try-on methods typically start by warping the in-shop garment image to match the given person image, then combine the warped image with the person image into a generator for synthesis. However, existing garment deformation approaches such as TPS [5], STN [14] and FlowNet [20] are not flexible enough to deal with challeging poses (Fig. 1). Additionally, images produced by GAN-based methods often lack a degree of realism and may fail to generate finer details.

Recently, diffusion models have garnered widespread applications in the field of image generation, demonstrating outstanding performance across various tasks such as super-resolution [18], image restoration [15] and text-guided image generation [30, 31]. Compared with GANs, diffusion models demonstrate enhanced stability [16] during training and excel in producing images with fine-grained realism such as clearer hands and arms. However, when applying diffusion models to virtual try-on task [1, 8, 24, 43], the controllability of the generated results, particularly preserving complex textures and patterns of target garment, remains challenging (Fig. 1). Furthermore, the generation of high-fidelity images via diffusion models requires a considerable number of sampling steps, thereby limiting their application in real-time virtual try-on scenarios.

To enhance the controllability of diffusion models, we propose a Garment-Conditioned Diffusion Model (GC-DM). This model utilizes the ControlNet [47] architecture to provide more garment-agnostic person representations as control conditions. On the other hand, GC-DM improves the feature extraction of garment images, providing the model with more detailed garment information to better control the pattern generation in the try-on images. In addition, to ensure that the image areas outside of the garment region remain unchanged, GC-DM employs Poisson blending [26] to seamlessly integrate the original person images with the generated try-on images.

Building upon the foundation of GC-DM, we further propose a truncation-based acceleration strategy to accelerate the inference speed of diffusion models. Inspired by truncated diffusion probabilistic model (TDPM) [49], we employ an implicit distribution to provide initial samples for the reverse denoising process, instead of using the Gaussian noise as the starting point for reverse denoising, thereby significantly reducing the sampling steps. Unlike TDPM which learns an implicit distribution, we utilize a pre-trained GAN-based model to generate an initial try-on image, and then add noise to this image to obtain the implicit distribution.

In summary, the proposed CAT-DM is equipped with GC-DM, a new diffusion-based virtual try-on model, and a truncation-based acceleration strategy initialized by the coarse image generated by a pre-trained GAN. As shown in Fig. 1, CAT-DM capitalizes on the robust generative capabilities of diffusion models, as well as the controllability of GAN-based models, meanwhile significantly reducing the number of required sampling steps. Extensive experiments validate the effectiveness of each component of CAT-DM, which achieves state-of-the-art results on two widely used benchmarks [4, 23].

The main contributions of our work are: (1) We propose CAT-DM, a virtual try-on model, to create high-fidelity images using fewer sampling steps. (2) We propose GC-DM to improve the controllability of diffusion models by offering additional control conditions and enhancing the extraction of garment image features. (3) We introduce a truncation-based acceleration strategy to synthesize the advantages of both GAN-based models and diffusion models, and reduce the sampling steps.

2. Related Work

2.1. Image-based Virtual Try-On

A recent survey [34] has comprehensively reviewed previous SOTA methods in image-based virtual try-on. GAN-based models play a dominant role in this research area, which usually warp target clothes first and then synthesize try-on images conditioned with warped clothes and person image. Different strategies have been tried for clothing warping, such as thin-plate-spline (TPS) interpolation [9], spatial transformation network (STN) [19] and flow estimation [50]. However, as shown in Fig. 1, complex or unconventional postures are still challenging for GAN-based methods. In terms of generation quality, some high-resolution methods are explored, e.g., VITON-HD [4] introduces a misalignment-aware normalization and HR-VTON [17] further refines it by predicting both segmentation and flow simultaneously. During the training process, GANs [40] usually encounter issues such as mode collapse [16].

Diffusion models significantly improve the realism of generated images, and several approaches [3, 21, 42, 51] apply this advanced model to image-based virtual try-on. PBE [43] is a robust diffusion model for image generation, capable of semantically altering image content based on exemplar image. MGD [1] uses the multimodal data to guide the generation of fashion images. However, as shown in Fig. 1, the controllability of generated contents is not well addressed. Later LaDI-VTON [24] uses a textual inversion component to map the visual features of garments to the CLIP token embedding space. DCI-VTON [8] pastes the warped garment to the input of the diffusion model as the local condition to better retain the characteristics of the garments. Although the controllability is improved, diffusion-

based methods still suffer from redundant sampling steps. Therefore, in this paper we propose a controllable and accelerated model to both enhance the generation quality and speed.

2.2. Diffusion Models

Diffusion models are a class of generative models that learn the target distribution through an iterative denoising process. Compared to GANs, diffusion models can generate images with more intricate details, resulting in higher quality outputs. Denoising diffusion probabilistic models (DDPMs) [13] consist of a Markovian forward process that gradually corrupts the data sample $\mathbf{x}_0$ into the Gaussian noise $\mathbf{x}_T$, and a learnable reverse process that converts $\mathbf{x}_T$ back to $\mathbf{x}_0$ iteratively. Assuming the parameter of the diffusion model is denoted as θ, the training process of the diffusion model can be described as:

$$\mathbb{E}_{\mathbf{x}_0,t,\epsilon}\left[\left\|\epsilon-\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\epsilon,t)\right\|_2^2\right], \quad (1)$$

where $\mathbf{x}_0$ is from the truth data distribution, t follows a uniform distribution over the set $\{1,2,...,T\}$, $\epsilon \sim \mathcal{N}(0,\mathbf{I})$ represents randomly generated noise and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$ is a pre-defined variance schedule in t steps. Therefore, $\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\epsilon$ represents the noisy image after adding noise to $\mathbf{x}_0$. The diffusion model is trained by predicting the added noise ϵ from the noisy image.

After the diffusion model is trained, diffusion models can be used to synthesize new images by taking a random noise sample $\mathbf{x}_T \sim \mathcal{N}(0,\mathbf{I})$ and denoise it for $1 \le t \le T$ iteratively:

$$\mathbf{x}_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t,t)\right)+\sigma_t\mathbf{z}, \quad (2)$$

where $\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})$ and $\sigma_t^2=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$. DDPMs are designed to make the denoising process closely resemble a Gaussian distribution. To achieve this, they often set the parameter T to a large value, typically around 1000. As a result, DDPMs require a sequence of 1000 noise prediction steps for image generation.

2.3. Accelerated Sampling

While diffusion models can produce realistic images, the necessity for multiple sampling steps to generate a single image limits their application in real-time virtual try-on scenarios. Denoising diffusion implicit models (DDIMs) [35] introduce a non-Markovian diffusion process that accelerates sampling without the need for additional training. This approach has been widely adopted in the field. Progressive distillation [32] is a highly effective approach for boosting the sampling rate of a diffusion model through repeated distillation. However, this repeated distillation comes with considerable training expenses. Although consistency model [36] can achieve accelerated sampling by employing only one distillation, its performance is often inadequate for many scenarios [33], which hampers its broader practical application. TDPM [49] shortens the diffusion trajectory by learning an implicit distribution at step T_{trunc}, which initiates the reverse diffusion process. However, it is hard to learn the corrupted data at step T_{trunc}. On the other hand, these methods of accelerating sampling inevitably compromise the model's performance to some extent and are unable to resolve the issue of limited controllability in diffusion models. Unlike directly predicting the implicit distribution at step T_{trunc}, CAT-DM first generates an initial try-on image using a pre-trained GAN-based model, and then obtains the data distribution at step T_{trunc}, by adding noise. Our approach not only achieves accelerated sampling but also enhances the controllability of diffusion models.

3. Method

To address the issues faced by diffusion models in virtual try-on tasks, such as the loss of garment pattern details and the necessity for numerous sampling steps, we propose CAT-DM. It consists of GC-DM, a novel garment-conditioned diffusion model designed for virtual try-on to enhance the controllability, and the truncation-based acceleration strategy with a pre-trained GAN-based model as initialization for acceleration.

3.1. Garment-Conditioned Diffusion Model

The GC-DM employs a ControlNet [47] architecture, which introduces additional control conditions while preserving the generative capabilities of PBE[43]. It enhances the feature extractor to provide more detailed information about garment, thus improving control over the generation of garment areas. For areas outside of garment, GC-DM uses a Poisson blending [26] to ensure that the original person information remains unchanged. The specific designs will be elaborated below.

ControlNet architecture. Diffusion models excel in image generation but require substantial computational resources and GPU memory due to a large number of parameters, limiting their application and advancement. Additionally, they mainly provide semantic-level control, and improving this controllability without retraining billion-parameter model is challenging.

As shown in Fig. 2, GC-DM consists of PBE with locked-parameters and a trainable ControlNet. PBE is a robust image generation model capable of semantically altering image content based on exemplar image, and trained on millions of images. This model utilizes an U-Net [29] architecture, comprising multiple SD Encoder Blocks, sev-

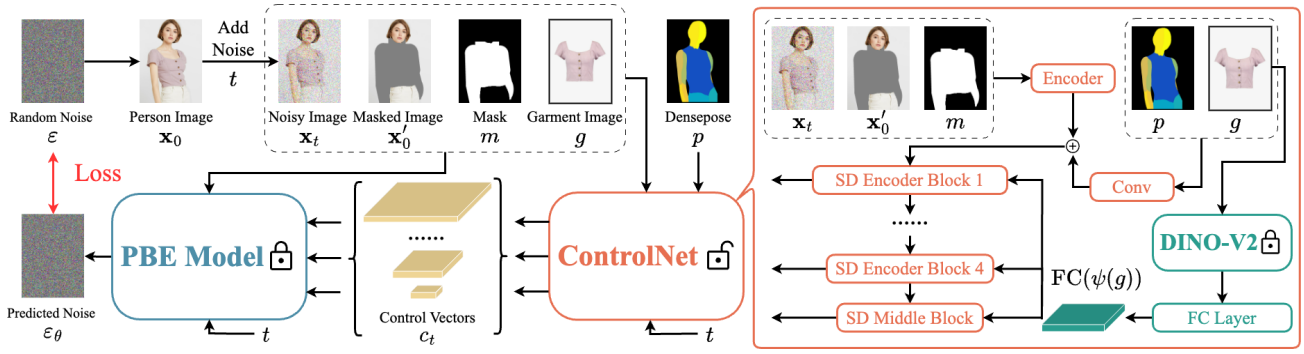

Figure 2. The training pipeline of the GC-DM in our method. GC-DM comprises a fixed-parameter PBE and a trainable ControlNet. Apart from the given noisy image $\mathbf{x}_t$, time steps t, mask m, masked image $\mathbf{x}_0'$ and garment image g, ControlNet generates a set of control vectors c_t by incorporating additional control conditions, such as densepose p. Control vectors are incorporated into the PBE to enhance the model's controllability while preserving the PBE's generative capabilities.

eral SD Decoder Blocks, and an SD Middle Block. The SD Encoder Blocks and SD Decoder Blocks are interconnected via skip-connections. However, PBE is not directly suited for the virtual try-on task, as this task requires the generated try-on images to remain pixel-consistent with the target garment images. Additionally, since PBE has hundreds of millions of parameters, the training of which requires significant computational resources.

We lock all parameters of PBE and copy the parameters of SD Encoder Blocks and SD Middle Block to ControlNet. During the training process, we perform gradient updates exclusively on the parameters of ControlNet. This approach accelerates training and conserves GPU memory, thereby reducing the diffusion model's demand for computational resources.

Given a person image $\mathbf{x}_0$, GC-DM progressively add noise to the image and produces a noisy image $\mathbf{x}_t$, with t being how many times the noise is added. Given a set of conditions including noisy image $\mathbf{x}_t$, time steps t, mask m, masked image $\mathbf{x}_0'$, garment image g as well as additional control conditions (such as densepose p), ControlNet generates a set of control vectors c_t. These vectors are integrated into the skip-connections and the SD Middle Block of PBE's U-Net architecture, thereby directing the generation process of the PBE. Similar to Eq. (1), GC-DM learns a network ϵ_θ to predict the noise added to the noisy image $\mathbf{x}_t$ with:

$$\mathbb{E}_{\mathbf{x}_0,\mathbf{x}_0',m,g,p,t,\epsilon}\left[\|\epsilon-\epsilon_\theta(\mathbf{x}_t,\mathbf{x}_0',m,g,p,t)\|_2^2\right] \quad (3)$$

By introducing more garment-agnostic person representations as control conditions through ControlNet, GC-DM can not only preserve the generative capabilities of the PBE but also enhance the controllability of the diffusion model.

Garment feature extraction. Although PBE is capable of generating realistic images based on example images, its lack of pixel-level control often results in inaccurate reconstruction of patterns on garment images in the virtual try-on task (as shown in Fig. 1). The underlying issue stems from PBE's use of CLIP [27] as an encoder ψ to extract feature information from garment images. While CLIP can align image information with corresponding textual descriptions in a shared space, it falls short in the virtual try-on task. The semantic information extracted by CLIP is insufficient to accurately describe the patterns and texture details of garment images.

To grant GC-DM pixel-level controllability, we employ DINO-V2 [25] as the feature extractor ψ for garment images g in ControlNet. Unlike CLIP, DINO-V2 not only encodes images into global tokens but also into patch tokens. This approach helps in preserving more pixel information of garment images g, offering a detailed representation. Additionally, we implement a fully connected layer (FC) to encode garment features $\psi(g)$ into the space where U-Net resides. Subsequently, these features FC($\psi(g)$) are integrated into U-Net through a cross-attention mechanism [37]. By enhancing GC-DM's feature extraction capabilities for garment images, we improve its controllability at the pixel level.

Poisson blending. PBE [43], as a Latent Diffusion Model (LDM)[28], uses pre-trained autoencoders[6] to convert images into latent space, thereby reducing computational demands. However, this conversion can cause pixel precision loss during image reconstruction, especially in complex images like faces, leading to noticeable differences from the original, as shown in Fig. 4.

In virtual try-on applications, we typically desire to replace the garment within a designated mask region, while keeping the area outside the mask unchanged. A straightforward approach is to concatenate the input image with the generated image using the mask. However, as shown in Fig. 4, this method can result in noticeable discontinuities at the junction of the two images.

To address the aforementioned issues, we adapt Poisson

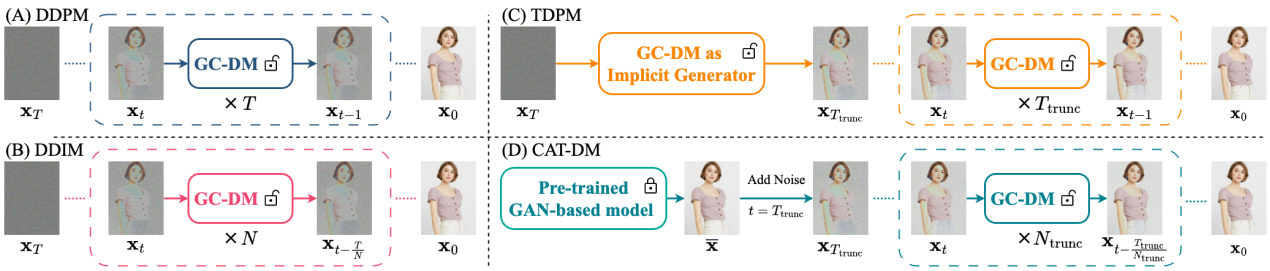

Figure 3. Illustration of different sampling methods in diffusion models. (A) The conventional DDPMs [13] denoise gradually with a large number of time steps T. (B) DDIMs [35] employ a class of non-Markovian diffusion processes to denoise gradually. Compared to DDPMs, DDIMs requires fewer sampling steps, that is, $N \ll T$. (C) TDPM [49] repurposes the parameter of the diffusion model to generate the implicit distribution at step T_{trunc}, using it as the initial sample for the reverse diffusion process. This approach accelerates sampling, resulting in $T_{\text{trunc}} \ll T$. (D) CAT-DM utilizes a pre-trained GAN-based model to generate an initial try-on image $\bar{\mathbf{x}}$, which is then subjected to noise addition, making the noisy image $\mathbf{x}_{T_{\text{trunc}}}$ as the starting point of the reverse diffusion process.

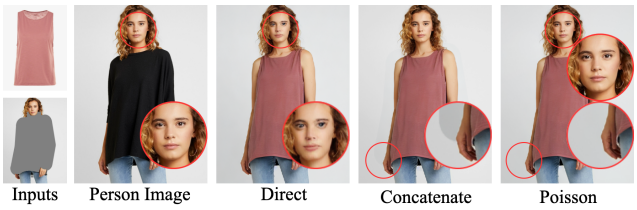

Figure 4. Results from different types of generation methods. The directly generated try-on images exhibit noticeable distortion in the face region. The results obtained through image concatenating have incongruities at the junctions of the images. This issue is resolved in the results obtained using Poisson blending.

blending [26] to seamlessly integrate the input image with the generated image. In particular, given a directly generated virtual try-on image f^*, the original person image h, and the non-clothing region Ω. The blending image f should satisfy the following equation:

$$\begin{cases} |N_p| f_p - \sum_{q \in N_p} f_q = \sum_{q \in N_p} v_{pq} & p \in \Omega \\ f_p = f_p^* & p \in \partial\Omega \end{cases} \tag{4}$$

where N_p is the set of 4-connected neighbors [46] for pixel p, $|N_p|$ denotes the number of pixels in the set N_p and $\partial\Omega$ represents of the boundary around Ω. The difference between pixel p and its neighboring pixel q is denoted as $v_{pq} = h_p - h_q$.

As shown in Fig. 4, the adoption of Poisson blending not only ensures that the areas outside the garment region remain unchanged, but also resolves the traces caused by image stitching.

3.2. Truncation-Based Acceleration Strategy

As shown in Fig. 3, for the conventional DDPMs [13], when the number of denoising steps T is reduced, the true denoising distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is not approximate to a Gaussian distribution and usually intractable. Although denoising diffusion implicit models (DDIMs) [35] introduce a non-Markovian diffusion process to accelerate sampling, it still requires dozens of samples to generate high-quality images. Hence, significant computational resources are needed even for inference with a trained model, limiting the research and application of diffusion models. On the other hand, we have observed that diffusion models, when generating patterns or text on garment, exhibit a clear disadvantage compared to GANs.

Also shown in Fig. 3, the core idea of TDPM [49] is to repurpose the diffusion model as an implicit generator to generate the starting point of the reverse diffusion chain. Compared to iteratively denoising starting from Gaussian-distributed noise, the reverse diffusion chain in TDPM is much shorter, i.e., $T_{\text{trunc}} \ll T$. However, the truncated chain of TDPM has an unknown corrupted data distribution at step T_{trunc}, and it is often challenging to learn the implicit distribution $\mathbf{x}_{T_{\text{trunc}}}$ at step T_{trunc} merely by reusing the parameters of the diffusion model. Besides, TDPM does not hold an advantage in terms of generation quality, especially when faced with garments containing complex patterns or text, i.e., the issue of uncontrollable generation still persists.

We introduce a truncation-based acceleration strategy to reduce the sample steps while effectively addressing the issue of uncontrollable generation of complex patterns. As shown in Fig. 3, CAT-DM consists of GC-DM and a pre-trained GAN-based virtual try-on model, integrated through the truncation-based acceleration strategy. It utilizes a pre-trained GAN-based model to generate the initial try-on image $\bar{\mathbf{x}}$. We achieve the implicit distribution $\mathbf{x}_{T_{\text{trunc}}}$ at step T_{trunc} for this image by adding noise through the following equation:

$$\mathbf{x}_{T_{\text{trunc}}} = \sqrt{\bar{\alpha}_{T_{\text{trunc}}}}\bar{\mathbf{x}} + \sqrt{1 - \bar{\alpha}_{T_{\text{trunc}}}}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$

In CAT-DM, $\mathbf{x}_{T_{\text{trunc}}}$ serves as the starting point of the

Method	DressCode-Upper						DressCode-Lower						DressCode-Dresses					
	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$
PBE [43]	20.32	7.01	18.79	6.64	0.872	0.1209	24.95	7.36	22.44	6.78	0.804	0.2108	31.25	19.09	30.04	18.44	0.761	0.2516
MGD [1]	17.30	5.11	15.03	5.54	0.912	0.0624	16.76	4.04	13.67	3.79	0.893	0.0689	15.11	3.36	12.14	2.41	0.844	0.1195
LaDI-VTON [24]	17.40	5.92	14.91	6.01	0.915	0.0649	17.90	5.45	13.76	4.61	**0.910**	**0.0596**	16.13	4.76	13.00	4.05	0.854	**0.1076**
GC-DM	**12.62**	**1.89**	**9.85**	**2.38**	**0.927**	**0.0507**	**14.83**	**2.82**	**10.25**	**1.81**	0.902	0.0621	**14.30**	**3.36**	**10.71**	**2.02**	**0.863**	0.1091

Table 1. Quantitative results on DressCode [23]. The subscripts 'u' and 'p' respectively represent the unpaired setting and paired setting.

reverse diffusion chain, followed by iterative denoising of the noisy image $\mathbf{x}_{T_{\text{trunc}}}$ via GC-DM. Unlike TDPM, we use DDIMs as sampler for generating high-quality samples more rapidly. The training process is consistent with Eq. (3), with the sole difference being that t no longer follows a uniform distribution over $\{1, 2, ..., T\}$, but rather over $\{1, 2, ..., T_{\text{trunc}}\}$.

By adjusting the size of T_{trunc} within CAT-DM, we can control the contribution ratio of the pre-trained GAN and GC-DM to the final generated image. Generally speaking, a larger T_{trunc} results in a greater influence of GC-DM on the final image, while a smaller T_{trunc} makes the final image lean more towards the result generated by the pre-trained GAN-based model.

4. Experiments

4.1. Experiments Setting

Datasets: In this work, we focus on evaluating virtual try-on tasks using two popular datasets: DressCode [23] and VITON-HD [4]. Both datasets contain high-resolution paired images of in-shop garments and their corresponding human models wearing the garments. The DressCode dataset includes three categories: upper-body, lower-body, and dresses. Test experiments are conducted under both paired and unpaired settings. In the paired setting, the input garment images and the garment worn by the model are the same item. Conversely, in the unpaired setting, a different garment is selected for the virtual try-on task.

Evaluation Metrics: To assess our model quantitatively, we use evaluation metrics to evaluate the coherence and realism of the generated output. In the paired and unpaired settings, we employ the Fréchet Inception Distance (FID) [12] and the Kernel Inception Distance (KID) [2] to evaluate the realism of the generated output. Furthermore, in the paired setting with available ground truth, we additionally employ the Learned Perceptual Image Patch Similarity (LPIPS) [48] and the Structural Similarity Index Measure (SSIM) [39] to evaluate the coherence of the generated image.

Implementation Details: During the experiments, we use an end-to-end training process. All experiments are conducted using two NVIDIA GeForce RTX 4090 GPUs with image resolutions of 512×384. We use the AdamW optimizer, set the learning rate to 2×10^{-5}.

Method	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$
VITON-HD [4]	14.64	6.10	12.81	5.52	0.848	0.1216
HR-VITON [17]	12.15	3.42	9.92	3.06	0.860	0.1038
GP-VTON [50]	10.49	2.23	7.71	2.01	0.857	0.0897
PBE [43]	15.77	6.22	14.32	5.44	0.763	0.2254
MGD [1]	13.34	3.93	11.12	3.38	0.827	0.1280
LaDI-VTON [24]	12.33	4.75	9.44	3.90	0.861	0.0968
DCI-VTON [8]	11.14	3.35	8.19	2.93	0.875	0.0816
GC-DM	9.67	**1.36**	7.11	1.12	0.862	0.0988
CAT-DM	**8.93**	1.37	**5.60**	**0.83**	**0.877**	**0.0803**

Table 2. Quantitative results on VITON-HD [4]. The subscripts 'u' and 'p' respectively represent the unpaired setting and paired setting.

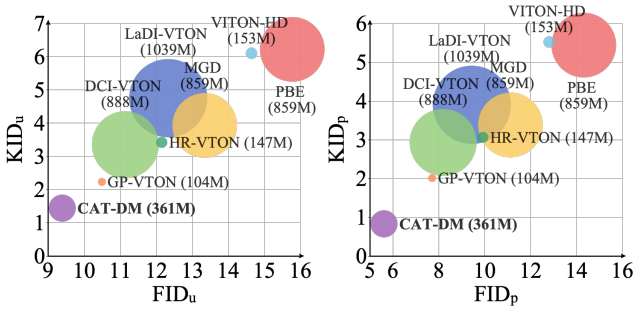

Figure 5. Comparative analysis of our method (CAT-DM) with other techniques using the VITON-HD dataset [4], focusing on the realism of results (better at the bottom left) and the number of trainable parameters (smaller is better). The unpaired setting is on the left, and the paired setting is on the right.

4.2. Quantitative Evaluation

We compare our method with previous virtual try-on methods, including GAN-based virtual try-on models such as VITON-HD [4], HR-VITON [17], and GP-VTON [50], as well as diffusion-based virtual try-on models including MGD [1], DCI-VTON [8], and LaDI-VTON [24]. Since our model utilizes PBE [43] as a locked-parameter network, we also include PBE in our comparison. The DressCode dataset was released recently, so we lack pre-trained GAN-based virtual try-on models that have been trained on the DressCode dataset. Therefore, for the quantitative results on the DressCode dataset, we only compare GC-DM instead of CAT-DM with other methods.

We employ two different methods for quantitative comparison of our model. For GC-DM, we use DDIMs sam-

Figure 6. Qualitative results. **Left**: Comparison results on VITON-HD [4]; **Right**: Comparison results on DressCode [23].

pling with the number of sampling steps set to 16. For CAT-DM, we employ a truncation-based acceleration strategy, utilize a pre-trained GAN-based model with GP-VTON, and set T_{trunc} to 100 and the number of sampling steps to 2. The generated results for both models are processed using Poisson blending.

As reported in Tab. 1 and Tab. 2, our GC-DM outperforms other methods on the majority of metrics, particularly in FID [12] and KID [2], demonstrating its effectiveness in image generation quality. CAT-DM utilizes a pre-trained GAN as a preconditioner to produce accurate and sharp images. CAT-DM significantly outperforms other methods across all metrics and, notably, can generate ideal images in just two steps, which greatly accelerates the sampling speed of diffusion models. As demonstrated in Fig. 5, compared to other diffusion models, CAT-DM also significantly reduces the number of trainable parameters while also offering a marked advantage in image generation quality.

4.3. Qualitative Evaluation

Fig. 6 displays the qualitative comparison of GC-DM and CAT-DM with the state-of-the-art baselines on VITON-HD dataset [4] and DressCode dataset [23] in the unpaired setting, respectively. Based on the test results of PBE, MGD, and LaDI-VTON, it is evident that they struggle to capture the details on the given garment images and cannot accurately reproduce the patterns on the garment. Although DCI-VTON can generate accurate garment patterns to some extent, it fails to detect changes in garment types. This leads to the residual traces of the original garment appearing in the generated virtual try-on images. GP-VTON shows commendable performance in generating accurate images and capturing details, but the resulting images contain some artifacts and lack a degree of realism. Compared to methods based on GANs, frozen CLIP and DINO-V2 benefit from large-scale datasets.

Compared to other diffusion methods, GC-DM shows advantages in both accuracy and realism of generation. The CAT-DM, created by integrating GP-VTON and GC-DM using the truncation-based acceleration strategy, not only rectifies the artifacts present in GP-VTON but also retains the generative capabilities of GC-DM. More importantly, in contrast to other diffusion-based virtual try-on models, CAT-DM can produce high-quality virtual try-on images in just two steps.

Extractor	Process	$FID_u \downarrow$	$KID_u \downarrow$	$FID_p \downarrow$	$KID_p \downarrow$	$SSIM_p \uparrow$	$LPIPS_p \downarrow$
DINO-V2 [25]	Direct Generation	10.76	2.53	8.25	2.09	0.835	0.1069
	Concatenation	10.57	2.59	8.18	2.42	0.854	0.1033
	Poisson Blending	**9.67**	**1.36**	**7.11**	**1.12**	**0.862**	**0.0988**
CLIP [27]	Poisson Blending	10.21	1.77	7.90	1.38	0.853	0.1111
IP-Adapter [45]	Poisson Blending	11.23	3.90	8.13	2.86	0.847	0.1127
SeeCoder [41]	Poisson Blending	9.94	1.66	7.13	1.58	0.856	0.1049

Table 3. Discussion about extractors and Poisson blending.

Figure 7. Visual ablation on garment feature extraction.

4.4. Discussion

Garment feature extraction: We explore the key factors for the garment feature extractor. We compare the results of GC-DM when using CLIP [27], DINO-V2 [25], IP-Adapter [45] and SeeCoder [41] as the garment feature extractor respectively. As reported in Tab. 3, for the VITON-HD dataset [4], with the integration of DINO-V2 [25], GC-DM has shown improvement across all metrics. As shown in Fig. 7, the GC-DM, utilizing DINO-V2 as a garment feature extractor, is capable of generating more accurate and realistic virtual try-on images. This demonstrates that DINO-V2 can enhance the model's capability to extract features from garment images, thereby also boosting the con-

trollability of the diffusion model.

Poisson blending: We examine the impact of various processing approaches on the quality of the generated images. As reported in Tab. 3, for the VITON-HD dataset [4], compared to using the frozen encoder-decoder of LDMs [28] to generate virtual try-on images directly, concatenating together the input person image with the generated try-on image can indeed improve the quality of the resulting image. However, the seams at the point of stitching can still impact the image quality. Employing Poisson blending can eliminate such issues, resulting in more realistic virtual try-on images.

Refinement function of the diffusion model: The diffusion model can be taken as a refined module. When the pre-trained GAN-based method generate a try-on result with over-distorted warped garment, diffusion model can adjust it. As shown in Fig. 8, the try-on images generated by GP-VTON lack an arm on one side, but CAT-DM is capable of rectifying it.

Pre-trained GAN-based model: We explore the impact of different GAN-based models on CAT-DM's performance and compare it with GC-DM, which does not use GAN-based models. The experimental results are shown in Fig. 9. For the VITON-HD dataset [4], the three dashed lines respectively represent the original performance of the three GAN-based methods: VITON-HD, HR-VITON, and GP-VTON. Although the performance of GC-DM surpasses that of all GAN-based models when the number of sampling steps is sufficient, the performance of GC-DM significantly degrades when the number of sampling steps is inadequate. CAT-DM leverages the rapid generation capabilities of GANs to significantly reduce the need for numerous sampling steps. Compared to GC-DM, CAT-DM avoids performance degradation when the number of sampling steps is low. Additionally, CAT-DM achieves higher performance compared to the GAN-based models it utilizes. Furthermore, we note that the performance of CAT-DM is, to a certain extent, reliant on the performance of GAN-based models. When the number of sampling steps is sufficient, CAT-DM, utilizing GP-VTON as its pre-trained GAN-based model, not only surpasses GP-VTON but also outperforms GC-DM.

Truncation step: As shown in Fig. 10, we conducte experiments with different truncation settings of $T_{\text{Trunc}} = 0$, 50, 100, 150 and 1000. On one hand, when T_{Trunc} is set to 0, CAT-DM and GP-VTON are essentially the same model. On the other hand, when T_{Trunc} is set to 1000, CAT-DM and GC-DM become identical models. When T_{Trunc} is set to 50, 100, and 150, we observe that the model tends to perform best when the number of sampling steps is 2.

Figure 8. CAT-DM can refine and adjust the try-on results generated by pre-trained GAN-based methods.

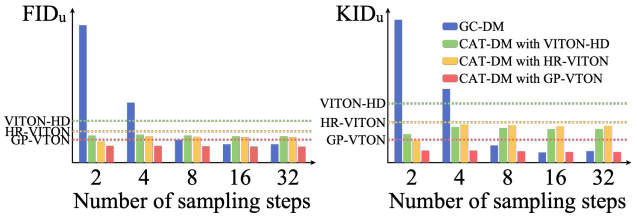

Figure 9. Discussion about the pre-trained GAN-based model. The bar chart represents the performance of GC-DM and CAT-DM using different GAN-based models across various sampling step counts, while the dashed lines indicate the performance of the different GAN-based models.

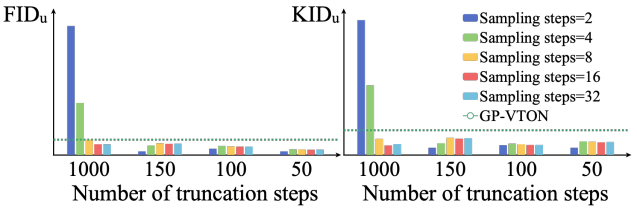

Figure 10. Discussion about the truncation step. The bar chart represents the performance of GC-DM and CAT-DM using different truncation steps across various sampling step counts.

5. Conclusion

To enhance the controllability of diffusion models in virtual try-on tasks and accelerate the sampling speed of these models, we introduce the CAT-DM. It combines a specially designed try-on model, GC-DM, with a pre-trained GAN model, utilizing an innovative truncation-based acceleration strategy. Specifically, to enhance the generation of detailed garment textures, GC-DM improves the feature extraction from garment images. Additionally, by adopting the ControlNet architecture, GC-DM introduces extra control conditions, thereby increasing the controllability of the diffusion model. To accelerate the sampling speed of diffusion models, CAT-DM initiates a reverse denoising process with an implicit distribution generated by a pre-trained GAN-based model. A substantial number of experiments demonstrate the superiority of our method in terms of image quality, controllability, and sampling speed. The limitation of our CAT-DM will be discussed in the supplement.

References

[1] Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Multimodal garment designer: Human-centric latent diffusion models for fashion image editing. *arXiv preprint arXiv:2304.02051*, 2023. 2, 6

[2] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. 6, 7

[3] Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, and Wen-Huang Cheng. 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*, pages 7513–7522, 2023. 2

[4] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 2, 6, 7, 8

[5] Jean Duchon. 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*, pages 85–100. Springer, 1977. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 4

[7] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 2

[8] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. *arXiv preprint arXiv:2308.06101*, 2023. 1, 2, 6

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[10] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.

[11] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479, 2022. 2

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6, 7

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015. 2

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 2

[16] Youssef Kossale, Mohammed Airaj, and Aziz Darouichi. Mode collapse in generative adversarial networks: An overview. In *2022 8th International Conference on Optimization and Applications (ICOA)*, pages 1–6. IEEE, 2022. 2

[17] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. *arXiv preprint arXiv:2206.14180*, 2022. 2, 6

[18] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[19] Kedan Li, Min Jin Chong, Jeffrey Zhang, and Jingen Liu. Toward accurate and realistic outfits visualization with attention to details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15546–15555, 2021. 2

[20] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 2

[21] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22788–22797, 2023. 2

[22] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, pages 10–14, 2020. 2

[23] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2231–2235, 2022. 2, 6, 7

[24] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023. 2, 6

[25] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4, 7

[26] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 577–582. Unknown, 2023. 2, 3, 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 7

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 8

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[32] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. 3

[33] Shitong Shao, Xu Dai, Shouyi Yin, Lujun Li, Huanran Chen, and Yang Hu. Catch-up distillation: You only need to train once for accelerating sampling. *arXiv preprint arXiv:2305.10769*, 2023. 3

[34] Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Mohan Kankanhalli Tong, Ruofeng, and An-An Liu. Image-based virtual try-on: A survey. *arXiv preprint arXiv:2311.04811*, 2023. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 5

[36] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 3

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[38] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 2

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[40] Mehdi Mirza Bing Xu-David Warde and Farley Sherjil Ozair Aaron Courville Ian. J. goodfellow, jean pouget-abadie and yoshua bengio. generative adversarial nets, 2014. 2

[41] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023. 7

[42] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17194–17204, 2023. 2

[43] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2, 3, 4, 6

[44] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 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*, pages 7850–7859, 2020. 2

[45] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 7

[46] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 231–240, 2020. 5

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[49] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models and diffusion-based adversarial auto-encoders. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 5

[50] Xie Zhenyu, Huang Zaiyu, Dong Xin, Zhao Fuwei, Dong Haoye, Zhang Xijin, Zhu Feida, and Liang Xiaodan. 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 (CVPR)*, 2023. 2, 6

[51] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira

Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 2