MV-VTON: Multi-View Virtual Try-On with Diffusion Models

Haoyu Wang[1,2,*], **Zhilu Zhang**[2,†] **Donglin Di**[3], **Shiliang Zhang**[1], **Wangmeng Zuo**[2]
[1]State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University
[2]Harbin Institute of Technology
[3]Space AI, Li Auto
cshy2002@gmail.com, cszlzhang@outlook.com, donglin.ddl@gmail.com, slzhang.jdl@pku.edu.cn, wmzuo@hit.edu.cn

Abstract

The goal of image-based virtual try-on is to generate an image of the target person naturally wearing the given clothing. However, existing methods solely focus on the frontal try-on using the frontal clothing. When the views of the clothing and person are significantly inconsistent, particularly when the person's view is non-frontal, the results are unsatisfactory. To address this challenge, we introduce **M**ulti-**V**iew **V**irtual **T**ry-**ON** (MV-VTON), which aims to reconstruct the dressing results from multiple views using the given clothes. Given that single-view clothes provide insufficient information for MV-VTON, we instead employ two images, *i.e.*, the frontal and back views of the clothing, to encompass the complete view as much as possible. Moreover, we adopt diffusion models that have demonstrated superior abilities to perform our MV-VTON. In particular, we propose a view-adaptive selection method where hard-selection and soft-selection are applied to the global and local clothing feature extraction, respectively. This ensures that the clothing features are roughly fit to the person's view. Subsequently, we suggest joint attention blocks to align and fuse clothing features with person features. Additionally, we collect a MV-VTON dataset MVG, in which each person has multiple photos with diverse views and poses. Experiments show that the proposed method not only achieves state-of-the-art results on MV-VTON task using our MVG dataset, but also has superiority on frontal-view virtual try-on task using VITON-HD and DressCode datasets.

Code — https://github.com/hywang2002/MV-VTON

Figure 1: Motivation of this work. Previous VTON methods, *e.g.*, StableVITON (Kim et al. 2023) can only be used for the frontal-view person, and fail when facing the person with multiple views. Our MV-VTON can faithfully present the try-on results for a person with various views.

Introduction

Virtual Try-On (VTON) is a classic yet intriguing technology. It can be applied in the field of fashion and clothes online shopping to improve user experience. VTON aims to render the visual effect of a person wearing a specified garment. The emphasis of this technology lies in reconstructing a realistic image that faithfully preserves personal attributes and accurately represents clothing shape and details.

Early VTON methods (Lee et al. 2022; Xie et al. 2023; Bai et al. 2022; He, Song, and Xiang 2022) are based on generative adversarial networks (Goodfellow et al. 2020) (GANs). They generally align the clothing to the person's pose, and then employ a generator to fuse the warped clothing with the person. However, it poses a challenge to ensure that the warped clothing fits the target person's pose, and inaccurate clothes features will easily lead to distortion results. Recently, diffusion models (Rombach et al. 2022) have made remarkable strides in the field of image generation (Ruiz et al. 2023). Leveraging its potent generative capabilities, some researchers (Morelli et al. 2023; Kim et al. 2023) have integrated it into virtual try-on fields, building upon previous work and achieving commendable results.

Although VTON has made great progress, most existing methods focus on performing the frontal try-on. In practical applications, such as online shopping for clothes, customers may expect to obtain the dressing effect on multiple views (*e.g.*, side or back). In this case, the pose of the garment may be seriously inconsistent with the person's posture, and the

*This work was done while the first author was an undergraduate at Harbin Institute of Technology.

†Corresponding Author.

single-view clothing may not be enough to provide complete try-on information. Thus, these methods will easily generate results with poorly deformed clothing, and lead to the loss of high-frequency details such as texts, patterns, and other textures on clothing, as shown in Figure 1.

To address these issues, we introduce Multi-View Virtual Try-ON (MV-VTON), which aims to reconstruct the appearance and attire of a person from multiple views. For example, for clothing in Figure 1, which may exhibit significant differences between frontal and back styles, MV-VTON should be able to display try-on results in various views, including front, back, and side ones. Thus, providing single clothing can't meet the needs of dressing up, as the clothing only has partial information. Instead, we utilize both the frontal and back views of the clothing, which covers approximately complete view with as few images as possible.

Given the frontal and back clothing, we utilize the popular diffusion method to achieve MV-VTON. It is natural but doesn't work well to simply concatenate two pieces of clothing together as conditions of diffusion models, as it is difficult for the model to learn how to assign two-view clothes to a person, especially when the person is sideways. Instead, we propose a view-adaptive selection mechanism, which picks appropriate features of two-view clothes based on the posture information of the person and clothes. Therein, the hard-selection module chooses one of the two clothes for global feature extraction, and the soft-selection module modulates the local features of two clothes. We utilize CLIP (Radford et al. 2021) and a multi-scale encoder to extract the global and local clothing features, respectively. Moreover, to enhance the preservation of high-frequency details in clothing, we present joint attention blocks. They independently align global and local features with the person features, and selectively fuse them to refine the local clothing details while preserving global semantic information.

Furthermore, we collect a multi-view virtual try-on dataset, named Multi-View Garment (MVG). It contains thousands of samples, and each sample contains 5 images under different views and poses. We conduct extensive experiments not only on MV-VTON task using the MVG dataset, but also on the frontal-view VTON task using VITON-HD (Lee et al. 2022) and DressCode (Morelli et al. 2022) datasets. The results demonstrate that our method outperforms existing methods on both tasks.

In summary, our contributions are outlined below:

- We introduce a novel Multi-View Virtual Try-ON (MV-VTON) task, which aims at generating realistic dressing-up results of the multi-view person by using the given frontal and back clothing.
- We propose a view-adaptive selection method, where hard-selection and soft-selection are applied to global and local clothing feature extraction, respectively. It ensures that the clothing features are roughly fit to the person's view.
- We propose joint attention blocks to align the global and local features of selected clothing with the person ones, and fuse them.
- We collect a multi-view virtual try-on dataset. Extensive experiments demonstrate that our method outperforms previous approaches quantitatively and qualitatively in both frontal-view and multi-view virtual try-on tasks.

Related Work

GAN-Based Virtual Try-On

Existing methods are aimed at the frontal-view VTON task. To reconstruct realistic results, these methods based on generative adversarial networks (GAN) (Goodfellow et al. 2020) are typically divided into two steps. Firstly, the frontal-view clothing is deformed to align with the target person's pose. Afterward, the warped clothing and target person are fused through a GAN-based generator. In the warping step, some methods (Yang et al. 2020; Ge et al. 2021a; Wang et al. 2018) use TPS transformation to deform the frontal-view clothing, and others (Lee et al. 2022; Ge et al. 2021b; Xie et al. 2023) predict the global and local optical flow required for clothing deformation. However, when the clothing possesses intricate high-frequency details and the person's pose is complex, the effectiveness of clothing deformation is often diminished. Moreover, GAN-based generators generally encounter challenges in convergence and are highly susceptible to mode collapse (Miyato et al. 2018), leading to noticeable artifacts at the junction between warped clothing and the target person in the final results. In addition, previous multi-pose virtual try-on methods (Dong et al. 2019; Wang et al. 2020; Yu et al. 2023) can change the person's pose, but are also limited by GAN-based generator and insufficient clothing information.

Diffusion-Based Virtual Try-On

Thanks to the rapid advancement of diffusion models, recent works have sought to utilize the generative prior of large-scale pre-trained diffusion models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022; Yang et al. 2023) to tackle frontal-view virtual try-on tasks. TryOnDiffusion (Zhu et al. 2023) introduces two U-Nets to encode target person and frontal-view clothing images respectively, and interacts with the features of the two branches through the cross-attention mechanism.

LaDI-VTON (Morelli et al. 2023) encodes the frontal-view clothing image through textual inversion (Gal et al. 2022; Wei et al. 2023) and serves as the conditional input of backbone. DCI-VTON (Gou et al. 2023) first conducts an initial deformation of frontal-view clothing by incorporating a pre-trained wrapping network (Ge et al. 2021b). Subsequently, it attaches the deformed clothing to the target person image and feeds it into the diffusion model. While their frontal-view virtual try-on results seem more natural compared to GAN-based methods, they face difficulties in preserving high-frequency details due to the loss of details from the CLIP image encoder (Radford et al. 2021). To address this problem, StableVITON (Kim et al. 2023) attempts to introduce an additional encoder (Zhang, Rao, and Agrawala 2023) to encode the features of frontal-view clothing, and align the obtained clothing features through the zero cross-attention block. However, due to the absence of adequate

clothing priors, the generated results often struggle to remain faithful to the original clothing. Therefore, we introduce joint attention blocks to extract the global and local features of clothing, and employ the view-adaptive selection to choose the clothing features from the two views.

Method

Preliminaries for Diffusion Models

Diffusion Models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) have demonstrated strong capabilities in visual generation, which transforms a Gaussian distribution into a target distribution by iterative denoising. In particular, Stable Diffusion (Rombach et al. 2022) is a widely used generative diffusion model, which consists of a CLIP text encoder $\mathcal{E}_T$, a VAE encoder $\mathcal{E}$ as well as decoder $\mathcal{D}$, and a time-conditional denoising model ϵ_θ. The text encoder $\mathcal{E}_T$ encodes the input text prompt y as conditional input. The VAE encoder $\mathcal{E}$ compresses the input image I into latent space to get the latent variable $z_0 = \mathcal{E}(I)$. In contrast, the VAE decoder $\mathcal{D}$ decodes the output of backbone from latent space to pixel space. Through the VAE encoder $\mathcal{E}$, at an arbitrary time step t, the forward process is performed:

$$\alpha := \prod_{s=1}^{t}(1-\beta_s)\,, \quad z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\epsilon, \quad (1)$$

where $\epsilon \sim \mathcal{N}(0,1)$ is the random Gaussian noise and β is a predefined variance schedule. The training objective is to acquire a noise prediction network that minimizes the disparity between the predicted noise and the noise added to ground truth. The loss function can be defined as,

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(I),y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}_T(y))\|_2^2], \quad (2)$$

where z_t represents the encoded image $\mathcal{E}(I)$ with random Gaussian noise $\epsilon \sim \mathcal{N}(0,1)$ added.

In our work, we use an exemplar-based inpainting model (Yang et al. 2023) as a backbone, which employs an image c rather than texts as the prompt and then encode c by the image encoder $\mathcal{E}_I$ of CLIP. Thus, the loss function in Eq. (2) can be modified as,

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(I),c,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}_I(c))\|_2^2]. \quad (3)$$

Method Overview

While existing virtual try-on methods are designed solely for frontal-view scenarios, we present a novel approach to handle both frontal-view and multi-view virtual try-on tasks, along with a multi-view virtual try-on dataset MVG comprising try-on images captured from five different views. Examples of it are shown in Figure 2(b). Formally, given a person image x in an arbitrary view, along with a frontal view clothing c_f and a back view clothing c_b, our goal is to generate the result of the person wearing the clothing in its view. Considering the substantial differences between the front and back of most clothing, another challenge is to make informed decisions regarding the two provided clothing images based on the target person's pose, ensuring a natural try-on result across multiple views.

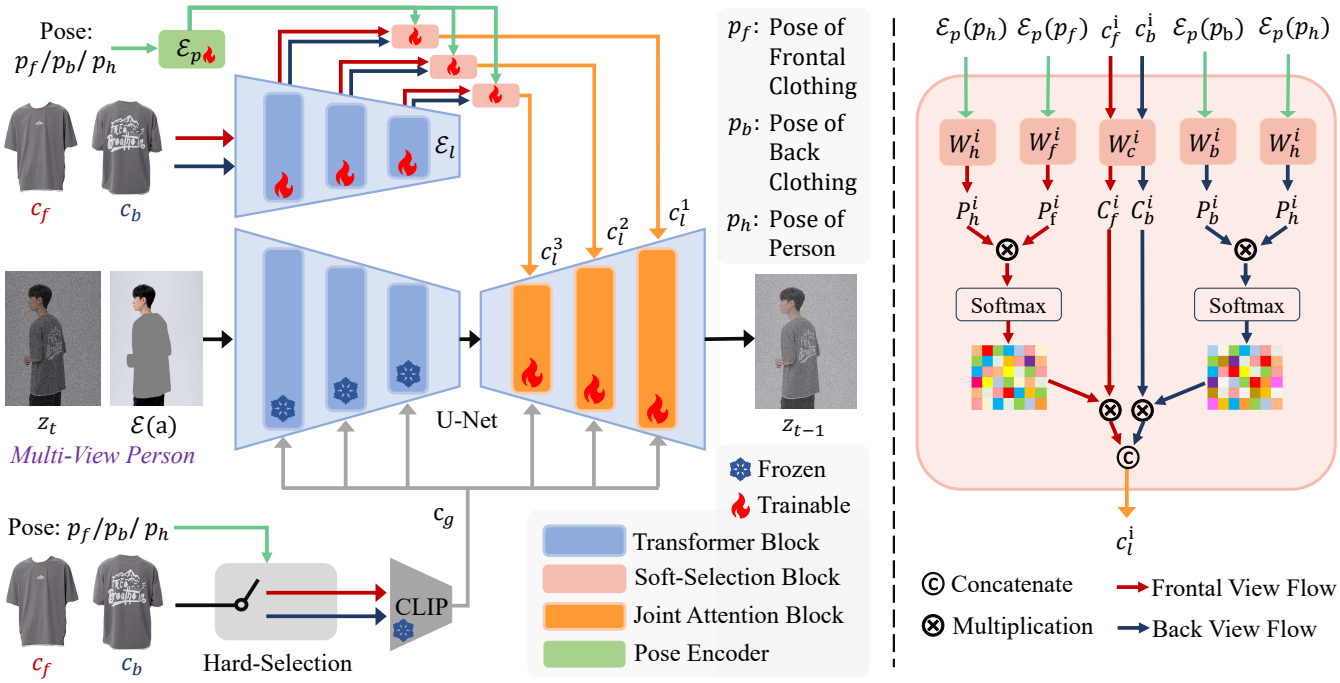
In this work, we use an image inpainting diffusion model (Yang et al. 2023) as our backbone. Denote by M the inpainting mask, and denote by a the masked person image x. The model concatenates z_t ($z_0 = \mathcal{E}(x)$), the encoded clothing-agnostic image $\mathcal{E}(a)$, and the resized clothing-agnostic mask m in the channel dimension, and feeds them into the backbone as spatial input. Besides, we use an existing method to pre-warp the clothing and paste it on a. While utilizing CLIP image encoder to encode clothing as the global condition of the diffusion model, we also introduce an additional encoder (Zhang, Rao, and Agrawala 2023) to encode clothing to provide more refined local conditions. Since both the frontal and back view clothing need to be encoded, directly sending both into the backbone as conditions may result in confusion of clothing features. To alleviate this problem, we propose a view-adaptive selection mechanism. Based on the similarity between the poses of the person and two clothes, it conducts hard-selection when extracting global features and soft-selection when extracting local features. To preserve semantic information in clothing and enhance high-frequency details in global features using local ones, we introduce joint attention blocks. They first independently align global and local features to the person ones and then selectively fuse them. Figure 3(a) depicts an overview of our proposed method.

Figure 2: Comparison between previous datasets and our proposed MVG dataset. (a) is the dataset used by the previous work, which only have clothing and person in the frontal-view. In contrast, our dataset (b) offers images from five different views.

View-Adaptive Selection

For multi-view virtual try-on task, given the substantial differences between the frontal and back views, as illustrated in Figure 2(b), it's imperative to extract and assign the features of frontal and back view clothing for the person tendentiously. Actually, based on the pose of the target person, we can determine which view of clothing should be given more attention during the try-on process. For example, if the target pose resembles the pose in the fourth column of Figure 2(b), it's evident that we should rely more on the characteristics of the back view clothing to generate the try-on result. Specifically, we propose a view-adaptive selection mechanism to achieve this purpose, including hard- and soft-selection.

Hard-Selection for Global Clothing Features. We deploy a CLIP image encoder to extract global features of clothing. During this process, we perform hard-selection on the frontal and back view clothing based on the similarity be-



(a) Overview of MV-VTON (b) Soft-Selection Block

Figure 3: (a) Overview of MV-VTON. It encodes frontal and back view clothing into global features using the CLIP image encoder and extracts multi-scale local features through an additional encoder $\mathcal{E}_l$. Both features act as conditional inputs for the decoder of backbone. Besides, both features are selectively extracted through view-adaptive selection mechanism. (b) Soft-selection modulates the clothing features on frontal and back views, respectively, based on the similarity between the clothing's pose and the person's pose. Then the features from both views are concatenated in the channel dimension.

tween the garments' pose and the person's pose. It means that we only select one piece of clothing that is closest to the person's pose as the input of the image encoder, since it is enough to cover global semantic information. When generating pre-warped clothing for $\mathcal{E}(a)$, the selection is also performed. Implementation details of hard-selection can be found in the supplementary material.

Soft-Selection for Local Clothing Features. We utilize an additional encoder $\mathcal{E}_l$ to extract the multi-scale local features of frontal and back view clothing, which in the i-th scale are denoted as c_f^i and c_b^i, respectively. When reconstructing the try-on results, it may be insufficient to rely solely on the clothing from either frontal or back view under certain specific scenes, such as the third column shown in Figure 2(b). In these cases, it may be necessary to incorporate clothing features from both views. However, simply combining the two may lead to confusion of features. Instead, we introduce soft-selection block to modulate their features, respectively, as shown in Figure 3(b).

First, the person's pose p_h, frontal-view clothing's pose p_f, and back view clothing's pose p_b are encoded by the pose encoder $\mathcal{E}_p$ to obtain their respective features $\mathcal{E}_p(p_h)$, $\mathcal{E}_p(p_f)$, and $\mathcal{E}_p(p_b)$. Details of the pose encoder can be found in the supplementary material. When processing frontal-view clothing, in i-th soft-selection block, we map $\mathcal{E}_p(p_h)$ and $\mathcal{E}_p(p_f)$ to P_h^i and P_f^i through a linear layer with weights W_h^i and W_f^i, respectively. We also map c_f^i to C_f^i through a linear layer with weights W_c^i. Then, we calculate the similarity between the person's pose and frontal-view clothing's pose to get the selection weights of frontal-view clothing, *i.e.*,

$$weights = softmax(\frac{P_h^i (P_f^i)^T}{\sqrt{d}}), \tag{4}$$

where $weights$ represents the selection weights of frontal-view clothing, and d represents the dimension of these matrices. Assuming that the person's pose is biased towards the front, as depicted in the second column of Figure 2(b), the similarity between the person's pose and the front view clothing's pose will be higher. Consequently, the corresponding clothing features will be enhanced by $weights$, and vice versa. The features of back view clothing c_b^i undergo similar processing. Finally, the two selected clothing features are concatenated along the channel dimension as the local condition c_l^i of backbone.

Joint Attention Blocks

Global clothing features c_g provide identical conditions for blocks at each scale of U-Net, and multi-scale local clothing features c_l allow for reconstructing more accurate details. We present joint attention blocks to align c_g and c_l with the current person features, as shown in Figure 4. To retain most of the semantic information in global features c_g, we use

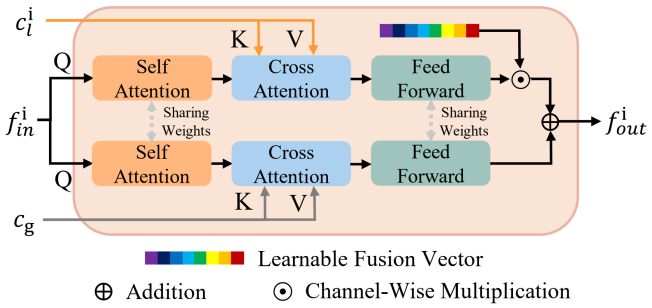


Figure 4: Overview of the proposed joint attention blocks.

local features c_l to refine some lost and erroneous detailed texture information in c_g by selective fusion.

Specifically, in the i-th joint attention block, we first calculate self-attention for the current features f_{in}^i. Then, we deploy a double cross-attention. The queries (Q) come from f_{in}^i and global features c_g serve as one set of keys (K) and values (V), while local features c_l^i serve as another set of keys (K) and values (V). After aligning to the person's pose through cross-attention, the clothing features c_g and c_l^i are selectively fused in channel-wise dimension, *i.e.*,

$$f_{out}^i = softmax(\frac{Q_g^i(K_g^i)^T}{\sqrt{d}})V_g^i + \lambda \odot softmax(\frac{Q_l^i(K_l^i)^T}{\sqrt{d}})V_l^i \,, \tag{5}$$

where Q_g^i, K_g^i, V_g^i represent the Q, K, V of global branch, Q_l^i, K_l^i, V_l^i represent the Q, K, V of local branch, λ is the learnable fusion vector, $\odot$ represents channel-wise multiplication, and f_{out}^i represents the clothing features after selective fusion. By engaging and fusing the global and local clothing features, we can enhance the retention of high-frequency garment details, *e.g.*, texts and patterns.

Training Objectives

As stated in preliminaries, diffusion models learn to generate images from random Gaussian noise. However, the training objective in Eq. (3) is performed in latent space, and does not explicitly constrain the generated results in visible image space, resulting in slight differences in color from the ground truth. To alleviate the problem, we additionally employ ℓ_1 loss $\mathcal{L}_1$ and perceptual loss (Johnson, Alahi, and Fei-Fei 2016) $\mathcal{L}_{perc}$. The $\mathcal{L}_1$ loss is calculated by

$$\mathcal{L}_1 = \|\hat{x} - x\|_1 \,, \tag{6}$$

where $\hat{x}$ is the reconstructed image using Eq. (1). The perceptual loss is calculated as,

$$\mathcal{L}_{perc} = \sum_{k=1}^{5} \|\phi_k(\hat{x}) - \phi_k(x)\|_1 \,, \tag{7}$$

where ϕ_k represents the k-th layer of VGG (Simonyan and Zisserman 2014). Totally, the overall training objective can be written as,

$$\mathcal{L} = \mathcal{L}_{LDM} + \lambda_1 \mathcal{L}_1 + \lambda_{perc}\mathcal{L}_{perc} \,, \tag{8}$$

where λ_1 and λ_{perc} are the balancing weights.

Experiments

Experiments Settings

Datasets: For the proposed multi-view virtual try-on task, we collect MVG dataset containing 1,009 samples. Each sample contains five images of the same person wearing the same garment from five different views, for a total of 5,045 images, as shown in Figure 2(b). The image resolution is about 1K. We'll explain how the datasets are collected and how they're used for MV-VTON in the supplementary material. The proposed method can also be applied to frontal-view virtual try-on task. Our frontal-view experiments are carried out on VITON-HD (Lee et al. 2022) and DressCode (Morelli et al. 2022) datasets. They contain more than 10,000 frontal-view person and upper-body clothing image pairs. We follow previous work for the use of them.

Evaluation Metrics. Following previous works (Kim et al. 2023; Morelli et al. 2023), we use four metrics to evaluate the performance of our method: Structural Similarity (SSIM) (Wang et al. 2004), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), Frechet Inception Distance (FID) (Heusel et al. 2017) and Kernel Inception Distance (KID) (Bińkowski et al. 2018). Specifically, for paired test setting, which means directly using the paired data in the dataset, we utilize the above four metrics for evaluation. For unpaired test setting, which means that the given garment is different from the garment originally worn by target person, we use FID and KID for evaluation, and in order to distinguish them from the paired setting, we named them FID_u and KID_u respectively.

Implementation Details. We use Paint by Example (Yang et al. 2023) as the backbone of our method and copy the weights of its encoder to initialize $\mathcal{E}_l$. The hyper-parameter λ_1 is set to 1e-1, and λ_{perc} is set to 1e-4. We train our model on 2 NVIDIA Tesla A100 GPUs for 40 epochs with a batch size of 4 and a learning rate of 1e-5. We use AdamW (Loshchilov and Hutter 2017) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Comparison Settings. We compare our method with Paint By Example (Yang et al. 2023), PF-AFN (Ge et al. 2021b), GP-VTON (Xie et al. 2023), LaDI-VTON (Morelli et al. 2023), DCI-VTON (Gou et al. 2023), StableVITON (Kim et al. 2023) and IDM-VTON (Choi et al. 2024) on both frontal-view and multi-view virtual try-on tasks. For multi-view virtual try-on, we compare these methods on the proposed MVG dataset. For the sake of fairness, we fine-tune the previous methods on the MVG dataset according to its original training settings. Since previous methods can only input a single clothing image, we input frontal and back view clothing respectively and select the best result. For frontal-view virtual try-on, we compare these methods on VITON-HD and DressCode datasets. Following previous works' settings, the proposed MV-VTON only inputs one frontal-view garment during training and inference.

Quantitative Evaluation

Table 1 reports the quantitative results on the paired setting, and Table 2 shows the unpaired setting's results. On the multi-view virtual try-on task, as can be seen, thanks

Methods	Reference	MVG				VITON-HD				DressCode - Upper Body			
		LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓	LPIPS↓	SSIM↑	FID↓	KID↓
Paint by Example	CVPR23	0.120	0.880	54.38	14.95	0.150	0.843	13.78	4.48	0.078	0.899	15.21	4.51
PF-AFN	CVPR21	0.139	0.873	49.47	12.81	0.141	0.855	7.76	4.19	0.091	0.902	13.11	6.29
GP-VTON	CVPR23	-	-	-	-	0.085	0.889	6.25	0.77	0.236	0.781	19.37	8.07
LaDI-VTON	MM23	0.069	0.921	29.14	4.39	0.094	0.872	7.08	1.49	0.063	0.922	11.85	3.20
DCI-VTON	MM23	0.062	0.929	25.71	0.95	0.074	0.893	5.52	0.57	0.043	0.937	11.87	1.91
StableVITON	CVPR24	0.063	0.929	23.52	0.46	0.073	0.888	6.15	1.34	**0.040**	0.937	10.18	1.70
IDM-VTON	ECCV24	0.095	0.896	34.66	5.33	0.135	0.826	14.36	8.63	0.066	0.912	13.88	5.39
Ours	-	**0.050**	**0.936**	**22.18**	**0.35**	**0.069**	**0.897**	**5.43**	**0.49**	**0.040**	**0.941**	**8.26**	**1.39**

Table 1: Quantitative comparison with previous work on paired setting. For multi-view virtual try-on task, we show results on our proposed MVG dataset. For frontal-view virtual try-on task, we show results on VITON-HD dataset (Lee et al. 2022) and DressCode dataset (Morelli et al. 2022). The best results have been bolded. Note that all previous works have been finetuned on our proposed MVG dataset when comparing on multi-view virtual try-on task.

Figure 5: Qualitative comparisons on multi-view virtual try-on task with MVG dataset.

Figure 6: Qualitative comparisons on frontal-view virtual try-on task with VITON-HD and DressCode datasets.

to the view-adaptive selection mechanism, our method can reasonably select clothing features according to the person's pose, so it is better than existing methods in various metrics, especially on LPIPS and SSIM. Furthermore, owing to joint attention blocks, our approach excels in preserving high-frequency details of the original garments across both frontal-view and multi-view virtual try-on scenarios, thus achieving superior performance in these metrics.

Qualitative Evaluation

Multi-View Virtual Try-On. As shown in Figure 5, MV-VTON generates more realistic multi-view results compared to the previous five methods. Specifically, in the first row, due to the lack of adaptive selection of clothes, previous methods have difficulty in generating hoods of the original cloth. Moreover, in the second row, previous methods often struggle to maintain fidelity to the original garments. In contrast, our method effectively addresses the aforementioned problems and generates high-fidelity results. We provide more results of multi-view virtual try-on in the supplementary materials.

Frontal-View Virtual Try-On. As shown in Figure 6, our method also demonstrates superior performance over existing methods on frontal-view virtual try-on task, particularly in retaining clothing details. Specifically, our method not only faithfully generates complex patterns (in the first row),

Method	MVG		VITON-HD	
	$FID_u\downarrow$	$KID_u\downarrow$	$FID_u\downarrow$	$KID_u\downarrow$
Paint by Example	43.79	5.92	17.27	4.56
PF-AFN	47.38	7.04	21.18	6.57
GP-VTON	-	-	9.11	1.21
LaDI-VTON	36.61	3.39	9.55	1.83
DCI-VTON	36.03	3.79	8.93	1.07
StableVITON	35.85	4.22	9.86	1.09
IDM-VTON	40.73	5.74	18.27	10.43
Ours	**33.44**	**2.69**	**8.67**	**0.78**

Table 2: Unpaired setting's quantitative results on our MVG dataset and VITON-HD dataset. The best results have been bolded.

Hard	Soft	LPIPS↓	SSIM↑	FID↓	KID↓	$FID_u\downarrow$	$KID_u\downarrow$
×	×	0.068	0.925	25.13	0.77	35.28	3.24
×	√	0.064	0.928	24.58	0.62	34.67	3.05
√	×	0.052	0.934	**22.18**	0.43	33.47	2.74
√	√	**0.050**	**0.936**	**22.18**	**0.35**	**33.44**	**2.69**

Table 3: Ablation study of our proposed view-adaptive selection mechanism on MVG dataset.

	Global	Local	LPIPS↓	SSIM↑	FID↓	KID↓	$FID_u\downarrow$	$KID_u\downarrow$
MVG	√	×	0.062	0.929	25.71	0.95	36.01	3.78
	×	√	0.058	0.931	26.16	1.21	36.29	3.91
	√	√	**0.050**	**0.936**	**22.18**	**0.35**	**33.44**	**2.69**
VITON-HD	√	×	0.074	0.893	5.52	0.57	8.93	1.07
	×	√	0.070	0.896	5.76	0.81	9.15	1.09
	√	√	**0.069**	**0.897**	**5.43**	**0.49**	**8.67**	**0.78**

Table 4: Ablation study of joint attention blocks on MVG and VITON-HD datasets.

but also better preserves the literal 'Wrangler' in the clothing (in the second row). We provide more qualitative comparisons in the supplementary materials, as well as dressing results under complex human pose conditions.

Ablation Studies

Effect of View-Adaptive Selection. We investigate the effect of view-adaptive selection on the multi-view virtual try-on task. Specifically, no hard-selection represents that we directly concatenate two garments' features encoded by CLIP, and no soft-selection means that two clothing features are concatenated without passing soft-selection blocks. Comparison results are shown in Table 3 and Figure 7. As can be seen, the performance is greatly reduced without hard-selection and soft-selection. No hard-selection will confuse two view's cloth features, as shown by the blurriness of the 'POP' text in Figure 7. In addition, no soft-selection causes the model to lose some cloth information when processing the side view situation, such as the missing white hood and cuffs in Figure 7.

Effect of Joint Attention Blocks. In order to demonstrate the effectiveness of fusing global and local features through joint attention blocks, we discard the global feature extraction branch and the local feature extraction branch respectively. Results are shown in Table 4 and Figure 8. As can be seen, relying solely on global features may lead to loss of details, such as the distorted text 'VANS' in the first row and the missing letter 'C' in the second row. Moreover, if only local features are provided, the results may also have unfaithful textures, such as artifacts on the person's chest. Compared to them, we fuse global and local features through joint attention blocks, which can refine details in garments while preserving semantic information.

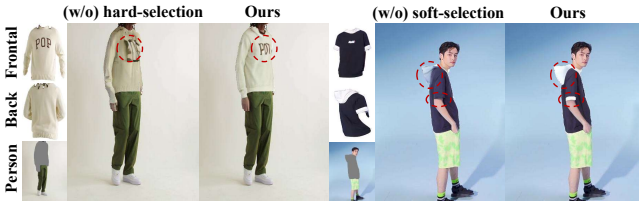


Figure 7: Visualization of view-adaptive selection's effect.

Figure 8: Visualization of joint attention blocks' effect.

Conclusion

We introduce a novel and practical Multi-View Virtual Try-ON (MV-VTON) task, which aims at using the frontal and back clothing to reconstruct the dressing results of a person from multiple views. To achieve the task, we propose a diffusion-based method. Specifically, the view-adaptive selection mechanism exacts more reasonable clothing features based on the similarity between the poses of a person and two clothes. The joint attention block aligns the global and local features of the selected clothing to the target person, and fuse them. In addition, we collect a multi-view garment dataset for this task. Extensive experiments demonstrate that the proposed method achieves state-of-the-art performance both on frontal-view and multi-view virtual try-on tasks, compared with existing methods.

Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFA1004100).

References

Bai, S.; Zhou, H.; Li, Z.; Zhou, C.; and Yang, H. 2022. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, 409–425. Springer.

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

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

Dong, 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 9026–9035.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 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.

He, S.; Song, Y.-Z.; and Xiang, T. 2022. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3470–3479.

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

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

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

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*.

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

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

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

Morelli, D.; Fincato, M.; Cornia, M.; Landi, F.; Cesari, F.; and Cucchiara, R. 2022. Dress code: high-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2231–2235.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

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

Wang, J.; Sha, T.; Zhang, W.; Li, Z.; and Mei, T. 2020. Down to the last detail: Virtual try-on with fine-grained details. In *Proceedings of the 28th ACM International Conference on Multimedia*, 466–474.

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, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Xie, Z.; Huang, Z.; Dong, X.; Zhao, F.; Dong, H.; Zhang, X.; Zhu, F.; and Liang, X. 2023. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23550–23559.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18381–18391.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7850–7859.

Yu, F.; Hua, A.; Du, C.; Jiang, M.; Wei, X.; Peng, T.; Xu, L.; and Hu, X. 2023. VTON-MP: Multi-Pose Virtual Try-On via Appearance Flow and Feature Filtering. *IEEE Transactions on Consumer Electronics*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.