WildVidFit: Video Virtual Try-On in the Wild via Image-Based Controlled Diffusion Models

Zijian He[1], Peixin Chen[1], Guangrun Wang[1], Guanbin Li[1,2*], Philip H.S. Torr[3], and Liang Lin[1,2]

[1] Sun Yat-sen University
[2] Peng Cheng Laboratory
[3] University of Oxford
{hezj39,chenpx28}@mail2.sysu.edu.cn, wanggrun@gmail.com, liguanbin@mail.sysu.edu.cn, philip.torr@eng.ox.ac.uk, linliang@ieee.org

Fig. 1: Examples of our virtual try-on results on real-life TikTok videos.

Abstract. Video virtual try-on aims to generate realistic sequences that maintain garment identity and adapt to a person's pose and body shape in source videos. Traditional image-based methods, relying on warping and blending, struggle with complex human movements and occlusions, limiting their effectiveness in video try-on applications. Moreover, video-based models require extensive, high-quality data and substantial computational resources. To tackle these issues, we reconceptualize video try-on as a process of generating videos conditioned on garment descriptions and human motion. Our solution, WildVidFit, employs image-based controlled diffusion models for a streamlined, one-stage approach. This model, conditioned on specific garments and individuals, is trained on still images rather than videos. It leverages diffusion guidance from pre-trained models including a video masked autoencoder for segment smoothness improvement and a self-supervised model for feature alignment of adjacent frame in the latent space. This integration markedly boosts the model's ability to maintain temporal coherence, enabling

* Corresponding Author

more effective video try-on within an image-based framework. Our experiments on the VITON-HD and DressCode datasets, along with tests on the VVT and TikTok datasets, demonstrate WildVidFit's capability to generate fluid and coherent videos. The project page website is at `wildvidfit-project.github.io`.

Keywords: Video Virtual Try-on · In the wild · Image-Based Video synthesis

1 Introduction

In video virtual try-on, the objective is to generate seamless videos that preserve the appearance of a specific garment while accurately adapting to the pose and body shape of the individual in the source video. This domain has garnered significant attention due to its potential applications in e-commerce and the burgeoning short-form video sector.

Recent advancements in video virtual try-on have evolved from initial two-stage image-based approaches, involving flow-based warping and blending, to incorporating an additional temporal module for ensuring frame consistency. Notably, FW-GAN [8] introduced an optical flow-guided fusion module, utilizing past frame warping results for current frame prediction. MV-TON [47] employed garment-to-person flow estimation for each frame, coupled with a memory module for refining frames using space-time information. ClothFormer [22] achieved realistic, spatio-temporally consistent results with its anti-occlusion warping, appearance-flow tracking, and dual-stream transformer. However, these methods face challenges in "in-the-wild" video applications due to two primary obstacles. Firstly, the collection of robust video data is costly, and developing a temporal module requires extensive, high-quality videos and computational resources. These methods, trained on specific datasets [8, 22], have limited generalization ability. Secondly, limb occlusions and significant garment deformation, more prevalent in videos than still images, lead to misalignment issues in current methodologies.

Addressing the video virtual try-on challenge hinges on generating images that adhere to given conditions, such as garment descriptions and human motion sequences. Particularly for videos captured in uncontrolled environments, a method must robustly handle intricate motions and limb occlusions. An image-based approach, leveraging extensive image foundation model knowledge and abundant image data, is particularly beneficial. This leads to decomposing the video try-on task into two subtasks: developing a fine-grained image try-on model for complex movements and occlusions, and extending it to video while maintaining frame coherence.

In response, we introduce WildVidFit, a novel, video training-free virtual try-on framework. WildVidFit utilizes image-based controlled diffusion models for realistic video try-on results. It bypasses explicit warping limitations in occlusion handling with a detail-focused, one-stage image try-on network, synthesizing

outputs based on unified representations of garments and individuals. It incorporates implicit warping inspired by TryOnDiffusion [48] for naturalistic outcomes and features a diffusion guidance module. This module enhances the temporal consistency of videos by improving segment smoothness with a pre-trained video masked autoencoder and aligning features of adjacent frames in the latent space through a self-supervised model. Crucially, WildVidFit streamlines the process by using editing and content consistency cues from pre-trained models, eliminating the need for additional fine-tuning or new temporal modules. Our contributions can be summarized as follows:

- We present WildVidFit, a video training-free virtual try-on framework capable of handling complex limb occlusions and actions in wild videos with a straightforward process.
- We introduce a diffusion guidance module to enhance temporal consistency, employing pre-trained video models and image self-supervised models to establish frame feature correspondence in the latent space.
- Our experiments on the TikTok dataset demonstrate WildVidFit's effectiveness in dynamic, real-world scenarios, highlighting its practicality and versatility.

2 Related Work

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. Most of these methods [1, 6, 9, 11, 14, 15, 20, 25, 28, 32, 38, 41–44] decompose the try-on task into two generation stages, i.e., warping and blending. The pioneering work, VITON [14], introduced a coarse-to-fine pipeline that was guided by the thin-plate-spline (TPS) warping of the target garment. ClothFlow [13] advanced the warping process by directly estimating the flow field using a neural network instead of the TPS. VITON-HD [6] released a high-resolution virtual try-on dataset and increased the resolution of generated images from 256×192 to 1024×768 with an alignment-aware generator. GP-VITON [40] developed an innovative Local-Flow-Global-Parsing warping module to preserve the semantic information of different parts of the garment. Moreover, [19] integrated geometric priors of 3D human bodies, enabling a more nuanced handling of pose and viewpoint variations. Although these methods have made significant progress, explicit warping still struggles to cope with complex poses and occlusions due to pixel misalignment.

Recently, diffusion models [17,34,36] have risen to prominence as the leading family of generative models. As a result, there is a growing interest in leveraging diffusion models as an alternative to GANs to achieve more realistic outcomes. LaDI-VTON [29] incorporated the latent diffusion model [33] into the blending stage of virtual try-on and introduced a textual inversion module to enhance the texture on garments. DCI-VTON [12] proposed an exemplar-based inpainting approach that leveraged a warping module to guide the diffusion model's

generation. Both methods follow the previous two-stage approach. TryOnDiffusion [48] presented a diffusion-based architecture, enabling the preservation of garment details and the ability to warp the garment to accommodate significant pose and body changes within a single network. However, the network design of two parallel UNets followed by a super-resolution module will bring huge computation cost when extending TryOnDiffusion to video synthesis.

Video Virtual Try On. Researches extended the two-stage approach used in image virtual try-on to video applications by integrating a specially designed temporal module. FW-GAN [8] successfully applied a video generation framework to the task of virtual try-on by incorporating relevant factors like warped garments and human postures. MVTON [47] introduced a try-on module for garment warping using pose alignment and regional pixel displace, and a memory refinement module that embedded prior generated frames into a latent space, serving as external memory for subsequent frame generation. ClothFormer [22], on the other hand, refined flow predictions using inter-frame information and employed a Dual-Stream Transformer to produce the video try-on result from warping results of multiple frames. Despite great advancements in walking scenarios, these methods face challenges when applied to wild videos featuring complex human movements. This is attributed to the high cost of labeled video data and the inherent limitations of explicit warping.

3 Method

Fig. 2 provides an overview of our proposed WildVidFit for video virtual try-on. Given a reference person video sequence $\mathbf{I} := \{I_1, ..., I_N\} \in \mathbb{R}^{3\times H\times W}$ and a target garment image $G \in \mathbb{R}^{3\times H\times W}$, where H and W denote height and width of the image, and N is the frame length of the sequence, WildVidFit aims to synthesis a realistic video sequence $\tilde{\mathbf{I}} := \{\tilde{I}_1, ..., \tilde{I}_N\} \in \mathbb{R}^{3\times H\times W}$. This video showcases the person wearing the target garment G, while maintaining the integrity of all other elements. WildVidFit successfully accomplishes the video try-on task through an image-based approach with two core modules: a one-stage virtual try-on network conditioned on both human motions and garment texture, and a diffusion guidance module for temporal coherence. We start with the preprocessing procedures, followed by a brief introduction on diffusion models. Subsequent subsections further elaborate on the designed one-stage try-on network (Sec. 3.1) and diffusion guidance module (Sec. 3.2).

Preprocessing of Inputs. Drawing inspiration from [5, 25], we propose a method to construct separate representations for the person and the garment, aiming to preserve the individual's identity and accurately reproduce the intricate textures of the garment. Specifically, we obtain the human segmentation map sequence $\mathbf{S} := \{S_1, ..., S_N\}$ and pose maps $\mathbf{P} := \{P_1, ..., P_N\}$ using off-the-shelf methods [3, 26]. Then we produce cloth-agnostic RGB images $\mathbf{A} := \{A_1, ..., A_N\}$ following the progress described in VITON-HD [25]. This process utilizes $\mathbf{P}$ and $\mathbf{S}$ to effectively remove the original clothing but retains the person identity. Finally the cloth-agnostic RGB images $\mathbf{A}$ and the pose maps $\mathbf{P}$

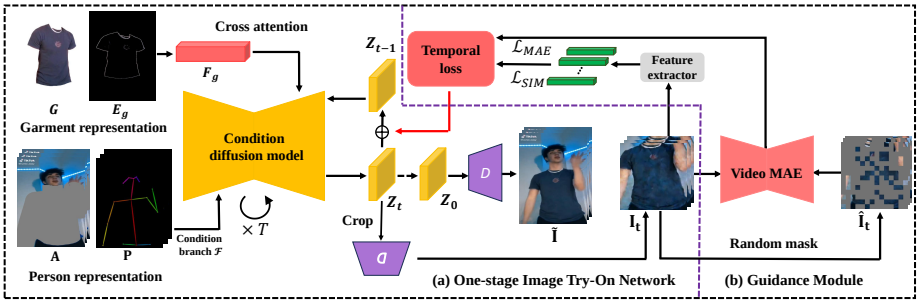

Fig. 2: Overview of our WildVidFit framework. Our method contains two modules, i.e., a one-stage image try-on network and a guidance module. In timestep t, we crop the garment area and decode the latent Z_t into sequence $\mathbf{I_t}$. The similarity loss L_{SIM} is calculated between adjacent frames I_t^{j+1} and I_t^j using spherical distance. Additionally, we randomly mask the sequence $\mathbf{I_t}$ into $\hat{\mathbf{I}}_t$, which is then inputted into VideoMAE for reconstruction. L_{MAE} represents the distance between the sequences $\mathbf{I_t}$ and $\hat{\mathbf{I}}_t$. We assume that a lower reconstruction loss will result in a smoother sequence. L_{SIM} and L_{MAE} together constitute the temporal loss, which controls the sampling process from Z_t to Z_{t-1}.

together form our person representation. For garment representation, in addition to the original garment image G, we introduce low-level information represented by edge map E_g. E_g is detected by Sobel [23]. We utilized DINO-V2 [31] for feature extraction from both garment image G and edge map E_g and then concatenate them into a vector $F_g \in \mathbb{R}^{257\times2048}$. The dimension of 257 refers to the concatenation of a global token and 256 patch tokens.

Controlled Diffusion Model. Diffusion models [17,34] are a class of generative models that learn the target distribution via an iterative denoising procedure. They consist of a Markovian forward process that progressively corrupts the data sample $\mathbf{x}$ into the Gaussian noise $\mathbf{z}_T$, and a learnable reverse process that converts $\mathbf{z}_T$ back to $\mathbf{x}$ iteratively. Importantly, diffusion models can be conditioned on various signals like texts or images. A conditional diffusion model $\hat{\mathbf{x}}_\theta$ can be trained with a weighted denoising score matching objective:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2], \tag{1}$$

where $\mathbf{x}$ is the target data sample, $\mathbf{c}$ is the conditional input, $\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{E})$ is the noise term. Here, $\mathbf{E}$ is used to denote the identity matrix. α_t,σ_t,w_t are functions of the timestep t according to the formulation of diffusion models. In practice, $\hat{\mathbf{x}}_\theta$ is reparameterized as $\hat{\boldsymbol{\epsilon}}_\theta$ to predict the noise that corrupts $\mathbf{x}$ into $\mathbf{z}_t := \alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon}$. During inference, data samples can be generated from Gaussian noise $\mathbf{z}_T\sim\mathcal{N}(\mathbf{0},\mathbf{E})$ using DDIM [35] sampler.

To enable our diffusion model training and inference on limited computational resources without compromising quality and flexibility, we use the pre-trained autoencoder to compress data sample $\mathbf{x}$ into the latent space.

3.1 One-stage Image Try-On Network

The try-on task requires to make controllable image generation where the person wears the target garment while maintaining the original motion. We extend the diffusion model into video try-on task in the form of a conditional image generation task under the joint restriction of the person representation and the garment representation.

As shown in Fig. 3, we take the concatenation of the cloth-agnostic image A and the pose image P as the input in condition branch. A and P are critical for preserving the identity of the person as well as the background. The garment representation including the garment image G and its edge map E_g, is not aligned with the try-on results. Unlike using the warped garment as the input condition to remove this misalignment [12,29], our network adopts a one-stage paradigm, applying implicit warping via the cross attention between the reference person and the extracted garment feature F_g inspired by TryOnDiffusion [48]. The edge map emphasizes the garment details that need to be maintained. Beneficial in avoiding reliance on the explicit optical flow estimation, our network learns how the garment naturally fits on the person, rather than relying on strict pixel-level transformation, which is essential for generalizing to in-the-wild video images.

We pick Stable Diffusion [33] as our base architecture but add a condition branch and make cross attention on the garment instead of text. Both the encoder and decoder of the main UNet consist of four blocks with different scale. The architecture of the condition branch is the same as UNet encoder except the first convolution. We inject the condition signal into the main UNet via convolution. To preserve the prior knowledge essential for improved generation quality, feature aggregation is only performed in the UNet decoder. Formally, the condition branch $\mathcal{F}$ extracts multi-scale features $\mathbf{F_c} = \{\mathbf{F_c^1}, \mathbf{F_c^2}, \mathbf{F_c^3}, \mathbf{F_c^4}\}$ from the input condition $\mathbf{c} = \{\mathbf{A}, \mathbf{P}\}$, $\mathbf{F_c}$ is corresponding to the output of four blocks. We inject the condition features $\mathbf{F_c}$ into the decoder feature $\mathbf{F}_{dec} = \{\mathbf{F}_{dec}^1, \mathbf{F}_{dec}^2, \mathbf{F}_{dec}^3, \mathbf{F}_{dec}^4\}$:

$$\mathbf{F_c} = \mathcal{F}(\mathbf{A}, \mathbf{P}), \tag{2}$$

$$\hat{\mathbf{F}}_{dec}^i = Conv(\mathbf{F}_{dec}^i, \mathbf{F_c^{5-i}}),\ i \in \{1, 2, 3, 4\}, \tag{3}$$

The objective function in training is the same as Eq (1).

3.2 Temporal Coherent Editing using Diffusion Guidance

Generating videos on a frame-by-frame basis will lead to inconsistencies, arising from discrepancies between individual frames. One way to enhance temporal consistency without training a specific temporal module is to leverage the priors in foundational video models.

Diffusion Guidance. One of the diffusion models' notable strengths is their capacity to tailor outputs according to auxiliary information by guiding the sampling process, without fine-tuning the network. Inspired by classifier guidance [7,37], we propose updating the intermediate representation of the sampling

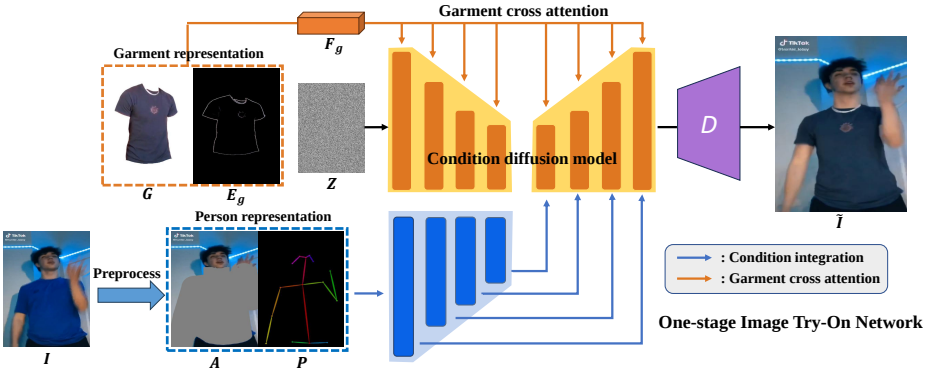

Fig. 3: Overview of the proposed one-stage image try-on network. First, we extract the person representation and garment representation during preprocessing. The person representation includes the cloth-agnostic image A and the human pose P while the garment representation includes the garment image G and the edge map E_g. Then two representations condition the diffusion model in the way of hierarchical fusion in UNet decoder and cross attention respectively.

process by introducing pre-trained models into the gradients through score functions, thereby achieving coherent video generation.

As illustrated in Fig. 2, we introduce the self-supervised video model VideoMAE [10] to enhance the coherence of video clips and a self-supervised image model DINO-V2 [31] to prevent excessive feature distance between adjacent frames. The video masked autoencoder (VideoMAE), which takes masked videos as input and attempts to reconstruct them by leveraging inter-frame relationships, has been proven to learn strong spatio-temporal representations effectively. This guidance is based on the assumption that smoother videos facilitate easier restoration of masked areas by the autoencoder using information from adjacent frames, resulting in lower reconstruction loss. We incorporate the score functions into the DDIM [35] process, as detailed in the following formulation:

$$\begin{aligned}\hat{\epsilon}_t = \epsilon_\theta(z_t; t, \mathbf{c}) - w_1 \nabla_{z_t} \mathcal{L}_{MAE}(z_t) \\ - w_2 \nabla_{z_t} \mathcal{L}_{SIM}(z_t),\end{aligned} \tag{4}$$

$$\mathcal{L}_{MAE} = \frac{1}{\Omega} \sum_{p \in \Omega} ||\mathbf{I_t}(p) - \hat{\mathbf{I}}_\mathbf{t}(p)||_2, \ \mathbf{I_t} = D(z_t), \tag{5}$$

$$\mathcal{L}_{SIM} = \frac{1}{L-1} \sum_{j=1}^{L-1} dist(f(I_t^{j+1}) - f(I_t^j)), \ \mathbf{I_t} = D(z_t), \tag{6}$$

where $z_t \in \mathbb{R}^{L \times H \times W}$ represents video noise at the timestep t, L is length of the video clip, D is the decoder in autoencoder. $\mathcal{L}_{MAE}$ is the reconstruction loss of the masked decoded image sequence $\hat{\mathbf{I}}_\mathbf{t}$, p denotes the token index and Ω denotes

the set of masked tokens in image sequence $\mathbf{I_t}$. $\mathcal{L}_{SIM}$ is the similarity loss that represents the average distance between two adjacent decoded frames I_t^{j+1} and I_t^j at the timestep t. Here f represents the feature extraction function DINO-V2 and we use spherical distance to measure the feature similarity. w_1 and w_2 are the guidance weights. $\mathcal{L}_{SIM}$ and $\mathcal{L}_{MAE}$ together constitute the temporal loss to guide the iterative denoising procedure. In implementation, we set the masking ratio to 0.7, $w_1 = 2000$ and $w_2 = 1000$. Due to the memory limitation, the loss is computed only on the garment area.

Long Video Generation. The length of video clip is fixed in VideoMAE [10]. The naive approach to generate long videos is sequential generation, but this approach tends to perform poorly at the junctions of individual clips. In our framework, we adopt a temporal co-denoising strategy to generate longer videos and ensure temporal smoothness. Specifically, we divide the complete reference video into overlapping short video clips, each differing by stride s, where s is typically $L//2$ or $L//4$. The co-denoising process can then be represented as follows: at the timestep t, the latent z_t^j according the j^{th} frame is the average of all M clips $z_{t,k}, k = 1, ..., M$ including the j^{th} frame:

$$z_t^j = \frac{1}{M} \sum_{k=1}^{M} z_{t,k}^j, \tag{7}$$

3.3 Other module for Enhanced Performance

Autoencoder with Enhanced Mask-Aware Skip Connections. The autoencoder directly impacts the image quality. In order to preserve fine details better outside the garment region, we fine-tune the autoencoder using the mask-aware skip connection module (EMASC) proposed in [29]. The EMASC module is defined as follows, taking the garment mask M from the segmentation map S:

$$D_i = D_{i-1} + f(E_i) * \neg m_i, \tag{8}$$

where f is a learned non-linear function, E_i is the i-th feature map coming from the encoder in autoencoder, D_i is the corresponding i-th decoder feature map, and m_i is obtained by resizing the mask M to adapt the spatial dimension. Here, $\neg m_i$ yields the logical negation of m_i, i.e., obtaining the unmasked region.

Fully Cross-frame Attention. We replace self-attention by fully cross-frame attention in the UNet while making sequential inference to increase the spatial-temporal coherency as proposed in [46].

$$\begin{aligned} \text{Attention}(Q, K, V) &= \text{softmax}(\frac{QK^T}{\sqrt{d}})V, \\ \text{where } Q = W^Q z_t,\ K &= W^K z_t,\ V = W^V z_t, \end{aligned} \tag{9}$$

Here $z_t = \{z_t^i\}_{i=1}^L$ denotes all L latent frames of the video clip at the timestep t, while W^Q, W^K, and W^V project z_t into query, key, and value, respectively.

4 Experiments

Our experiments are divided into four parts. Firstly, we demonstrate the superiority of our one-stage virtual try-on network through image-based virtual try-on experiments in Sec. 4.2. Secondly, we validate the efficacy of our image-based approach on the public VVT dataset [8] in Sec. 4.3. Thirdly, to showcase the robustness and generalizability of the proposed WildVidFit framework, we conduct video virtual try-on on the TikTok dataset [21], which is introduced in Sec. 4.4. Finally, ablation studies are conducted in Sec. 4.5.

4.1 Experiment Setup

Datasets. Our image virtual try-on experiments are conducted on two existing high-resolution virtual try-on benchmarks VITON-HD [6] and DressCode [30]. VITON-HD contains 13679 garment-person pairs, 11647 for training while remaining 2023 for testing. For DressCode dataset, we use the upper subset of it and 15365 image pairs are split into 13564/1801 training/testing pairs.

In the video try-on task, we evaluate our WildVidFit framework on the VVT [8] and TikTok dataset [21]. The VVT dataset [8] includes 791 videos, each with a resolution of 256×192. It is divided into a training set with 159,170 frames and a test set with 30,931 frames. However, the VVT dataset [8] primarily features monotonous and simple human poses against predominantly white backgrounds. In contrast, the TikTok dataset [21] comprises over 300 dance videos that captures a single person performing complex dance moves with intricate limb occlusions and dynamic postures. We selected 165 videos with clear upper body views from this collection. Garment-person pairs are created from TikTok frames using Grounded-SAM [24]. The training set includes 130 videos and 34,933 frames, while the test set contains 35 videos and 9816 frames.

Training and Testing. The main UNet of our one-stage try-on network inherits the parameter of Stable Diffusion [33]. When adapting the network for the try-on task, we train only the UNet decoder and the condition branch $\mathcal{F}$, while keeping the encoder frozen. This training strategy preserves the priors and avoids overfitting. We train the network for 100K iterations with a batch size of 16 using AdamW optimizer [27]. The learning rate is set to $5e^{-5}$. The resolution is 512×384 on the VITON-HD, DressCode and TikTok Dataset, while the VVT dataset maintains its original resolution of 256×192.

At inference time, We use DDIM [35] as the sample method, and the total steps is 30. Classifier-free guidance sample [18] is able to strength the influence of conditions on the generated images. We use the conditional classifier-free guidance on the garment feature F_g, the guidance sacle is set to 2.

4.2 Image Try-on Results

Evaluation Metrics. Following previous studies [25], we make quantitative evaluation on both paired and unpaired setting. In the paired setting, we employ

SSIM [39] and LPIPS [45] as evaluation metrices. In the unpaired setting, where ground truth is unavailable, we evaluate realism using the Fréchet Inception Distance (FID) [16], Kernel Inception Distance (KID) [2] scores and user study. For user study, 100 samples are randomly selected and 50 volunteers are asked to select the one of best quality among different methods.

Fig. 4: Qualitative comparison on VITON-HD dataset. Zoom in for best view.

Comparison with State-of-the-Art Models. We compare our method against CP-VTON [38], HR-VITON [25], LaDI-VTON [25] and DCI-VTON [12] using their official codes and checkpoints. Since no available checkpoint or code for DCI-VTON on DressCode dataset, we skip this comparison.

Qualitative comparison on VITON-HD [6] is exhibited in Fig. 4. Our method consistently preserves essential clothing characteristics, setting it apart from other methods that often falter with inadequate feature retention and evident blurring, as illustrated in rows 2 and 3. Notably, for garments with intricate folds, our model adeptly retains the complex textures, while competing methods tend to produce overly smoothed results, as observed in rows 1 and 4. Such distinc-

Table 1: Quantitative comparison with baselines on VITON-HD dataset.

Methods	SSIM↑	LPIPS↓	FID↓	KID↓	User↑
CP-VTON [38]	0.785	0.2871	48.86	4.42	3.86%
HR-VTON [25]	0.878	0.0987	11.80	0.37	6.62%
Ladi-VTON [29]	0.871	0.0941	13.01	0.66	16.02%
DCI-VTON [12]	0.882	0.0786	11.91	0.51	12.18%
WildVidFit(Ours)	**0.883**	**0.0773**	**8.67**	**0.10**	**61.32%**

Table 2: Quantitative comparison with baselines on DressCode-Upper dataset.

Methods	SSIM↑	LPIPS↓	FID↓	KID↓	User↑
CP-VTON [38]	0.820	0.2764	57.70	4.56	0.00%
HR-VTON [25]	0.924	0.0605	13.80	0.28	5.16%
Ladi-VTON [29]	0.915	0.0620	16.71	0.61	26.20%
WildVidFit(Ours)	**0.928**	**0.0432**	**12.48**	**0.19**	**68.64%**

tions underscore our model's capacity to discern the nuanced interplay between human and garment. Examples on DressCode are presented in the Appendix.

Table 1 and Table 2 show quantitative comparison with previous methods, confirming the superiority of our method in image visual quality under both paired and unpaired evaluation. This reveals that our method achieves state-of-the-art performance in the image-level virtual try-on task.

4.3 Video Try-On Results on VVT Dataset

Evaluation Metrics. We use Video Frechet Inception Distance (VFID) to measure the generation quality and temporal consistency following [8]. VFID is a variant of FID, extracting feature vector of video clips for metric computation by pre-trained video backbone I3D [4]. Each video clip includes 36 frames. Also we adds a user survey for subjective evaluation, with settings consistent with image try-on evaluation above.

Comparison with State-of-the-Art Models. We compare our method with video-based method ClothFormer [13] and imaged-based methods HR-VTON [25] and LaDI-VTON [29]. The quantitative experiment, as shown in Table 3, demonstrates that our method outpaces image-based approaches and matches the performance of the video-based ClothFormer. This underscores the robustness of our one-stage image virtual try-on network and the effectiveness of diffusion guidance in maintaining temporal consistency. The visual comparison can be seen in the Appendix.

4.4 In-the-Wild Video Virtual Try-On

Table 3: Quantitative comparison on the VVT and TikTok dataset.

Methods	Dataset	VFID↓	User↑
HR-VTON [25]	VVT	4.852	9.46%
LaDI-VTON [29]	VVT	4.442	4.24%
ClothFormer [22]	VVT	**4.192**	**46.44%**
WildVidFit(Ours)	VVT	4.202	39.86%
HR-VTON [25]	TikTok	25.43	0.00%
LaDI-VTON [29]	TikTok	14.24	26.90%
WildVidFit(Ours)	TikTok	**9.87**	**73.10%**

Fig. 5: Cross-dataset video try-on results, given a reference video from TikTok dataset and a garment item from DressCode (1st row) and VITON-HD (2nd row) dataset. Zoom in for optimal viewing.

Joint Training on Multiple DataSets. Dance videos from TikTok dataset [21] can effectively evaluate the capability of our method in handling wild videos. To enhance the model's generalization ability for processing the TikTok videos, we conduct joint training using three datasets: VITON-HD [6], DressCode [30], and TikTok [21]. Benefit from this, we are able to transfer the garments from VITON-HD and DressCode onto the TikTok videos as shown in Fig. 5. This to some extent demonstrates the robustness of our method.

Comparison with State-of-the-Art Models. Since there is no public source or commercial software for video try-on, we are compelled to compare our method with image-based methods, i.e, HR-VTON [25] and LaDI-VTON [29]. For fair comparison, HR-VTON and LaDI-VTON also adopt the strategy of joint training on the three datasets mentioned above. Fig. 6 visualizes the comparison results. It can be observed that the GAN-based method HR-VTON completely fails, whereas LaDI-VTON, despite leveraging the foundational capabilities from Stable Diffusion [33], performs poorly in cases of limb occlusion. This is primarily due to the challenges in warping. Our method, on the other hand, accurately

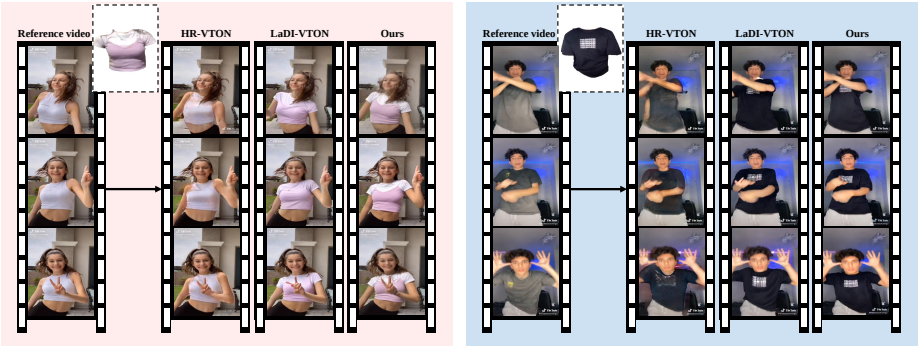

Fig. 6: Qualitative comparison on the TikTok dataset. Our approach can reproduce the details of clothing under dance movements, while other methods perform poorly in cases of limb occlusion. Zoom in for optimal viewing.

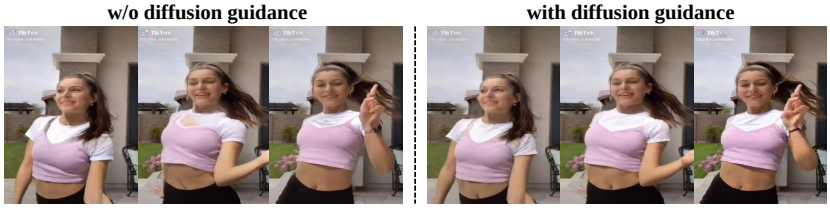

Fig. 7: Effects of diffusion guidance. The guidance module enhances the smoothness and mitigates artifacts on the garment by incorporating overall information.

reproduces the details of the garments, ensuring that the garment fits well with the person's motions. Table 3 also shows the clear superiority of our method.

4.5 Ablation Study

In this section, we analyze the effectiveness the edge map as well as the classifier-free guidance scale (CFG) and the contribution of each module to the temporal consistency.

Effectiveness of Edge Maps and Guidance Scale. In the virtual try-on task, we aim to enhance the preservation of garment textures. We conducted an ablation experiment on the guidance scale of garment feature F_g and the effectiveness of the edge map E_g. As shown in Table 4, the introduction of edge maps using cross-attention has resulted in improvement. And our method achieved the best results when the guidance scale was set to 2, yielding a KID score of 8.67 and a FID score of 0.10.

Effectiveness of Temporal Module. We conducted an ablation study to analyze the designed guided diffusion module and other temporal techniques. The

Table 4: Ablation study for edge maps and CFG on VITON-HD dataset.

Edge maps	Guidance scale	FID↓	KID↓
✗	2	8.93	0.12
✓	1	9.47	0.17
✓	2	**8.67**	**0.10**
✓	3	8.68	0.10

Table 5: Ablation study for temporal modules on TikTok dataset.

Methods	VFID↓
Image-based	13.45
+ Fully cross-frame attention	12.14
+ Guidance with $\mathcal{L}_{MAE}$	10.64
+ Guidance with $\mathcal{L}_{MAE}$ and $\mathcal{L}_{SIM}$	10.28
+ Temporal co-denoising strategy	**9.87**

baseline is the direct prediction of image sequences, and then we sequentially incorporate fully cross-frame attention, guidance with $\mathcal{L}_{MAE}$, guidance with $\mathcal{L}_{SIM}$ and co-denoising strategy, in order to analyze the effectiveness of each module. It can be seen that each module can bring improvement. Fig. 7 shows that the images generated without diffusion guidance exhibit flaws in the garment. This supports the idea that the propose diffusion guidance module can not only enhance the smoothness of videos, but also use the overall video information to rectify some inconsistencies in single images.

5 Conclusions

To effectively tackle the complexities of video virtual try-on in the wild , we introduce WildVidFit, an innovative image-based virtual try-on framework. Specifically designed to manage the challenged poses by frequent movement and significant occlusions common in wild video footage, WildVidFit employs a one-stage, detail-oriented image diffusion model conditioned on both garment and person. The training with a large number of image pairs endows our model with robust performance. Moreover, WildVidFit achieves inter-frame consistency through the technique of diffusion guidance, thereby enabling successful video try-on within a predominantly image-based framework. Our comprehensive experiments reveal that our method not only achieves state-of-the-art performance in image try-on task but also marks a significant foray into video try-on in the wild.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NO. 62322608, NO. 62325605), in part by the Fundamental Research

Funds for the Central Universities under Grant 22lgqb25, in part by the CAAI-MindSpore Open Fund, developed on OpenI Community, and in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2023A01).

References

1. Bai, S., Zhou, H., Li, Z., Zhou, C., Yang, H.: Single stage virtual try-on via deformable attention flows. In: European Conference on Computer Vision. pp. 409–425. Springer (2022)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
5. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023)
6. Choi, S., Park, S., Lee, M., Choo, J.: Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14131–14140 (2021)
7. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
8. Dong, H., Liang, X., Shen, X., Wu, B., Chen, B.C., Yin, J.: Fw-gan: Flow-navigated warping gan for video virtual try-on. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1161–1170 (2019)
9. Dong, X., Zhao, F., Xie, Z., Zhang, X., Du, D.K., Zheng, M., Long, X., Liang, X., Yang, J.: Dressing in the wild by watching dance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3480–3489 (2022)
10. Feichtenhofer, C., Li, Y., He, K., et al.: Masked autoencoders as spatiotemporal learners. Advances in neural information processing systems **35**, 35946–35958 (2022)
11. Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P.: Parser-free virtual try-on via distilling appearance flows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8485–8493 (2021)
12. Gou, J., Sun, S., Zhang, J., Si, J., Qian, C., Zhang, L.: Taming the power of diffusion models for high-quality virtual try-on with appearance flow. arXiv preprint arXiv:2308.06101 (2023)
13. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10471–10480 (2019)
14. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7543–7552 (2018)
15. He, S., Song, Y.Z., Xiang, T.: Style-based global appearance flow for virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3470–3479 (June 2022)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)

18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
19. Huang, Z., Li, H., Xie, Z., Kampffmeyer, M., Liang, X., et al.: Towards hard-pose virtual try-on via 3d-aware global correspondence learning. Advances in Neural Information Processing Systems **35**, 32736–32748 (2022)
20. Issenhuth, T., Mary, J., Calauzènes, C.: Do not mask what you do not need to mask: a parser-free virtual try-on. In: European Conference on Computer Vision. pp. 619–635. Springer (2020)
21. Jafarian, Y., Park, H.S.: Self-supervised 3d representation learning of dressed humans from social media videos. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
22. Jiang, J., Wang, T., Yan, H., Liu, J.: Clothformer: Taming video virtual try-on in all module. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10799–10808 (2022)
23. Kanopoulos, N., Vasanthavada, N., Baker, R.L.: Design of an image edge detection filter using the sobel operator. IEEE Journal of solid-state circuits **23**(2), 358–367 (1988)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
25. Lee, S., Gu, G., Park, S., Choi, S., Choo, J.: High-resolution virtual try-on with misalignment and occlusion-handled conditions. In: Proceedings of the European conference on computer vision (ECCV) (2022)
26. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020). https://doi.org/10.1109/TPAMI.2020.3048039
27. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
28. Men, Y., Mao, Y., Jiang, Y., Ma, W.Y., Lian, Z.: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5084–5093 (2020)
29. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladivton: Latent diffusion textual-inversion enhanced virtual try-on. arXiv preprint arXiv:2305.13501 (2023)
30. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2231–2235 (2022)
31. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
32. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13535–13544 (2022)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)

35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
36. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems **32** (2019)
37. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
38. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: Proceedings of the European conference on computer vision (ECCV). pp. 589–604 (2018)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
40. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: 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. pp. 23550–23559 (2023)
41. Yang, H., Yu, X., Liu, Z.: Full-range virtual try-on with recurrent tri-level transform. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3460–3469 (June 2022)
42. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: 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. pp. 7850–7859 (2020)
43. Yu, R., Wang, X., Xie, X.: 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. pp. 10511–10520 (2019)
44. Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7982–7990 (2021)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
46. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
47. Zhong, X., Wu, Z., Tan, T., Lin, G., Wu, Q.: Mv-ton: Memory-based video virtual try-on network. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 908–916 (2021)
48. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4606–4615 (2023)