PEMF-VTO: Point-Enhanced Video Virtual Try-on via Mask-free Paradigm

Tianyu Chang[1,2*] Xiaohao Chen[4*] Zhichao Wei[4] Xuanpu Zhang[3]
Qing-Guo Chen[4] Weihua Luo[4] Peipei Song[1] Xun Yang[1†]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Tianjin University [4] Alibaba International Digital Commerce

Abstract

Video Virtual Try-on aims to seamlessly transfer a reference garment onto a target person in a video while preserving both visual fidelity and temporal coherence. Existing methods typically rely on inpainting masks to define the try-on area, enabling accurate garment transfer for simple scenes (e.g., in-shop videos). However, these mask-based approaches struggle with complex real-world scenarios, as overly large and inconsistent masks often destroy spatial-temporal information, leading to distorted results. Mask-free methods alleviate this issue but face challenges in accurately determining the try-on area, especially for videos with dynamic body movements. To address these limitations, we propose PEMF-VTO, a novel Point-Enhanced Mask-Free Video Virtual Try-On framework that leverages sparse point alignments to explicitly guide garment transfer. Our key innovation is the introduction of point-enhanced guidance, which provides flexible and reliable control over both spatial-level garment transfer and temporal-level video coherence. Specifically, we design a Point-Enhanced Transformer (PET) with two core components: Point-Enhanced Spatial Attention (PSA), which uses frame-cloth point alignments to precisely guide garment transfer, and Point-Enhanced Temporal Attention (PTA), which leverages frame-frame point correspondences to enhance temporal coherence and ensure smooth transitions across frames. Extensive experiments demonstrate that our PEMF-VTO outperforms state-of-the-art methods, generating more natural, coherent, and visually appealing try-on videos, particularly for challenging in-the-wild scenarios.

1. Introduction

Video Virtual Try-On, which aims to transfer the provided garment into a specific area in the source person video while maintaining the inter-frame coherence, has garnered significant attention especially for the E-Commerce and fashion design fields. This technology greatly reduces related costs and brings more convenient shopping and work experience to consumers and fashion designers.

Recently, based on powerful diffusion models [26, 37, 43, 54] and existing image virtual try-on training paradigm [7, 30, 38, 51, 52, 61], many video virtual try-on methods [15, 24, 47, 49, 58] have been proposed to take advantage of the inpainting mask to ensure the try-on area and the temporal attention module to keep the video coherence, obtaining natural and impressive try-on results on simple in-shop videos. However, in realistic scenarios, person videos often exhibit complex body movements and significant scene motion (*e.g.*, street dance videos). The pre-acquired agnostic mask of such challenging videos will lead to 1) the loss of spatial information on human postures and 2) temporal inconsistency in try-on areas between adjacent frames, causing mask-based methods to generate distorted and incoherent try-on video, as Fig. 1 shown. Therefore, to learn a more general video virtual try-on model, it is necessary to design a more reasonable framework to eliminate the inherent deficiencies of mask-based try-on methods.

In the image virtual try-on area, several attempts have been made to alleviate the negative impact of the inpainting mask by either 1) correcting the initial inaccurate agnostic mask [51, 56] or 2) constructing the large-scale paired training data to learn a mask-free try-on model [16, 23, 39, 57]. However, the correlation of the agnostic mask does not eliminate the dependencies of the mask guidance. In addition, as Fig. 1 shows, directly transferring the mask-free paradigm to a video virtual try-on task will bring confusing inconsistencies between reference garments and generated video frames. The core reason is that person video data contain more complex actions or movements at the temporal level, compared to image data. Based on the above analysis, *it is necessary to design an innovative video virtual try-on paradigm that can avoid the deficiencies of agnostic masks while providing explicit and flexible guidance on the specific garment try-on area.*

In this work, we propose a novel **P**oint-**E**nhanced

*Equal contribution.
†Corresponding author.

Figure 1. Comparison of three virtual try-on paradigms. Rows from top to bottom denote the (a) image try-on results in the wild, (b) video try-on results in the shop and (c) video try-on results in the wild. The testing data are respectively from open-sourced StreetVTON [10], ViViD [15] and TikTok [29] datasets. *Best viewed with Acrobat Reader. Click the images to play the video clips.*

Mask-**F**ree **V**ideo virtual **T**ry-**O**n (PEMF-VTO) framework, which exploits the sparse frame-cloth and frame-frame point alignments to explicitly guide the desirable try-on area while enhancing the coherence of the generated try-on video. Specifically, we first leverage the pre-trained mask-based try-on method ViViD [15] to construct large-scale paired data, thus learning a mask-free virtual try-on model that eliminates the negative impact of the unreliable agnostic mask. Inadequately, the pure mask-free model can not determine the consistent and correct try-on area at the temporal level, especially for realistic videos with diverse body movements. To this end, we introduce sparse frame-cloth and frame-frame correspondences (matching point pairs) to enhance the consistency and coherence of garment transfer to source video. Concretely, a **P**oint-enhanced **S**patial **A**ttention (PSA) is designed to strengthen the garment transfer to a specific area. Besides, a **P**oint-enhanced **T**emporal **A**ttention (PTA) is also proposed to increase the coherence of the generated video. In this way, our PEMF-VTO can simultaneously meet the three important and challenging requirements of the video virtual try-on task: 1) the accurate transfer of reference garment 2) the preservation of the non-try-on area and 3) the continuity of generated video frames. Extensive qualitative and qualitative experiments clearly show the effectiveness of our method.

Our contributions can be briefly summarized as follows:

- We investigate the deficiencies of current learning paradigms in virtual try-on methods and propose a more flexible and generalizable point-enhanced mask-free paradigm compared to prior approaches.
- We design the **P**oint-enhanced **S**patial **A**ttention (PSA) module and the **P**oint-enhanced **T**emporal **A**ttention (PTA) module to enhance the garment transfer ability and temporal coherence of the generated try-on video.
- Extensive experiments illustrate that our method achieves higher-quality and more coherent results for video virtual try-on, especially in challenging in-the-wild scenarios.

2. Related Work

Image Virtual Try-On. Given a source person image and a reference garment image, image virtual try-on aims to synthesize an identity and background preserved and cloth changed image. Previous methods [6, 13, 14, 16, 16, 21–23, 46, 50, 53, 59] mainly leveraged Generative Adversarial Networks (GANs) [17] to first warp the reference garment to fit the person's body and then transfer the deformed garment into the source person. However, the limited generational capacity of GANs significantly influences the quality of try-on images for these GAN-based methods.

Recently, based on the amazing performance of diffusion models in generating high-quality images at high resolutions, many diffusion-based virtual try-on methods [1, 5, 7, 18, 30, 34, 60] have been proposed to generate natural and realistic try-on images. For instance, StableVITON [30] employed a ControlNet-like [54] encoder to ensure the fine-grained garment transfer to person data. IDM-VTON [7] designed a dual U-Net architecture to respectively encode the person feature and cloth feature, then conducted the cross attention between them to achieve high-quality garment transfer in wild realistic scenarios. While significant progress has been made, these methods heavily depended on the quality of the inpainting mask to determine the try-on area. When evaluating more complex try-on data that

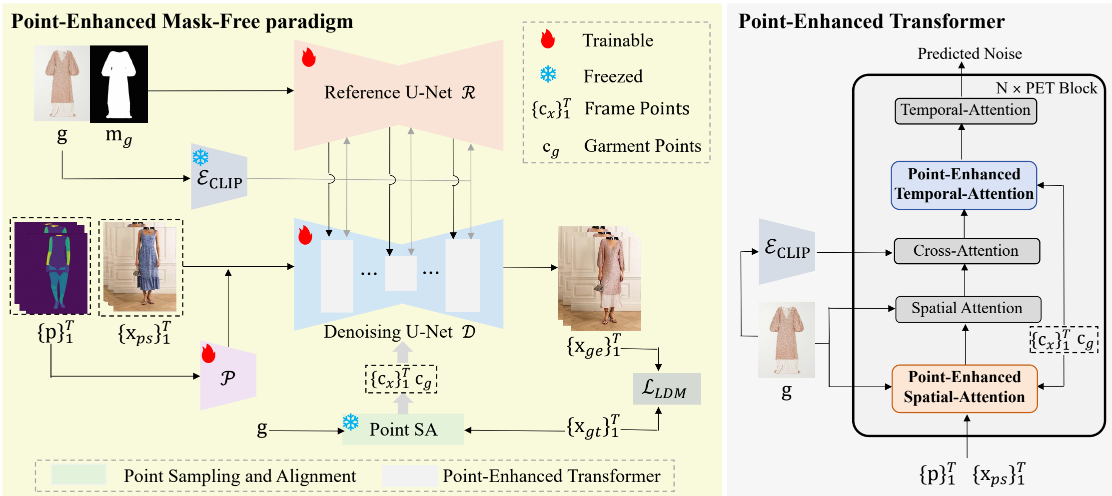


Figure 2. The pipeline of our PEMF-VTO framework. It leverages the paired pseudo-person data ($\{\mathbf{x}_{ps}\}_1^T$, $\{\mathbf{x}_{gt}\}_1^T$) to train a mask-free model, thereby avoiding the loss of spatial-temporal information in the try-on area. Besides, based on the pre-acquired alignments between frame points $\{\mathbf{c}_x\}_1^T$ and garment points $\mathbf{c}_g$, a novel point-enhanced transformer is proposed to respectively improve the garment transfer ability and coherence in the try-on area by the point-enhanced spatial attention and point-enhanced temporal attention modules.

contains diverse foreground occlusions and person poses, they always fail to restore the non-try-on area.

To alleviate the above issue, TPD [51] and Betterfit [56] proposed mask prediction or correlation modules to dynamically identify precise try-on areas. However, these methods do not completely eliminate the dependencies of the mask guidance. Furthermore, BooW-VTON [57] and AnyDesign [39] adopted a mask-free virtual try-on training paradigm. Though achieving impressive progress, they sometimes struggle to accurately identify try-on areas, especially when handling ambiguous or diverse clothing types.

Video Virtual Try-On. Recently, inspired by the training paradigm of image virtual try-on, several diffusion-based video virtual try-on methods have been proposed [15, 24, 47, 49, 58]. They utilized the inpainting mask to ensure the try-on area and employed the temporal attention module to ensure video coherence. However, these methods suffer from more severe drawbacks of the mask-based training paradigm due to the increased complexity of video data. Designing an effective and versatile video virtual try-on paradigm remains both critical and challenging.

In contrast to the aforementioned mask-based paradigm, we propose a point-enhanced mask-free video virtual try-on method to simultaneously achieve: 1) precise control over the try-on area, 2) accurate preservation of non-try-on regions and 3) temporal coherence of video frames, thus synthesizing more realistic and coherent virtual try-on videos.

3. The Approach

3.1. Preliminary

Stable Diffusion. Our PEMF-VTO leverages the Stable Diffusion (SD) [41], one of the most widely applied generation models based on the Latent Diffusion Model (LDM). LDM performs the denoising process in the latent space to conduct a more effective image generation. Specifically, a VAE encoder $\mathcal{E}(\cdot)$ first converts the image $\mathbf{x}$ into a latent embedding $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$. Then the forward diffusion process is exploited by adding the noise to the latent embedding:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad (1)$$

where $t \in \{1, ..., T\}$ represents the number of diffusion timesteps and $\{\bar{\alpha}_i\}_{i=1}^t$ determines the diffusion schedule. Finally, the denoising model is trained to predict the added noise of the noisy latent $\mathbf{z}_t$ by the loss constraints of LDM:

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

where ϵ_θ represents the denoising model and $\mathbf{y}$ is the conditional embedding to control the content of generation. *In our task, the feature embeddings derived from the reference garment image should serve as the condition to generate identity-preserved and cloth-changed try-on results.*

3.2. Existing Virtual Try-on Paradigms.

Given a source person video and a reference garment image, our target is to fluently transfer the garment to a desirable try-on area in the source video, thus synthesizing a natural and coherent try-on video. Current virtual try-on mainly includes two implementation paradigms: 1) Mask-based paradigm [7, 15, 27, 30, 47] and 2) Mask-free paradigm [16, 23, 28, 39, 57]. The former considers the virtual try-on process as a mask-inpainting task, which leverages the pre-acquired agnostic mask to determine the try-on area. However, when dealing with more complex realistic person videos, these pre-acquired agnostic masks not only destroy the original spatial information of person actions and postures, but also lead to the temporal inconsistency of the try-on area, which has a high risk of generating conflicting and disjointed try-on videos. The latter is constructed by learning from paired training data which is generated by the pre-trained mask-based model, avoiding the destruction of essential spatial-temporal information in original person data. However, due to the lack of mask guidance, it is more difficult for the mask-free model to identify the accurate try-on area in each frame of the video, leading to inconsistencies in the try-on area of the generated videos.

To alleviate issues that brought by above two paradigms, it is worth considering the other explicit guidance to simultaneously achieve 1) *guidance on the garment try-on area*, 2) *preservation of video spatial-temporal information* and 3) *coherence of generated video virtual try-on results*.

3.3. Point-Enhanced Mask-Free paradigm

Overview. In this work, we propose a new point-enhanced mask-free paradigm to conduct the video virtual try-on task. Specifically, we first exploit a pre-trained mask-based try-on model to construct paired pseudo-person training samples, thus learning a mask-free virtual try-on model. Then, to further boost the try-on performance, we leverage the pre-acquired sparse frame-cloth and frame-frame point alignments and integrate a novel Point-Enhanced Transformer (PET) into the mask-free model. In PET module, the designed Point-enhanced Spatial Attention (PSA) and Point-enhanced Temporal Attention (PTA) perform explicit feature alignments of garment image and video frames, as well as alignments between video frames, respectively, thereby greatly enhancing the garment transfer ability and temporal coherence on more complex realistic human videos. The rest of this section will introduce our proposed method in detail. The pipeline of our PEMF-VTO and PET is shown in Fig. 2. For the pre-trained mask-based model, we adopt the open-sourced video virtual try-on method ViViD [15].

3.3.1. Pseudo-Person Data Preparation

Given a person video $\{\mathbf{x}\}_1^T$ and a garment image $\mathbf{g}$, we should first obtain the cloth-agnostic video $\{\mathbf{a}\}_1^T$, agnostic mask sequence $\{\mathbf{m}\}_1^T$ and human pose $\{\mathbf{p}\}_1^T$ of $\{\mathbf{x}\}_1^T$. Following [15], the $\{\mathbf{a}\}_1^T$ and $\{\mathbf{m}\}_1^T$ are extracted through the human parsing model SCHP [2, 32]. The pose detection model DensePose [19] is employed to acquire the pose information $\{\mathbf{p}\}_1^T$. Besides, the garment mask $\mathbf{m}_g$ of $\mathbf{g}$ is segmented by semantic segment model SAM [31]. After acquiring above inputs, we leverage the pre-trained mask-based model to perform the virtual try-on operation on the publicly available image and video virtual try-on datasets with randomly selected same-type clothes. Then, following [39], the large-scale paired training data is applied to train a powerful mask-free video virtual try-on model.

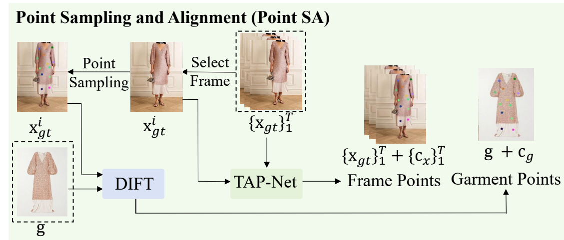


Figure 3. The pipeline of the construction for point alignments between video frames and garment images.

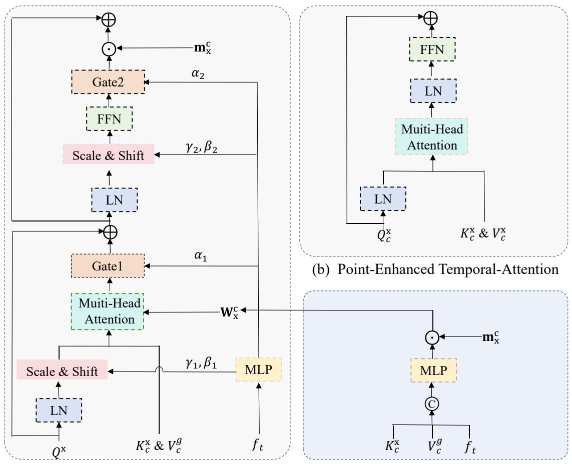
3.3.2. Point-Enhanced Transformer

Motivated by recent studies[4, 35, 36, 42], we choose flexible and powerful matching point pairs to enhance both garment transfer accuracy and video coherence. Specifically, we first acquire the sparse frame-cloth and frame-frame point alignments through the diffusion-based matching model DIFT and point tracking model TAP-Net. Then, as shown inFig. 2, a **P**oint-**E**nhanced **T**ransformer (PET), which adds the **P**oint-enhanced **S**patial **A**ttention (PSA) and **P**oint-enhanced **T**emporal **A**ttention (PTA) compared to the transformer layer of denoising U-Net $\mathcal{D}$ in baseline model, is designed to fully leverage the guidance provided by these point alignments.

Point Sampling and Alignment. The pipeline of our Point Sampling and Alignment (Point SA) is shown in Fig. 3. In the training stage, we first randomly select a frame $\mathbf{x}_{gt}^i$ from the realistic ground truth video $\{\mathbf{x}_{gt}\}_1^T$. Then M sparse points are randomly sampled from the agnostic mask (try-on area) of $\mathbf{x}_{gt}^i$. In this work, the number of selected points $M \leqslant K$, where K denotes the maximum number of matching correspondences. After that, DIFT will calculate the semantic-aware frame-cloth point alignments of these sparse points from the garment image $\mathbf{g}$. Finally, we exploit the point tracking model TAP-Net to acquire the frame-frame correspondences based on the select points in $\mathbf{x}_{gt}^i$. In the inference stage, after selecting a frame $\mathbf{x}^i$ from the source video $\{\mathbf{x}\}_1^T$, users can mark the matching points of $\mathbf{x}^i$ and $\mathbf{g}$ to achieve a spatially controllable and temporally smooth virtual try-on process.

After acquiring the point alignments, we construct two $0-1$ binary masks $\{\mathbf{c}_{\mathrm{x}}\}_1^T$ and $\mathbf{c}_g$ to respectively represent the positions of frame points in $\{\mathbf{x}_{gt}\}_1^T$ and garment points in $\mathbf{g}$. The values at the selected point positions on $\{\mathbf{c}_{\mathrm{x}}\}_1^T$ and $\mathbf{c}_g$ are set to 1. A max pooling operation MaxPool is then applied to obtain the masks in lower resolutions, which align with the denoising U-Net $\mathcal{D}$. Besides, a 3×3 channel-wise convolution is applied to $\{\mathbf{c}_{\mathrm{x}}\}_1^T$ and $\mathbf{c}_g$ to expand the receptive field of matching points, thus obtaining a soft alignment mask $\mathbf{m}_{\mathrm{x}}^c$.

Point-enhanced Spatial Attention. Referring to the powerful Diffusion Transformer (DiT) [40], in Fig. 4 (a), we design our point-enhanced spatial attention (PSA) to pro-



(a) Point-Enhanced Spatial-Attention (b) Point-Enhanced Temporal-Attention (c) Point-aware Attention Bias

Figure 4. Point-enhanced Spatial Attention (PSA) and Point-enhanced Temporal (PTA) Attention modules.

vide explicit guidance of the try-on area without destroying any spatial information on human movements and postures. Specifically, the input of j-th PET layer $\{\mathbf{d}_j\}_1^T$ is employed as the query $Q^{\mathrm{x}} \in \mathbb{R}^{T\times N\times C}$ of PSA, where N is the pixel number of $\mathbf{d}_j$ and C is the number of latent channel. Then, based on the point position representations $\{\mathbf{c}_{\mathrm{x}}\}_1^T$ and $\mathbf{c}_g$, we acquire the sparse frame-level point latents $\{\mathbf{d}_j^c\}_1^T$ and garment point latents $\mathbf{r}_j^c$ as the key $K_c^{\mathrm{x}} \in \mathbb{R}^{1\times M\times C}$ and value $V_c^g \in \mathbb{R}^{T\times M\times C}$ of PSA. Before entering the PSA, following [40], we exploit the timestep embedding f_t to regress the dimension-wise scale parameters γ_1, γ_2 and shift parametersβ_1 and β_2 of query $\{\mathbf{d}_j\}_1^T$ and the dimension-wise scale parameters α_1 and α_2 of residual connections by MLP layers. The formula of PSA is given as:

$$\begin{aligned} \mathrm{Attn} &= \mathrm{softmax}((\gamma_1 \cdot \mathrm{LN}(Q^{\mathrm{x}}) + \beta_1) \cdot K_c^{\mathrm{x}} + W_{\mathrm{x}}^c) \\ Q^{\mathrm{x}} &= Q^{\mathrm{x}} + \alpha_1 \cdot \mathrm{Attn} \cdot V_c^g, \end{aligned} \tag{3}$$

where the $\mathbf{W}_{\mathrm{x}}^c \in \mathbb{R}^{T\times N\times M}$ is a point-wise attention bias to adaptively adjust the similarities between $\{\mathbf{d}_j\}_1^T$ and $\mathbf{r}_j^c$ in the try-on area. The $\mathbf{W}_{\mathrm{x}}^c = \mathbf{m}_{\mathrm{x}}^c \cdot \mathrm{FFN}(\mathrm{Cat}(K_c^{\mathrm{x}}, V_c^g, f_t))$ as Fig. 4 (c) shown. Besides, to further enhance the controllability of the try-on area, the soft alignment mask $\mathbf{m}_{\mathrm{x}}^c$ explicitly constrains the updated residual to only work in the surrounding area of point pairs by the following formula:

$$Q^{\mathrm{x}} = Q^{\mathrm{x}} + m_{\mathrm{x}}^c \cdot \alpha_2 \cdot \mathrm{FFN}(\gamma_2 \cdot \mathrm{LN}(Q^{\mathrm{x}}) + \beta_2) \tag{4}$$

Point-enhanced Temporal Attention. In [20, 27], it employed a temporal attention module to ensure the coherence of the generated video. However, this method assumes that the pixels between different frames with the same coordinates are matching points, which is inexact or even wrong, especially for human videos with complex actions or movements. To this end, as shown in Fig. 4 (b), we leverage the pre-acquired frame-frame alignments $\{\mathbf{c}_{\mathrm{x}}\}_1^T$ to achieve a more reasonable point-enhanced temporal attention (PTA) to further improve the coherence of the try-on video. Concretely, we employ the sparse frame-level point latents $\{\mathbf{d}_j^c\}_1^T$ as the query Q_c^{x} of the PTA, while utilizing the sparse frame-level point latents $\{\mathbf{d}_j^c\}_1^T$ in conjunction with garment point latents $\mathbf{r}_j^c$ as the key K_c^{x} and value V_c^{x}. Then, the multi-head attention is employed to Q_c^{x}, K_c^{x} and V_c^{x} to enhance the consistency between frames of the surrounding area of these matching points $\{\mathbf{c}_{\mathrm{x}}\}_1^T$, thus promoting the overall video smoothness. The formula is given as:

$$Q_c^{\mathrm{x}} = Q_c^{\mathrm{x}} + \mathrm{FFN}(\mathrm{LN}(\mathrm{Softmax}(\mathrm{LN}(Q_c^{\mathrm{x}}) \cdot K_c^{\mathrm{x}}) \cdot V_c^{\mathrm{x}})) \tag{5}$$

Remark. The implementation of PEMF-VTO is mainly motivated by the deficiencies of existing virtual try-on paradigms in 3.2, which can be clearly illustrated by Fig. 1. From Fig. 1, we observe that the mask-based method can not restore the crucial spatial-temporal details from the agnostic mask, and the mask-free method struggles to accurately perceive the try-on regions, especially for realistic in-the-wild human videos. To this end, it is natural and intuitive to leverage more flexible and reliable point alignments [4, 11, 44] to guide the virtual try-on model, thereby obtaining more reasonable and coherent video virtual try-on results. Specifically, our designed PSA explicitly enhances the precise garment transfer to the try-on area with the cross-attention operation between sparse frame-cloth point pairs. Besides, the PTA leverages inter-frame matching point pairs to conduct a enhanced temporal attention, thus achieving superior coherence of the try-on video. The detailed ablation experiments in 4.4 will show the effectiveness of our proposed PET module.

3.4. Training and Inference

Training scheme. The training scheme of PEMF-VTO comprises three stages: **(1) Stage 1**: the denoising U-Net $\mathcal{D}$ is initialized as a 2D inpainting model, and only a single frame of pseudo-person data is taken as training data, which enables the model to perceive and transfer the reference garment to a reasonable try-on area. The parameters of $\mathcal{P}$, $\mathcal{R}$ and $\mathcal{D}$ are updated in this stage. **(2) Stage 2**: we initialize the temporal attention module with the parameters in [20] and exploit both image and video data to only finetune it, thus enhancing the temporal coherence of generated try-on results. **(3) Stage 3**: since most of the training samples have been well learned through the first and second stages, we first leverage the mask-free model of the second stage to construct hard-paired pseudo-person training samples with lower generation performance (*i.e.* SSIM < 0.75). Then, to ensure the PSA and PTA can really promote the temporal consistent and controllable garment transfer, we only train the PSA module and the PTA module with these hard training samples. The learning objectives of the three training stages are the same LDM loss in Eq. (2).

Method	Unpaired						Paired					
	VVT		ViViD		TikTok		VVT			ViViD		
	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	$\text{VFID}_I\downarrow$	$\text{VFID}_R\downarrow$	SSIM↑	LPIPS↓	$\text{VFID}_I\downarrow$	SSIM↑	LPIPS↓	$\text{VFID}_I\downarrow$
PF-AFN* [16]	5.12	0.125	43.22	1.87	43.38	7.16	0.856	0.233	7.33	0.776	0.183	39.28
Flow-Style* [23]	4.79	0.097	41.93	1.69	42.28	11.38	0.854	0.241	7.64	0.781	0.169	38.64
StableVITON* [30]	-	-	36.90	0.91	-	-	0.902	0.078	3.54	0.802	0.134	34.24
IDM-VTON* [7]	2.77	0.027	25.50	0.72	38.62	5.97	0.896	0.079	3.91	0.823	0.116	20.08
CatVTON* [8]	2.49	0.016	22.65	1.14	39.67	6.53	0.899	0.082	3.75	0.834	0.089	17.73
ViViD† [15]	3.99	0.041	21.80	0.82	46.73	6.94	0.822	0.107	3.78	0.803	0.122	17.29
CatV²TON† [9]	1.90	0.014	19.51	0.53	43.62	6.18	0.900	0.039	1.78	0.873	0.064	13.60
PEMF-VTO	**0.95**	**0.007**	**16.67**	**0.34**	**31.62**	**2.05**	**0.915**	**0.035**	**0.87**	**0.911**	**0.040**	**7.54**

Table 1. Quantitative results on the VVT, ViViD and TikTok datasets. The best are marked in bold. The $*$ and † respectively denote image and video virtual try-on methods.

Method	VITON-HD			DressCode		
	SSIM↑	LPIPS↓	FID↓	SSIM↑	LPIPS↓	FID↓
PF-AFN* [16]	0.857	0.142	17.28	0.878	0.102	20.51
Flow-Style* [23]	0.860	0.133	16.84	0.882	0.094	19.24
LaDI-VTON* [34]	0.871	0.094	13.01	0.915	0.062	16.71
StableVITON* [30]	0.888	0.073	-	-	-	-
IDM-VTON* [7]	0.870	0.102	**6.29**	0.920	0.062	8.64
GPD-VVTO† [47]	0.891	0.070	8.57	0.924	0.045	4.18
ViViD† [15]	0.881	0.089	8.67	0.907	0.070	8.28
PEMF-VTO	**0.894**	**0.062**	6.86	**0.927**	**0.034**	**3.41**

Table 2. Quantitative results on the VITON-HD and DressCode datasets. The best are marked in bold. The $*$ and † respectively denote image and video virtual try-on methods.

Inference scheme. In the inference stage, users can manually click on the matching point pairs between a randomly selected single video frame and reference garment image, thereby acquiring a more coherent and natural generation, especially for the video with complex actions and postures.

4. Experiments

4.1. Dataset and Experimental Setting

Datasets. Following our baseline backbone [15], we train our model on two image datasets, VITON-HD [6] and DressCode [33], and one video dataset, ViViD [15]. For the video virtual try-on, following CatV^{2}TON [9], the VVT [12] dataset and a testset of ViViD [15] are selected as the evaluation datasets. Besides, referring to [24], a more realistic and challenging TikTok [29] dataset is chosen to further illustrate the superiority and generalization of our method. For the image virtual try-on, comparative experiments are conducted on the test sets of VITON-HD, DressCode and StreetVTON [10] datasets. All datasets are open-sourced for wide research purposes. The detailed introduction of all datasets is shown in the supplementary materials.
Metrics. Following previous methods [9, 15, 47], both frame-level metrics LPIPS [55], SSIM [48], FID [25] and video-level metric VFID [45] metric with I3D [3] and ResNext are adopted for comprehensive evaluation.
Implementation Details. We initialize the denoising U-Net $\mathcal{D}$ and reference U-Net $\mathcal{R}$ with the weights from Stable Diffusion-1.5. The temporal module is initialized with the weights from the motion module of [20], and the CLIP image encoder is the same as the baseline model [15]. During training, all data is resized to a uniform resolution of 512×512. Experiments are conducted on 8 Nvidia H100 GPUs with a learning rate of 5e-5. The numbers of training iterations in the three stages are 40000, 40000 and 20000 respectively. Since images can be considered as a single frame of videos, both image and video data are utilized during the three training stages, with the proportions of video data being 0.3, 0.9 and 0.5 for each stage respectively. In each iteration, only one form of training data is chosen, either video or image. With the selection of image datasets, training is conducted with a batch size of 128. In contrast, the selection of video datasets involves training the model with 16-frame video sequences and a batch size of 8. To enable a fair comparison with mask-based methods, the agnostic mask can be treated as a specific garment to participate in the training of our PEMF-VTO, accompanied by 20% ratios. The maximum number of points K is set to 16.

4.2. Quantitative Results

Video Virtual Try-on. To achieve a fair and comprehensive comparison with previous methods, we evaluate our method on two test settings for the video virtual try-on task. The unpaired setting means that the virtual try-on model should substitute the garment of the input source person video with a different garment. The paired setting is defined as reconstructing the person video by providing an agnostic person video and its original garment. In Tab. 1, the two open-sourced video try-on methods ViViD [15] and CatV^{2}TON [9] are chosen as the video-based comparative methods. Besides, many relatively high-performing image-based virtual try-on methods are also exploited to enrich the experimental comparisons. As Tab. 1 shown, our PEMF-VTO significantly outperforms all image-based and video-based methods in all metrics of two different settings. The superior performance on two simple model video datasets VVT and ViViD and one realistic in-the-wild dataset TikTok further demonstrates the generalization ability of our method.
Image Virtual Try-on. To further illustrate the effective-

Figure 5. Qualitative comparison on StreetVTON, ViViD and TikTok datasets. *Best viewed with Acrobat Reader. Click the images to play the video clips.*

ness and robustness of our proposed method, for the image virtual try-on task, we compare our method with both image-based and video-based virtual try-on methods in the paired setting. As shown in Tab. 2, although our approach is mainly designed for video virtual try-on, it outperforms two video-based try-on methods across all metrics and achieves comparable performance with SOTA image-based virtual try-on methods, This is primarily due to the lower quality of video frames used during the training stage of our PEMF-VTO, which negatively impacts the model's image generation capability. These experimental results show that our method can also successfully generalize to the image virtual try-on task to show competitive performance.

4.3. Qualitative Results

As shown in Fig. 5, we show the visualization comparisons of our method with the SOTA image-based method CatVTON [8] and two video-based method CatV^{2}TON [9] and ViViD [15]. Specifically, our generated results achieve significant visual fidelity to the reference garment in the spatial dimension and content consistency in the temporal dimension in all datasets. It is worth noting that our model is only trained on simple model images and videos and obtains superior try-on results on more challenging StreetVTON and TikTok datasets, which demonstrates the effectiveness and generalization of our method.

Besides, in Fig. 1 and Fig. 6 (a), it is clear that existing mask-based and mask-free virtual try-on paradigms have obvious deficiencies, especially for realistic video data with complex scene changes and body movements. Concretely, the mask-based method can not recover the important spatial-temporal information due to the inaccurate agnostic mask. The mask-free method may get confused about how to determine a reasonable and coherent try-on area. Differently, our PEMF-VTO can simultaneously achieve the precise determination of try-on area and the video coherence without the guidance of the agnostic mask. Furthermore, as Fig. 6 (b) shown, our method has the potential to conduct a more flexible and controllable try-on process through the explicit guidance of the point alignments. We provide all video results in the supplementary materials.

4.4. Ablation Studies

In this section, we design different variants to perform a detailed analysis of our PET module. To increase the credibility, in Tab. 3, we also synthesized 430 pseudo video pairs with the points alignments from the ViViD testing dataset. Thus, the LPIPS and SSIM metrics can be applied to further evaluate the generation quality.

Point-Enhanced Transformer. As shown in Tab. 3, compared to the baseline (Mask-Free variant), our PEMF-VTO obtains obvious performance improvement to acquire more natural and coherent try-on results.

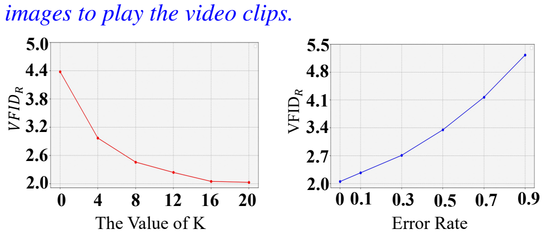
Point-Enhanced Spatial Attention. Based on the frame-cloth point alignments, the PSA is designed to provide explicit guidance to transfer the garment into the desirable try-on area, which significantly improves the performance of our baseline model as Tab. 3 shown. Besides, in the upper part of Tab. 3, we analyze the effectiveness of soft alignment mask $\mathbf{m}_x^c$ and point-wise attention bias $\mathbf{W}_x^c$. Our PSA obtains performance decrease when removing the two designs, which illustrates the reasonability of PSA.

Point-Enhanced Temporal Attention. The complex human actions at the temporal level always lead to the discontinuity of the try-on video. To this end, we design the PTA to enhance the coherence of the try-on area between different frames. As Tab. 3 shown, the PTA brings clear improvements to the baseline model. Besides, the performance of

(a) Ablation visual results about PET

(b) Controllability of our method

Figure 6. Qualitative results for the (a) Ablation visual results about PET and (b) Controllability of our method. The top row and bottom row of each subfigure are respectively from StreetV-TON and ViViD datasets. *Best viewed with Acrobat Reader. Click the images to play the video clips.*



(a) Influence of K (b) Robustness Analysis of PET

Figure 7. (a) Ablation studies for different K and (b) Robustness Analysis for different error rates of point alignments. We conduct the experiments on the TikTok dataset.

PTA decreases when only selecting the frame-level point latents $\{\mathbf{d}_j^c\}_1^T$ (*i.e.* *w/o* $\mathbf{r}_j^c$) as the key K_c^{x} and value V_c^{x} of PTA, which further verifies the effectiveness of our PTA.

Different maximum number of points. In Fig. 7 (a), we investigate the effects of the maximum number of points K on the performance of the TikTok dataset. As K increases, the FID and VFID metrics first improve and then stabilize. It can be attributed to the repetition of more sampling points. Therefore, we choose $K = 16$ in this paper.

Robustness Analysis of PET. The pre-acquired point alignments have the probability of error which may lead to a negative impact on the try-on performance. Therefore, to verify the robustness of our PET, we first randomly perturb a specific proportion of the pre-acquired point alignments to obtain the point pair guidance with different error

Variants	Realistic				Pseudo	
	ViViD		TikTok		ViViD	
	VFID$_I$↓	**VFID$_R$↓**	**VFID$_I$↓**	**VFID$_R$↓**	**SSIM↑**	**LPIPS↓**
ViViD	21.80	0.82	46.73	6.94	0.840	0.114
Mask-Free	19.32	0.57	38.96	4.38	0.857	0.083
PEMF-VTO	**16.67**	**0.34**	**31.62**	**2.05**	**0.867**	**0.073**
base + PSA	**17.86**	**0.42**	**33.69**	**2.68**	**0.863**	**0.075**
w/o $\mathbf{m}_{\mathrm{x}}^c$	18.44	0.47	34.72	3.14	0.860	0.078
w/o $\mathbf{W}_{\mathrm{x}}^c$	18.23	0.45	34.17	2.91	0.862	0.076
base + PTA	**18.03**	**0.44**	**34.78**	**2.91**	**0.864**	**0.076**
w/o $\mathbf{r}_j^c$	18.52	0.48	35.61	3.17	0.861	0.079

Table 3. Ablation studies for our PSA and PTA modules on realistic ViViD and TikTok datasets and pseudo ViViD dataset.

rates. In Fig. 7 (b), if the error rate is low (*i.e.* < 20%), our method shows impressive robustness to achieve similar improvement with all correct variants. Besides, even though the error rate is nearly 70%, the VFID$_R$ of our PEMF-VTO on the TikTok dataset also outperforms the base mask-free model, thereby significantly reflecting the reasonability and robustness of our proposed PET module.

5. Limitation

Since the quality and diversity of the person-pseudo training data is limited, our method is still insufficient in maintaining and transferring garment details, such as the bag strap on the shoulder of the person in the bottom of Fig. 6 (a). Besides, the ability to control and edit the try-on area should be further improved. In the future, we will construct more diverse training data to continually train our model, thereby promoting the continuous upgrading and evolution of our video virtual try-on model to achieve better performance for more realistic and difficult try-on data.

6. Conclusion

To alleviate the deficiencies of existing virtual try-on paradigms and synthesize more realistic and coherent video try-on results, in this work, we propose a novel Point-Enhanced Mask-Free Video Virtual Try-On method (PEMF-VTO). Concretely, we first leverage the pre-trained try-on model to construct paired pseudo-person training samples to learn a mask-free try-on model. Then, based on the pre-acquired sparse alignments, the Point-enhance Spatial Attention (PSA) and Point-enhance Temporal Attention (PTA) are designed to improve the garment transfer ability and coherence of more complex realistic human video. Our PEMF-VTO can simultaneously achieve 1) the accurate transfer of reference garment 2) the preservation of non-try-on areas and 3) the continuity of generated video frames. Extensive quantitative and qualitative experimental results clearly show the effectiveness of our method.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In *CVPR*, 2017. 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[4] Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. *arXiv preprint arXiv:2403.12965*, 2024. 4, 5

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 2

[6] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021. 2, 6

[7] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 1, 2, 3, 6

[8] Zheng Chong, Xiao Dong, Haoxiang Li, Shiyue Zhang, Wenqing Zhang, Xujie Zhang, Hanqing Zhao, Dongmei Jiang, and Xiaodan Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 6, 7

[9] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025. 6, 7

[10] Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, and Svetlana Lazebnik. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. *arXiv preprint arXiv:2311.16094*, 2023. 2, 6

[11] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, and Andrew Zisserman. BootsTAP: Bootstrapped training for tracking-any-point. *arXiv preprint arXiv:2402.00847*, 2024. 5

[12] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019. 6

[13] Xue Dong, Xuemeng Song, Na Zheng, Jianlong Wu, Hongjun Dai, and Liqiang Nie. Tryoncm2: Try-on-enhanced fashion compatibility modeling framework. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):246–257, 2022. 2

[14] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022. 2

[15] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. 1, 2, 3, 4, 6, 7

[16] 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. 1, 2, 3, 6

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[18] 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. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 2

[19] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense Human Pose Estimation In The Wild. In *CVPR*, 2018. 4

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 5, 6

[21] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 2

[22] 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.

[23] 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. 1, 2, 3, 6

[24] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. *arXiv preprint arXiv:2407.10625*, 2024. 1, 3, 6

[25] 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

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[27] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 3, 5

[28] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 2020. 3

[29] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 2, 6

[30] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024. 1, 2, 3, 6

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[32] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3260–3271, 2020. 4

[33] 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 European Conference on Computer Vision*, 2022. 6

[34] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. In *Proceedings of the ACM International Conference on Multimedia*, 2023. 2, 6

[35] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. *arXiv preprint arXiv:2402.02583*, 2023. 4

[36] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 4

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1

[38] Shuliang Ning, Duomin Wang, Yipeng Qin, Zirong Jin, Baoyuan Wang, and Xiaoguang Han. Picture: Photorealistic virtual try-on from unconstrained designs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6976–6985, 2024. 1

[39] Yunfang Niu, Lingxiang Wu, Dong Yi, Jie Peng, Ning Jiang, Haiying Wu, and Jinqiao Wang. Anydesign: Versatile area fashion editing via mask-free diffusion. *arXiv preprint arXiv:2408.11553*, 2024. 1, 3, 4

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4, 5

[41] 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. 3

[42] Yujun Shi, Chuhui Xue, Jun Hao Liew, Jiachun Pan, Hanshu Yan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8849, 2024. 4

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[44] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 5

[45] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[46] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 2

[47] Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. Gpd-vvto: Preserving garment details in video virtual try-on. In *ACM Multimedia 2024*, 2024. 1, 3, 6

[48] 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

[49] Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos. *arXiv preprint arXiv:2404.17571*, 2024. 1, 3

[50] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3460–3469, 2022. 2

[51] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7026, 2024. 1, 3

[52] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8372–8382, 2024. 1

[53] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021. 2

[54] 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. 1, 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[56] Xuanpu Zhang, Dan Song, Pengxin Zhan, Qingguo Chen, Kuilong Liu, and Anan Liu. Better fit: Accommodate variations in clothing types for virtual try-on. *arXiv preprint arXiv:2403.08453*, 2024. 1, 3

[57] Xuanpu Zhang, Dan Song, Pengxin Zhan, Qingguo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Anan Liu. Boowvton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. *arXiv preprint arXiv:2408.06047*, 2024. 1, 3

[58] Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint*, 2024. 1, 3

[59] Na Zheng, Xuemeng Song, Qingying Niu, Xue Dong, Yibing Zhan, and Liqiang Nie. Collocation and try-on network: Whether an outfit is compatible. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 309–317, 2021. 2

[60] 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 (CVPR)*, pages 4606–4615, 2023. 2

[61] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 1