Fashion-VDM: Video Diffusion Model for Virtual Try-On

JOHANNA KARRAS, Google Research, University of Washington, USA
YINGWEI LI, Google Research, USA
NAN LIU, Google Research, USA
LUYANG ZHU, Google Research, University of Washington, USA
INNFARN YOO, Google Research, USA
ANDREAS LUGMAYR, Google Research, USA
CHRIS LEE, Google Research, USA
IRA KEMELMACHER-SHLIZERMAN, Google Research, University of Washington, USA

Fig. 1. **Fashion-VDM.** Given an input garment image and a person video, Fashion-VDM generates a video of the person virtually trying on the given garment, while preserving their original identity and motion.

We present Fashion-VDM, a video diffusion model (VDM) for generating virtual try-on videos. Given an input garment image and person video, our method aims to generate a high-quality try-on video of the person wearing the given garment, while preserving the person's identity and motion. Image-based virtual try-on has shown impressive results; however, existing video virtual try-on (VVT) methods are still lacking garment details and temporal consistency. To address these issues, we propose a diffusion-based architecture for video virtual try-on, split classifier-free guidance for increased control over the conditioning inputs, and a progressive temporal training strategy for single-pass 64-frame, 512px video generation. We also demonstrate the effectiveness of joint image-video training for video try-on, especially when video data is limited. Our qualitative and quantitative experiments show that our approach sets the new state-of-the-art for video virtual try-on.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Computer vision**.

ACM Reference Format:
Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. 2024. Fashion-VDM: Video Diffusion Model for Virtual Try-On. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3680528.3687623

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12.
https://doi.org/10.1145/3680528.3687623

1 INTRODUCTION

With the popularity of online clothing shopping and social media marketing, there is a strong demand for virtual try-on methods. Given a garment image and a person image, virtual try-on aims to

show how the person would look wearing the given garment. In this paper, we explore *video* virtual try-on, where the input is a garment image and person video. The benefit of a video virtual try-on (VVT) experience is that it would depict how a garment looks at different angles and how it drapes and flows in motion.

VVT is a challenging task, as it requires synthesizing realistic try-on frames from different viewpoints, while generating realistic fabric dynamics (e.g. folds and wrinkles) and maintaining temporal consistency between frames. Additional difficulty arises if the person and garment poses vary significantly, as this creates occluded garment and person regions that need to be hallucinated. Another challenge is the scarcity of try-on video data. Perfect ground truth data (i.e. two videos of different people wearing the same garment and moving in the exact same way) is difficult and expensive to acquire. In general, available human video data, such as UBC Fashion [Zablotskaia et al. 2019], are much more scarce and less diverse than image data, such as LAION 5B [Schuhmann et al. 2022].

Past approaches to virtual try-on typically leverage dense flow fields to explicitly warp the source garment pixels onto the target person frames [Dong et al. 2022; Haoye Dong and Yin 2019; Jiang et al. 2022; Wen-Jiin Tsai 2023; Zhong et al. 2021]. However, these flow-based approaches can introduce artifacts due to occlusions in the source frame, large pose deformations, and inaccurate flow estimates. Moreover, these methods are incapable of producing realistic and fine-grained fabric dynamics, such as wrinkling, folding, and flowing, as these details are not captured by appearance flows. A recent breakthrough in image-based virtual try-on uses a diffusion model [Zhu et al. 2023], which implicitly warps the input garment under large pose gaps and heavy occlusion using spatial cross-attention. However, directly applying [Zhu et al. 2023] or other image-based try-on methods for VVT in a frame-by-frame manner creates severe flickering artifacts and temporal inconsistencies.

Diffusion models [Dhariwal and Nichol 2021; Ho et al. 2020; Sohl-Dickstein et al. 2015; Song et al. 2020; Song and Ermon 2019] have shown promising results on various video synthesis tasks, such as text-to-video generation [Ho et al. 2022b] and image-to-video generation [Guo et al. 2023; Hu et al. 2023; Karras et al. 2023]. However, a key challenge is generating longer videos, while maintaining temporal consistency and adhering to computational and memory constraints. Previous works use cascaded approaches [Ho et al. 2022a], sliding windows inference [Ho et al. 2022b; Xu et al. 2023], past-frame conditioning [Harvey et al. 2022; Lee et al. 2023; Mei and Patel 2023], and transitions or interpolation [Chen et al. 2023a; Wang et al. 2023b]. Yet, even with such schemes, longer videos are temporally inconsistent, contain artifacts, and lack realistic textures and details. We argue that, similar to context modeling for LLM's [Chen et al. 2023b], short-video generation models can be naturally extended for long-video generation by a temporally progressive finetuning scheme, without introducing additional inference passes or multiple networks.

A potential option for diffusion-based VVT is to apply an animation model to a single try-on image generated by an image try-on model. However, as this is not an end-to-end trained system, any image try-on errors will accumulate throughout the video. We argue that a single VVT model would overcome this issue by 1) injecting explicit person and garment conditioning information into the model and 2) having an end-to-end training objective.

We present *Fashion-VDM*, the first VVT method to synthesize temporally consistent, high-quality try-on videos, even on diverse poses and difficult garments. Fashion-VDM is a single-network, diffusion-based approach. To maintain temporal smoothness, we inflate the M&M VTO [Zhu et al. 2024] architecture with 3D-convolution and temporal attention blocks. We maintain temporal consistency in videos up to 64-frames long with a single network by training in a temporally progressive manner. To address input person and garment fidelity, we introduce split classifier-free guidance (split-CFG) that enables increased control over each input signal. In our experiments, we also show that split-CFG increases realism, temporal consistency, and garment fidelity, compared to ordinary or dual CFG. Additionally, we increase garment fidelity and realism by training jointly with image and video data. Our results show that Fashion-VDM surpasses benchmark methods by a large margin and synthesizes state-of-the-art try-on videos.

2 RELATED WORKS

2.1 Video Diffusion Models

Many early video diffusion models [Ho et al. 2022b] (VDMs) adapt text-to-image diffusion models to generate batches of consecutive video frames, often employing temporal blocks within the denoising UNet architecture to learn temporal consistency [Ho et al. 2022a,b]. Latent VDM's [Andreas Blattmann 2023; Blattmann et al. 2023; Gu et al. 2023; Guo et al. 2023; He et al. 2022b; Karras et al. 2023; Mei and Patel 2023; Wang et al. 2023a] reduce the computational complexity of standard VDM's by performing diffusion in the latent space.

To achieve longer videos and increased spatial resolution, [Ho et al. 2022a] proposes a cascade of temporal and spatial upsampling UNets. Other methods employ similar schemes of cascaded models for long video generation [Wang et al. 2023a]. However, cascaded strategies require multiple networks and inference runs. Another strategy is to synthesize sparse keyframes, then use frame interpolation [Mei and Patel 2023], past-frame conditioning [He et al. 2022b], temporally overlapping frames [Xu et al. 2023], and predicting transitions between frames [Chen et al. 2023a; Wang et al. 2023b] to achieve longer, temporal-consistent videos. Unlike past long-video VDM's, Fashion-VDM is a unified (non-cascaded) diffusion model that generates a long video up to 64 frames long in a single inference run, thereby reducing memory requirements and inference time.

2.2 Image and Pose Guidance

Many VDM's are text-conditioned [Andreas Blattmann 2023; Blattmann et al. 2023; Ho et al. 2022a; Mei and Patel 2023] and there is increasing interest in image-conditioned VDM's [Guo et al. 2023; Hu et al. 2023; Karras et al. 2023]. To maintain the exact details of input images, some methods require inference-time finetuning [Andreas Blattmann 2023; Guo et al. 2023; Karras et al. 2023]. In contrast, Fashion-VDM requires no additional finetuning during test time to maintain high-quality details of the input person and garment.

Some recent diffusion-based animation methods are both image- and pose-conditioned [Girdhar et al. 2023; Guo et al. 2023; Hu et al.

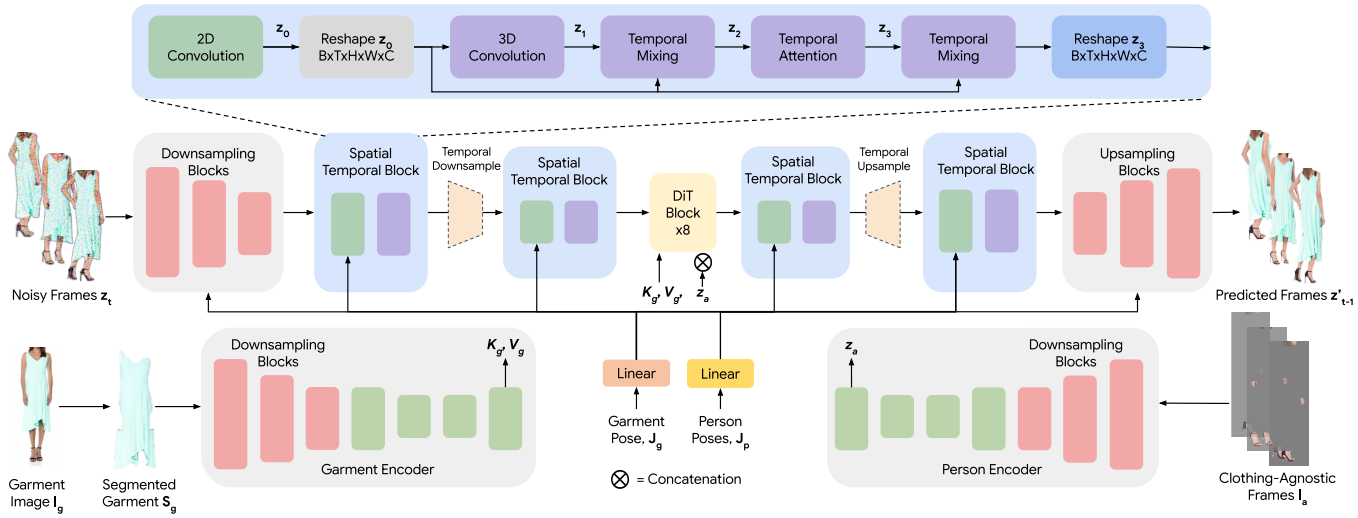

Fig. 2. **Fashion-VDM Architecture.** Given a noisy video z_t at diffusion timestep t, a forward pass of Fashion-VDM computes a single denoising step to get the denoised video z'_{t-1}. Noisy video z_t is preprocessed into person poses J_p and clothing-agnostic frames I_a, while the garment image I_g is preprocessed into the garment segmentation S_g and garment poses J_g (Section 3.3). The architecture follows [Zhu et al. 2024], except the main UNet contains 3D-Conv and temporal attention blocks to maintain temporal consistency. Additionally, we inject temporal down/upsampling blocks during 64-frame temporal training. Noisy video z_t is encoded by the main UNet and the conditioning signals, S_g and I_a, are encoded by separate UNet encoders. In the 8 DiT blocks at the lowest resolution of the UNet, the garment conditioning features are cross-attended with the noisy video features and the spatially-aligned clothing-agnostic features z_a and noisy video features are directly concatenated. J_g and J_p are encoded by single linear layers, then concatenated to the noisy features in all UNet 2D spatial layers.

2023; Karras et al. 2023; Xu et al. 2023]. DreamPose uses a pre-trained (latent) Stable Diffusion model without temporal layers to generate videos in a frame-by-frame manner [Karras et al. 2023]. More recently, Animate Anyone [Hu et al. 2023] encodes the image using ReferenceNet and their diffusion model incorporates spatial, cross, and temporal attention layers to maintain consistency and preserve details, while MagicAnimate [Xu et al. 2023] introduces an appearance encoder to maintain the fidelity across the frames and generates a long video using temporally overlapping segments. In contrast, Fashion-VDM is a non-latent, temporally-aware video diffusion model, capable of synthesizing up to 64 consecutive frames in a single inference pass.

2.3 Virtual Try-On

Traditional image virtual try-on approaches first warp the target garment onto the input person, then refine the resulting image [Bai et al. 2022; Choi et al. 2021; Cui et al. 2023; Han et al. 2018; He et al. 2022a; Lee et al. 2022; Men et al. 2020; Ren et al. 2022; Yang et al. 2020; Yu et al. 2019; Zhang et al. 2021]. Similarly, for video virtual try-on (VVT), past methods often rely on multiple networks to predict intermediate values, such as optical flow, background masks, and occlusion masks, to warp the target garment to the person in each frame of the video [Dong et al. 2022; Haoye Dong and Yin 2019; Jiang et al. 2022; Wen-Jiin Tsai 2023; Zhong et al. 2021]. However, inaccuracies in these intermediate values lead to artifacts and misalignment. Some image try-on approaches incorporate optical flow estimation to alleviate this misalignment [Bai et al. 2022; Lee et al. 2022; Lewis et al. 2021; Xintong Han and Scott 2020]. For VVT, MV-TON [Zhong et al. 2021] proposes a memory refinement module to correct inaccurate details in the generated frames by encoding past frames into latent space, then using this as external memory to generate new frames. ClothFormer [Jiang et al. 2022] estimates an occlusion mask to correct for flow inaccuracies. Current state-of-the-art VVT methods achieve improved results by utilizing attention modules in the warping and fusing phases [Jiang et al. 2022; Wen-Jiin Tsai 2023].

In contrast to earlier flow-based methods, TryOnDiffusion [Zhu et al. 2023] leverages a diffusion-based method conditioned with pose and garment for image virtual try-on. WarpDiffusion [Zhang et al. 2023] tries to reduce the computational cost and data requirements by bridging warping and diffusion-based virtual try-on methods. StableVITON [Kim et al. 2023] avoids warping by finetuning pre-trained latent diffusion [Rombach et al. 2022] encoders for input person and garment conditioning via cross-attention blocks. Mix-and-match (M&M) VTO [Zhu et al. 2023] extends single tryon task for mixmatch tryon application with a novel person embedding finetuning strategy.

2.4 Image and Video Training

Video datasets are often smaller and less diverse, compared to image datasets, as images are more abundant online. To alleviate this problem, [Ho et al. 2022a,b; Xu et al. 2023] propose jointly leveraging image and video data for training. VDM [Ho et al. 2022b] and Imagen Video [Ho et al. 2022a] implement joint training by applying

a temporal mask to image batches. MagicAnimate [Xu et al. 2023] applies joint training during the pretraining stage of their appearance encoder and pose ControlNet. We improve upon existing joint training schemes (see Section 3.7), ultimately demonstrating the benefit of joint image and video training for video try-on.

3 METHOD

We propose Fashion-VDM, a unified video diffusion model for synthesizing state-of-the-art virtual try-on (VTO) videos up to 64 frames long at 512px resolution. Our method introduces an end-to-end diffusion-based VVT architecture based on [Zhu et al. 2024] (Section 3.4), split classifier-free guidance (split-CFG) for increased garment fidelity (Section 3.5), progressive temporal training for long-video generation (Section 3.6), and joint image-video training for improved garment fidelity (Section 3.7).

3.1 Problem Formulation

In video virtual try-on, the input is a video $\{I_p^0, I_p^1, ..., I_p^{N-1}\}$ of a person p consisting of N frames and a single garment image I_g of another person wearing garment g. The goal is to synthesize a video $\{I_{tr}^0, I_{tr}^1, ..., I_{tr}^{N-1}\}$, where I_{tr}^i denotes the i-th **tr**y-on video frame that preserves the identity and motion of the person p wearing the garment g.

3.2 Preliminary: M&M VTO

Our VTO-UDiT network architecture is inspired by [Zhu et al. 2024], a state-of-the-art multi-garment *image* try-on diffusion model that also enables text-based control of garment layout. VTO-UDiT is represented by

$$\hat{x_0} = x_\theta(z_t, t, c_{tr}) \tag{1}$$

where $\hat{x_0}$ is the predicted try-on image by the network x_θ, parameterized by θ, at diffusion timestep t, z_t is the noisy image, and c_{tr} is the conditioning inputs. VTO-UDiT is parameterized in v-space, following [Salimans and Ho 2022]. Each conditioning input is encoded separately by fully convolutional encoders and processed at the lowest resolution of the main UNet via DiT blocks [Peebles and Xie 2022], where conditioning features are processed with self-attention or cross-attention modules. However, while it shows impressive results for image try-on, VTO-UDiT cannot reason about temporal consistency when applied to video inputs.

3.3 Input Preprocessing

From the input video frames, we compute the clothing-agnostic frames $I_a = \{I_a^0, I_a^1, ..., I_a^{N-1}\}$, person poses $J_p = \{J_p^0, J_p^1, ..., J_p^{N-1}\}$, and person masks $\{M_p^0, M_p^1, ..., M_p^{N-1}\}$. The clothing-agnostic frames mask out the entire bounding box area of the person in the frame, except for the visible body regions (head, hands, legs, and shoes), following TryOnDiffusion [Zhu et al. 2023]. Optionally, the clothing-agnostic frames can keep the original bottoms, if doing top try-on only. From the input garment image I_g, we extract the garment segmentation image S_g, garment pose J_g, and garment mask M_g. The garment pose refers to the pose keypoints of the person wearing the garment before segmentation. We channel-wise concatenate M_p^i to I_a^i and M_g to I_g. Poses, masks, and segmentations are computed using an in-house equivalent of Graphonomy [Gong et al. 2019]. Both person and garment pose keypoints are preprocessed to be spatially aligned with the person frames and garment image, respectively.

3.4 Architecture

Our overall architecture is depicted in Figure 2. We adapt the VTO-UDiT architecture [Zhu et al. 2023] by inflating the two lowest-resolution downsampling and upsampling blocks with temporal attention and 3D-Conv blocks, as shown in Figure 2. To be specific, after the 2D-Conv layers, we add a 3D-Conv block, a temporal attention block, and a temporal mixing block to linearly combine spatial and temporal features, as proposed in [Blattmann et al. 2023]. In the temporal mixing blocks, processed features after the spatial attention layer z_s are linearly combined with processed features after the temporal attention layer z_t via learned weighting parameter α:

$$z_t' = \alpha \cdot z_s + (1 - \alpha) \cdot z_t \tag{2}$$

During 64-frame training (see Section 3.6), we further inflate the model with temporal downsampling and upsampling blocks with factor 2, to reduce the memory footprint of the model. These blocks are added before and after the lowest-resolution spatial blocks, respectively.

Image and Pose Conditioning The person and garment poses are encoded and used to condition all 2D spatial layers in the UNet. The 8 Diffusion Transformer (DiT) blocks [Peebles and Xie 2022] between the UNet encoder and decoder condition our model on the segmented garment and clothing-agnostic image features, as proposed by [Zhu et al. 2024]. In each block, the garment images are cross-attended with the noisy target features, while the agnostic input images are concatenated to the noisy target features.

3.5 Split Classifier-Free Guidance

Standard classifier-free guidance (CFG) [Ho and Salimans 2022] is a sampling technique that pushes the distribution of inference results towards the input conditioning signal(s); however, it does not allow for disentangled guidance towards separate conditioning signals. Instruct-Pix2Pix [Brooks et al. 2023] introduces dual-CFG, which separates the CFG weights for text and image conditioning signals, drawing inspiration from Composable Diffusion [Liu et al. 2022].

We introduce *split*-CFG, a generalization of dual-CFG which allows independent control over multiple conditioning signals. See Algorithm 1. The inputs to `Split-CFG` are the trained denoising UNet ϵ_θ, the list of all conditioning signal sets C, and the respective conditioning weights W. For each subset of conditioning signals $c_i \in C$, containing one or more conditional inputs, the algorithm computes the conditional result ϵ_i given c_i. Then, the weighted difference between the conditional result ϵ_i from the past conditional result ϵ_{i-1} is added to the prediction. In this way, the prediction is pushed in the direction of c_i.

Split-CFG is naturally dependent on the order of the conditioning signals. Intuitively, the first conditional output will have the largest distance from the null output, thus most affecting the final result. In our implementation, our conditioning groups C consist of (1) the empty set (unconditional inference), (2) the clothing-agnostic images ($\{I_a^0, ..., I_a^{N-1}\}$), (3) all clothing-related inputs (S_g, J_g, M_g), and (4) lastly, all remaining conditioning inputs ($\{J_p^0, ..., J_p^{N-1}\}$. We

Algorithm 1: Split Classifier-Free Guidance

Split-CFG(ϵ_θ, C, W)
- $c \leftarrow \emptyset$ ▷ current conditioning signals;
- $\hat{\epsilon}_\theta(z_t, C) \leftarrow w_\emptyset \epsilon_\theta(z_t, \emptyset)$ ▷ initialize prediction;
- $\hat{\epsilon}_0 \leftarrow \hat{\epsilon}_\theta(z_t, C)$ ▷ store past prediction;
- **for** c_i *in* C **do**
  - $c \leftarrow c \cup \{c_i\}$ ▷ update c;
  - $\hat{\epsilon}_i \leftarrow \epsilon_\theta(z_t, c)$ ▷ store new prediction;
  - $\hat{\epsilon}_\theta(z_t, C) \leftarrow \hat{\epsilon}_\theta(z_t, C) + w_i(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})$;
  - $\hat{\epsilon}_{i-1} \leftarrow \hat{\epsilon}_i$ ▷ update $\hat{\epsilon}_{i-1}$
- **end**
- **return** $\hat{\epsilon}_\theta(z_t, C)$

denote the respective weights of each term as $(w_\emptyset, w_p, w_g, w_{\text{full}})$. Empirically, we find this ordering yields the best results.

Overall, we find that controlling sampling via split-CFG not only enhances the frame-wise garment fidelity, but also increases photo-realism (FID) the inter-frame consistency of video (FVD), compared to ordinary CFG.

3.6 Progressive Temporal Training

Our novel progressive temporal training enables up to 64-frame video generation in a single inference run. We first train a base image model from scratch on image data at 512px resolution and image batches of shape $BxTxHxWxC$, with batch size $B = 8$ and length $T = 1$, for 1M iterations. Then, we inflate the base architecture with temporal blocks and continue training the same spatial layers and new temporal layers with image and video batches with batch size $B = 1$ and length $T = 8$. Video batches are consecutive frames of length T from the same video. After convergence, we double the video length T to 16. This process is repeated until we reach the target length of 64 frames. Each temporal phase is trained for 150K iterations. The benefit of such a progressive process is a faster training speed and better multi-frame consistency. Additional details are provided in the Supplementary.

3.7 Joint Image and Video Training

Training the temporal phases solely with video data, which is much more limited in scale compared to image data, would disregard the image dataset entirely after the pretraining phase. We observe that video-only training in the temporal phases sacrifices image quality and fidelity for temporal smoothness. To combat this issue, we train the temporal phases jointly with 50% image batches and 50% video batches. We implement joint training via conditional network branching [Huang et al. 2016], i.e. for image batches, we skip updating the temporal blocks in the network. Unlike temporal masking strategies[Ho et al. 2022a,b], using conditional network branching allows us to include other temporal blocks (Conv-3D, temporal mixing) in addition to temporal attention. Critically, we also train with either image-only or video-only batches, rather than batches of video with appended images [Ho et al. 2022a,b]. This improves data diversity and training stability by not constraining the possible batches by the number of available video batches. We observe that improved garment fidelity and multi-view realism, especially for synthesized details in occluded garment regions with

Fig. 3. **Joint Training Ablation.** Joint image and video training improves the realism of occluded views.

Fig. 4. **Split-CFG Ablation.** We compare different split-cfg weights, where $(w_\emptyset, w_p, w_g, w_{\text{full}})$ correspond to the unconditional guidance, person-only guidance, person and cloth guidance, and full guidance terms, respectively.

joint image-video training compared to video-only training (see Figure 3).

4 EXPERIMENTS

In this section, we describe our datasets (Section 4.1), evaluation metrics (Section 4.2), and results (Section 4.3). We provide training and inference details in the Supplementary.

4.1 Datasets

Our image dataset is a collection of publicly-crawled online fashion images, containing 17M paired images of people wearing the same garment in different poses. We also collect a video dataset of over 52K publicly-available fashion videos totalling 3.9M frames, which we use for the temporal training phases. During training, the garment image and person frames are randomly sampled from the same video. For evaluation, we collect a separate dataset of 5K videos, containing person videos paired with garment images from a *different* video. Our custom image and video datasets contain a diverse range of skin tones, body shapes, garments, genders, and motions. We also evaluate on the UBC test dataset [Zablotskaia et al. 2019] of 100 videos. For both test datasets, we randomly pair a garment frame from each video clip with three distinct other video clips to get swapped try-on datasets.

4.1.1 Reproducibility. To promote future work in this area and allow fair comparisons with our method, we plan to release a benchmark dataset, including sample paired person videos, garment images, and corresponding preprocessed inputs. We also analyze a version of our model trained and tested exclusively on publicly-available UBC video data [Zablotskaia et al. 2019] in Section 4.3.1.

	UBC Test Dataset			Our Test Dataset		
	FID ↓	FVD ↓	CLIP ↑	FID ↓	FVD ↓	CLIP ↑
w/o Split-CFG	145	687	0.745	78	450	0.663
w/o Joint Training	106	579	0.744	96	565	0.651
w/o Prog. Training	102	631	0.736	87	824	0.651
w/o Temporal Blocks	94	1019	0.739	95	565	0.642
Ours (Full)	**86**	**515**	**0.752**	**71**	**377**	**0.669**

Table 1. Quantitative Ablation Studies. For each ablated version of our model, we compute FID, FVD, and CLIP scores using both UBC and our test videos with randomly paired garments. Bolded values indicate the best score in each column.

4.2 Metrics

We evaluate our method using FID [Heusel et al. 2017], FVD [Unterthiner et al. 2018], and CLIP [Radford et al. 2021] scores in Tables 1 and 2. FID measures the similarity of the distributions of the predicted and ground-truth frames, which gives a measure of the realism of the generated video frames. FVD measures temporal consistency of video frames. We compute the CLIP image similarity between the segmented garments of the input garment image and predicted frames. In this way, the CLIP score gives us a measure of try-on garment fidelity.

4.3 Results

We showcase qualitative results of our full method in Figure 7 and provide more qualitative results in the Supplementary. Fashion-VDM is capable of synthesizing smooth, photorealistic try-on videos on a variety of input garment types, patterns, skin tones, genders, and motions.

4.3.1 UBC-Only Model. In order to provide a fair comparison to other methods [Andreas Blattmann 2023; Guo et al. 2023; Karras et al. 2023], we train a version of Fashion-VDM using video data only from the UBC dataset. Similar to other methods, we leverage a pretrained image try-on diffusion models and further train using the publicly-available UBC dataset for the video stages. We show the quantitative results in Table 2 and provide further details, discussion, and qualitative examples in the Supplementary.

5 ABLATION STUDIES

We ablate each of our design choices with respect to garment fidelity, temporal smoothness, and photorealism. We report quantitative results for each ablated version in Table 1. All components are essential to improving realism (FID), temporal consistency (FVD), and garment fidelity (CLIP). Qualitatively, we find that split-CFG and joint training have the largest effect on person/garment fidelity and overall quality (Figure 5), while progressive training and temporal blocks affect the temporal smoothness (Figure 6). We discuss these effects in detail in the remainder of this section.

5.1 Split Classifier-Free Guidance

Split-CFG improves per-frame person and garment fidelity, thereby improving overall inter-frame temporal consistency and photrealism. In Figure 4, we compare results generated with different split-CFG weights at inference time. By increasing the person guidance weight w_p from 0 to 1, the realism and identity of the input person are improved. Increasing the full-conditional weight w_{full} improves the garment fidelity, but not as much as by increasing the garment weight w_g alone, as in the last column. We provide quantitative split-CFG ablation results in the Supplementary. In the Supplementary, we demonstrate that increasing w_g also increases fine-grain garment details when using a version of our model trained on limited video data. This suggests split-CFG does not require extensive training to be useful and can be impactful in low-resource settings.

5.2 Joint Image-Video Training

We find that training with video data only in the temporal phases sacrifices garment fidelity compared to the base image model. Training jointly with images and videos increases the fidelity to garment details, even compared to the image baseline, as shown by the improved FID and CLIP scores in Table 1. The increased access to diverse data with joint image-video training also enables the model to synthesize more plausible occluded regions. For example, as shown in Figure 3, the jointly trained model is able to generate a hood with more realism than the video-only model.

5.3 Temporal Blocks

As seen in prior works [Ho et al. 2022a,b], interleaving 3D-convolution and temporal attention blocks into the 2D UNet greatly improves temporal consistency. Removing temporal blocks entirely causes large temporal inconsistencies. For instance, in the top row of Figure 6, the ablated model without temporal blocks swaps the pants and body shape in each frame.

5.4 Progressive Temporal Training

To ablate our progressive training scheme, we train our image base model directly with 16-frame video batches for the same total number of iterations, but skipping the 8-frame training phase entirely. Progressive training enables more temporally smooth results with the same number of training iterations. This is supported by our quantitative findings in Table 1, which indicates worse FVD when not doing progressive training. Qualitatively, in Figure 6, the non-progressively trained model in the middle row exhibits temporal artifacts in the pants region and intermittently merges the pant legs into a skirt. We hypothesize that, given limited training iterations, it is easier to learn temporal consistency well across a small number of frames. Then, to transfer that knowledge to larger temporal windows only requires minimal additional training.

6 COMPARISONS TO STATE-OF-THE-ART

We qualitatively and quantitatively compare our method to the state-of-the-art in diffusion-based try-on and animation, as no previous diffusion-based video try-on baselines with publicly-available code currently exist: (1) TryOn Diffusion [Zhu et al. 2023] (2) MagicAnimate [Xu et al. 2023], and (3) Animate Anyone [Hu et al. 2023]. For (1), we generate try-on results in a frame-by-frame manner for each input frame to generate a video. For (2) and (3), we first generate a single try-on image from the first input frame and garment image using TryOn Diffusion, then use the extracted poses from the input frames to animate the result. In addition, we provide user survey results in the Supplementary.

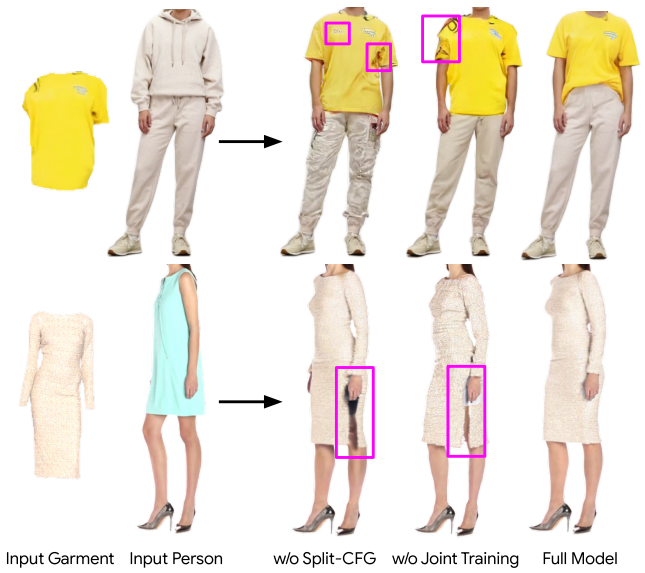

Fig. 5. **Garment Fidelity Ablations.** We compare our full model with ablated versions without split-CFG and without joint image-video training in terms of garment fidelity. Both split-CFG and joint image-video training improve fine-grain garment details (top row) and novel view generation (bottom row).

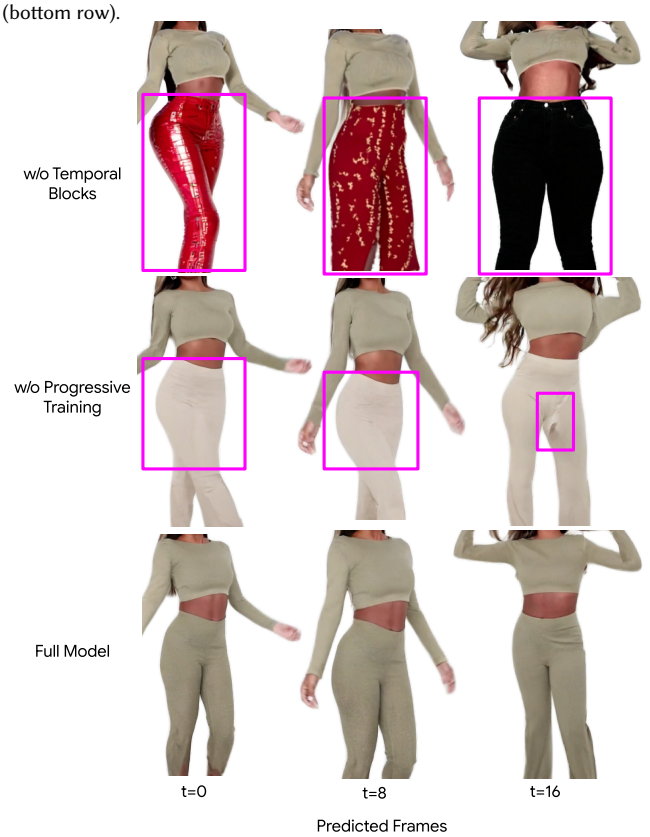

Fig. 6. **Temporal Smoothness Ablations.** We compare video frames generated by our ablated model without temporal blocks (top row) and without progressive training (middle row) to our full model (bottom row). Both ablated versions exhibit large frame-to-frame inconsistencies and artifacts.

	UBC Test Dataset			Our Test Dataset		
	FID ↓	FVD ↓	CLIP* ↑	FID ↓	FVD ↓	CLIP* ↑
TryOn Diffusion	94	1019	0.739	95	960	0.663
Magic Animate	155	1861	0.702	97	694	0.642
Animate Anyone	118	819	0.727	112	468	0.629
Ours (Full)	86	515	0.752	**71**	**377**	**0.669**
Ours (UBC-Only)	**39**	**172**	**0.749**	129	949	0.657

Table 2. Quantitative Comparisons. We compare Fashion-VDM to the baseline methods using the UBC test dataset [Zablotskaia et al. 2019] and our test dataset of internet videos. Fashion-VDM quantitatively outperforms other methods on all metrics.

6.1 Qualitative Results:

We qualitatively compare Fashion-VDM to the baseline methods in Figure 8. In the top and bottom rows, we show how other methods exhibit large artifacts with large pose changes. In these examples, baseline methods struggle to preserve garment details and hallucinate plausible occluded views. Plus, both MagicAnimate and Animate Anyone create an overall cartoon-like appearance.

In our supplementary video results, we observe that frame-by-frame TryOn Diffusion results exhibit lots of flickering and garment inconsistencies. MagicAnimate fails to preserve the correct background and also does not maintain a consistent garment appearance througout the video. Animate Anyone also exhibits garment temporal inconsistency, especially with large viewpoint changes, and the human motion has an unrealistic, warping effect. Overall, Fashion-VDM synthesizes more natural-looking garment motion, such as folding, wrinkling, and flow, and better preserves garment appearance.

6.2 Quantitative Results:

We compute FID scores on from 300 16-frame videos of the UBC dataset and on 300 16-frame videos of our custom video test dataset. For both datasets, we compute FVD scores and CLIP on 100 distinct 16-frame videos. The results are displayed in Table 2. In our experiments, Fashion-VDM surpasses all baselines in both image quality (FID), video quality (FVD), and garment fidelity (CLIP). Although the UBC-only model excels in terms of all UBC metrics, we qualitatively observe over-smoothing and worse garment detail preservation, compared to the full version trained on our larger, more diverse video dataset.

7 LIMITATIONS & FUTURE WORK

The main limitations of Fashion-VDM include inaccurate body shape, artifacts, and incorrect details in occluded garment regions. See examples and further discussion in the Supplementary. Improbable details may be hallucinated in unseen garment regions, because the input image only shows one view of the garment. Future work might consider multi-view conditioning and individual person customization for improved garment and person fidelity. Other errors include minor aliasing for fine-grained patterns. Finally, our method does not simulate exact physical cloth dynamics, but rather realistic video try-on visualization. Establishing physics could be a great next step.

8 DISCUSSION

We present Fashion-VDM, a diffusion-based video try-on model. Given an input garment image and person video, Fashion-VDM synthesizes a try-on video with the input garment fitted to the person in motion, maintaining realistic details and fabric dynamics. We show qualitatively and quantitatively that our method significantly surpasses existing state-of-the-art diffusion-based image try-on and animation methods.

9 ETHICS STATEMENT

While we believe our research creates a positive contribution to the research community by advancing the state-of-the-art in generative video diffusion, we also condemn its potential for misuse, including any spreading misinformation or manipulating human content for malicious purposes. While our method is trained on public data containing identifiable humans, we will not release any images or videos containing personally identifiable features, such as faces, tattoos, or logos to protect the privacy of these individuals.

REFERENCES

Sumith Kulal Daniel Mendelevitch Maciej Kilian Dominik Lorenz Yam Levi Zion English Vikram Voleti Adam Letts Varun Jampani Robin Rombach Andreas Blattmann, Tim Dockhorn. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets.

Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. 2022. Single Stage Virtual Try-On Via Deformable Attention Flows. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 409–425.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 2023. Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22563–22575.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18392–18402.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending Context Window of Large Language Models via Positional Interpolation. arXiv:arXiv:2306.15595

Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023a. SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. arXiv:2310.20700

Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. VITON-HD: High-Resolution Virtual Try-On via Misalignment-Aware Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14131–14140.

Aiyu Cui, Jay Mahajan, Viraj Shah, Preeti Gomathinayagam, and Svetlana Lazebnik. 2023. Street TryOn: Learning In-the-Wild Virtual Try-On from Unpaired Person Images. arXiv:2311.16094

Prafulla Dhariwal and Alex Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. arXiv:arXiv:2105.05233

Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K. Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. 2022. Dressing in the Wild by Watching Dance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3480–3489.

Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. 2023. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. arXiv:arXiv:2311.10709

Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. 2019. Graphonomy: Universal Human Parsing via Graph Transfer Learning. arXiv:arXiv:1904.04536

Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. 2023. Reuse and Diffuse: Iterative Denoising for Text-to-Video Generation. arXiv:arXiv:2309.03549

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2023. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. arXiv:arXiv:2307.04725

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaohui Shen B. Wu Bing cheng Chen Haoye Dong, Xiaodan Liang and J. Yin. 2019. FW-GAN: Flow-Navigated Warping GAN for Video Virtual Try-On. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Yunlin, Taiwan, 1161––1170.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. 2022. Flexible Diffusion Modeling of Long Videos. arXiv:arXiv:2205.11495

Sen He, Yi-Zhe Song, and Tao Xiang. 2022a. Style-Based Global Appearance Flow for Virtual Try-On. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3470–3479.

Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. 2022b. Latent Video Diffusion Models for High-Fidelity Long Video Generation. arXiv:arXiv:2211.13221

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022a. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:arXiv:2210.02303

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. arXiv:arXiv:2006.11239

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. arXiv:arXiv:2207.12598

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022b. Video Diffusion Models. arXiv:arXiv:2204.03458

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. 2023. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. arXiv:arXiv:2311.17117

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. 2016. Deep Networks with Stochastic Depth. arXiv:arXiv:1603.09382

Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. 2022. ClothFormer: Taming Video Virtual Try-On in All Module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10799–10808.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. 2023. DreamPose: Fashion Video Synthesis with Stable Diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22680–22690.

Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2023. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. arXiv:2312.01725

Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 204–219.

Seung Hyun Lee, Sieun Kim, Innfarn Yoo, Feng Yang, Donghyeon Cho, Youngseo Kim, Huiwen Chang, Jinkyu Kim, and Sangpil Kim. 2023. Soundini: Sound-Guided Diffusion for Natural Video Editing. arXiv:2304.06818

Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. 2021. TryOnGAN: body-aware try-on via layered interpolation. *ACM Trans. Graph.* 40, 4, Article 115 (jul 2021), 10 pages. https://doi.org/10.1145/3450626.3459884

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*. Springer, 423–439.

Kangfu Mei and Vishal Patel. 2023. VIDM: Video Implicit Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (Jun. 2023), 9117–9125.

Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. 2020. Controllable Person Image Synthesis With Attribute-Decomposed GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

William Peebles and Saining Xie. 2022. Scalable Diffusion Models with Transformers. arXiv:arXiv:2212.09748

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:arXiv:2103.00020

Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H. Li. 2022. Neural Texture Extraction and Distribution for Controllable Person Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13535–13544.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 10684–10695.

Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. arXiv:arXiv:2202.00512

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:arXiv:2210.08402

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. arXiv:arXiv:1503.03585

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. arXiv:arXiv:2010.02502

Yang Song and Stefano Ermon. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. arXiv:arXiv:1907.05600

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards Accurate Generative Models of Video: A New Metric Challenges. arXiv:arXiv:1812.01717

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. 2023b. Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising. arXiv:2305.18264

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. 2023a. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. arXiv:arXiv:2309.15103

Yi-Cheng Tien Wen-Jiin Tsai. 2023. Attention-based Video Virtual Try-On. ACM, Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, 209–216.

Weilin Huang Xintong Han, Xiaojun Hu and Matthew R Scott. 2020. Clothflow: A flow-based model for clothed person generation. Proceedings of the IEEE/CVF international conference on computer vision, 139–144,.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2023. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. arXiv:arXiv:2311.16498

Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. 2019. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. Proceedings of the IEEE/CVF international conference on computer vision, 10511–10520.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. DwNet: Dense warp-based network for pose-guided human video generation. arXiv:arXiv:1910.09139

Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. 2021. PISE: Person Image Synthesis and Editing With Decoupled GAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7982–7990.

Xujie Zhang, Xiu Li, Michael Kampffmeyer, Xin Dong, Zhenyu Xie, Feida Zhu, Haoye Dong, and Xiaodan Liang. 2023. WarpDiffusion: Efficient Diffusion Model for High-Fidelity Virtual Try-on. arXiv:2312.03667

Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. 2021. MV-TON: Memory-based Video Virtual Try-on network. (2021). arXiv:arXiv:2108.07502

Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. 2024. MM VTO: Multi-Garment Virtual Try-On and Editing.

Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. 2023. TryOnDiffusion: A Tale of Two UNets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4606–4615.

Fig. 7. **Qualitative Results.** We showcase video try-on results generated by Fashion-VDM using randomly paired person-garment test videos from the UBC dataset [Zablotskaia et al. 2019] and our own collected test dataset. Note that the input garment image and input person frames come from different videos.

Fig. 8. **Qualitative Comparisons.** Fashion-VDM outperforms past methods in garment fidelity and realism. Especially in cases of large disocclusion, our method synthesizes more realistic novel views.