OOTDiffusion: Outfitting Fusion Based Latent Diffusion for Controllable Virtual Try-On

Yuhao Xu, Tao Gu, Weifeng Chen, Arlene Chen

Xiao-i Research
{levihsu958, gutao2049, weifeng.chen1024}@gmail.com, arlenecc@xiaoi.com

Abstract

We present OOTDiffusion, a novel network architecture for realistic and controllable image-based virtual try-on (VTON). We leverage the power of pretrained latent diffusion models, designing an outfitting UNet to learn the detailed garment features. Without a redundant warping process, the garment features are precisely aligned with the target human body via the proposed outfitting fusion in the self-attention layers of the denoising UNet. In order to further enhance the controllability, we introduce outfitting dropout to the training process, which enables us to adjust the strength of the garment features through classifier-free guidance. Our comprehensive experiments on the VITON-HD and Dress Code datasets demonstrate that OOTDiffusion efficiently generates high-quality try-on results for arbitrary human and garment images, which outperforms other VTON methods in both realism and controllability, indicating a breakthrough in virtual try-on.

Code — https://github.com/levihsu/OOTDiffusion

Introduction

Image-based virtual try-on (VTON) is a popular and promising image-synthesis technology for the current e-commerce industry, which is able to dramatically improve the shopping experience of consumers and reduce the advertising cost of clothing merchants. As its name suggests, the VTON task aims to generate an outfitted image of a target human wearing a given garment. It has taken tremendous efforts from numerous researchers (Han et al. 2018; Wang et al. 2018; Han et al. 2019; Dong et al. 2019; Minar et al. 2020; Yang et al. 2020; Ge et al. 2021b,a; He, Song, and Xiang 2022; Xie et al. 2023; Kim et al. 2023) for more natural and accurate virtual try-on results in the past few years.

Image-based VTON is currently facing two main challenges. First, the generated images should be realistic and natural enough to avoid dissonance. Most of recent researches on virtual try-on leverage generative adversarial networks (Goodfellow et al. 2020) (GANs) or latent diffusion models (Rombach et al. 2022) (LDMs) for image generation. Previous GAN-based methods (Han et al. 2018, 2019; Choi et al. 2021; He, Song, and Xiang 2022; Lee et al. 2022; Xie et al. 2023) usually have difficulty in generating correct garment folds, natural light and shadow, or realistic human bodies. Hence more recent work favors LDM-based methods (Zhu et al. 2023; Morelli et al. 2023; Gou et al. 2023; Kim et al. 2023), which effectively improve the realism of outfitted images. The second critical challenge is how to preserve as much as possible the detailed garment features, such as complicated text, textures, colors, patterns and lines, etc. Previous researches (Choi et al. 2021; Lee et al. 2022; Morelli et al. 2023; Xie et al. 2023; Gou et al. 2023) usually perform an explicit warping process to align the garment features with the target human body, and then feed the warped garment into generative models (i.e., GANs and LDMs, etc.). Thus the performance of such VTON methods is extremely dependent on the efficacy of the independent warping process which is prone to overfitting the training data. On the other hand, some LDM-based methods (Morelli et al. 2023; Gou et al. 2023; Kim et al. 2023) attempt to learn garment features via CLIP textual-inversion (Gal et al. 2022), which fail to preserve fine-grained garment details.

Figure 1: Outfitted images (1024 × 768) generated by our OOTDiffusion trained on the VITON-HD (Choi et al. 2021) (1st row; supporting upper-body garments) and Dress Code (Morelli et al. 2022) (2nd row; supporting upper-body garments, lower-body garments and dresses) datasets, with various input human and garment images. Please zoom in for more details.

Motivated by the aforementioned prospects and challenges of image-based VTON, we present a novel LDM-based virtual try-on method, namely Outfitting over Try-on Diffusion (OOTDiffusion; see Figure 2). First, we make full use of the advantages of pretrained latent diffusion models (Rombach et al. 2022) to ensure high realism of generated images and natural try-on effects, and design an out-

fitting UNet to learn the detailed features of garments in the latent space in a single step. Then, we propose an outfitting fusion process to precisely align the garment features with the noisy human body in the self-attention layers (Vaswani et al. 2017) of the denoising UNet. In this way, the garment features are smoothly adapted to various target human body types and postures, without suffering information loss or feature distortion caused by an independent warping process. Furthermore, we perform an outfitting dropout operation, randomly dropping a handful of garment latents in training to enable classifier-free guidance (Ho and Salimans 2022) with respect to the garment features. Through this approach, the strength of garment control over the generated result can be simply adjusted by a guidance scale, which further enhances the controllability of our VTON method.

Our contributions are summarized as follows:

- We present OOTDiffusion, an LDM-based network architecture with a novel outfitting UNet for realistic and controllable virtual try-on.
- We propose outfitting fusion to efficiently align the garment features with the target human body in the self-attention layers without redundant warping.
- We introduce outfitting dropout to the training process, which further improves the controllability of the outfitting UNet.
- We train our OOTDiffusion on two broadly-used high-resolution benchmark datasets, i.e., VITON-HD (Choi et al. 2021) and Dress Code (Morelli et al. 2022), respectively. Extensive qualitative and quantitative evaluations demonstrate our superiority over the state-of-the-art VTON methods in both realism and controllability for various target human and garment images (see Figure 1), implying an impressive breakthrough in image-based virtual try-on.

Related Work

Image-based Virtual Try-on. Image-based virtual try-on has been investigated for many years as a promising and challenging task (Zhu et al. 2023; Wang et al. 2018; Issenhuth, Mary, and Calauzenes 2020; Han et al. 2018; Fenocchi et al. 2022; Gou et al. 2023; Morelli et al. 2022; Xie et al. 2023; Kim et al. 2023; Lee et al. 2022; Morelli et al. 2023; Choi et al. 2021). Aiming at more natural and accurate results, recent researches are mainly based on generative adversarial networks (Goodfellow et al. 2020) (GANs) or latent diffusion models (Rombach et al. 2022) (LDMs) for image generation. Among the GAN-based VTON methods (Choi et al. 2021; Lee et al. 2022; Xie et al. 2023), VITON-HD (Choi et al. 2021) collected a high-resolution dataset and proposed ALIAS normalization and generator to address the misalignment between warped clothes and target regions. HR-VITON (Lee et al. 2022) simultaneously performed warping and segmentation to handle the body occlusion and garment misalignment. GP-VTON (Xie et al. 2023) proposed an LFGP warping module to generate deformed garments and introduced a DGT training strategy for the warping network. As introduced above, GAN-based methods usually rely on an explicit warping process neglecting realistic garment folds and natural light and shadow, which seriously degrades the fidelity and realism of outfitted images. Meanwhile, GAN-based methods are prone to overfitting the training data and causing severe performance degradation on out-of-distribution images.

With respect to the LDM-based approaches (Morelli et al. 2023; Gou et al. 2023; Kim et al. 2023), LaDI-VTON (Morelli et al. 2023) and DCI-VTON (Gou et al. 2023) also require an explicit warping process. In specific, LaDI-VTON (Morelli et al. 2023) performed textual-inversion to map the visual garment features to the CLIP (Radford et al. 2021) token embedding space and condition the latent diffusion model along with the warped input. DCI-VTON (Gou et al. 2023) directly combined the warped clothes with the masked person image to get a coarse result, and then refined it by the diffusion model. Neither of these methods succeeded in fully preserving garment details like complicated patterns and text due to the information loss caused by the CLIP encoder. More recently, StableVITON (Kim et al. 2023) discarded independent warping and proposed a zero cross-attention block to learn semantic correlation between the clothes and human body. However, information loss remains in the cross-attention layers, and the extra zero-initialized blocks heavily increase the training and inference cost. In contrast, again without warping, our LDM-based OOTDiffusion finetunes the pretrained outfitting UNet to learn garment details in one step and efficiently incorporates them into the denoising UNet via our outfitting fusion with negligible information loss.

LDM-based Controllable Image Generation. Latent diffusion models (Rombach et al. 2022) have achieved great success in text-to-image (Podell et al. 2023; Betker et al. 2023; Saharia et al. 2022b; Ruiz et al. 2023; Kumari et al. 2023) and image-to-image (Saharia et al. 2022a; Kawar et al. 2023; Saharia et al. 2022c; Tumanyan et al. 2023; Parmar et al. 2023) generation in recent years. For the purpose of more controllable generated results, Prompt-to-Prompt (Hertz et al. 2022) and Null-text Inversion (Mokady et al. 2023) controlled the cross-attention layers to finely edit images by modifying the input captions without extra model training. InstructPix2Pix (Brooks, Holynski, and Efros 2023) created paired data to train diffusion models that generate the edited image given an input image and a text instruction. Paint-by-Example (Yang et al. 2023) trained image-conditioned diffusion models in a self-supervised manner to offer fine-grained image control. ControlNet (Zhang, Rao, and Agrawala 2023) and T2I-Adapter (Mou et al. 2023) incorporated additional blocks into pretrained diffusion models to enable spatial conditioning controls. IP-Adapter (Ye et al. 2023) adopted a decoupled cross-attention mechanism for text and image features to enable controllable generation with image prompt and additional structural conditions. In this paper, we focus on the image-based VTON task, employing outfitting fusion in the self-attention layers of the denoising UNet and performing outfitting dropout at training time to enable latent diffusion models to generate more controllable outfitted images with respect to the garment features.

Method

Preliminary

Stable Diffusion. Our OOTDiffusion is an extension of Stable Diffusion (Rombach et al. 2022), which is one of the most commonly-used latent diffusion models. Stable Diffusion employs a variational autoencoder (Kingma and Welling 2013) (VAE) that consists of an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ to enable image representations in the latent space. And a UNet (Ronneberger, Fischer, and Brox 2015) ϵ_θ is trained to denoise a Gaussian noise ϵ with a conditioning input encoded by a CLIP text encoder (Radford et al. 2021) τ_θ. Given an image $\mathbf{x}$ and a text prompt $\mathbf{y}$, the training of the denoising UNet ϵ_θ is performed by minimizing the following loss function:

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

where $t \in \{1, ..., T\}$ denotes the time step of the forward diffusion process, and $\mathbf{z}_t$ is the encoded image $\mathcal{E}(\mathbf{x})$ with the added Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ (i.e., the noise latent). Note that the conditioning input $\tau_\theta(\mathbf{y})$ is correlated with the denoising UNet by the cross-attention mechanism (Vaswani et al. 2017).

OOTDiffusion

Overview. Figure 2 illustrates the overview of our method. Given a target human image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ and an input garment image $\mathbf{g} \in \mathbb{R}^{3\times H\times W}$, OOTDiffusion is capable of generating a realistic outfitted image $\mathbf{x_g} \in \mathbb{R}^{3\times H\times W}$. We employ OpenPose (Cao et al. 2019; Simon et al. 2017; Cao et al. 2017; Wei et al. 2016) and HumanParsing (Li et al. 2020) to generate a masked human image $\mathbf{x_m} \in \mathbb{R}^{3\times H\times W}$, and use a VAE encoder $\mathcal{E}$ to transform it into the latent space as $\mathcal{E}(\mathbf{x_m}) \in \mathbb{R}^{4\times h\times w}$, where $h = \frac{H}{8}$ and $w = \frac{W}{8}$. Then we concatenate $\mathcal{E}(\mathbf{x_m})$ with a Gaussian noise $\epsilon \in \mathbb{R}^{4\times h\times w}$ as the input latent $\mathbf{z}_T \in \mathbb{R}^{8\times h\times w}$ for the denoising UNet. Note that we add 4 zero-initialized channels to the first convolutional layer of the denoising UNet to support our input with 8 channels.

On the other side, we feed the encoded garment latent $\mathcal{E}(\mathbf{g}) \in \mathbb{R}^{4\times h\times w}$ into an (i) outfitting UNet to learn the garment features in a single step, and integrate them into the denoising UNet via our (ii) outfitting fusion. And we perform (iii) outfitting dropout for $\mathcal{E}(\mathbf{g})$ particularly in the training process. In addition, we also conduct CLIP textual-inversion (Gal et al. 2022) for the garment image $\mathbf{g}$, and optionally concatenate it with a text embedding of the garment label $\mathbf{y} \in \{\text{“}upperbody\text{”}, \text{“}lowerbody\text{”}, \text{“}dress\text{”}\}$ as an auxiliary conditioning input, which is fed into both outfitting and denoising UNets via the cross-attention mechanism (Vaswani et al. 2017). Finally, after multiple steps of the denoising process, we use a VAE decoder $\mathcal{D}$ to transform the denoised latent $\mathbf{z}_0 \in \mathbb{R}^{4\times h\times w}$ back into the image space as the output image $\mathbf{x_g} = \mathcal{D}(\mathbf{z}_0) \in \mathbb{R}^{3\times H\times W}$. We will elaborate the key technologies (i.e., (i) outfitting UNet, (ii) outfitting fusion, and (iii) outfitting dropout) of our OOTDiffusion in the following sections.

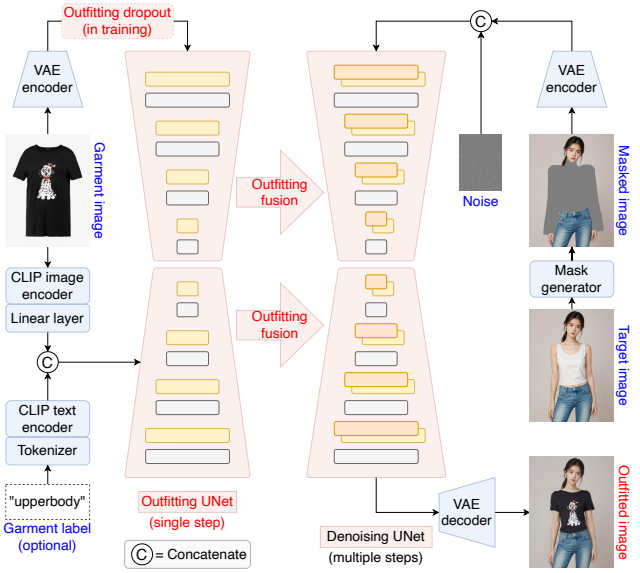

Figure 2: Overview of our proposed OOTDiffusion model. On the left side, the garment image is encoded into the latent space and fed into the outfitting UNet for a single step process. Along with the auxiliary conditioning input generated by CLIP encoders, the garment features are incorporated into the denoising UNet via outfitting fusion. Outfitting dropout is performed for the garment latents particularly in training to enable classifier-free guidance. On the right side, the input human image is masked with respect to the target region and concatenated with a Gaussian noise as the input to the denoising UNet for multiple sampling steps. After denoising, the feature map is decoded back into the image space as our try-on result.

Outfitting UNet. As introduced above, we propose an outfitting UNet to efficiently learn the detailed features of the garment image $\mathbf{g}$. The left side of Figure 2 shows the architecture of our outfitting UNet, which is essentially identical to the denoising UNet of Stable Diffusion. The encoded garment latent $\mathcal{E}(\mathbf{g}) \in \mathbb{R}^{4\times h\times w}$ is fed into the outfitting UNet $\omega_{\theta'}$, and then incoporated into the denoising UNet ϵ_θ via our outfitting fusion (see the next section). Along with the aforementioned auxiliary conditioning input, the outfitting and denoising UNets are jointly trained by minimizing the following loss function:

$$\mathcal{L}_{OOTD} = \mathbb{E}_{\mathcal{E}(\mathbf{x_m}),\mathcal{E}(\mathbf{g}),\psi,\epsilon\sim\mathcal{N}(0,1),t} \left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \omega_{\theta'}(\mathcal{E}(\mathbf{g}), \psi), \psi)\|_2^2\right], \tag{2}$$

where $\psi = \tau_g(\mathbf{g})$ ⓒ $\tau_y(\mathbf{y})$ represents the auxiliary conditioning input for both $\omega_{\theta'}$ and ϵ_θ. While τ_g and τ_y refer to the pretrained CLIP image encoder and text encoder respectively, and ⓒ denotes concatenation.

In practice, we directly duplicate the pretrained UNet weights of Stable Diffusion (Rombach et al. 2022) for the initialization of both our outfitting and denoising UNets (except for the zero-initialized channels added to the first convolutional layer), and jointly finetune them on the high-resolution VTON datasets (Choi et al. 2021; Morelli et al. 2022). Note that $\omega_{\theta'}$ and ϵ_θ do not share any weights in the

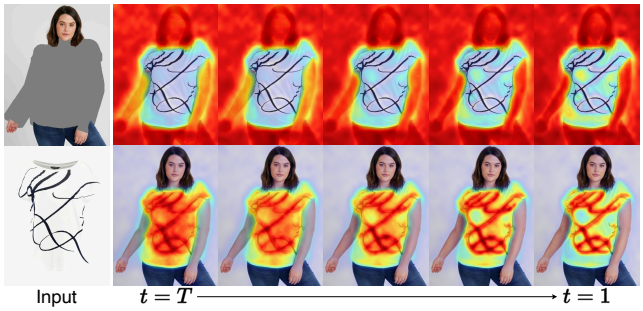

Figure 3: Visualization of the attention maps with respect to the human body (1st row) and garment features (2nd row) aligned by our outfitting fusion.

training process. We claim that our tactical utilization of the pretrained models dramatically improves the training efficiency and reduces the training cost. Moreover, compared with the denoising UNet, a significant difference in our outfitting UNet is that it requires only one step forward process before the multiple denoising steps in inference, causing a minimal amount of extra computational cost to the original Stable Diffusion (Rombach et al. 2022).

Outfitting Fusion. Based on our proposed outfitting UNet and inspired by the spatial-attention mechanism (Vaswani et al. 2017; Hu et al. 2023), we propose an outfitting fusion process to incorporate the learned garment features into the denoising UNet. First, we dive into the transformer blocks (Vaswani et al. 2017) of two UNets, finding each pair of feature maps used as the input to the corresponding self-attention layers (Vaswani et al. 2017). Given the nth pair of the feature maps $\mathbf{g}_n, \mathbf{x}_n \in \mathbb{R}^{c_n \times h_n \times w_n}$, we concatenate them in the spatial domain as:

$$\mathbf{x}_{\mathbf{g}_n} = \mathbf{x}_n \copyright \mathbf{g}_n \in \mathbb{R}^{c_n \times h_n \times 2w_n}. \tag{3}$$

And we replace $\mathbf{x}_n$ with the concatenated feature map $\mathbf{x}_{\mathbf{g}_n}$ as the input to the self-attention layer of the denoising UNet. Then we crop out the fist half of the output feature map as the final output of the self-attention layer. Figure 3 visualizes the attention maps learned in our modified self-attention. We observe that the unmasked region focuses attention on the human body (1st row), and the masked pixels are attracted to the garment features (2nd row). Meanwhile, during the denoising process, the attention to the human body gradually includes part of the masked region like the neck and arms, and the attention to the garment features gradually increases in the region of the complicated patterns. Through outfitting fusion in the self-attention layers, the garment features are implicitly warped and effectively correlated with the target human body with negligible information loss. Hence the denoising UNet is made capable of learning the precise features from the outfitting UNet for preserving garment details and naturally adapting them to the target human body in the generated image.

Outfitting Dropout. In order to further enhance the controllability of our VTON method, we employ an outfitting dropout operation in training to enable classifier-free guidance (Ho and Salimans 2022) with respect to the garment features. Classifier-free guidance has been broadly used in conditional image generation (Nichol et al. 2021; Yu et al. 2022; Saharia et al. 2022b; Brooks, Holynski, and Efros 2023) for trading off the quality and diversity of images generated by latent diffusion models. Specifically in the training process of our outfitting UNet, we randomly drop the input garment latent as $\mathcal{E}(\mathbf{g}) = \varnothing$, where $\varnothing \in \mathbb{R}^{4\times h\times w}$ refers to an all-zero latent. In this way, the denoising UNet is trained both conditionally and unconditionally, i.e., with and without the outfitting fusion. Then at inference time, we simply use a guidance scale $s_{\mathbf{g}} \geq 1$ to adjust the strength of conditional control over the predicted noise $\hat{\epsilon}_\theta$ as:

$$\begin{aligned}\hat{\epsilon}_\theta(\mathbf{z}_t, &\omega_{\theta'}(\mathcal{E}(\mathbf{g}))) = \epsilon_\theta(\mathbf{z}_t, \varnothing) + \\ &s_{\mathbf{g}} \cdot (\epsilon_\theta(\mathbf{z}_t, \omega_{\theta'}(\mathcal{E}(\mathbf{g}))) - \epsilon_\theta(\mathbf{z}_t, \varnothing)),\end{aligned} \tag{4}$$

where we omit some minor terms compared with Equation 2 for the sake of brevity.

In practice, we empirically set the outfitting dropout ratio to 10% in training, i.e., 10% of garment latents $\mathcal{E}(\mathbf{g})$ are set to $\varnothing$. And the optimal value of the guidance scale $s_{\mathbf{g}}$ is usually around $1.5 \sim 2.0$ according to our ablation study. Figure 4 and Table 1 demonstrate the effects of our outfitting dropout and different guidance scale values.

Experiments

Experimental Setup

Datasets. Our experiments are performed on two high-resolution (1024×768) virtual try-on datasets, i.e., VITON-HD (Choi et al. 2021) and Dress Code (Morelli et al. 2022). The VITON-HD dataset consists of 13,679 image pairs of frontal half-body models and corresponding upper-body garments, where 2032 pairs are used as the test set. The Dress Code dataset consists of 15,363/8,951/29,478 image pairs of full-body models and corresponding upper-body garments/lower-body garments/dresses, where 1,800 pairs for each garment category are used as the test set.

Compared Methods. On the VITON-HD dataset (Choi et al. 2021), we compare our OOTDiffusion with multiple state-of-the-art VTON methods, including the GAN-based VITON-HD (Choi et al. 2021), HR-VITON (Lee et al. 2022) and GP-VTON (Xie et al. 2023), as well as the LDM-based LaDI-VTON (Morelli et al. 2023) and StableVITON (Kim et al. 2023).

While for the evaluation on the Dress Code dataset (Morelli et al. 2022), since VITON-HD (Choi et al. 2021), HR-VITON (Lee et al. 2022) and StableVITON (Kim et al. 2023) are not designed for the entire dataset beyond upper-body garments, we select two VTON methods (i.e., GP-VTON (Xie et al. 2023) and LaDI-VTON (Morelli et al. 2023)) and another LDM-based inpainting method (i.e., Paint-by-Example (Yang et al. 2023)) for fair comparison.

Evaluation Metrics. We evaluate the results in both the paired and unpaired settings, where the paired setting provides the target human and the corresponding garment images for reconstruction, and the unpaired setting provides the different garment images for virtual try-on. Specifically for

Figure 4: Qualitative comparison of outfitted images generated by OOTDiffusion models trained without/with outfitting dropout and using different values of the guidance scale $s_{\mathbf{g}}$. Please zoom in for more details.

Dress Code (Morelli et al. 2022), we note that the evaluation is performed on the entire dataset rather than being limited to upper-body garments. This more effectively validates the feasibility of each method in real-world applications with various garment types.

In the quantitative evaluation, though our OOTDiffusion supports higher-resolution (1024×768) virtual try-on, all the experiments are conducted at the resolution of 512×384 for fair comparison with previous VTON methods. For the paired setting, we use LPIPS (Zhang et al. 2018) and SSIM (Wang et al. 2004) to measure the quality of the generated image in terms of restoring the original image. For the unpaired setting, we employ FID (Heusel et al. 2017) and KID (Bińkowski et al. 2018) for realism and fidelity assessment. We follow the previous work (Detlefsen et al. 2022; Parmar, Zhang, and Zhu 2022; Morelli et al. 2023) to implement all of these metrics.

Implementation Details

In our experiments, we initialize the OOTDiffusion models by inheriting the pretrained weights of Stable Diffusion v1.5 (Rombach et al. 2022). Then we finetune the outfitting and denoising UNets using an AdamW optimizer (Loshchilov and Hutter 2018) with a fixed learning rate of 5e-5. Note that we train four types of models on VITON-HD (Choi et al. 2021) and Dress Code (Morelli et al. 2022) datasets at resolutions of 512×384 and 1024×768, separately. All the models are trained for 36,000 iterations on a single NVIDIA A100 GPU, with a batch size of 64 for the 512×384 resolution and 16 for the 1024×768 resolution. At inference time, we run our OOTDiffusion on a single NVIDIA RTX 4090 GPU for 20 sampling steps using the UniPC sampler (Zhao et al. 2024).

Ablation Study

We investigate the effects of our proposed outfitting dropout as well as the different values of the guidance scale $s_{\mathbf{g}}$ on the VITON-HD dataset (Choi et al. 2021). First, we train two variants of our OOTDiffusion models without/with outfitting dropout, respectively. Then for the model trained with outfitting dropout, we set $s_{\mathbf{g}} = 1.0, 1.5, 2.0, 2.5, 3.0, 5.0$ for classifier-free guidance. At inference time, we guarantee all of other parameters (including the random seed) are consistent for fair comparison. As Figure 4 shows, without outfitting dropout, classifier-free guidance is not supported and the generated result is obviously the worst. While for the model trained with outfitting dropout, when $s_{\mathbf{g}} = 1.0$, the inference process is identical to the model without outfitting dropout (see Equation 4), which gets a similarly bad result. When $s_{\mathbf{g}} > 1.0$, we see that the fine-grained garment features become clearer as $s_{\mathbf{g}}$ increases. However, color distortion occurs when $s_{\mathbf{g}} \geq 2.5$ and becomes extremely significant when $s_{\mathbf{g}} = 5.0$ (see the flower patterns). Furthermore, Table 1 quantitatively proves the efficacy of our outfitting dropout which enables classifier-free guidance with respect to the garment features, and finds the optimal guidance scale value is around $1.5 \sim 2.0$ in most cases. According to this study, we consistently conduct outfitting dropout for OOTDiffusion, and empirically set $s_{\mathbf{g}} = 1.5$ for the VITON-HD dataset (Choi et al. 2021) and $s_{\mathbf{g}} = 2.0$ for the Dress Code dataset (Morelli et al. 2022) in the following experiments.

Outfitting Dropout	Guidance Scale	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
✗	-	0.0750	0.8699	8.91	0.89
✓	1.0	0.0749	0.8705	8.99	0.89
✓	1.5	**0.0705**	**0.8775**	<u>8.81</u>	**0.82**
✓	2.0	<u>0.0708</u>	<u>0.8766</u>	**8.80**	<u>0.86</u>
✓	2.5	0.0746	0.8691	8.84	0.89
✓	3.0	0.0753	0.8684	8.95	0.96
✓	5.0	0.0788	0.8640	9.28	1.22

Table 1: Ablation study of outfitting dropout and different guidance scale values on the VITON-HD dataset (Choi et al. 2021). The best and second best results are reported in **bold** and <u>underline</u>, respectively.

Experimental Results

Qualitative Results. Figure 5 visually shows some example results of our method and other VTON methods on the test set of VITON-HD (Choi et al. 2021). We observe that compared with other methods, our OOTDiffusion consistently achieves the best try-on effects for various upper-body garments. More specifically, GAN-based methods like GP-VTON (Xie et al. 2023) often fail to generate realistic human bodies (1st and 4th rows) or natural garment folds (2nd and 3rd rows), making the outfitted images look unrealistic. While other LDM-based methods including LaDI-VTON (Morelli et al. 2023) and StableVITON (Kim et al. 2023) tend to lose some garment details such as complicated text (2nd and 4th rows) or patterns (1st and 3rd rows). In contrast, our OOTDiffusion not only generates realistic images but also preserves most of the fine-grained garment details.

Regarding the more complicated Dress Code dataset (Morelli et al. 2022), which consists of full-body models and various garment categories, our OOTDiffusion still visually outperforms other VTON methods. As illus-

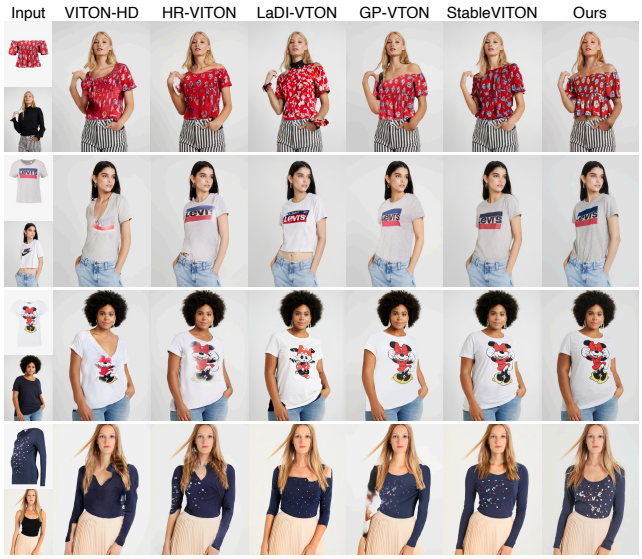

Figure 5: Qualitative comparison on the VITON-HD dataset (Choi et al. 2021) (half-body models with upper-body garments). Please zoom in for more details.

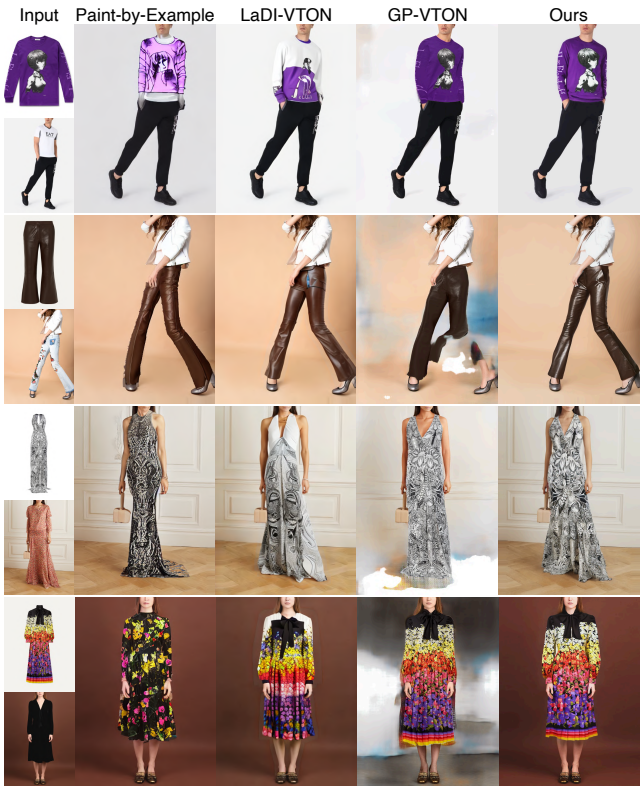

Figure 6: Qualitative comparison on the Dress Code dataset (Morelli et al. 2022) (full-body models with upper-body garments/lower-body garments/dresses). Please zoom in for more details.

trated in Figure 6, Paint-by-Example (Yang et al. 2023) and LaDI-VTON (Morelli et al. 2023) fail to preserve the garment features, and GP-VTON (Xie et al. 2023) tends

Figure 7: Qualitative results of the cross-dataset evaluation. The models are trained on the VITON-HD dataset (Choi et al. 2021) and tested on the Dress Code dataset (Morelli et al. 2022). Please zoom in for more details.

to cause severe body and background distortion. On the contrary, our OOTDiffusion consistently shows very stable performance on different garment categories including upper-body garments (1st row), lower-body garments (2nd row) and dresses (3rd and 4th rows).

In order to evaluate the generalization ability of our method, we conduct an additional cross-dataset experiment, i.e., training on one dataset and testing on the other. Figure 7 demonstrates that among all the models trained on the VITON-HD dataset (Choi et al. 2021), our OOTDiffusion is optimally adapted to the test examples in the Dress Code dataset (Morelli et al. 2022), generating more realistic outfitted images and preserving much more garment details. In summary, the observations above qualitatively prove the superiority and generalization capability of our OOTDiffusion in generating natural and accurate try-on results for various human and garment images.

Quantitative Results. Table 2 presents the quantitative evaluation results on the VITON-HD dataset (Choi et al. 2021). We find that some GAN-based models like HR-VITON (Lee et al. 2022) and GP-VTON (Xie et al. 2023) achieve relatively high SSIM scores, indicating that they are able to retain the structural information of the original images. However, their generated images lack detail fidelity, and thus drop behind ours on LPIPS. The previous LDM-based methods including LaDI-VTON (Morelli et al. 2023) and StableVITON (Kim et al. 2023) generate more realistic images according to their FID and KID scores, but they fail to restore the detailed features due to their lossy feature fusion. In comparison, our OOTDiffusion not only generates realistic outfitted images but also preserves the precise details, and thus substantially outperforms other methods on the other three metrics (LPIPS, FID and KID) while obtaining comparable SSIM scores to the GAN-based methods.

Table 3 demonstrates the state-of-the-art performance of our method on the Dress Code dataset (Morelli et al. 2022), which outperforms others on all the metrics for all the garment categories (upper-body/lower-body/dresses), confirming our feasibility in more complicated cases. Note that

Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
VITON-HD (Choi et al. 2021)	0.116	0.863	12.13	3.22
HR-VITON (Lee et al. 2022)	0.097	0.878	12.30	3.82
LaDI-VTON (Morelli et al. 2023)	0.091	0.875	9.31	1.53
GP-VTON (Xie et al. 2023)	0.083	**0.892**	9.17	0.93
StableVITON (Kim et al. 2023)	0.084	0.862	9.13	1.20
OOTDiffusion (Ours)	**0.071**	0.878	**8.81**	**0.82**

Table 2: Quantitative results on the VITON-HD dataset (Choi et al. 2021). The best and second best results are reported in **bold** and underline, respectively.

Method	All				Upper-body		Lower-body		Dresses	
	LPIPS ↓	SSIM ↑	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
PSAD* (Morelli et al. 2022)	0.058	0.918	10.61	6.17	17.51	7.15	19.68	8.90	17.07	6.66
Paint-by-Example (Yang et al. 2023)	0.142	0.851	9.57	3.63	18.63	4.81	15.89	4.12	19.15	5.88
LaDI-VTON (Morelli et al. 2023)	0.067	0.910	5.66	1.21	12.30	1.30	13.38	1.98	13.12	1.85
GP-VTON (Xie et al. 2023)	0.051	0.921	5.88	1.28	12.20	1.22	16.65	2.86	12.65	1.84
OOTDiffusion (Ours)	**0.045**	**0.927**	**4.20**	**0.37**	**11.03**	**0.29**	**9.72**	**0.64**	**10.65**	**0.54**

Table 3: Quantitative results on the Dress Code dataset (Morelli et al. 2022). The best and second best results are reported in **bold** and underline, respectively. The * marker refers to the results reported in previous work.

Train/Test	VITON-HD/Dress Code				Dress Code/VITON-HD			
Method	LPIPS ↓	SSIM ↑	FID ↓	KID ↓	LPIPS ↓	SSIM ↑	FID ↓	KID ↓
VITON-HD* (Choi et al. 2021)	0.187	0.853	44.26	28.82	-	-	-	-
HR-VITON* (Lee et al. 2022)	0.108	0.909	19.97	7.35	-	-	-	-
LaDI-VTON (Morelli et al. 2023)	0.154	0.908	14.58	3.59	0.235	0.812	29.66	20.58
GP-VTON (Xie et al. 2023)	0.291	0.820	74.36	80.49	0.266	0.811	52.69	49.14
StableVITON (Kim et al. 2023)	0.065	0.914	13.18	2.26	-	-	-	-
OOTDiffusion (Ours)	**0.061**	**0.915**	**11.96**	**1.21**	**0.123**	**0.839**	**11.22**	**2.72**

Table 4: Quantitative results of the cross-dataset evaluation. Each model is trained on one of the VITON-HD (Choi et al. 2021) and Dress Code (Morelli et al. 2022) datasets, and evaluated on the other. The best and second best results are reported in **bold** and underline, respectively. The * marker refers to the results reported in previous work.

GP-VTON (Xie et al. 2023) applies extra data modifications such as background removal and pose normalization to Dress Code, and only provides part of their test data. Despite this, our OOTDiffusion still achieves the best results on the more challenging original test dataset.

Furthermore, the generalization capability of our method is quantitatively verified by the results of the cross-dataset evaluation listed in Table 4. We find that GP-VTON (Xie et al. 2023) falls far behind other methods on all the metrics since its warping module severely overfits the training data. While our method leads again on all the metrics for the out-of-distribution test dataset. Overall, the observations above further demonstrate that our OOTDiffusion significantly outperforms previous VTON methods in both realism and controllability in all kinds of scenarios and conditions.

Limitations

Despite the state-of-the-art performance achieved in the image-based virtual try-on task, limitations still exist in our OOTDiffusion which demand further improvement. First, since our models are trained on paired human and garment images, it may fail to get perfect results for cross-category virtual try-on, e.g., to put a T-shirt on a woman in a long dress, or to let a man in pants wear a skirt. This issue can be partially solved in the future by collecting datasets of each person wearing different clothes in the same pose. Another limitation is that some details in the original human image might be altered after virtual try-on, such as muscles, watches or tattoos, etc. The reason is that the relevant body area is masked and repainted by the diffusion model. Thus more practical pre- and post-processing methods are required for addressing such problems.

Conclusion

In this paper, we present OOTDiffusion, a novel LDM-based network architecture for image-based vitrual try-on. The proposed outfitting UNet efficiently learns the garment features and incorporates them into the denoising UNet via the proposed outfitting fusion process with negligible information loss. Classifier-free guidance for the garment features is enabled by the proposed outfitting dropout in training, which further enhances the controllability of our method. Extensive experiments on high-resolution datasets show our superiority over other VTON methods in both realism and controllability, indicating that our OOTDiffusion has broad application prospects for virtual try-on.

Acknowledgments

We sincerely thank our colleagues including Yilan Ye, Bin Fu, Wei Du, Xuping Su, and Chi Zhang, etc., for kindly supporting and promoting our work. Special thanks to Minh-Duc Vo for his helpful advice.

References

Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8.

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 18392–18402.

Cao, Z.; Hidalgo Martinez, G.; Simon, T.; Wei, S.; and Sheikh, Y. A. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 14131–14140.

Detlefsen, N. S.; Borovec, J.; Schock, J.; Jha, A. H.; Koker, T.; Di Liello, L.; Stancl, D.; Quan, C.; Grechkin, M.; and Falcon, W. 2022. Torchmetrics-measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70): 4101.

Dong, H.; Liang, X.; Shen, X.; Wang, B.; Lai, H.; Zhu, J.; Hu, Z.; and Yin, J. 2019. Towards multi-pose guided virtual try-on network. In *ICCV*, 9026–9035.

Fenocchi, E.; Morelli, D.; Cornia, M.; Baraldi, L.; Cesari, F.; and Cucchiara, R. 2022. Dual-branch collaborative transformer for virtual try-on. In *CVPR*, 2247–2251.

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

Ge, C.; Song, Y.; Ge, Y.; Yang, H.; Liu, W.; and Luo, P. 2021a. Disentangled cycle consistency for highly-realistic virtual try-on. In *CVPR*, 16928–16937.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021b. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 8485–8493.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

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.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 10471–10480.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. Viton: An image-based virtual try-on network. In *CVPR*, 7543–7552.

He, S.; Song, Y.-Z.; and Xiang, T. 2022. Style-based global appearance flow for virtual try-on. In *CVPR*, 3470–3479.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*.

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.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, L.; Gao, X.; Zhang, P.; Sun, K.; Zhang, B.; and Bo, L. 2023. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*.

Issenhuth, T.; Mary, J.; and Calauzenes, C. 2020. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 619–635. Springer.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 6007–6017.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2023. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. *arXiv preprint arXiv:2312.01725*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*, 1931–1941.

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

Li, P.; Xu, Y.; Wei, Y.; and Yang, Y. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 3260–3271.

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam.

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

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 6038–6047.

Morelli, D.; Baldrati, A.; Cartella, G.; Cornia, M.; Bertini, M.; and Cucchiara, R. 2023. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. *arXiv preprint arXiv:2305.13501*.

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *CVPR*, 2231–2235.

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

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

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–11.

Parmar, G.; Zhang, R.; and Zhu, J.-Y. 2022. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 11410–11420.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

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.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. 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*, 234–241. Springer.

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 *CVPR*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.

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. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022c. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4713–4726.

Simon, T.; Joo, H.; Matthews, I.; and Sheikh, Y. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 1921–1930.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Zheng, H.; Liang, X.; Chen, Y.; Lin, L.; and Yang, M. 2018. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 589–604.

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.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *CVPR*.

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 *CVPR*, 23550–23559.

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 *CVPR*, 18381–18391.

Yang, H.; Zhang, R.; Guo, X.; Liu, W.; Zuo, W.; and Luo, P. 2020. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 7850–7859.

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, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*, 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 *CVPR*, 586–595.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2024. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. TryOnDiffusion: A Tale of Two UNets. In *CVPR*, 4606–4615.