Efficient and Robust Video Virtual Try-On via Enhanced Multi-garment Alignment

Zijian He†, Peixin Chen†, Guolin Zheng, Guangrun Wang *Member, IEEE*, Xiaonan Luo, Liang Lin *Fellow, IEEE*, Guanbin Li *Member, IEEE*

***Abstract*—Video virtual try-on aims to generate realistic sequences where garments maintain their identity and adapt accurately to a person's pose and body shape in source video. This task can be regarded as video inpainting, whereas previous methods focus primarily on the specific try-on region while simply "copying" the remaining parts of the person. However, this approach limits the degrees of freedom and heavily relies on precise human parsing. In complex in-the-wild scenarios, dynamic blurring and limb occlusions can introduce errors and discontinuities in the inpainting regions, adversely affecting the video try-on results. Our solution, VidClothEditor, adopts a relaxed editing approach that allows for full-body inpainting and treats non-edited regions as a reconstruction task. It utilizes multiple garment alignment with a proposed region guidance to enhance the naturalness of video try-on results. Additionally, we employ garment-augmented video consistency learning, which significantly reduces the inference time and increases the practical potential for video editing. Comprehensive experiments on the VITON-HD and TikTok datasets confirm VidClothEditor's ability to generate high-quality images and smooth videos. The project website is at video-tryon.github.io.**

***Index Terms*—Video try-on in the wild, multiple garment alignment, controlled diffusion model, video consistency learning**

Fig. 1. Achieving precise human parsing in in-the-wild videos is challenging, often leading to inconsistent and erroneous inpainting region sequences. Previous methods, such as AnimateAnyone, which rely on such strict conditional sequences, are inherently prone to failure (shown in the red boxes). We instead employ a full-body inpainting strategy, incorporating a novel region-guidance multi-garment alignment module, to ensure consistent and accurate video virtual try-on results.

I. Introduction

VIDEO virtual try-on is a video editing task aimed at generating seamless videos that maintain the appearance of a particular garment while precisely conforming it to the pose and body shape of the human in the source video. This task has garnered increasing attention due to its wide-ranging potential applications in e-commerce, digital avatar live streaming, and short-form video editing.

Research on virtual try-on primarily falls into two categories: GAN-based [1]–[24] and diffusion-based approaches [25]–[29]. GAN-based methods typically employ a two-stage process consisting of warping and blending. These methods entail a complicated and protracted workflow, and exhibit high sensitivity to occlusions. On the other hand, the recent diffusion-based methods leverage the foundational capabilities of pretrained diffusion models to facilitate a single-stage, end-to-end virtual try-on process. These methods integrate garment features directly into the denoising process through cross-modal attention mechanisms, implicitly learning the warping process.

Despite significant advancements in virtual try-on, extending these techniques to in-the-wild videos remains challenging due to two key issues: 1) The inconsistent conditions of compact inpainting regions for generation. Previous methods, including both GAN-based and diffusion-based methods [1]–[29], consider virtual try-on as an inpainting task and hope to construct this region strictly to avoid changes to the remaining parts. However, in in-the-wild scenes (such as short-video editing), due to unusual clothing pairings and dynamic blurring, human parsing is prone to failure. Strictly constructing this inpainting region will result in discontinuous input conditions. As illustrated in Fig. 1, the inconsistent inpainting region (highlighted in red boxes) ultimately gives rise to incorrect editing results. 2) The suboptimal trade-off between performance and inference speed. GAN-based methods [1], [2], [8], [12], [22] are time efficient but suffer from limited generalization capability, preventing them from effectively handling complex, real-world scenarios characterized by dynamic actions. Conversely, while diffusion-

Zijian He, Peixin Chen, Guolin Zheng, Guangrun Wang, Liang Lin and Guanbin Li are with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, and are also with the Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, 510006, China. Xiaonan Luo is with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China. E-mail:hezj39@mail2.sysu.edu.cn, chenpx28@mail2.sysu.edu.cn, zhengglin@mail2.sysu.edu.cn, wanggrun@gmail.com, luoxn@guet.edu.cn, linliang@ieee.org, liguanbin@mail.sysu.edu.cn. Corresponding author: Guanbin Li.

†The first two authors share equal contributions.

based methods [25]–[31] achieve much more impressive try-on effects, their iterative reverse sampling process results in slow generation speeds, limiting their applicability in real-time scenarios. Some researchers have introduced consistency models [32], [33], distilled from pretrained diffusion models for faster generation. However, these methods have not adequately preserved image texture details.

To tackle the aforementioned challenges in complex natural scenes, we introduce VidClothEditor, an efficient and robust virtual video try-on framework. VidClothEditor extends the inpainting region to encompass the entire body and treats non-edited regions as a reconstruction task. At the same time, both non-edited and target try-on garments are regarded as weak video conditions, requiring the model to extract their features to produce a comprehensive try-on result across the whole body region. This approach relaxes the strict condition associated with precise inpainting regions, offering two main advantages: 1) Consistent inpainting regions can be more easily obtained even in in-the-wild scenarios, ensuring the correct condition for video virtual try-on. 2) The generative model no longer processes edited and non-edited garments separately, enabling it to learn the integration of multiple garment features, thereby enhancing the naturalness of the try-on effect. As shown in Fig. 1, our method can not only edit the upper garment but also reconstruct the bottoms successfully. Inspired by Animate Anyone [34], VidClothEditor employs spatial attention to implement implicit warping. We observed that attentions on upper and lower regions may interfere with each other (as seen in Fig. 3), leading to unsatisfactory results. To address this, we propose a region-guidance multiple garment fusion strategy, where coarse human parsing is introduced for guidance to assist in the learning of appropriate attention regions. This "soft" guidance from human parsing enables the model to learn the structural priors of the human body during training while avoiding the strict definition of inpainting regions during inference. Additionally, recognizing the substantial time consumption of diffusion-based frameworks, we develop a garment-augmented video consistency learning approach for fast inference. During the consistency distillation process [32], [33], we enhance the control over garment features to mitigate the loss of clothing details through distillation. The video consistency model significantly reduces the required sampling steps, thereby enhancing its potential for real-time video editing applications. Our contributions can be summarized as follows:

- We present VidClothEditor, an efficient and robust diffusion-based video virtual try-on framework.
- VidClothEditor employs weak conditions and full-body inpainting to increase flexibility, mitigate the impact of inaccurate human parsing, and enhance the naturalness of try-on results. Additionally, region guidance for multiple garment alignment addresses potential interference between garments by using coarse parsing results to guide the learning of appropriate attention regions.
- We propose a garment-augmented video consistency learning approach that enables VidClothEditor to perform efficient and high-quality video-level editing.
- Extensive experiments and evaluations conducted on the VITON-HD and real-life TikTok datasets demonstrate that our method achieves the state-of-the-art video try-on results.

II. Related Work

A. Image Virtual Try On

Given a pair of images (reference person, target garment), image virtual try-on methods aim to generate the appearance of the reference person wearing the target garment. A majority of virtual try-on methods [1]–[24] decompose the task into two stages, (1) deforming the clothing to fit garment region on the human body and (2) blending the warped clothing into the target human via try-on generator. Prior approaches [1], [7], [8], [13], [35] utilize trainable networks to estimate dense flow maps for precise clothing deformation. Furthermore, researchers make attempts to mitigate the misalignment between the warped clothing and the human body, including parser-free strategies and the integration of additional information such as local flow [1], semantic maps [21], [22] or 3D geometric priors [36]. Despite significant advancements, these methods still face challenges in handling complex poses and occlusions caused by pixel misalignment. [23], [24] address self-occlusion using semantic parsing or human keypoints; however, their performance is limited in in-the-wild images.

In recent years, diffusion models [37]–[39] have emerged as prominent contenders of generative models. LaDI-VTON [25] and DCI-VTON [26] use the latent diffusion model as the generator in the blending stage, replacing generative adversarial networks (GANs). To overcome the limitations of the two-stage approach, researchers have explored implicit warping and single-step generation. TryOnDiffusion [28] introduced a diffusion-based architecture with two parallel UNets, preserving intricate garment details and enabling implicit garment warping to adapt to significant pose and body variations within a single network. StableVITON [27] learns the semantic correspondence between garment and person within the latent space via zero cross-attention blocks. OOTDiffusion [29] proposes outfitting fusion in the self-attention layers of the UNet to align the garment feature with the target human body.

B. Video Virtual Try On

Researchers have extended image-based virtual try-on methods to video applications, using specially designed temporal modules. FW-GAN [40] pioneered this field by adapting a video generation framework to video virtual try-on, incorporating warped garments and human postures as conditions. MVTON [41] introduced a try-on module for garment warping via pose alignment and regional pixel displacement, along with a memory refinement module that embeds prior frames into latent space as external memory for subsequent frame generation. ClothFormer [42] improved flow predictions using inter-frame data and employed a Dual-Stream Transformer to derive video try-on outcomes from multiple frame warpings. Despite significant advancements, GAN-based methods rely on parsing and warping pipelines, limiting their applicability to simple scenarios. For instance, the VVT dataset [40]

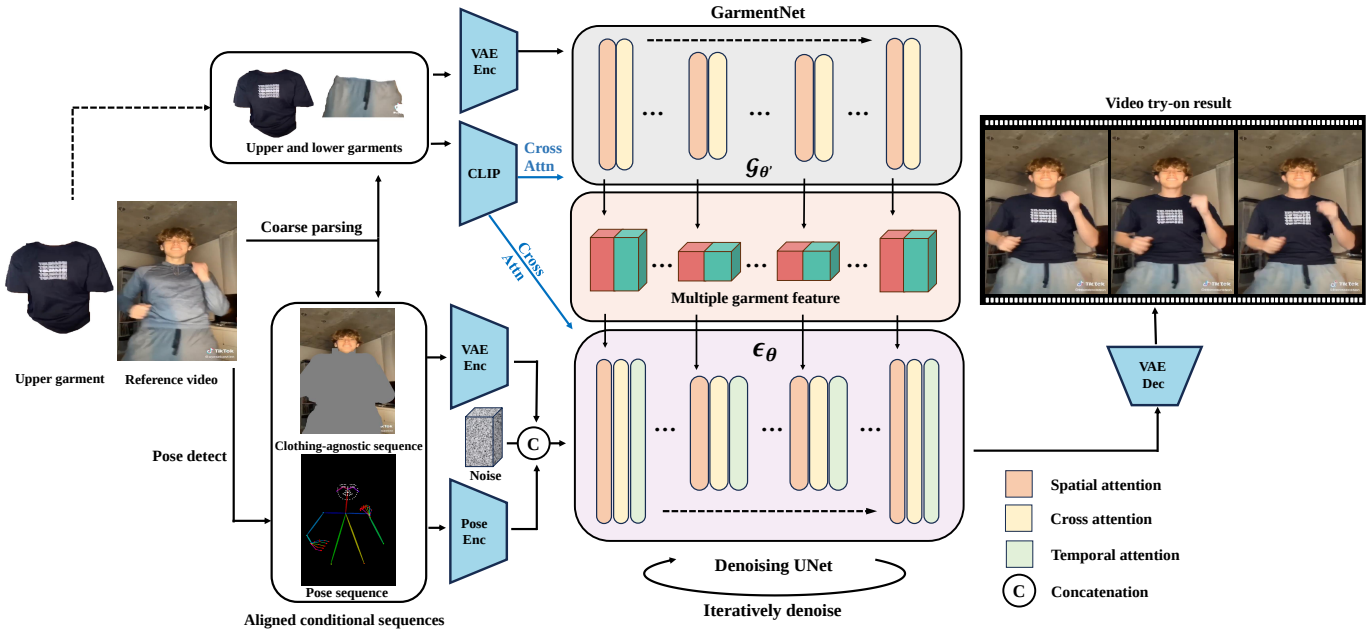

Fig. 2. Overview of the VidClothEditor framework: The reference video and target upper garment are initially preprocessed through segmentation and pose detection, yielding a clothing-agnostic sequence and a pose sequence. Full-body masking is applied to derive the lower garment image from the first frame. Next, the GarmentNet extracts hierarchical features from the target upper garment and the extracted lower garment. Finally, the UNet integrates aligned garment features and the conditional sequence into the denoising process to generate vivid video try-on results.

primarily includes simple textures, tight-fitting T-shirts, and repetitive human movements. To exploit the potentials of pre-trained image-based diffusion models, Tunnel Try-On [43] and ViViD [30] utilize a pre-trained inpainting U-Net as the main branch and introduce a reference U-Net to capture detailed clothing features. They augment temporal consistency by inserting standard temporal attention into the main UNet. However, existing methods are sensitive to complex human movements and dynamic blurs in real-life videos due to their intricate module designs. Our work leverages the foundational generation ability of pre-trained diffusion models and relaxes conditions to achieve robust and coherent video try-on results.

C. Controlled Diffusion Model for Visual Generation

In the field of visual generation, methods based on diffusion models have recently become the mainstream. Latent Diffusion [44] is a pioneer work that integrates text and images to achieve controlled image generation. ControlNet [45] and T2I-Adapter [46] explore the controllability of visual generation by integrating additional encoding layers, enabling controlled generation under multiple conditions such as pose, mask, edge, and depth. Anydoor [47] transforms specific image objects into identity features, merging these images into different surroundings while preserving the texture details of the objects. SGDM [48] proposes a style guidance module to equip the diffusion model with adaptive style personalization capability. DiffFashion [49] transfers a natural apperance image to a given clothing image for designing a new fashion. AnimateDiff [50] utilizes a vast amount of video data to train a motion module independently, which can be integrated into personalized Text-to-Image (T2I) models, bringing the generated images into videos with motion. VideoBooth [51] and Microcinema [52], in addition to text, introduce images to guide video generation, employing a proposed attention injection module to feed image embedding into the diffusion progress. In our work, we employ spatial attention to integrate the features of garment into the diffusion UNet model, effectively preserving the details of multiple garments.

III. Preliminaries

A. Diffusion Models

Using a predefined variance schedule β_t, we can establish a forward diffusion process in the latent space as outlined by denoising diffusion probabilistic models (DDPM) [39]:

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

where $t \in \{1, ..., T\}$, T represents the total number of forward diffusion steps, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \Pi_{s=1}^{t}\alpha_s$. As $N \to \infty$, the discrete Markov chain converges to the following stochastic differential equation (SDE),

$$\mathrm{d}\mathbf{z} = \mathbf{f}(\mathbf{z}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{2}$$

where $\mathbf{f}(\mathbf{z}, t) = -\frac{1}{2}\mathbf{z}\beta(t)$ and $g(t) = \sqrt{\beta(t)}$.

A notable property of this SDE is the existence of reverse-time ordinary differential equation (ODE). Moreover, the Probability flow ODE (PF-ODE) [53] indicates the presence of a correspoinding deterministic process whose solution trajectories at time t still follow the same noisy distribution $p_t(\mathbf{x})$ as ODE:

$$\mathrm{d}\mathbf{z} = \left[\mathbf{f}(\mathbf{z}, t) - g^2(t)\nabla_{\mathbf{z}} \log p_t(\mathbf{z})\right] \mathrm{d}t. \tag{3}$$

The score function $\nabla_{\mathbf{z}} \log p_t(\mathbf{z})$ can be predicted using a neural network $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)$, which then enables the removal of noise from the data point during the reverse process.

B. Latent Consistency Model

Consistency model (CM) [32] is a new family of generative models, designed to facilitate efficient generation through one-step or few-step sampling. Building on this foundation, the Latent Consistency Model (LCM) [33] extends CM into the latent space, enabling both fast and conditional generation. At its core, LCM expects to learn a function $f:(z_t,\mathbf{c},t)\longmapsto z_\epsilon$ which maps any point along a PF-ODE trajectory back to the origin; here $\mathbf{c}$ is the given conditions and ϵ is a small fixed number:

$$f(z_t,\mathbf{c},t)=f(z_{t'},\mathbf{c},t'),\forall t,t'\in[\epsilon,T] \tag{4}$$

To ensure the boundary condition, LCM is parameterized as:

$$\boldsymbol{f_\theta}(z,\mathbf{c},t)=c_{\text{skip}}(t)z+c_{\text{out}}(t)\boldsymbol{F_\theta}(z,\mathbf{c},t), \tag{5}$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are differentiable functions with $c_{\text{skip}}(\epsilon)=1$ and $c_{\text{out}}(\epsilon)=0$, and $\boldsymbol{F_\theta}(z,\mathbf{c},t)$ is a deep neural network. Then we can define the consistency loss as Eq. 6 to enforce the self-consistency property. Exponential moving average (EMA) is used to enhances the stability during training and the target model $\boldsymbol{\theta}^-$ is updated by $\boldsymbol{\theta}^-=\mu\boldsymbol{\theta}^-+(1-\mu)\boldsymbol{\theta}$.

$$\begin{aligned}&\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}^-;\Phi)\\&=\mathbb{E}_{z,\mathbf{c},t}\left[d\left(\boldsymbol{f_\theta}(z_{t_{n+1}},\mathbf{c},t_{n+1}),\boldsymbol{f_{\theta^-}}(\hat{z}^\phi_{t_n},\mathbf{c},t_n)\right)\right],\end{aligned} \tag{6}$$

where $d(\cdot,\cdot)$ is a distance measuring function and $\hat{\mathbf{z}}^\phi_{t_n}$ can be estimated by a one discretization step of a numerical ODE solver $\boldsymbol{\Phi}$ with a teacher diffusion model $\boldsymbol{\phi}$,

$$\hat{z}^\phi_{t_n}=z_{t_{n+1}}+(t_n-t_{n+1})\Phi(z_{t_{n+1}},t_{n+1},t_n,\mathbf{c},;\boldsymbol{\phi}). \tag{7}$$

IV. Method

Problem Statement: Given a reference person video sequence, denoted as $\mathbf{I}:=\{I_1,...,I_N\}\in\mathbb{R}^{3\times H\times W}$ and a target upper garment image $G_u\in\mathbb{R}^{3\times H\times W}$, where H and W represent height and width of the image respectively, and N is the number of frames in the sequence, video try-on requires to synthesis a realistic video sequence $\tilde{\mathbf{I}}:=\{\tilde{I}_1,...,\tilde{I}_N\}\in\mathbb{R}^{3\times H\times W}$. This sequence should depict the person wearing the target garment $\mathbf{g}_u$, while preserving the integrity of all other visual elements including the lower garment $\mathbf{g}_l$.

A. Preprocessing of Inputs

Video try-on requires editing the region of the garment while preserving other elements in the video. We expect to mask the full body of the human and the training progress can be regarded as self-reconstruction. Specifically, we obtain the sequences of human segmentation maps $\mathbf{S}:=\{S_1,...,S_N\}$ and pose maps $\mathbf{P}:=\{P_1,...,P_N\}$ using off-the-shelf methods [54], [55]. Following the precedures described in [12], [27], we then generate clothing-agnostic images $\mathbf{A}:=\{A_1,...,A_N\}$ where the full body is masked. Similarly, the upper and lower masks can also be obtained through parse maps. These masks cover the clothes and the corresponding limb regions, which reduces the impact of the discrepancy between the original and target garments.

We construct the video-garment pairs by extracting the reference garments from a clear frame in the source video. We utilize resizing and translation techniques to ensure that the garment is positioned at the center of a white background. Since our architecture does not require a template image, it can be trained using a large dataset of in-the-wild videos. During the inference phase for upper garment editing, the first frame is selected for extracting the lower garment image $\mathbf{g}_l$.

B. Architecture of VidClothEditor

In this work, we address the virtual try-on task as an exemplar-based inpainting problem, aiming to fill the agnostic map with given garments, as illustrated in Fig. 2. For tops, we extend the inpainting region from the upper body to the full body. This expansion alleviates constraints and prevents segmentation errors between upper and lower garments. Our framework, VidClothEditor, is adapted from the conditional video diffusion model utilized in Animate Anyone [34], with several modifications specifically tailored to the virtual try-on task. 1) As the input of UNet, we concatenate two aligned conditions with the Gaussian noise $\epsilon\in\mathbb{R}^{N\times4\times h\times w}$, latent agnostic map $\mathcal{E}(\mathbf{A})\in\mathbb{R}^{N\times4\times h\times w}$ and latent pose map $\mathcal{E}'(\mathbf{P})\in\mathbb{R}^{N\times4\times h\times w}$. $\mathcal{E}$ and $\mathcal{E}'$ represent the VAE encoder and the PoseEncoder, respectively; both encoders convert their respective images into the latent space. h and w denote the height and width in the latent space, here $h=H/8$ and $w=W/8$. The input channel of UNet is expanded from 4 to 12. 2) We introduce GarmentNet $\mathcal{G}$, a specialized network that processes latent clothing $\mathcal{E}(\mathbf{g})\in\mathbb{R}^{N\times4\times h\times w}$ and extract garment features $\mathcal{G}_{\theta'}(\mathcal{E}(\mathbf{g}))$. The architecture of GarmentNet is the same as the denoising UNet. To preserve the fine details of multiple garments and alleviate feature interference, spatial attention is used along with a specific designed region guidance. 3) For exemplar conditioning, we input the garment image into CLIP to obtain global embedding ψ, and then the embedding ψ is injected into UNet and GarmentNet via cross-attention. Finally, we distill the UNet via the proposed garment-augmented video consistency learning for boosting the inference process as shown in Fig. 5.

Formally, along with the aforementioned conditions, our proposed VidClothEditor can be trained by minimizing the following loss function.

$$\mathcal{L}_{ldm}=\mathbb{E}_{z_t,\eta,\mathcal{E}(\mathbf{g}_u),\mathcal{E}(\mathbf{g}_l),\psi,\epsilon,t}\left[\|\epsilon-\epsilon_\theta(z_t,t,\eta,\psi,\zeta))\|_2^2\right], \tag{8}$$

where $\zeta=[\mathcal{G}_{\theta'}(\mathcal{E}(\mathbf{g}_u));\mathcal{G}_{\theta'}(\mathcal{E}(\mathbf{g}_l))]$ represents the extracted garment features, $\eta=[\mathcal{E}(\mathbf{A});\mathcal{E}(\mathbf{P})]$ denotes the aligned condition sequences.

C. Region-guidance Multiple Garment Fusion

Drawing inspiration from the spatial-attention mechanism [34], as shown in the left of Fig. 4, we introduce a multiple garment fusion process to incorporate the learned garment features into the denoising UNet with specially designed region guidance. Specially, we reshape the feature map z_i from the i-th layer of the denoising UNet into vector v_i. Similarly, for the upper garment and lower garment, we obtain

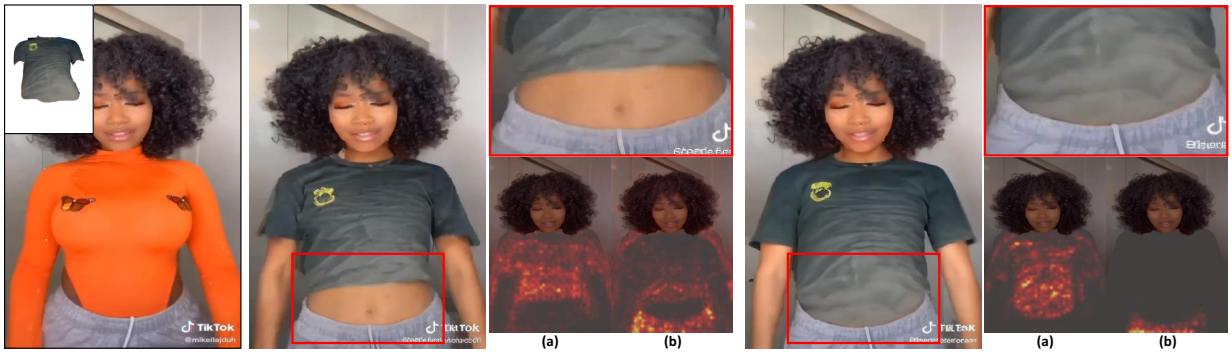

Fig. 3. Visualization of attention maps and generated results. (a) denotes the self-attention on upper region while (b) denotes the self-attention on lower region. Our proposed region-guidance multiple garment fusion strategy enhances the learning of effective attention maps.

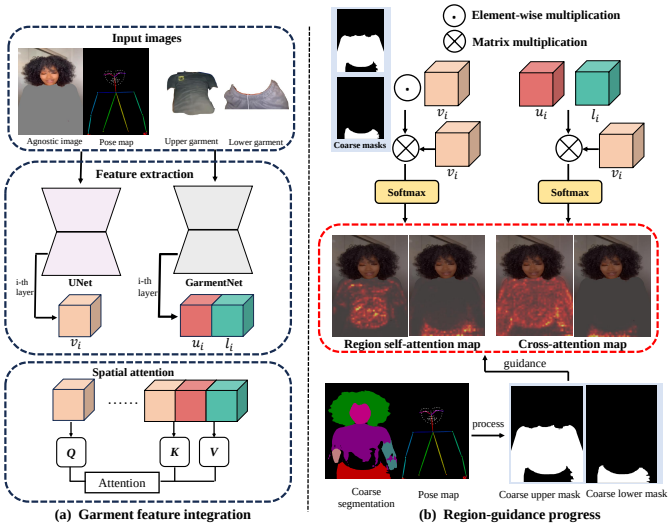

Fig. 4. Overview of the region-guidance multiple garment fusion. We use (a) spatial attention to integrate garment features into the denoising UNet. To prevent feature interference, we employ (b) region guidance to properly align the attention maps.

the corresponding features u_i and l_i, respectively. We then use these features to compute the output of the spatial-attention layer by the following formulas:

$$\begin{aligned} \text{Attention}(Q,K,V) &= \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \\ \text{where } v_i' &= [v_i \oplus u_i \oplus l_i], \\ Q = W_i^Q v_i,\ K &= W_i^K v_i',\ V = W_i^V v_i', \end{aligned} \tag{9}$$

Here $\oplus$ indicates matrix concatenation.

Fig. 3 visualizes the attention maps learned in the spatial-attention mechanism. While inpainting the entire body reduces the stringent precision requirements for segmentation and yields more natural outcomes, the attention focused on upper and lower garments may interfere with each other (see (a) and (b) under the setting of without region guidance). This interference potentially lead to overfitting specific clothing sets during training.

To address this issue, we introduce region guidance to enhance the learning of effective attention maps. As shown in the right of Fig. 4, this guidance for the spatial-attention mechanism comprises two components: regional self-attention guidance and cross-attention guidance. For regional self-attention, we introduce an upper region mask m_i^u and a lower region mask m_i^l. This design aims to specifically direct self-attention to concentrate within the designated upper or lower regions. In the case of cross-attention guidance, we ensure that the garment features u_i and l_i align precisely with the corresponding locations on the attention maps. Formally, we define the region-guidance loss as follows:

$$\begin{aligned} \mathcal{L}_{self} &= ((m_i^u \odot v_i) \otimes v_i) \odot m_i^u + ((m_i^l \odot v_i) \otimes v_i) \odot m_i^l \\ \mathcal{L}_{cross} &= (u_i \otimes v_i) \odot m_i^u + (l_i \otimes v_i) \odot m_i^l \\ \mathcal{L}_{rg} &= \mathcal{L}_{self} + \lambda_1 \mathcal{L}_{cross} \end{aligned} \tag{10}$$

where $\otimes$ means matrix multiplication, $\odot$ indicates element-wise multiplication and λ is a weight hyper parameter. Unlike directly using segmentation results to constrain the inpainting areas, our guidance approach is "soft". It facilitates the generation of natural virtual video try-on results instead of imposing a sequence of strict conditions. Finally, we train our network by adding $\mathcal{L}_{rg}$ to Eq. 8:

$$\mathcal{L} = \mathcal{L}_{ldm} - \lambda_2 \mathcal{L}_{rg} \tag{11}$$

D. Garment-augmented Video Consistency learning

The video diffusion model is notably time-consuming due to its iterative sampling process. To address this, we propose a garment-augmented consistency learning framework for few-step sampling, which is illustated by Fig. 5. Considering the problem statement of the video try-on task, given the trained GarmentNet $\mathcal{G}$ and the teacher diffusion model ϕ, we expect to distill ϕ into the student model $\boldsymbol{\theta}$. This process requires that the function $\boldsymbol{f_\theta}$ satisfy the consistency property and the model $\boldsymbol{\theta}$ can also effective handles the features extracted by $\mathcal{G}$.

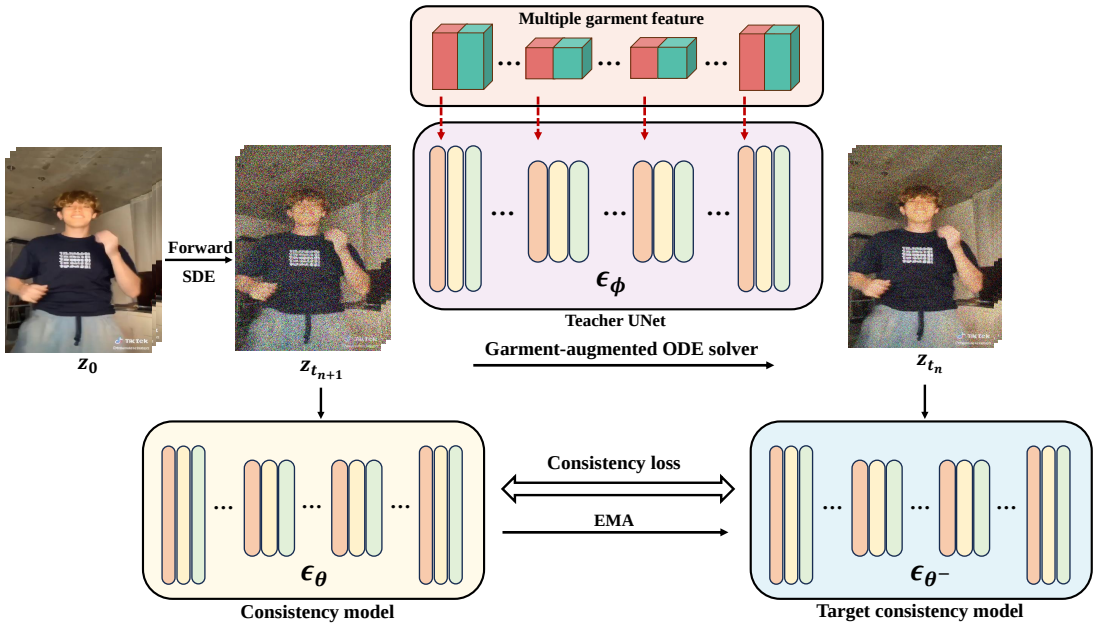

Fig. 5. Overview of the video consistency learning. Given a source video z_0 in the latent space, we perform a forward SDE to add noise on it, resulting in $z_{t_{n+1}}$. z_{t_n} is the one-step estimation from $z_{t_{n+1}}$, which is obtained by the garment-augmented PF-ODE solver. $z_{t_{n+1}}$ and z_{t_n} should be enforced to satisfy the self-consistency property.

TABLE I
QUANTITATIVE COMPARISON WITH BASELINES ON BOTH VITON-HD AND TIKTOK DATASETS.

Dataset	VITON-HD					TikTok			
Method	SSIM↑	LPIPS↓	FID↓	KID↓	User↑	SSIM↑	LPIPS↓	FVD↓	User↑
LADI-VTON [25]	0.873	0.0941	12.190	0.563	6.24%	0.868	0.0956	5.603	10.26%
DCI-VTON [26]	0.892	0.0719	12.105	0.544	8.12%	0.873	0.0836	6.336	6.46%
StableVITON [27]	0.890	0.0761	10.013	0.255	22.86%	0.846	0.1462	7.008	1.02%
Animate Anyone* [34]	**0.893**	**0.0705**	9.965	0.179	28.22%	**0.885**	**0.0678**	5.294	31.28%
ViViD [30]	-	-	-	-	-	0.843	0.1088	5.408	14.46%
VidClothEditor	0.879	0.0823	**9.186**	**0.113**	**34.56%**	0.873	0.0722	**5.002**	**38.98%**

We could reparametrize the consistency function $\boldsymbol{f_\theta}$ in Eq. 5 as follows:

$$\begin{aligned}&\boldsymbol{f_\theta}(z,\eta,\psi,\zeta,t)\\&=c_{\text{skip}}(t)z+c_{\text{out}}(t)\left(\frac{z-\sigma_t\boldsymbol{\epsilon}_\theta(z,t,\eta,\psi,\zeta)}{\alpha_t}\right)\end{aligned}\tag{12}$$

where α_t and σ_t are the function to specify the noise schedule and $\boldsymbol{\epsilon}_\theta(z,t,\eta,\psi,\zeta)$ is the noise prediction function via our proposed VidClothEditor.

Classifier-free guidance (CFG) [56] is crucial for controllability enhancement of the given condition. To ensure the detailed preservation of garment features during the distillation process, we intend to adopt the principles of CFG to accentuate these features. Employing a guidance scale $w_{\mathbf{g}} \geq 1$, we can modify the noise prediction function as follows:

$$\hat{\epsilon}_\theta(z,t,\eta,\psi,\zeta)=(1+w_{\mathbf{g}})(\boldsymbol{\epsilon_\theta}(z,t,\eta,\psi,\zeta))-w_{\mathbf{g}}\boldsymbol{\epsilon_\theta}(z,t,\eta,\varnothing)\tag{13}$$

where $\varnothing$ denotes we drop the garment features, specially, we use a zero vector to replace the embedding ψ extracted by CLIP and use self-attention to replace the spatial-attention. Based on this, we apply the garment-augmented ODE solver to run a one discretization step (denotes the one-step ODE solver applied to PF-ODE):

$$\begin{aligned}\Phi_{w_{\mathbf{g}}}=&(1+w_{\mathbf{g}})\Phi(\mathbf{z}_{t_n},\eta,\psi,\zeta,t_n,t_{n+1};\phi)\\&-w_{\mathbf{g}}\Phi(\mathbf{z}_{t_n},\eta,\varnothing,t_n,t_{n+1};\phi)\end{aligned}\tag{14}$$

Eventually, the consistence loss can be modified as follows:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\theta}^-;\Phi)=&\mathbb{E}_{\mathbf{z},\eta,\psi,\zeta,t}\Big[d\Big(\boldsymbol{f_\theta}(z_{t_{n+1}},\eta,\psi,\zeta,t_{n+1}),\\&\boldsymbol{f_{\theta^-}}(\hat{z}^\phi_{t_n},\eta,\psi,\zeta,t_n)\Big)\Big],\\\hat{z}^\phi_{t_n}=&z_{t_{n+1}}+(t_n-t_{n+1})\Phi_{w_{\mathbf{g}}}(\mathbf{z}_{t_n},\eta,\psi,\zeta,t_n,t_{n+1};\phi).\end{aligned}\tag{15}$$

E. Training Strategy

The training process is divided into three stages. The first stage is image-level training, where training is conducted on individual video frames. At this stage, we temporarily remove the temporal layers from the network. Both GarmentNet $\mathcal{G}$ and PoseEncoder $\mathcal{E}'$ are trained concurrently. The purpose of this stage is to develop effective feature extraction networks

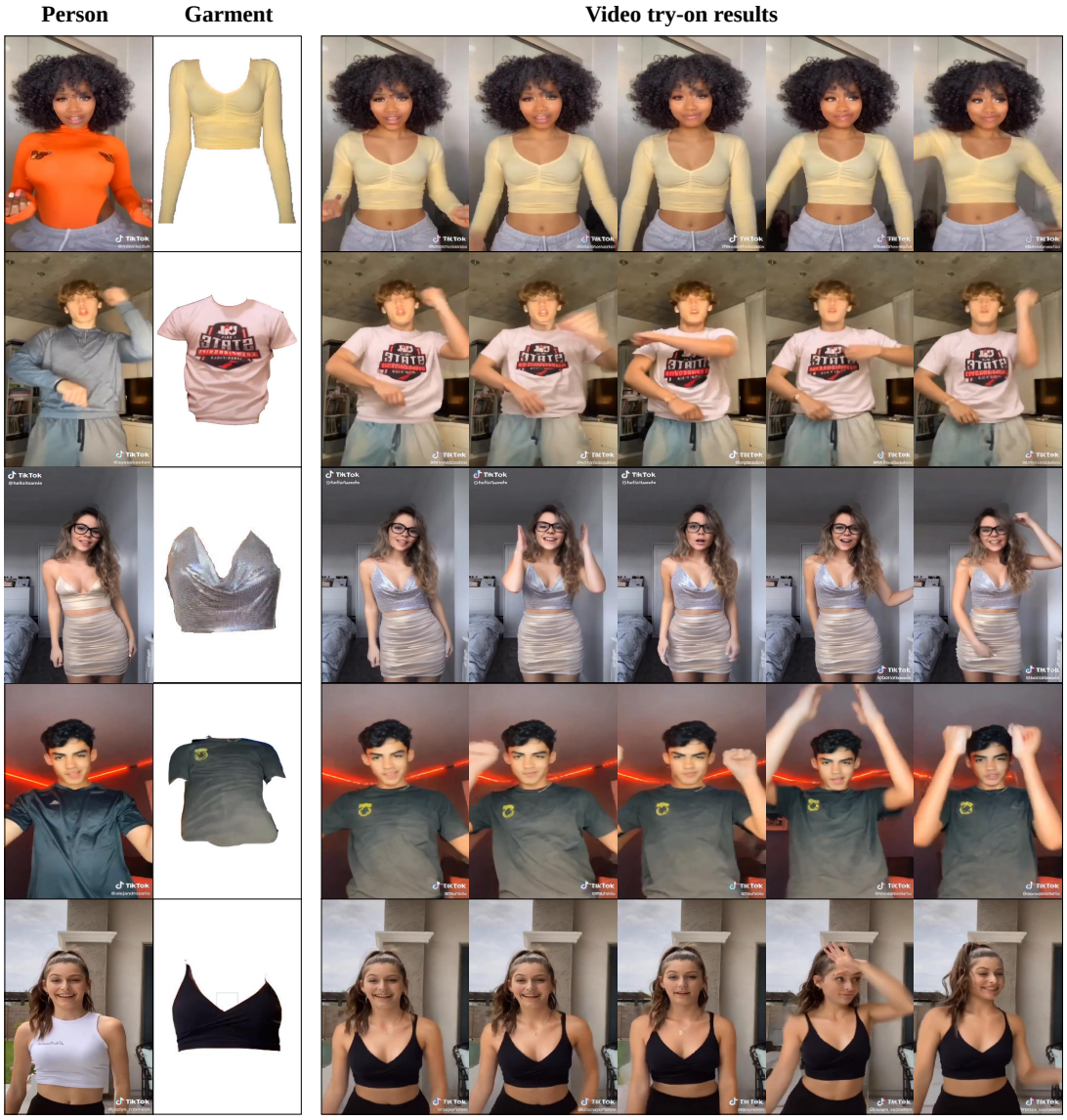

Fig. 6. Examples of our virtual try-on results on real-life TikTok videos. Our method produces coherent and natural video try-on results.

for garments and poses, and to enable the model to generate high-quality virtual try-on images under specified conditions. The second stage is video-level training, where the data sampling is changed to 18-frame video clips. We reintroduce the temporal layers into the UNet and train only this module, while keeping all other parameters fixed. The third stage involves video consistency distillation. Similar to the second stage, we treat the PoseEncoder $\mathcal{E}'$ and GarmentNet $\mathcal{G}$ as fixed feature extractors and focus solely on distilling the UNet.

V. Experiments

In this section, we first introduce the datasets and implementation details. We then proceed to compare VidClothEditor with state-of-the-art virtual try-on methods through qualitative and quantitative evaluations. The comparison comprises two parts, the video-level comparison on the in-the-wild TikTok dataset [57] and the image-level comparison on the well-collected VITON-HD and Dresscode datasets. The first part is to demonstrate the effectiveness of our framework on the video try-on task while the second part is to demonstrate that the full-body inpainting strategy will not degrade the image editing result with the proposed region guidance module on high-quality image data. Thirdly, we provide an ablation study on the proposed module. Finally, we analyze the limitation on videos with dynamic blurs.

A. Datasets

Our experiments are conducted on the TikTok [57], VITON-HD [7] and DressCode [58] datasets. TikTok comprises real-life single-person dancing videos with intricate limb occlusions and postures. From a total of 350 videos, we utilize 213 videos for training and 54 for testing, discarding 83 due to unclear views. Additionally, we perform an image-level comparison on the high-resolution virtual try-on benchmark VITON-HD [7] and Dresscode [58] to prove the robustness and versatility. All evaluations and visualizations are conducted on the testing set.

Fig. 7. Qualitative comparison with state-of-the-art try-on methods on the TikTok dataset. Our method effectively captures the details of clothing during dance movements and exhibits outstanding full-body results.

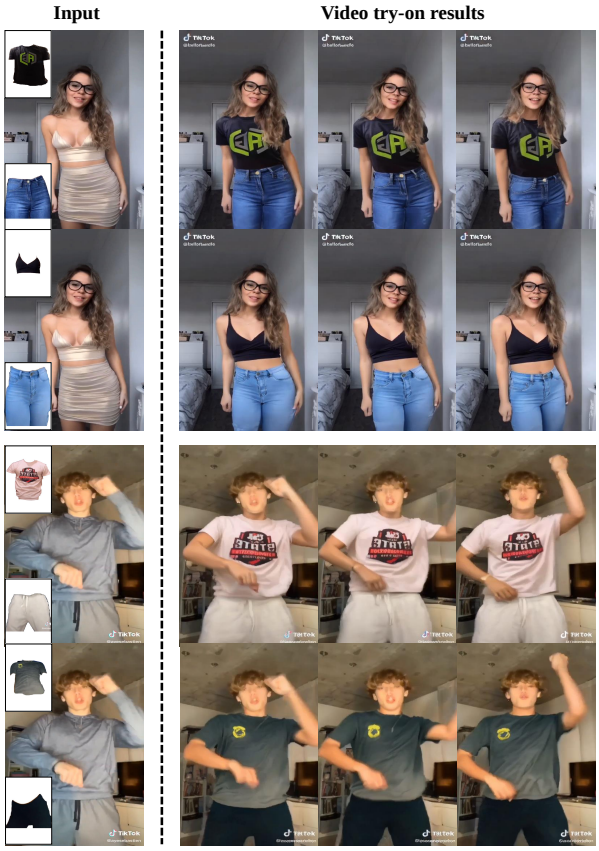

Fig. 8. Multi-garment video virtual try-on results on the TikTok dataset.

B. Implementation Details

1) Architecture: We initialize the GarmentNet and the denoising UNet by inheriting the pretrained weights of Stable Diffusion v1.5 [44]. The pose guider contains 4 convolution layers followed by a linear projection to align the dimension in the latent space. Each convolution layer contains a two convolution operations (one 1×1 stride and the other 2×2 stride). All the kernel size is 3. The CLIP global embedding $\psi \in \mathbb{R}^{50\times768}$ is the concatenation of a global token and 49 patch tokens of the CLIP output.

2) Training: Experiments are conducted on 4 NVIDIA A100 GPUs. Images or video frames are resized to a resolution of 512×384. We apply scale and translation augmentations to the reference garment to enhance generalization. The hyper-parameter weight λ_1 and λ_2 and for region guidance are set 2 and 5e-5 respectively. Our training includes 3 stages. In the first stage, individual video frames are sampled and the training step is 30,000 at a batch size of 24. In the second stage, we train the denoising UNet for 10000 steps using 18-frame video sequences. We use grad accumulation to increase the batch size from 4 to 24 in this stage. The final distillation stage is conducted with a batch size of 4 over 2,000 training steps. All the stages use AdamW [59] as optimizer and the learning rate is set 5e-5.

3) Testing: At inference time, we use LCM [33] scheduler as the sample method and the total sample steps is 10. The lower garment image for the test video is extracted from the first frame in the reference clips. We adopt the temporal aggregation method [60], integrating results across batches to produce long video sequences.

C. Results and Comparison with State-of-the-Art Methods

1) Baselines: We perform video-level comparison with LaDI-VTON [25], DCI-VTON [26], StableVITON [27], ViViD [30] and Animate Anyone [34] on TikTok [57] datasets. Moreover, to demonstrate the robustness and versatility of our method, we process image-level comparison with HR-VTON [12], LaDI-VTON [25], DCI-VTON [26], StableVI-TON [27] and Animate Anyone [34] on VITON-HD [7]

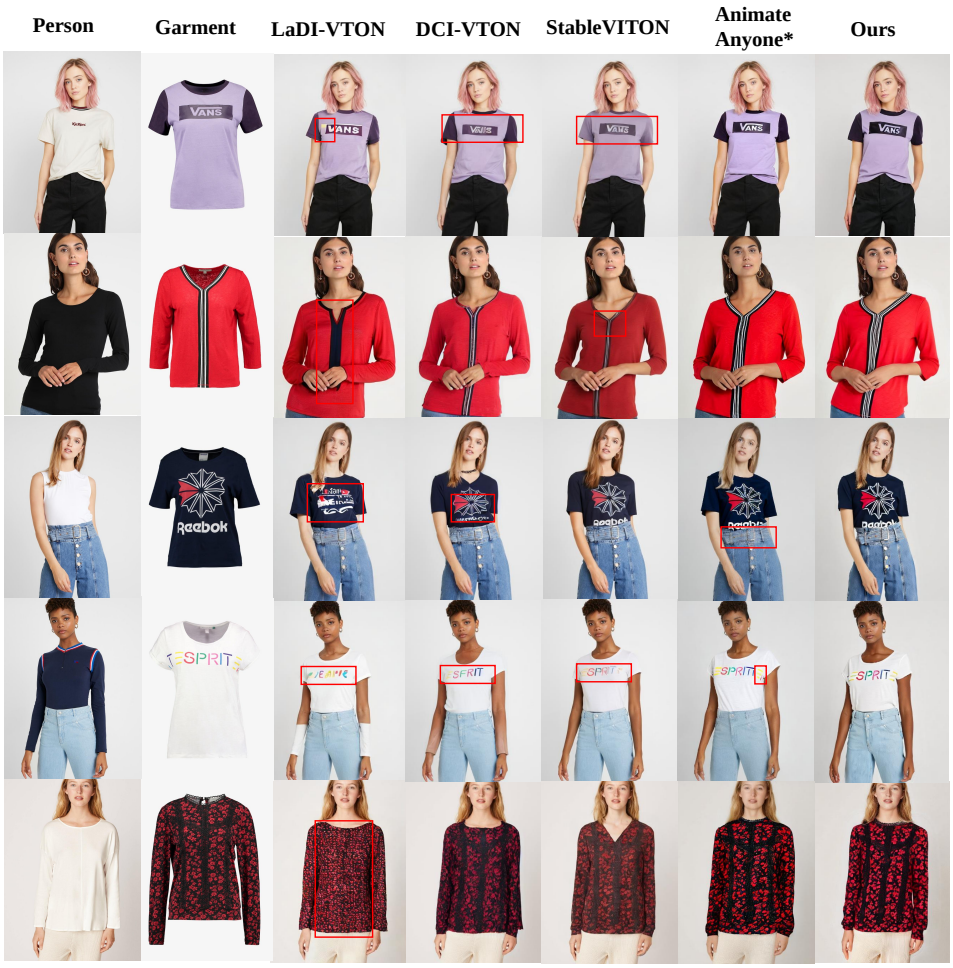

Fig. 9. Qualitative comparison on the VITON-HD dataset. The red boxes highlight the artifacts.

TABLE II
QUANTITATIVE COMPARISON WITH BASELINES ON THE DRESSCODE DATASET.

Dataset	DressCode-upper					DressCode-lower				
Method	SSIM↑	LPIPS↓	FID↓	KID↓	User↑	SSIM↑	LPIPS↓	FID↓	KID↓	User↑
HR-VTON [12]	0.907	0.0892	18.42	1.036	19.92%	**0.910**	**0.0620**	17.41	0.626	26.98%
LADI-VTON [25]	0.906	0.0983	16.71	0.610	17.24%	0.902	0.0834	16.58	0.608	14.68%
Animate Anyone* [34]	0.909	0.0634	16.03	0.524	24.48%	0.892	0.0976	16.12	0.568	23.82%
VidClothEditor	**0.913**	**0.0602**	**13.62**	**0.258**	**38.36%**	0.908	0.0656	**13.88**	**0.304**	**34.52%**

and DressCode Dataset [58]. We adapt the Animate Anyone method for the try-on task, denoted as "Animate Anyone*" in our experimental tables and figures to indicate this modification. For other baseline methods, we use publicly available checkpoints or the released code.

2) Metrics: Following previous studies [12], [27], we make quantitative evaluation on both paired and unpaired setting. In the paired setting, we employ SSIM [61] and LPIPS [62] as evaluation metrices. In the unpaired setting, where ground truth is unavailable, we evaluate realism using the Fréchet Inception Distance (FID) [63] and Kernel Inception Distance (KID) [64] scores for image comparison, and Fréchet Video Distance (FVD) [65] scores for video comparison. Additionally, we incorporate a human perception study for subjective evaluation, including 100 image samples from the VTION-HD dataset, 100 image samples from the DressCode dataset, and 20 video samples from the TikTok dataset in the survey.

3) Results on the TikTok Dataset: Fig. 7 illustrates the comparison results. Since LaDI-VTON, DCI-VTON and StableVITON are image-based methods, they lack coherence in the video-based clothing change task. Although ViViD enhances the temporal consisitency by introducing the temporal attention, however, it still fails to reproduce clothing patterns well. Animate Anyone* successfully maintain the patterns, yet the rigid parsing preprocess adversely affects the connection between the upper and lower garments, resulting in unnatural outcomes. Furthermore, as demonstrated in Fig. 1, the inaccuracies in the inpainting region sequences result in significant differences between frames, thereby compromising temporal consistency. In contrast, our approach utilizes full-body inpainting and region-guidance multiple garment fusion to achieve more realistic video try-on effects. Quantitative results are shown in Table I. VidClothEditor outperforms than others on KID and FID metrics. Since VidClothEditor reconstructs the unedited parts of a person, it performs slightly

Fig. 10. Qualitative comparison on the DressCode dataset. The first three rows and the last three rows represent the VTON of upper and lower garment respectively. The red boxes highlight the artifacts.

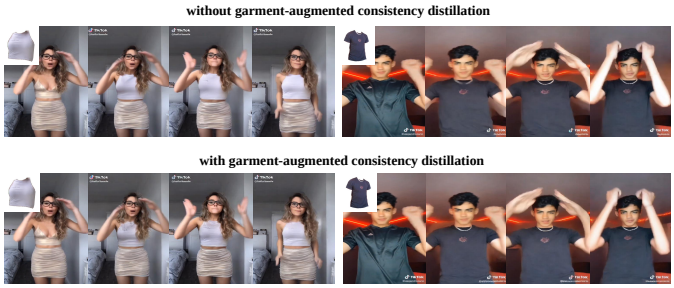

Fig. 11. Effects of the garment-augmented video consistency distillation. The appearance features of garments are still well preserved after distillation.

worse than Animate Anyone* on the pixel-based metrics SSIM and LPIPS.

Fig. 6 illustrates more results generated by our method, demonstrating its ability to effectively adapt to various human movements and garments. This results in high-detail preservation and temporal consistency in the generated try-on videos. A byproduct of our method is full-body try-ons. Fig. 8 displays the results of these transformations, further demonstrating the generalization ability of our approach.

4) Results on the VITON-HD and DressCode Dataset: Fig. 9 and Fig. 10 illustrate the qualitative comparison on VITON-HD and DressCode dataset respectively. Warping-based methods, HR-VTON, LaDI-VTON and DCI-VTON, tend to fail in the presence of occlusions or complex textures while StableVITON demonstrates some color deviations. In the third row of Fig. 9, a part of the clothing text is obscured by the jean. All baseline methods fail because they focus only on the upper region that needs editing, which reduces the degree of freedom in generation. Similar cases also appear in the virtual try on of lower garments. For example, in the fourth row of Fig. 10, the editing of the pants interferes with the reconstruction of the upper clothes. On the contrary, VidClothEditor can produce natural results with the proposed region-guidance multi-garment fusion module.

Quantitative results are shown in Tab. I and II. Since our method inpaints the whole body, the two metrics, SSIM and LPIPS, are adversely affected to some extent. Nevertheless, our method performs comparably to the baseline methods on both the VITON-HD and DressCode datasets. Correspondingly, due to its higher flexibility, our method outperforms other baseline algorithms in terms of FID and KID, which are metrics for evaluating the authenticity of images. This demonstrates that the full-body inpainting strategy does not degrade the image generation quality, although it is designed for video VTON.

D. Ablation Study

We conduct an ablation study for VidClothEditor to investigate the effects of the proposed region-guidance multiple garment fusion (Section IV-C) and garment-augmented video consistency distillation (Section IV-D).

TABLE III
QUANTITIES ABLATIONS FOR THE CORE COMPONENTS. THE TIME COST IS CALCULATED BASED ON THE GENERATION OF 18-FRAME VIDEOS.

Method	SSIM↑	LPIPS↓	FVD↓	Time cost(s)↓
AnimateAnyone*	0.885	0.0678	5.294	32.24
+ Full-body inpaint	0.864	0.0749	5.176	32.24
+ Region-guidance fusion	0.879	0.0658	4.802	32.86
+ Consistency distillation	0.873	0.0722	5.002	7.42

1) Analysis of Region-guidance Multiple Garment Fusion: In Table. III, we provide quantitative metrics related to the ablation experiments. Building upon AnimateAnyone*, the utilization of a full-body inpainting strategy has been proven to enhance the FVD metric, signifying an uplift in video quality. However, this approach inversely impacts the pixel-based metrics, such as SSIM and LPIPS, due to the amplification of the edited areas. Leveraging the region-guidance multiple garment fusion module, we've not only improved the FVD score further but also mitigated the adverse effects on SSIM and LPIPS metrics. A representative visualization sample is depicted in Fig. 3.

2) Analysis of Garment-augment Video Consistency Distillation: As shown in Fig. 5, distillation results in a minor quality degradation while preserving the appearance of the clothing. In Table III, the time cost is calculated based on the generation 18-frame videos. We employ the DDIM [66] sampler with 50 sample steps when distillation is not used, and switch to the LCM [33] sampler with 10 sample steps when distillation is implemented. Quantitative comparisons show that the SSIM, LPIPS, and FVD metrics decrease to some extent after distillation. However, the significant reduction in generation time greatly enhances the potential for practical applications. Additionally, we conduct an analysis of the sampling steps, as detailed in Table IV. We set the total sample steps as 10 to achieve a balance between quality and speed.

TABLE IV
ABLATION STUDY FOR SAMPLE STEPS ON THE CONSISTENCY MODEL.

Sample steps	SSIM↑	LPIPS↓	FVD↓	Time cost(s)↓
6	0.864	0.0789	5.476	5.82
8	0.870	0.0758	5.032	6.58
10	0.873	0.0722	5.002	7.44
12	0.874	0.0702	4.098	8.40

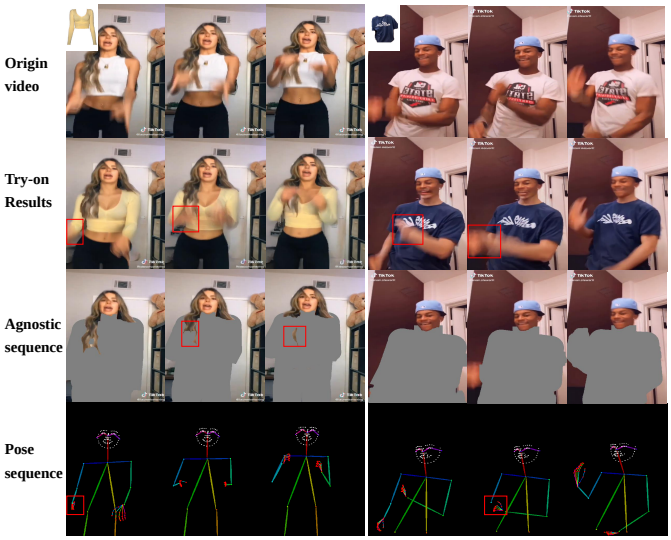

Fig. 12. Limitations of the limbs and long hair reconstruction under dynamic blur. The red boxes highlight the errors. Dynamic blur adversely affects training and results in erroneous segmentation and pose sequences, inhibiting the model's ability to effectively reconstruct limbs and long hair during inference.

E. Limitations

In real-world videos, high-speed motion of human subjects often results in dynamic blur. Under the influence of such dynamic blur, although our algorithm is robust for editing clothing, there are discernible shortcomings in the reconstruction of human limbs. This is predominately due to the reliance of the limb reconstruction on the conditions provided. As shown in Fig. 12, the limbs overlap with the upper body and we need to reconstruct the limbs. Due to dynamic blur, pose detection algorithms tend to miss detections especially around the hands and elbows. Under these conditions, the model fails to perfectly restore the limbs of the human due to the lack of reference. Additionally, the accurate segmentation of long female hair poses a challenge, further increasing the difficulty of human restoration. Furthermore, using these blurred limbs as training targets can adversely affect the training of the model. Gathering higher quality video data or providing close-up shots of the hands as an additional condition could help to mitigate this issue.

VI. CONCLUSIONS

To facilitate video virtual try-on in real-world scenarios, we propose VidClothEditor, a robust and efficient framework. VidClothEditor utilizes a full-body inpainting approach, relaxing strict inpainting boundaries and mitigating the impact of segmentation errors on the human body. It achieves multiple garments alignment and utilizes coarse parsing results to guide attention training, thereby enhancing the naturalness of the try-on results. Additionally, we integrate a video consistency model to expedite the try-on process. Extensive evaluations clearly demonstrate VidClothEditor's significant superiority over state-of-the-art methods.

REFERENCES

[1] Z. Xie, Z. Huang, X. Dong, F. Zhao, H. Dong, X. Zhang, F. Zhu, and X. Liang, "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*, 2023, pp. 23 550–23 559.

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

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

[4] R. Yu, X. Wang, and X. Xie, "Vtnfp: An image-based virtual try-on network with body and clothing feature preservation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 511–10 520.

[5] T. Issenhuth, J. Mary, and C. Calauzènes, "Do not mask what you do not need to mask: a parser-free virtual try-on," in *European Conference on Computer Vision*. Springer, 2020, pp. 619–635.

[6] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7850–7859.

[7] S. Choi, S. Park, M. Lee, and J. Choo, "Viton-hd: High-resolution virtual try-on via misalignment-aware normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 131–14 140.

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

[9] X. Dong, F. Zhao, Z. Xie, X. Zhang, D. K. Du, M. Zheng, X. Long, X. Liang, and J. Yang, "Dressing in the wild by watching dance videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3480–3489.

[10] S. He, Y.-Z. Song, and T. Xiang, "Style-based global appearance flow for virtual try-on," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3470–3479.

[11] H. Yang, X. Yu, and Z. Liu, "Full-range virtual try-on with recurrent tri-level transform," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 3460–3469.

[12] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo, "High-resolution virtual try-on with misalignment and occlusion-handled conditions," in *Proceedings of the European conference on computer vision (ECCV)*, 2022.

[13] S. Bai, H. Zhou, Z. Li, C. Zhou, and H. Yang, "Single stage virtual try-on via deformable attention flows," in *European Conference on Computer Vision*. Springer, 2022, pp. 409–425.

[14] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed gan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5084–5093.

[15] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7982–7990.

[16] Y. Ren, X. Fan, G. Li, S. Liu, and T. H. Li, "Neural texture extraction and distribution for controllable person image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 535–13 544.

[17] N. Fang, L. Qiu, S. Zhang, Z. Wang, and K. Hu, "Pg-vton: A novel image-based virtual try-on method via progressive inference paradigm," *IEEE Transactions on Multimedia*, 2024.

[18] J. Xu, Y. Pu, R. Nie, D. Xu, Z. Zhao, and W. Qian, "Virtual try-on network with attribute transformation and local rendering," *IEEE Transactions on Multimedia*, vol. 23, pp. 2222–2234, 2021.

[19] S. Zhang, X. Han, W. Zhang, X. Lan, H. Yao, and Q. Huang, "Limb-aware virtual try-on network with progressive clothing warping," *IEEE Transactions on Multimedia*, 2023.

[20] Y. Liu, W. Chen, L. Liu, and M. S. Lew, "Swapgan: A multistage generative approach for person-to-person fashion style transfer," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2209–2222, 2019.

[21] B. Hu, P. Liu, Z. Zheng, and M. Ren, "Spg-vton: Semantic prediction guidance for multi-pose virtual try-on," *IEEE Transactions on Multimedia*, vol. 24, pp. 1233–1246, 2022.

[22] C. Du, F. Yu, M. Jiang, A. Hua, X. Wei, T. Peng, and X. Hu, "Vton-scfa: A virtual try-on network based on the semantic constraints and flow alignment," *IEEE Transactions on Multimedia*, vol. 25, pp. 777–791, 2022.

[23] Z. Xing, Y. Wu, S. Liu, S. Di, and H. Ma, "Virtual try-on with garment self-occlusion conditions," *IEEE Transactions on Multimedia*, 2022.

[24] Z. Yang, J. Chen, Y. Shi, H. Li, T. Chen, and L. Lin, "Occlumix: Towards de-occlusion virtual try-on by semantically-guided mixup," *IEEE Transactions on Multimedia*, 2023.

[25] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, and R. Cucchiara, "Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on," *arXiv preprint arXiv:2305.13501*, 2023.

[26] J. Gou, S. Sun, J. Zhang, J. Si, C. Qian, and L. Zhang, "Taming the power of diffusion models for high-quality virtual try-on with appearance flow," *arXiv preprint arXiv:2308.06101*, 2023.

[27] J. Kim, G. Gu, M. Park, S. Park, and J. Choo, "Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on," *arXiv preprint arXiv:2312.01725*, 2023.

[28] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, and I. Kemelmacher-Shlizerman, "Tryondiffusion: A tale of two unets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4606–4615.

[29] Y. Xu, T. Gu, W. Chen, and C. Chen, "Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on," *arXiv preprint arXiv:2403.01779*, 2024.

[30] Z. Fang, W. Zhai, A. Su, H. Song, K. Zhu, M. Wang, Y. Chen, Z. Liu, Y. Cao, and Z.-J. Zha, "Vivid: Video virtual try-on using diffusion models," *arXiv preprint arXiv:2405.11794*, 2024.

[31] Z. He, P. Chen, G. Wang, G. Li, P. H. Torr, and L. Lin, "Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models," in *European Conference on Computer Vision*. Springer, 2025, pp. 123–139.

[32] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *arXiv preprint arXiv:2303.01469*, 2023.

[33] S. Luo, Y. Tan, L. Huang, J. Li, and H. Zhao, "Latent consistency models: Synthesizing high-resolution images with few-step inference," *arXiv preprint arXiv:2310.04378*, 2023.

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

[35] X. Han, X. Hu, W. Huang, and M. R. Scott, "Clothflow: A flow-based model for clothed person generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 471–10 480.

[36] Z. Huang, H. Li, Z. Xie, M. Kampffmeyer, X. Liang *et al.*, "Towards hard-pose virtual try-on via 3d-aware global correspondence learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 736–32 748, 2022.

[37] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[38] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[39] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[40] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "Fw-gan: Flow-navigated warping gan for video virtual try-on," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1161–1170.

[41] X. Zhong, Z. Wu, T. Tan, G. Lin, and Q. Wu, "Mv-ton: Memory-based video virtual try-on network," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 908–916.

[42] J. Jiang, T. Wang, H. Yan, and J. Liu, "Clothformer: Taming video virtual try-on in all module," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 799–10 808.

[43] Z. Xu, M. Chen, Z. Wang, L. Xing, Z. Zhai, N. Sang, J. Lan, S. Xiao, and C. Gao, "Tunnel try-on: Excavating spatial-temporal tunnels for high-quality virtual try-on in videos," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 3199–3208.

[44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[45] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

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

[47] X. Chen, L. Huang, Y. Liu, Y. Shen, D. Zhao, and H. Zhao, "Anydoor: Zero-shot object-level image customization," *arXiv preprint arXiv:2307.09481*, 2023.

[48] Y. Xu, X. Xu, H. Gao, and F. Xiao, "Sgdm: An adaptive style-guided diffusion model for personalized text to image generation," *IEEE Transactions on Multimedia*, 2024.

[49] S. Cao, W. Chai, S. Hao, Y. Zhang, H. Chen, and G. Wang, "Diffashion: Reference-based fashion design with structure-aware transfer by diffusion models," *IEEE Transactions on Multimedia*, 2023.

[50] Y. Guo, C. Yang, A. Rao, Y. Wang, Y. Qiao, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[51] Y. Jiang, T. Wu, S. Yang, C. Si, D. Lin, Y. Qiao, C. C. Loy, and Z. Liu, "Videobooth: Diffusion-based video generation with image prompts," *arXiv preprint arXiv:2312.00777*, 2023.

[52] Y. Wang, J. Bao, W. Weng, R. Feng, D. Yin, T. Yang, J. Zhang, Q. D. Z. Zhao, C. Wang, K. Qiu *et al.*, "Microcinema: A divide-and-conquer approach for text-to-video generation," *arXiv preprint arXiv:2311.18829*, 2023.

[53] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[54] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[55] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[56] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[57] Y. Jafarian and H. S. Park, "Self-supervised 3d representation learning of dressed humans from social media videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[58] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2231–2235.

[59] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[60] J. Tseng, R. Castellon, and K. Liu, "Edge: Editable dance generation from music," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 448–458.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

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

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

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

[65] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *arXiv preprint arXiv:1812.01717*, 2018.

[66] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

Zijian He received a B.E. degree from the School of Mathematical Sciences, South China Normal University, Guangzhou, China, in 2017 and a M.S. degree from the School of Information, Xiamen University, Xiamen, China, in 2020. He is currently pursuing a Ph.D. degree in computer science at Sun Yat-sen University. His research interests include multimedia and computer vision.

Peixin Chen received the B.E. degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2024, where he is currently pursuing a M.S degree at in computer science. His research interests include computer vision and machine learning.

Guolin Zheng received a B.E degree and MA.Eng degree from the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, in 2020 and 2022. His research interests include speech and computer vision.

Guangrun Wang (Member, IEEE) is currently an Associate Professor (a Ph.D. Supervisor) at Sun Yat-Sen University under the Track of Overseas Talent Recruitment. Previously, he was a Postdoctoral Researcher in the Department of Engineering Science at the University of Oxford. He received two B.E. degrees and one Ph.D. degree from SYSU in 2014 and 2020. He was a visiting scholar at the Chinese University of Hong Kong (CUHK). His research interest is machine learning. He is a Distinguished Senior Program Committee member for IJCAI and a top / highlighted / outstanding reviewer of NeurIPS, ICLR and ICCV for Six Times. He is the recipient of the 2018 Pattern Recognition Best Paper Award, two ESI Highly Cited Papers, Top Chinese Rising Stars in Artificial Intelligence, and Wu Wen-Jun Best Doctoral Dissertation.

Xiaonan Luo is currently a Professor with the School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China. His research interests include computer vision, image processing, computer graphics, and CAD.

Liang Lin (Fellow, IEEE) is a full professor at Sun Yat-sen University. From 2008 to 2010, he was a postdoctoral fellow at the University of California, Los Angeles. From 2016–2018, he led the SenseTime R&D teams to develop cutting-edge and deliverable solutions for computer vision, data analysis and mining, and intelligent robotic systems. He has authored and coauthored more than 100 papers in top-tier academic journals and conferences (e.g., 15 papers in TPAMI and IJCV and 60+ papers in CVPR, ICCV, NeurIPS, and IJCAI). He has served as an associate editor of IEEE Transactions on Multimedia, IEEE Transactions on Neural Networks and Learning Systems, and as an area/session chair for numerous conferences, such as CVPR, ICCV, AAAI, ICME, and ICMR. He was the recipient of the Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, the Best Paper Diamond Award at IEEE ICME 2017, the Best Paper Runner-Up Award at ACM NPAR 2010, Google Faculty Award in 2012, the Best Student Paper Award at IEEE ICME 2014, and the Hong Kong Scholars Award in 2014. He is a Fellow of IEEE, IAPR, AAIA and IET.

Guanbin Li (Member, IEEE) received the Ph.D. degree from The University of Hong Kong in 2016. He is currently a full Professor with the School of Computer Science and Engineering, Sun Yat-sen University. He has authored and coauthored more than 150 papers in top-tier academic journals and conferences. His current research interests include computer vision, image processing, and deep learning. He was a recipient of the ICCV 2019 Best Paper Nomination Award. He serves as an Area Chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, International Journal of Computer Vision, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CYBERNETICS, CVPR, ICCV, ECCV, and NeurIPS.