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

Technical Preview

Jun-Kun Chen¹ Aayush Bansal² Minh Phuoc Vo² Yu-Xiong Wang¹

¹University of Illinois Urbana-Champaign ²SpreeAI

{junkun3,yxw}@illinois.edu {aayush.bansal,minh.vo}@spreeai.com

immortalco.github.io/VirtualFittingRoom

Abstract

We introduce the 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

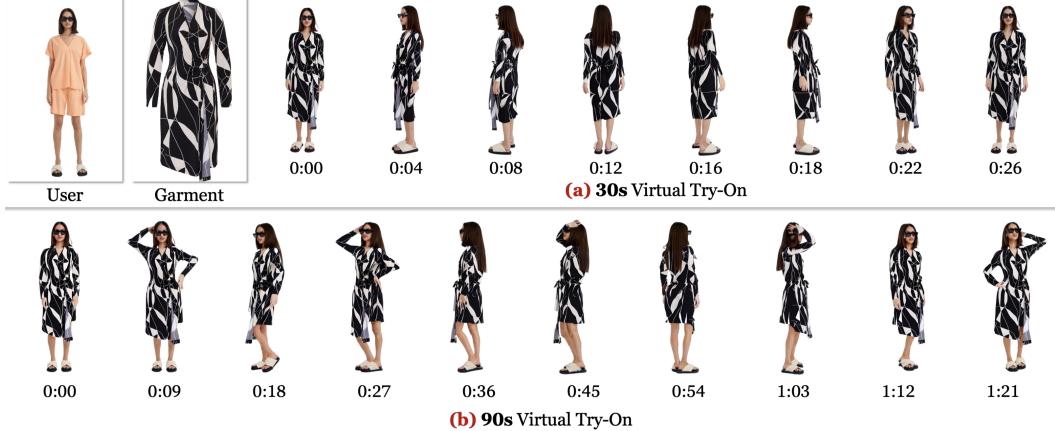


Figure 1: We generate two arbitrarily long videos: **(a)** a 30s video; and **(b)** a 90s video, for a user interacting with a given garment. Our approach preserves accessories – glasses and slippers, and allows a desirable user-garment interaction. Please refer to the [project page](#) for full streaming.

videos is a two-step process that involves: (1) creating an “outline” or an “anchor” to guide the 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) **360° 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. Our observation indicates that 3D *implicitly emerges*

while enforcing temporal consistency. Interestingly, we can *reclot* and *remotion* any human subject from a single image, and view them from any viewpoint.

Our Contributions. (1) We introduce VFR, a method to generate arbitrarily long, high-resolution (720 × 1152 resolution at 24FPS) human videos of virtual try-on from a single image. To our knowledge, no previous work has demonstrated these results. (2) We also introduce an evaluation protocol to assess the overall quality of virtual try-on methods. Finally, (3) we observe that the proposed method implicitly learns 3D consistency, enabling us to perform free viewpoint rendering.

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 also diminishes the availability of previous extensive databases necessary for learning. Our goal is

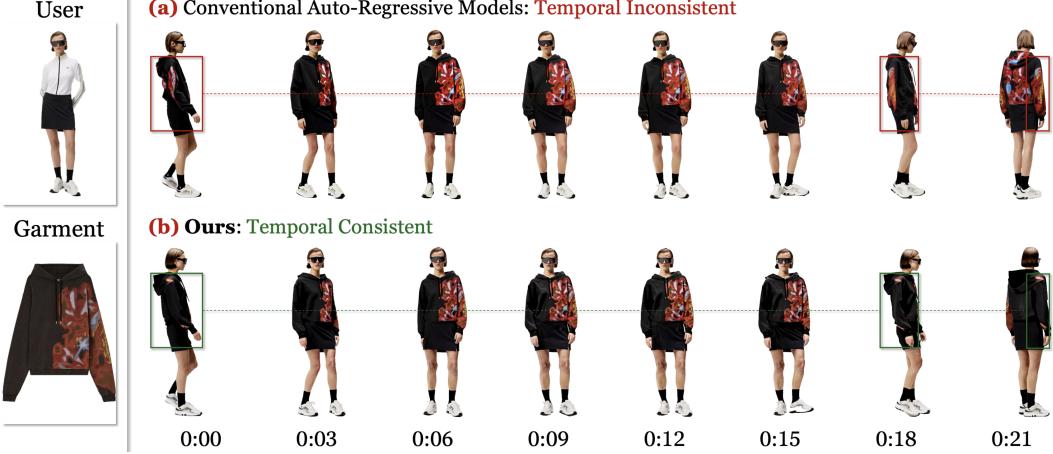


Figure 2: Given a user and target garment, (a) conventional auto-regressive video generators suffer from temporal inconsistency issues between distant frames. Note different patterns of sleeves in red bounding boxes across the time. (b) Our VFR generates temporally consistent try-on videos.



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.

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 [16–41]. 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 [42] or latent diffusion models [43–45]. A few exceptions [20, 33, 46] 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 [47–52]. However, due to error propagation, the garment’s appearance becomes inconsistent. Consequently, various methods [53–57] 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 [58–63]. An important distinction is ensuring that the synthesized garments and humans are temporally consistent and accurate. He et al. [64] 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 [65–77]. 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, 78]. 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, 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.

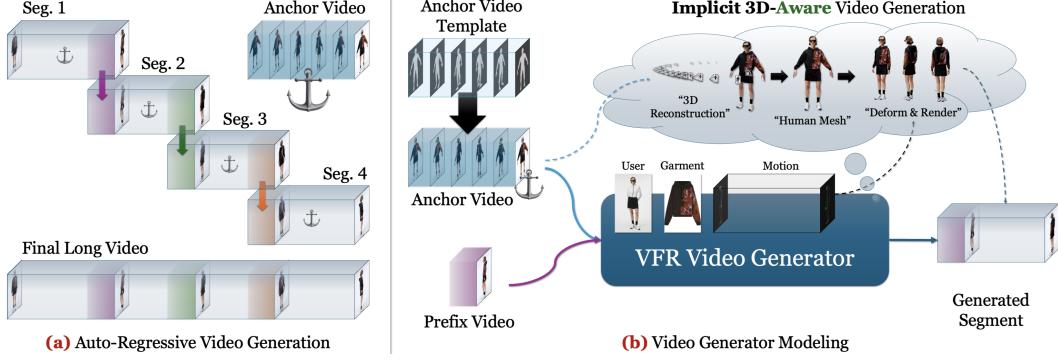


Figure 4: **(a)** Our VFR is an auto-regressive framework that generates a long video segment-by-segment. **(b)** The video generator model takes both an anchor video and a prefix video as input, and generates a new segment that continues the prefix video while maintaining consistency with the anchor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



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.

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. These approaches can also generate *smooth* long videos. However, their *temporal consistency* is not guaranteed and often violated due to ineffective history tracking or over-compressed memory. Our VFR method tackles this problem by designing the anchor video conditioning *tailored* for virtual try-on tasks to promote temporal consistency across the long video.

Free Viewpoint Rendering. Our ability to generate arbitrarily long videos implicitly allows us to perform free-viewpoint rendering of humans [79, 80]. 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 Overview

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.

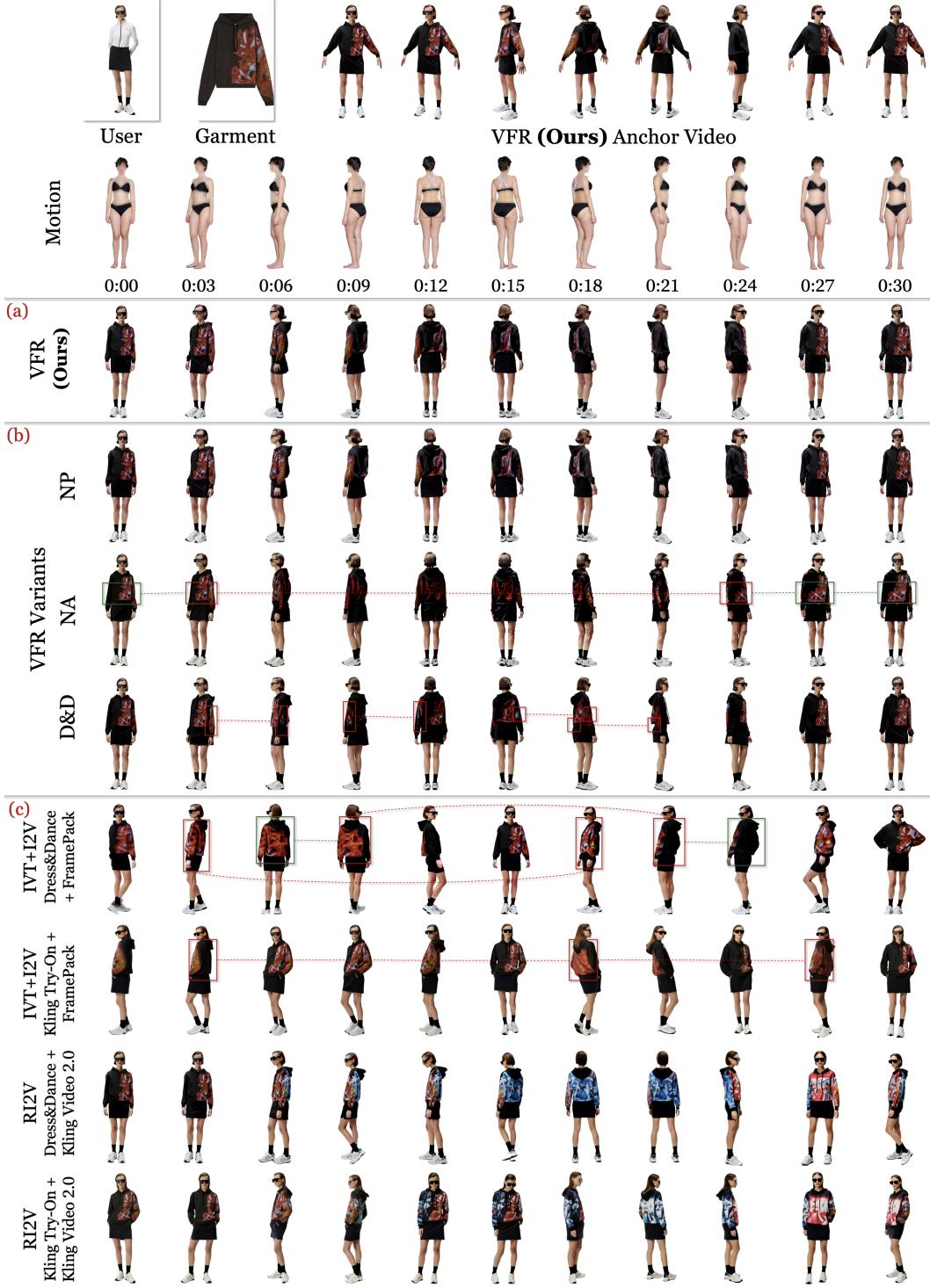


Figure 6: In the virtual try-on with 360° dynamic motion, (a) our VFR generates high quality, long virtual try-on videos, while (b) removing either anchor video or prefix conditioning results in noticeable degradations. On the contrary, (c) the baselines suffer from smoothness and temporal consistency issues. Please refer to the [project page](#) for full streaming.

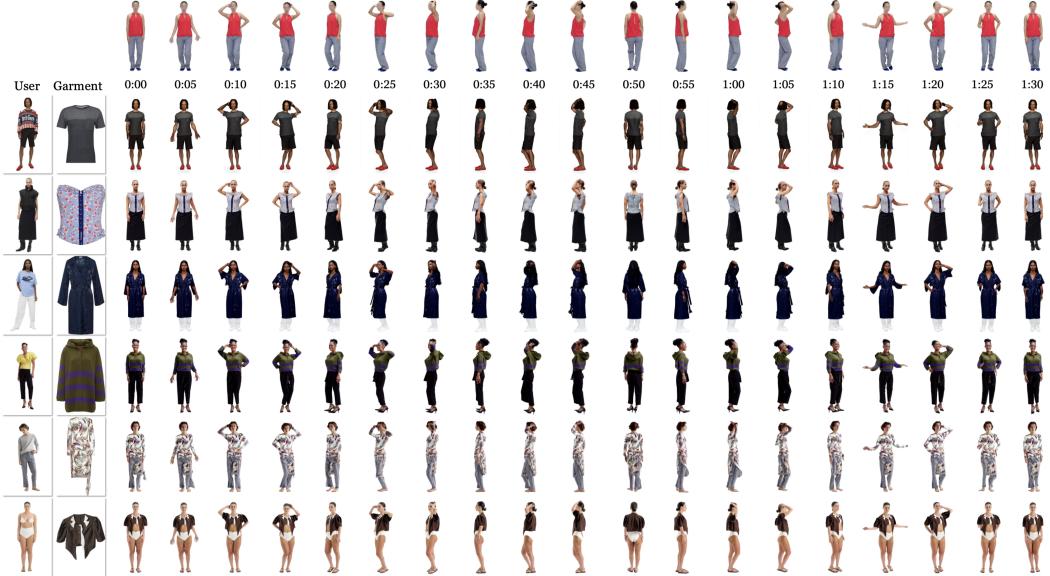


Figure 7: In the virtual try-on with the 90s hand-body interaction motion, our VFR generates temporally consistent try-on videos. Please refer to the [project page](#) for full streaming.

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. As mentioned in Sec. 1, we have 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.

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** [81] 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** [81], to generate the first frame of the video.

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 [82] 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) **“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. (4) **“No Refine” (NR)**, which does not use of the immediate refiner to refine each segment’s output.

Metrics. Consistent with Dress&Dance [8] and FramePack [14], we utilize GPT [83]-based scores and VBench [84–86] 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.



Figure 8: In the virtual try-on with a ~ 50 s arbitrary motion, our VFR faithfully preserves consistent garment details and human appearance, showcasing various poses with high quality. These results are shown as videos in our [project page](#).

4.2 Experimental Results and Analysis

We present our qualitative results in Figs. 6, 7, 8 as images, and in our [project page](#) as videos. We provide the quantitative results in Tab. 1.

360° Human+Garment Consistency (30s) – Fig. 6. As shown Fig. 6-(a), our VFR produces high-quality, long virtual try-on videos. In Fig. 6-(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 in 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. 6-(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 [81]-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 Faithfulness (90s) – Fig. 7. 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 (~ 50 s) – Fig. 8. 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 Experiments. The quantitative evaluation comparisons are provided in Table 1. We use the “Subject Consistency,” “Background Consistency,” and “Motion Smoothness” from VBench [84] 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.

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 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.

Method	Subject Consistency	Background Consistency	Motion Smoothness	GPT _{Try-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+H2V D&D + FramePack [14]	88.33	90.98	98.24	86.96	85.35	77.51	79.33	81.86
RI2V D&D + Kling [81]	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 [84], while achieving comparable try-on and visual quality as Dress&Dance [8].

We further observe that our NP variant achieves performance very similar to that of our full model, reflecting the strong control provided by the anchor video. Without the anchor video, however, the NA and D&D variants suffer a 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 [81], 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 1~2 hours to generate a 30s video, which is not efficient enough to produce long videos in nearly real-time – 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.

References

- [1] Zhaotong Yang, Zicheng Jiang, Xinzhe Li, Huiyu Zhou, Junyu Dong, Huaidong Zhang, and Yong Du. D 4-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. In *European Conference on Computer Vision*, 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 *ArXiV*, pages 23550–23559, 06 2023. doi: 10.1109/CVPR52729.2023.02255.
- [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.
- [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, Innfarm 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] Anonymous Author(s). Dress&Dance: Dress up and dance as you like it. In *Under Review*, January 2025. URL https://anonymous.4open.science/r/Dress_and_Dance_paper. [1, 3, 6, 7, 8](#)
- [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. URL <https://arxiv.org/abs/2504.05298>. [1, 4, 6](#)
- [10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025. [6](#)
- [11] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025. URL <https://arxiv.org/abs/2502.06764>. [6](#)
- [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. [6](#)
- [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. [6](#)
- [14] Lvmin Zhang and Maneesh Agrawala. Packing input frame contexts in next-frame prediction models for video generation. *Arxiv*, 2025. [1, 4, 6, 7, 8](#)
- [15] Matthew Tancik, Ethan Weber, Evonne Ng, Rui long 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](#)
- [16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 619–635. Springer, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7850–7859, 2020.

- [20] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021.
- [22] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022.
- [24] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022.
- [26] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3470–3479, 2022.
- [27] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022.
- [28] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3460–3469, 2022.
- [29] Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, and Alex C Kot. Virtual try-on with pose-garment keypoints guided inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22788–22797, 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 *Proceedings of the 31st ACM international conference on multimedia*, pages 8580–8589, 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23550–23559, 2023.
- [32] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. Linking garment with person via semantically associated landmarks for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17194–17204, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4606–4615, 2023. 3
- [34] Kedan Li, Jeffrey Zhang, and David Forsyth. Povnet: Image-based virtual try-on through accurate warping and residual. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12222–12235, 2023. doi: 10.1109/TPAMI.2023.3283302.
- [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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7017–7026, 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 *European Conference on Computer Vision*, pages 206–235. Springer, 2024.
- [37] Jeffrey Zhang, Kedan Li, Shao-Yu Chang, and David Forsyth. Acdg-vton: Accurate and contained diffusion generation for virtual try-on, 2024. URL <https://arxiv.org/abs/2403.13951>.

- [38] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8176–8185, 2024.
- [39] Kedan Li, Jeffrey Zhang, Shao-Yu Chang, and David Forsyth. Controlling virtual try-on pipeline through rendering policies. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5866–5875, 2024.
- [40] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8996–9004, 2025.
- [41] 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 *The Thirteenth International Conference on Learning Representations*. 3
- [42] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 3
- [43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3
- [44] 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.
- [45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3
- [46] Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1346–1356, 2024. 3
- [47] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017. 3
- [48] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8340–8348, 2018.
- [49] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8857–8866, 2018.
- [50] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018.
- [51] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020.
- [52] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 717–734. Springer, 2020. 3
- [53] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. Swapnet: Image based garment transfer. In *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII 15*, pages 679–695. Springer, 2018. 3
- [54] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10471–10480, 2019.
- [55] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 596–613. Springer, 2020.

- [56] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14638–14647, October 2021.
- [57] 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
- [58] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 191–200, 2021. 3
- [59] Xiaojing Zhong, Zhonghua Wu, Taizhe Tan, Guosheng Lin, and Qingyao Wu. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 908–916, 2021.
- [60] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10799–10808, 2022.
- [61] 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 *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3199–3208, 2024.
- [62] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024.
- [63] Yuanbin Wang, Weilun Dai, Long Chan, Huanyu Zhou, Aixi Zhang, and Si Liu. Gpd-vvto: Preserving garment details in video virtual try-on. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7133–7142, 2024. 3
- [64] 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 *European Conference on Computer Vision*, pages 123–139. Springer, 2024. 3
- [65] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3
- [66] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.
- [67] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5819, June 2021.
- [68] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.
- [69] 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. 2024.
- [70] 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 *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=Fx2SbBgcte>.
- [71] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. 2023.
- [72] 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.
- [73] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. Make-an-animation: Large-scale text-conditional 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15039–15048, 2023.

- [74] 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.
- [75] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022.
- [76] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.
- [77] 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 *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 3
- [78] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019. 3
- [79] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 4
- [80] 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. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 4
- [81] Kling AI. Kling ai: Next-generation ai creative studio, 2024. URL <https://klingai.com/>. 6, 7, 8
- [82] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *ArXiv*, abs/2210.11427, 2022. URL <https://api.semanticscholar.org/CorpusID:253018768>. 6
- [83] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6
- [84] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6, 7, 8
- [85] 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.
- [86] 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. 6