Virtual Fitting Room: Generating Arbitrarily Long Videos of Virtual Try-On from a Single Image

Jun-Kun Chen[1,2*] **Aayush Bansal**[1] **Minh Phuoc Vo**[1] **Yu-Xiong Wang**[2]
[1]SpreeAI [2]University of Illinois Urbana-Champaign
immortalco.github.io/VirtualFittingRoom

Abstract

We introduce Virtual Fitting Room (VFR), a novel video generative model that produces arbitrarily long virtual try-on videos. Our VFR models long video generation tasks as an auto-regressive, segment-by-segment generation process, eliminating the need for resource-intensive generation and lengthy video data, while providing the flexibility to generate videos of arbitrary length. The key challenges of this task are twofold: ensuring local smoothness between adjacent segments and maintaining global temporal consistency across different segments. To address these challenges, we propose our VFR framework, which ensures smoothness through a prefix video condition and enforces consistency with the anchor video—a 360° video that comprehensively captures the human's whole-body appearance. Our VFR generates minute-scale virtual try-on videos with both local smoothness and global temporal consistency under various motions, making it a pioneering work in long virtual try-on video generation.

1 Introduction

Imagine being in a fitting room, trying on a garment, when a hurried knock interrupts you. Would that allow you to truly experience the garment before buying it? No. To truly understand a garment, one may want to interact with it in various ways. The computational methods for virtually trying on a garment enable a user to see themselves, but only in an image [1–6] or a short 5∼10s video [7, 8], limiting the user's ability to fully experience a garment. We introduce *Virtual Fitting Room* (VFR) to enable a user to study the interaction of garments with their body as long as they like. Unlike existing image or video try-on methods, VFR allows a user to create arbitrarily long videos (720×1152 resolution at 8 FPS and can be further refined to 24 FPS) of themselves, given a single user image, a desired garment, and a reference video performing the desired try-on motion. Fig. 1 shows a 30s and a 90s video generated using our method.

Generating a 5s-long video is already a computationally demanding task [7, 8]. Naively extending these methods requires even more computational resources and a large-scale video dataset containing long videos for learning. To overcome these limitations, one may generate multiple short segments of a long video one by one "auto-regressively" in timestamp order, and then merge them to create a long video, as visualized in Fig. 4-(a). Inspired by common approaches in general long video generation [9–14], we allow each segment to slightly overlap with the previous segment, and pass the overlapped "prefix" of the current segment as a condition, ensuring a locally smooth transition between each adjacent segment pair. However, these generated videos lack global temporal consistency, as shown in Fig. 2-(a), which is difficult to fix after once they are generated. In this work, we draw inspiration from the process of writing an essay with an outline as an "anchor." We posit that creating long videos is a two-step process that involves: (1) creating an "outline" or an "anchor" to guide the

*Work done during an internship at SpreeAI.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

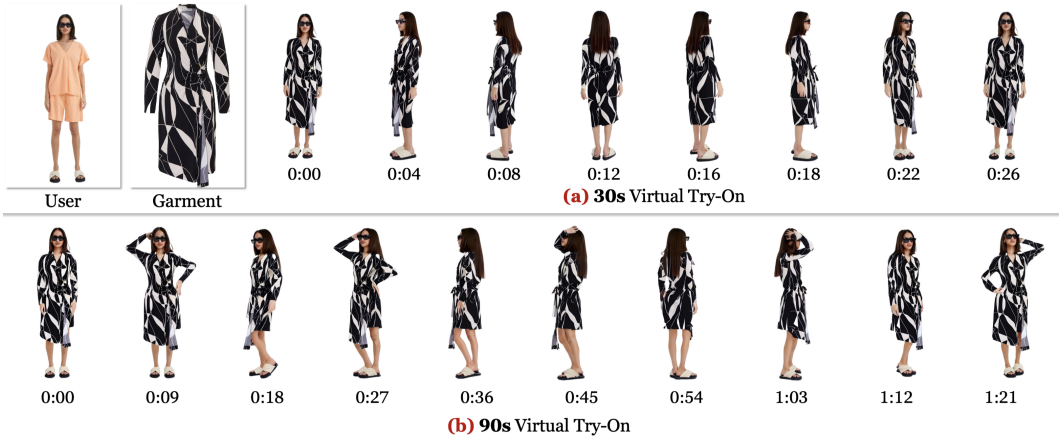

Figure 1: We generate two arbitrarily long videos: **(a)** a 30s video; and **(b)** a 90s video, for a user interacting with a given garment. Our approach preserves accessories – glasses and slippers, and allows a desirable user-garment interaction. Please refer to the project page for full streaming.

generation; and (2) generating multiple short videos that are consistent with the anchor. We observe that a short 360° video like Fig. 3-(a) of the human subject in a simple "A" pose serves as a reasonable anchor, allowing the model to comprehensively design the whole-body appearance of the human. The multiple short video segments are consistent with the anchor video and, therefore, are also consistent with each other. With the anchor video, our generated long video achieves temporal consistency (Fig. 2-(b)) *without* requiring long videos for training.

Evaluating the quality of long video virtual try-on. With the flexibility to generate arbitrarily long videos, we introduce an evaluation protocol to assess performance across four aspects with varying difficulty levels: (1) **360° Garment Consistency** – a 360° video of a stationary human subject in "A" pose, which allows us to study the quality of the generated garment (Fig. 3-(a)); (2) **30s Human+Garment Consistency** – a 30s video of a human subject casually moving around a point in front of a stationary camera, which enables us to assess the quality of the generated human and garment (Fig. 1-(a)); (3) **Hand-Body Interaction Faithfulness** – a 90s video of a human subject performing a fixed set of poses in front of a stationary camera, which facilitates the evaluation of the robustness of the virtual try-on method in controlled settings (Fig. 1-(b)); and (4) **Capability for Arbitrary Poses** - a 30∼60s video of a human subject freely interacting with their body, which allows us to investigate robustness in various poses and orientations. We believe that this evaluation protocol will enable us to comprehensively assess the quality of virtual try-on methods.

***Free* viewpoint rendering is for *free*.** A by-product of learning temporally consistent video is that we can render a human subject in any pose and viewpoint. Fig. 3-(a) shows a generated 360° anchor video ("A" pose) of a user wearing the target garment. The output is 3D consistent, enabling us to reconstruct it into a 3D mesh, as visualized in Fig. 3-(b) in a NeRFStudio [15] viewer. We observe that 3D implicitly emerges while enforcing temporal consistency. Interestingly, we can *recloth* and *reanimate* any human subject from a single image, and view them from any viewpoint.

Our Contributions. (1) We introduce VFR, the first method to generate arbitrarily long, high-resolution (720×1152 resolution at 24FPS) human videos of virtual try-on from a single image. (2) We also introduce an evaluation protocol to assess the overall quality of virtual try-on methods. (3) We observe that the proposed method implicitly learns 3D consistency, enabling us to perform free-viewpoint rendering. Though designed for long video virtual try-on, our approach also provides insights for general long video generation, where the anchor mechanism offers a simple way to maintain global consistency over time.

2 Related Work

We believe *static 2D imagery isn't enough for a realistic virtual try-on experience*. The challenge of achieving high-resolution virtual try-on escalates as we progress from images to videos, and ultimately to 4D. This progression not only heightens the demand for computational resources, but

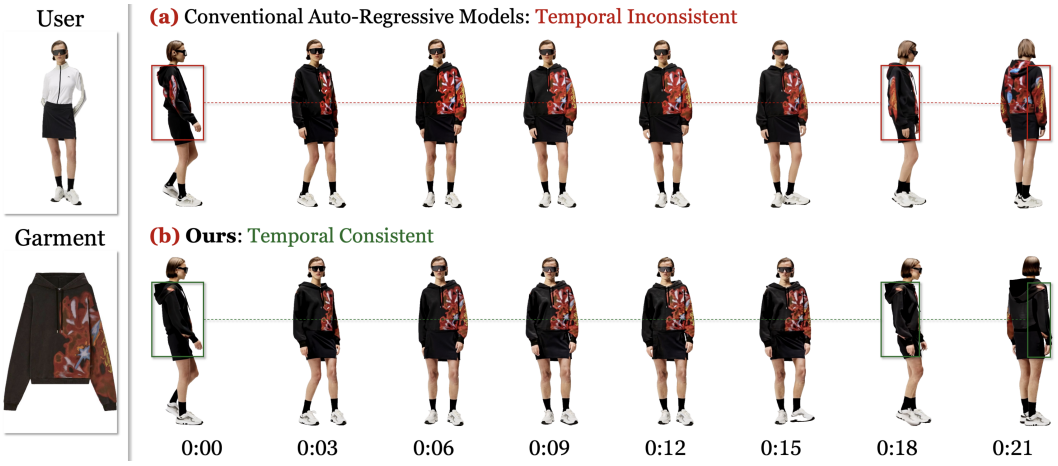

Figure 2: Given a user and target garment, **(a)** conventional auto-regressive video generators suffer from temporal inconsistency issues between distant frames. Note different patterns of sleeves in red bounding boxes across the time. **(b)** Our VFR generates temporally consistent try-on videos.

(a) 360°-View Anchor Video **(b)** Reconstructed 3D Human

Figure 3: **(a)** VFR produces a 360° anchor video ("A" pose) of a user for a given garment. We observe that the outputs are 3D-consistent, allowing us to **(b)** reconstruct it into a 3D human mesh.

also diminishes the availability of previous extensive databases necessary for learning. Our goal is to identify computational methods that enable the generation of arbitrarily long videos, even with constrained resources.

Image-to-Image Try-on. Given an image of the user and a reference to the target garment, the goal here is to synthesize a new image of the user wearing the target garment [16–37, 4, 38–40]. Most methods take a two-step approach: (1) deform the garment for a given user, known as warping; and (2) generate a new image with the deformed garment using GAN [41] or latent diffusion models [42–44]. A few exceptions [20, 33, 45] generate the output without an intermediate warping step. The primary limitation in this field is the restricted experience a user can have with garments. One can only see them in exactly the same pose as in the input image. This issue can potentially be addressed by incorporating an additional module that can synthesize humans in different poses [46–51]. However, due to error propagation, the garment's appearance becomes inconsistent. Consequently, various methods [52–56] aim to jointly change the pose and the garment. These methods, however, lack temporal consistency and smoothness when applied across a series of poses.

Video-to-Video Try-on. Given a video of the user and a reference to the target garment, the objective here is to synthesize a new video of the user wearing the target garment [57–62]. An important distinction is ensuring that the synthesized garments and humans are temporally consistent and accurate. He et al. [63] employs an image-to-image try-on methodology, but incorporates an additional temporal loss during training to enforce consistency. Recent methods [7, 8] can generate high-resolution output, but they are limited by the duration of the generated video ($5s$). Naively increasing the duration of videos will require enormous computational resources.

Image-to-Video Try-on. In this work, we explore the generation of arbitrarily long videos from a single user image and garment images. A naive approach is to utilize a single image try-on method and animate it using an image-to-video creation module [64–76]. However, a modular approach leads to error propagation, which degrades the quality of the generated outputs. Therefore, we seek an end-to-end video generation pipeline that allows us to preserve the details of the garment [8, 77]. We observe that text conditioning cannot effectively capture long and subtle movements [8]. Instead, we use example videos as a reference to guide the creation of new videos. A notable prior work,

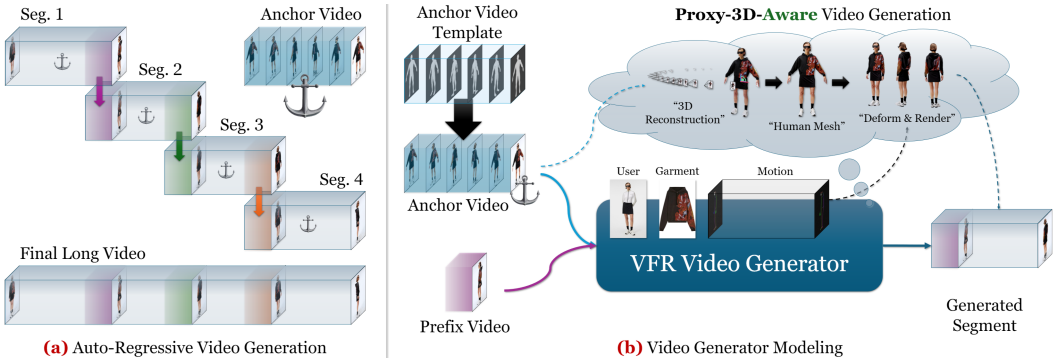

Figure 4: **(a)** Our VFR is an auto-regressive framework that generates a long video segment-by-segment. **(b)** The video generator takes both an anchor video and a prefix video as input, and generates a new segment that continues the prefix video while maintaining consistency with the anchor video.

Dress&Dance (DnD) [8], generates 5s videos with high resolution 720×1152 at 24FPS. In this work, we generate arbitrarily long videos from a single user image and the reference garments.

Long Video Generation. There is also a line of work [9–14] that investigates long video generation for text-to-video and image-to-video tasks. The common idea behind these methods is to introduce an additional "memory back" or "history tracking" mechanism to ensure consistency with the previous frames in a typical auto-regressive generation process. For example, a concurrent work, FramePack [14] designs a computationally efficient way to consider all the previous frames as conditions when generating a new frame. Similarly, StarGen [78] proposes a spatio-temporal autoregression framework for static scene synthesis, while LCT [79] introduces a long-context tuning strategy that leverages interleaved 3D positional embeddings and KV-cache mechanisms to extend temporal receptive fields. These approaches can generate smooth long videos; however, their temporal consistency is not guaranteed and often violated due to ineffective history tracking or over-compressed memory. In contrast, our VFR method tackles this problem by designing an anchor video conditioning mechanism tailored for virtual try-on tasks, providing a global appearance prior that promotes strong temporal consistency across the entire long video.

Free Viewpoint Rendering. Our ability to generate arbitrarily long videos implicitly allows us to perform free-viewpoint rendering of humans [80, 81]. We observe that a model trained to capture consistent temporal characteristics implicitly learns 3D consistency in its outputs. This analysis could potentially pave the way for 4D try-ons in the future.

3 VFR: Methodology

Given a user image, a reference garment image, an optional text prompt, and a long motion reference video, our VFR is a method that generates long, minute-scale, high-quality try-on videos of the user wearing the desired garment and performing the indicated motion, in an auto-regressive, segment-by-segment manner. In an auto-regressive generation framework, the core challenge is to achieve both (1) *local smoothness* such that the video transitions seamlessly without noticeable sudden changes or morphing (Fig. 5); and (2) *global temporal consistency* such that the appearance of both the user and the garments at all occurrences in the video is the same (Fig. 2).

To address these crucial challenges, our key insight is to (1) propose the "anchor video" generation to ensure a consistent appearance throughout the entire video, and (2) introduce the video prefix condition and immediate refiner to enhance video smoothness through strong conditioning. With both insights, our VFR achieves high-quality long try-on videos, while ensuring both smoothness and temporal consistency.

Auto-Regressive Video Generation. There are two approaches to generating long videos: one can either produce the entire long video in a single forward pass, just like generating a short video, or generate the video auto-regressively as multiple shorter segments (Fig. 4-(a)). The former method necessitates significant computational resources during the generation process, as well as extensive long video datasets; whereas the latter requires similar computational resources and training datasets to those used for short video generation.

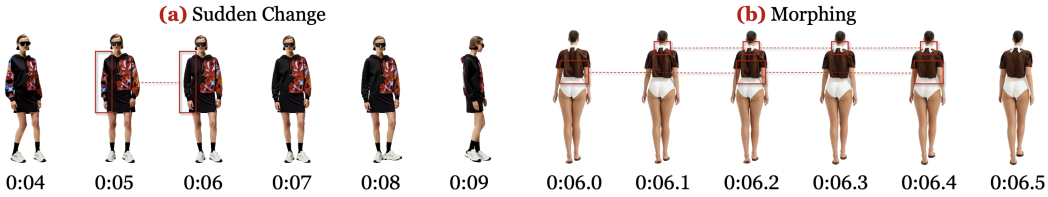

Figure 5: Without our prefix conditioning, the generated video may contain artifacts like **(a)** sudden changes or **(b)** morphing, as highlighted in the boxes, which violate the smoothness requirements.

We opt for the second approach and design VFR as an *auto-regressive* generative model $v \leftarrow G(c, v_{\text{prefix}})$. As seen in Fig. 4-(b), our model generates a short F-frame video v with input conditions c, such that the generated video has v_{prefix} as its prefix, and then seamlessly continues into the preceding frames. In long video generation, we first generate the initial F frames $v_0 \leftarrow G(c_0, \varnothing)$, and then generate each $(F-f)$-frame segment $v_i \leftarrow G\left(c_i, v_{i-1}^{[F-f+1,F]}\right)$, using the f overlapped frames $v_{i-1}^{[F-f+1,F]}$ from the previous segment as its prefix. The reason for using f frames instead of just one frame is to ensure that both the visual appearance and the motion, which requires multiple frames to represent, are smoothly continued into the next segment v_i. To obtain the different conditions c_i for each segment v_i, we only need to replace the full motion reference video with its corresponding segment while keeping all other conditions the same. The final generated video is $v = v_0 + \text{Concat}_{i=1}^{n} v_i^{[f+1,F]}$ (Fig. 4-(a)). Please refer to the Supplementary for the model details.

3.1 Anchor Video for Temporal Consistency

The issue of temporal consistency arises from the ambiguity of the unseen aspects of the garment or the user's appearance. For instance, given only a frontal image of the garment, its back and side views can be interpreted in multiple, yet plausible, ways. This leads to temporal inconsistencies when different appearances are generated in two non-overlapping segments, as shown in Fig. 2-(a). To ensure consistency, a naive approach is to use the prefix of the whole video $v_i^{\text{pre}} = v_0 + \text{Concat}_{j=1}^{i-1} v_j^{[f+1,F]}$ as part of the condition c_i to generate the next segment v_i, such that the generated v_i remains consistent with any frame that has appeared in v_i^{pre}. However, this approach negates all the advantages of the auto-regressive generative framework – it also incurs significant computational costs and requires long videos for training to effectively handle v_i^{pre}.

Our insight is to organize the appearance information of garment and user in a more effective representation than the whole v_i^{pre}. Therefore, we introduce the concept of the anchor video v_a, which is a short F_a-frame video of the user wearing the garment while performing a different motion, ensuring that any part of the user and the garment is visible in at least one of its frames. Given a motion template m_t, the generation of the anchor video can be modeled as $v_a \leftarrow G(c|_{\text{motion}=m_t}, \varnothing)$, i.e., generating a virtual try-on video with fixed motion m_t.

We choose m_t to be a stationary A-pose like Fig. 3-(a), with both arms slightly raised near the sides of the waist and legs naturally apart, under a 360°-view camera trajectory. As shown in Fig. 3-(b), such an anchor video generated from this m_t is already sufficient for reconstructing a 3D mesh representation with NeRFStudio's NeRFacto model [15], indicating that this video effectively captures a 3D-consistent appearance of the try-on human video. Conditioning the generator on this video anchor is sufficient to promote consistency across poses and viewpoints, ie., a proxy-3D-aware video generation, without any explicit 3D representation [82] or 3D motion modeling like SMPL [83]. Since consistency is an equivalence relationship, each v_i's consistency with v_a implies their consistency with one another. As a result, we achieve cross-segment global appearance consistency, which guarantees the temporal consistency of the final generated long video, as shown in Fig. 2-(b).

3.2 Auto-Regressive Prefix Conditioning for Smoothness

In our auto-regressive generation, we aim to generate the next segment v_i, such that its first f frames $v_i^{[1,f]}$ precisely match the given v_{prefix}. If we generate an arbitrary video v_i and then replace its first f frames, there may be a sudden change between the f-th and $(f+1)$-th frame, resulting in a discontinuous video. A straightforward approach is to model this task as "outpainting" the remaining

Figure 6: In the virtual try-on with the 90s hand-body interaction motion, our VFR generates temporally consistent try-on videos. Please refer to the project page for full streaming.

$(F - f)$ frames, given the first f frames. This allows us to apply a training-free inpainting/outpainting method, DiffEdit [84], leveraging the same video diffusion model to generate a visually smooth video that starts from the given prefix. However, such a generated video may still contain sudden changes (Fig. 5-(a)) and morphing between different appearances (Fig. 5-(b)).

To enforce the video diffusion model to generate smooth videos, we must explicitly consider the fixed prefix v_{prefix} during the generation process. Therefore, as demonstrated in Fig. 4-(b), we introduce the prefix as a conditioning input. This approach ensures that the model can learn to generate a smooth video segment that seamlessly continues the input prefix.

3.3 Training Long Video Generator with Short Video Data

The introduction of auto-regressive generation focuses on each F frame segment of the video, which only requires videos containing at least F frames for training. This approach eliminates the need for large-scale, lengthy video training data, which is challenging to obtain, especially for our virtual try-on videos. Each single generation can be modeled as $v \leftarrow G(c, v_{\text{prefix}})$, where $c = (c_{\text{try-on}}, v_a)$ represents both try-on conditions (garment and user image, motion reference video, etc.) and the anchor video. For each F-frame segment v in our dataset, we can set $v_{\text{prefix}} \leftarrow v^{[1,f]}$ to obtain the corresponding input. However, in most cases, we do not have the desired anchor video v_a that is exactly at the m_t we want. Generating m_t by ourselves may also be inconsistent with the video, e.g., with different back or side views.

To address this issue, our approach does not focus on the v_a at a specific m_t, but instead allows the model to learn to leverage v_a's under an *arbitrary* motion m_t. This ensures that, regardless of the pose of the v_a, the video generator can always optimize its usage to produce videos that are consistent with the given v_a. With this modeling, *any* video segment showing the same person wearing the same garment can be used as an anchor video for the target video v. More specifically, for each F_{raw}-frame raw video data, we can sample two intervals – with the same user and garments – to construct a training batch: a F-frame interval as target v, and another *arbitrary length* interval to temporally upsample or downsample it to F_a frames and use as v_a. Utilizing this training strategy, we can effectively train the model with only short ($F \sim 3F$) raw videos, which is not significantly longer than F, yet can already yield a sufficient variety of different pairs of v and v_a.

Figure 7: In the virtual try-on with a ∼50s arbitrary motion, our VFR faithfully preserves consistent garment details and human appearance, showcasing various poses with high quality. These results are shown as videos on our project page.

3.4 Immediate Refinement of the Generated Segments

The robustness of auto-regressive video generation is significantly more critical than that of end-to-end generation, as artifacts and errors can accumulate and amplify over time. Specifically, given that both anchor and prefix video conditions are complex, pixel-unaligned video conditions, we observe increasing artifacts in our initial generation results, which may accumulate to severe degradation.

Inspired by Dress&Dance [8], which leverages a standalone generator as a "refiner" to temporally upsample the video and remove artifacts, we also design a refiner for our generation. In our case, we do not aim to upsample the videos; instead, we focus entirely on artifact removal and video smoothing. Our refiner can be modeled as $v' \leftarrow G_{\text{refiner}}(c, v_{\text{prefix}}, v)$, which takes a F-frame video v as input, along with all the other conditions c, including v_a, obtaining a refined F-frame video v'. Such a refiner is directly applied to each v_i, immediately after its initial generation, so we call it an "immediate refiner." Additionally, since the input v and output v' have the same number of frames, we apply the refiner G_{refiner} multiple times to the same video to ensure complete artifact removal. The final output segment v_i can be expressed as $v_i \leftarrow G_{\text{refiner}}^{\circ k}\left(c, v_{i-1}^{[F-f+1,F]}, G\left(c, v_{i-1}^{[F-f+1,F]}\right)\right)$, where k is the number of iterations. With this immediate refinement technique, we resolve the issue of degradation over time in long video generation, maintaining high quality from beginning to end.

4 Experiments

4.1 Experimental Settings

Model Training Settings. Our VFR model is built on Dress&Dance [8] with the addition of "prefix video" and "anchor video" CondNets. We train VFR on both Internet and captured datasets from Dress&Dance for $10,000$ iterations. Specifically, the immediate refiner is initialized from the $5,000$-th iteration checkpoint of the base VFR, and trained for another $5,000$ iterations.

Evaluation Tasks. We evaluate VFR on four different parts: (1) **360° Garment Consistency**, to generate a 5s 360°-view "A" pose video, which also serves as the anchor videos for the other tasks; (2) **360° Human+Garment Consistency**, to generate a 30s 360° casually moving video; (3) **360° Hand-Body Interaction Faithfulness**, to generate a 90s video with a fixed motion; and (4) **Capability for Arbitrary Poses**, to generate a 30∼60s video with arbitrary motion. These benchmarks are available on our project page. Notably, existing public video try-on datasets VVT [85] and ViViD [61] are not suitable for evaluating our setting, as their spatial resolutions (256×192 and 832×624, respectively) and temporal lengths (typically only a few seconds) are insufficient to capture the high-resolution, minute-scale, and temporally consistent generation capabilities targeted by our benchmark.

Baselines. We compare our VFR against the baselines **FramePack** [14], utilizing an "image virtual try-on + image-to-video animation" (IVT+I2V) procedure that is aligned with Dress&Dance [8]. We also compare with **Kling Video 2.0** [86] in a "repeating image-to-video" (RI2V) manner mentioned in [14]. We are unable to compare with the following baselines: StreamT2V [13], as it only supports 16:9 landscape videos; CausVid [12], given that there is no available image-to-video checkpoint released; TTT [9], since it only supports Tom and Jerry videos; and DiffusionForcing [10] and HistoryGuidance [11], as they are restricted to videos from their respective trained datasets. As for the image try-on method, we mainly use the two state-of-the-art method, **Dress&Dance** image try-on [8] and **Kling Try-On** [86], to generate the first frame of the video. More combinations including other image virtual try-on methods are compared in our Supplementary.

Ablation studies. We also compare our full VFR with the following variants: (1) **"No Prefix" (NP)**, which does not use prefix conditioning, but directly utilizes DiffEdit [84] to outpaint the video with the prefix; (2) **"No Anchor" (NA)**, which does not generate and condition the segment generations on the anchor video; (3) **"No Refine" (NR)**, which skips the immediate refinements after each segment generation; (4) **"Dress&Dance" (D&D)**, which does not use either the anchor video or the prefix conditioning, making it equivalent to a training-free method that employs Dress&Dance with DiffEdit for long video generation.

Metrics. Consistent with Dress&Dance [8] and FramePack [14], we utilize GPT [87]-based scores and VBench [88–90] to evaluate our videos. GPT scores can effectively assess the try-on quality from various aspects, leveraging GPT's visual capabilities; VBench introduces a set of metrics that comprehensively evaluate the videos from both quality and semantic perspectives.

4.2 Experimental Results and Analysis

We present our qualitative results in Figs. 6,7,8 as images, and on our project page as videos.

360° Human+Garment Consistency (30s) – Fig. 8. As shown Fig. 8-(a), our VFR produces high-quality, long virtual try-on videos. In Fig. 8-(b), our "No Prefix" (NP) variant, due to the global anchor videos, generates results that are comparable to our full VFR, but exhibits some sudden changes, as illustrated on our project page. In contrast, our "No Anchor" (NA) variant's video displays long-term temporal inconsistencies, while our "Dress&Dance training-free" (D&D) variant exhibits even greater temporal inconsistencies in both the short and long term. This highlights that both designs in our VFR for local smoothness and global temporal consistency are effective and essential. In Fig. 8-(c), the baseline FramePack [14] produces overall smooth results with long-term inconsistencies, while the appearances deviate significantly in the results produced by the Kling Video 2.0 [86]-based RI2V method. This demonstrates that the virtual try-on tasks are non-trivial and challenging, underscoring our contribution in achieving high-quality results.

Hand-Body Interaction Faithfullness (90s) – Fig. 6. The motion in these evaluation tasks encompasses both human rotation and arm movements. Our VFR faithfully performs the same motion in different virtual try-on tasks, demonstrating the ability to depict the same motion across various garments – either pants, skirts, or dresses. Even for such a long-term video, our VFR still maintains high temporal consistency, which can be observed by comparing the first and the last frame.

Capability for Arbitrary Poses (∼50s) – Fig. 7. In these highly challenging tasks, the motions can be arbitrary, encompassing various arm and leg movements, which lead to even more diverse showcases and interactions between garments and users. We observe that our VFR effectively handles these motions and generates high-quality visualizations to depict them. This shows VFR has the capability to generalize to various long video virtual try-on tasks.

Quantitative Analysis. The quantitative evaluation comparisons are provided in Table 1. We use the "Subject Consistency," "Background Consistency," and "Motion Smoothness" from VBench [88] to evaluate how the two major challenges – temporal consistency and smoothness – are addressed, as well as the GPT metric in [8] to assess the overall virtual try-on quality. We provide the more quantitative evaluations in Supplementary.

As shown in Table 1, in each component of the evaluation protocols, our full VFR consistently outperforms all the baselines and variants. Furthermore, since our VFR is based on the previous work Dress&Dance [8], all these variants maintain a portion of its virtual try-on capability, achieving

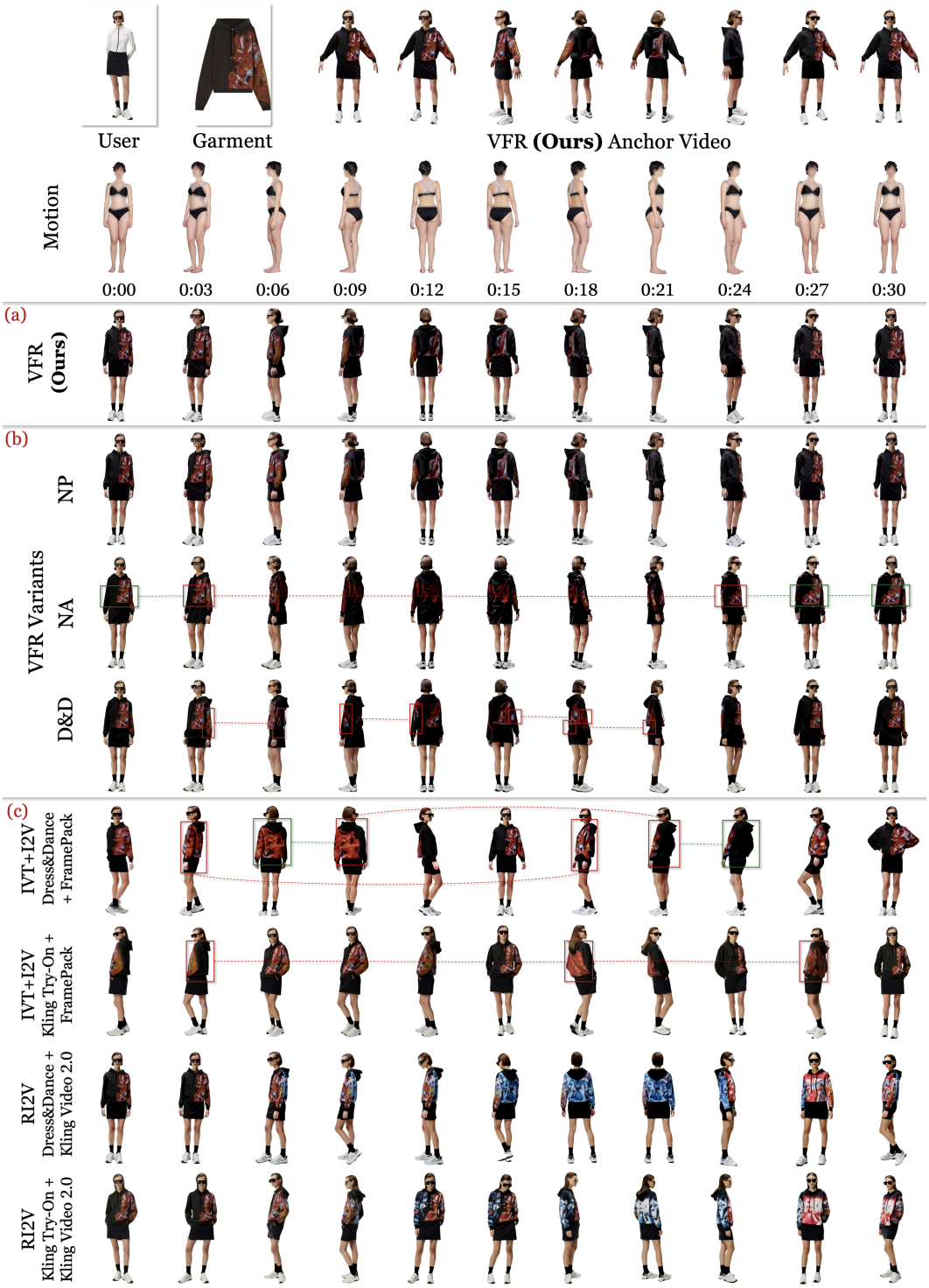

Figure 8: In the virtual try-on with 360° dynamic motion, (a) our VFR generates high quality, long virtual try-on videos, while (b) removing either anchor video or prefix conditioning results in noticeable degradations. On the contrary, (c) the baselines suffer from smoothness and temporal consistency issues. Please refer to the project page for full streaming.

comparable GPT evaluation scores, where the D&D variant, as a training-free long video generation pipeline utilizing D&D, preserves the majority of its capability.

We also observe that our NP variant performs similar to our full model, reflecting the strong control provided by the anchor video. However, without the anchor video the NA and D&D variants suffer a

Method		Subject Consistency	Background Consistency	Motion Smoothness	$GPT_{Try\text{-}On}$ ↑	GPT_{User} ↑	GPT_{Motion} ↑	GPT_{Visual} ↑	$GPT_{Overall}$ ↑
1. 360° Garment Consistency (5s)									
Ours		92.84	95.38	98.35	87.83	85.90	76.67	80.35	82.06
2. 360° Human+Garment Consistency (30s)									
Ours		**94.06**	**96.53**	**99.37**	**90.09**	**88.11**	84.08	**86.33**	**87.14**
Ours	NP	93.58	96.10	99.31	89.66	87.24	**84.20**	85.86	86.69
Ours	NA	92.77	95.55	99.22	88.47	84.91	81.75	81.66	84.04
Ours	D&D	93.70	96.01	99.23	89.72	86.16	83.20	86.05	86.10
Ours	NR	90.80	94.82	99.21	87.29	85.55	78.74	71.90	80.40
3. Hand-Body Interaction Faithfulness (90s)									
Ours		**92.13**	**94.02**	**99.24**	87.20	84.88	77.85	81.12	82.62
Ours	NP	91.88	93.91	99.15	89.65	86.28	80.15	83.22	84.62
Ours	NA	88.00	91.50	99.09	85.20	80.35	70.67	69.05	75.70
Ours	D&D	91.65	93.25	99.11	**89.95**	**87.25**	**82.53**	**85.15**	**86.15**
Ours	NR	86.02	89.84	99.01	85.22	84.83	72.25	65.10	75.03
4-Med. Capability for Arbitrary Poses – Medium (25s)									
Ours		**91.09**	**93.82**	**98.62**	87.11	84.45	77.83	79.77	81.93
Ours	NP	90.19	93.00	97.85	87.40	**85.48**	78.40	80.55	82.56
Ours	NA	88.46	92.10	97.92	86.87	84.39	77.92	77.56	81.31
Ours	D&D	89.27	92.81	97.83	**87.83**	85.23	**80.24**	**82.22**	**83.61**
Ours	NR	87.56	91.79	97.76	85.28	83.75	74.85	68.98	77.26
4-Hard. Capability for Arbitrary Poses – Hard (30-50s)									
Ours		**93.21**	**95.29**	**99.35**	87.03	85.83	69.99	78.84	79.05
Ours	NP	92.71	94.63	99.29	87.69	**86.38**	71.12	79.47	79.60
Ours	NA	91.17	93.31	99.22	87.14	85.03	71.42	77.11	78.99
Ours	D&D	92.27	94.09	99.25	**88.05**	84.97	71.90	**81.38**	80.39
Ours	NR	88.17	92.08	99.19	85.36	84.61	67.45	64.55	73.44
IVT+I2V	D&D + FramePack [14]	88.33	90.98	98.24	86.96	85.35	**77.51**	79.33	**81.86**
RI2V	D&D + Kling [86]	92.71	94.08	98.92	87.25	85.59	65.80	79.40	77.24

Table 1: Quantitative experiments show that our VFR consistently outperforms all variants and baselines in both consistency and smoothness metrics from VBench [88], while achieving comparable try-on and visual quality as Dress&Dance (D&D) [8]. `Notations:` NP is No Prefix; NA is No Anchor; and NR is No Refine.

significant drop in both (human) subject and background consistencies, and the degradation of virtual try-on quality even occurs for the NA variant, as reflected in the GPT evaluation metric. This shows that the anchor video controls not only the consistent appearance but also the try-on quality.

Finally, we compare the most powerful baselines, FramePack [14] and Kling [86], combined with the state-of-the-art virtual try-on method Dress&Dance. FramePack has a significantly lower consistency metric, indicating that its conditioning method, which takes into account the previous frames, is not sufficient to enforce consistency. Kling still achieves a slightly lower consistency metric compared to our VFR, but the quality of the virtual try-on degrades.

5 Conclusion

We propose VFR, a virtual try-on method that generates arbitrarily long, high-resolution videos from a single user image, garment, and motion reference. The key of our method is an anchor video-guided framework that ensures temporal consistency across segments and implicitly captures 3D structure, enabling free-viewpoint rendering without 3D supervision. Along with the prefix conditioning, we achieve both local smoothness and global temporal consistency in the long video generation. We further introduce a new evaluation protocol tailored to long video virtual try-on, covering garment fidelity, user appearance, and motion robustness. Experiments show that VFR produces realistic, smooth, temporally consistent, and garment-faithful results that significantly surpass the capabilities of prior methods. We believe that VFR opens up new avenues for interactive, personalized virtual try-on experiences—whether in e-commerce, virtual social platforms, or creative content generation.

Discussions. We made progress in generating arbitrarily long, high-resolution videos for virtual try-on. Our preliminary analysis shows that additional video data can help us further improve the quality of the generated videos. Secondly, while this work is the first of its kind, it takes 1$\sim$2 hours to generate a 30s video, which is not efficient enough. We leave this speed-up as an interesting future work. Finally, we believe that our work will pave the way for the transition from long videos to arbitrary 4D content, where a user can both change camera perspective and motion.

Potential Societal Impacts. The positive societal impacts of our VFR may include (1) revolutionizing the online shopping experience for clothing, (2) decreasing returns and replacements of clothes through improved pre-sale understanding, and (3) leading to an increase in both the number and volume of online clothing shops. On the other hand, VFR is inherently a model that produces human videos, and also brings the risks to produce biased, unethical, or unsafe results.

Acknowledgement: We thank Pixophilia Technologies Private Limited, India, and VietTechTools Co., LTD, Vietnam, for generously providing the visual data for our benchmark.

References

[1] Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D4-VTON: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *ECCV*, pages 36–52. Springer, 2024. 1

[2] Zhenyu Xiel, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. GP-VTON: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, pages 23550–23559, 06 2023.

[3] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on, 2024.

[4] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024. 3

[5] Rui Wang, Hailong Guo, Jiaming Liu, Huaxia Li, Haibo Zhao, Xu Tang, Yao Hu, Hao Tang, and Peipei Li. Stablegarment: Garment-centric generation via stable diffusion. *arXiv preprint arXiv:2403.10783*, 2024.

[6] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. 1

[7] Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. Fashion-vdm: Video diffusion model for virtual try-on. In *Proceedings of ACM SIGGRAPH Asia 2024*, December 2024. 1, 3

[8] Jun-Kun Chen, Aayush Bansal, Minh Phuoc Vo, and Yu-Xiong Wang. Dress&Dance: Dress up and dance as you like it - technical preview, 2025. URL https://arxiv.org/abs/2508.21070. 1, 3, 4, 7, 8, 10, 20, 22, 23

[9] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, Tatsunori Hashimoto, Sanmi Koyejo, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training, 2025. 1, 4, 8

[10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 37, 2025. 8

[11] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. 8

[12] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 8

[13] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. StreamingT2V: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 8

[14] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025. 1, 4, 8, 10

[15] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *SIGGRAPH*, 2023. 2, 5

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 3

[17] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018.

[18] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzenes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *ECCV*, 2020.

[19] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020.

[20] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACMTOG*, 40(4):1–10, 2021. 3

[21] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021.

[22] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021.

[23] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *CVPR*, 2022.

[24] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *ECCV*, 2022.

[25] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *CVPR*, 2022.

[26] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *CVPR*, 2022.

[27] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *ECCV*, 2022.

[28] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *CVPR*, 2022.

[29] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *ICCV*, 2023.

[30] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *ACMMM*, 2023.

[31] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *CVPR*, 2023.

[32] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *CVPR*, 2023.

[33] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *CVPR*, 2023. 3

[34] Kedan Li, Jeffrey Zhang, and David Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *TPAMI*, 45(10):12222–12235, 2023.

[35] Xu Yang, Changxing Ding, Zhibin Hong, Junhao Huang, Jin Tao, and Xiangmin Xu. Texture-preserving diffusion models for high-fidelity virtual try-on. In *CVPR*, 2024.

[36] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *ECCV*, 2024.

[37] Jeffrey Zhang, Kedan Li, Shao-Yu Chang, and David Forsyth. Acdg-vton: Accurate and contained diffusion generation for virtual try-on, 2024. 3

[38] Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In *WACV*, 2024. 3

[39] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. OOTDiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI*, 2025.

[40] Zheng Chong, Xiao Dong, Haoxiang Li, Wenqing Zhang, Hanqing Zhao, Dongmei Jiang, Xiaodan Liang, et al. Catvton: Concatenation is all you need for virtual try-on with diffusion models. In *ICLR*, 2025. 3, 23

[41] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 3

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[43] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3

[45] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *CVPR*, 2024. 3

[46] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, 2017. 3

[47] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018.

[48] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

[49] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018.

[50] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020.

[51] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, 2020. 3

[52] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *ECCV*, 2018. 3

[53] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019.

[54] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *CVPR*, 2020.

[55] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *ICCV*, 2021.

[56] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 3

[57] Gaurav Kuppa, Andrew Jong, Xin Liu, Ziwei Liu, and Teng-Sheng Moh. Shineon: Illuminating design choices for practical video-based virtual clothing try-on. In *WACV*, 2021. 3

[58] Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. MV-TON: Memory-based video virtual try-on network. In *ACMMM*, 2021.

[59] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *CVPR*, 2022.

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

[61] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models, 2024. 7

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

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

[64] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019. 3

[65] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019.

[66] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *CVPR*, 2021.

[67] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

[68] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 2024.

[69] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

[70] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. DreamPose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023.

[71] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

[72] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *ICCV*, 2023.

[73] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[74] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022.

[75] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024.

[76] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 3

[77] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019. 3

[78] Shangjin Zhai, Zhichao Ye, Jialin Liu, Weijian Xie, Jiaqi Hu, Zhen Peng, Hua Xue, Danpeng Chen, Xiaomeng Wang, Lei Yang, Nan Wang, Haomin Liu, and Guofeng Zhang. StarGen: A spatiotemporal autoregression framework with video diffusion model for scalable and controllable scene generation. In *CVPR*, 2025. 4

[79] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation, 2025. URL https://arxiv.org/abs/2503.10589. 4

[80] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 4

[81] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACMTOG*, 2021. 4

[82] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5

[83] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACMTOG*, 34, 2015. 5

[84] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, abs/2210.11427, 2022. 6, 8

[85] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. FW-GAN: Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019. 7

[86] Kling. Kling ai: Next-generation ai creative studio, 2024. URL https://klingai.com/. 8, 10, 23

[87] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

[88] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. 8, 10

[89] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024.

[90] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 8

[91] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 19

[92] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024. 21

[93] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2022. 22

[94] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 23

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. As claimed in Sec. 1, we solved and achieved state-of-the-art in long video virtual try-on.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of VFR in Sec. 5 and in our Supplementary.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper is an empirical work and does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the detailed experiment settings in our Supplementary. Even though our method incorporates inherent randomness, characteristic of typical diffusion models, it consistently yields high-quality results across nearly all random seeds.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case

of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide reference to our code and benchmark on our project page.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide detailed implementation details in our Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to the time-consuming training process, we are unable to perform the training multiple times to calculate the standard variances. We only provide the standard variances of the GPT-based metrics, which contain randomness, in our Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are clearly stated in the implementation details section in our Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

Justification: The research is conducted with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impacts are discussed in Sec. 5.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We will release our VFR with the safeguards used by open-sourced generative models like Stable Diffusion [91].

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We used the dataset from Dress&Dance [8] and have cited it in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper only uses assets from existing papers ([8]), and does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We describe the use of LLMs in our Supplementary. As a summary, we use LLMs to (1) generate the optional text captions for the try-on tasks, based on the other inputs, and (2) evaluate the model's performance with GPT-based VQAScore [92]. Both uses regard LLMs as black boxes, and they are not a part of our technical contributions.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Supplementary Material

A Supplementary Video (SV)

We provide a supplementary video (**SV**) as this YouTube video to visualize our videos and comparisons against baselines. Here is a table of contents of the **SV**:

- **SV**0:17 360° Garment Consistency (Anchor Videos).
- **SV**0:28 360° Human+Garment Consistency (30s).
- **SV**1:27 Hand-Body Interaction Faithfulness (90s).
- **SV**2:57 Capability for Arbitrary Poses – Medium (25s).
- **SV**3:23 Capability for Arbitrary Poses – Hard (30-50s).
- **SV**4:56 Comparison with VFR Variants (Ablation Study).
- **SV**6:49 Comparison with Baselines (speeded-up to the same duration).
- **SV**7:23 3D Reconstruction.

B Implementation Details

B.1 Base Model and Architecture

Our VFR is built on Dress&Dance (D&D) [8], a prior work that generates high-resolution, 5-second virtual try-on videos.

D&D proposes CondNets for unified conditioning. Our VFR introduces two different conditions: an anchor video and a prefix video. These conditions are passed to the diffusion transformer (DiT) [93] model with two additional CondNets.

B.2 Issues in CondNets

While the conditions in conventional D&D are either images or pixel-aligned videos (motion or refinement source), the conditions in VFR include the anchor video (pixel-*unaligned* videos) and the prefix video (pixel-aligned, but with a *shorter* duration). These conditions are significantly more challenging for conditioning – when generating each frame, the generator must consider the entire condition video instead of only the corresponding frame.

We observe that CondNets continue to perform well for the anchor video in an out-of-the-box manner. However, when the prefix video is present, it is common for the continued video to exhibit artifacts, such as floating dust, as demonstrated in the "No Refine" variant in **SV**4:56. This necessitates the use of the immediate refiner to remove such artifacts.

B.3 Training Procedure

We train our VFR on 8× NVIDIA H100 GPUs using a traditional PyTorch Distributed Data Parallel (DDP) framework. As mentioned in the "Experiments" section in the main paper, both the main generative model and the refiner are trained for 10,000 iterations on both Internet and captured datasets from D&D [8].

In order to learn how to remove the specific dusty artifacts caused by the prefix conditions in the refiner, we employ a new data augmentation method. We generate a number of ovals with random sizes, positions, orientations, time intervals, and velocities, and then manually create artifacts in these regions by either increasing or decreasing brightness, or replacing them with other regions to create residual artifacts. With this procedure, our immediate refiner is capable of effectively removing various and aggressive artifacts – which can be referred to as the artifacts in the "No Refine" variant's results in **SV**4:56.

B.4 Generation Setting

The total number of frames F generated by VFR is 41, at a rate of 8 frames per second, resulting in a 5s video. In the generation of non-first segments, we select $f = 9$, which corresponds to a 1s overlap between the current segment and the previous segment. Therefore, after generating the first segment of $F = 41$ frames (5s), we produce $F - f = 32$ new frames (4s) at each of the following segments. The cutting points of the video are at the $(4k + 5)$-th second, *i.e.*, 0:05, 0:09, 0:13, 0:17, *etc*. Without prefix condition, *i.e.*, the "No Prefix (NP)" variant, these positions are most likely to have sudden changes or morphings.

C Additional Experimental Results

C.1 Additional Image Try-On Model for Baselines

We primarily use Dress&Dance [8] image try-on as part of the baseline model for our experiments. Additionally, we also employ Kling try-on [86] and CatVTON [40] to establish some other baselines. As shown in **SV**6:49, although these methods generate reasonable virtual try-on images, the limitation of the animation models (FramePack or Kling in "repeat I2V" mode) still prevents them from achieving temporal consistency.

C.2 Ablation Study: Qualitative Experiments

The qualitative ablation studies are presented in **SV**4:56. From the video, we can clearly observe that the "No Prefix (NP)" variants exhibit highly noticeable sudden changes during the cut, highlighting the necessity of the prefix conditioning. Furthermore, for the "No Anchor (NA)" and "D&D" variants, the performance even degrades over time, emphasizing the importance of the anchor videos to stabilize the overall appearance and quality. Finally, without the immediate refiner, the NR variants show noticeable artifacts that accumulate over time. This demonstrates that the immediate refiner is an effective tool to prevent artifact accumulation and ensure robust long video generation. In practice, we adopt a fixed two-step lightweight refinement strategy, where each refinement round takes approximately 15–20 minutes on a single GPU, resulting in a total refinement time of about 30–40 minutes—roughly 25% of the overall generation time, with the main generation requiring around 80–90 minutes. Overall, the "No Refine (NR)'' variant exhibits mild background artifacts and boundary noise toward the end of long sequences while maintaining strong identity and garment consistency; the refinement effectively mitigates these degradations and enhances the temporal stability of the generated videos.

C.3 Limitations of Metrics

As shown in **SV**4:56, there are indeed notable smoothness and consistency issues present in the variants and baselines, even for the most powerful variant NP. However, these metrics – which report very similar numbers – do not adequately reflect such discrepancies. Additionally, the GPT metric, which evaluates only the try-on quality, are also not effectively assessing our long virtual try-on generation quality, and they do not capture the consistency of the garments.

D 3D Reconstruction

As mentioned in the main paper, the anchor videos generated by VFR are almost 3D consistent, and therefore can be reconstructed into 3D shapes. We perform some of this reconstructions. As shown in **SV**7:23, even if the videos are not perfectly 3D consistent, we can still obtain a noisy yet reasonable 3D shape. Therefore, with advanced 2D-to-3D distillation techniques like SDS [94], one might be able to actually obtain a precised a mesh and use it for rendering and deformation – which is implicitly achieved within our VFR.

Notably, being perfectly 3D consistent is *not* a requirement of VFR. In fact, in the "Methodology" section of the main paper, VFR is trained to be sufficiently robust to support an *arbitrary* video as the anchor video. It is also frequently the case that the generated anchor video has a stationary background while the person is rotating (**SV**0:17) – this kind of video also works well and is even

more desirable, as the background is an integral part of the anchor that needs to remain consistent throughout the video. The 3D reconstruction results in SV7:23 are constructed from the video after background removal.