TryOnCrafter: Unleashing Camera Trajectories for Realistic Video Virtual Try-on via a Renderable 4D Try-on Proxy

Hao Sun[1,2,3], Hao Yan[2], Mengting Chen[2†], Quanjian Song[2,4], Yu Li[1,3], Juan Cao[1,3], Jinsong Lan[2], Xiaoyong Zhu[2], Bo Zheng[2], and Sheng Tang[1,3*]

[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] Taobao & Tmall Group of Alibaba
[3] University of Chinese Academy of Sciences
[4] Xiamen University

ts@ict.ac.cn

Fig. 1: Examples synthesized by TryOnCrafter. We introduce a Renderable 4D Try-on Proxy (middle) as a geometric anchor to guide the Video Diffusion Transformer. This explicit 4D representation enables photorealistic try-on across unconstrained, novel camera trajectories (bottom) not present in the source video (top).

Abstract. While Video Virtual Try-on (VVT) has achieved remarkable progress in synthesizing realistic garment overlays on dynamic subjects, existing paradigms remains fundamentally constrained by a passive dependency on source camera trajectories, failing to accommodate the requisite interactive freedom for omnidirectional viewpoint exploration. To address this limitation, we define a pioneering research frontier: Camera-controllable Video Virtual Try-on (CaM-VVT). Unlike con-

† Project leader.
* Corresponding author.

ventional VVT, CaM-VVT not only necessitates viewpoint-agnostic texture hallucination but also strict structural synchronization between non-rigid human dynamics and background contexts under arbitrary, unconstrained camera movements. To tackle these challenges, we present TryOnCrafter, the first unified DiT-based framework specifically architected for the CaM-VVT task. Departing from implicit pixel-space manipulation, we introduce a Renderable 4D Try-on Proxy that explicitly decouples the human subject from the environment. This is achieved by distilling high-fidelity 2D try-on priors into a clothed 3DGS-based avatar, which is subsequently animated via SMPL-X sequences and metric-aligned into a reconstructed background point cloud. This proxy establishes a robust structural foundation with superior texture density and motion integrity. Our Proxy-Anchored Video DiT leverages this robust structural foundation as a primary geometric anchor, ensuring that the synthesized photorealistic videos are strictly constrained by prescribed trajectories and physically plausible deformations. Benefiting from the inherent editability of the 4D proxy, TryOnCrafter facilitates diverse downstream applications, including human relocalization, "bullet time" effects, and 360-degree orbital viewing. Extensive experiments on our established CaM-VVTBench demonstrate that TryOnCrafter significantly outperforms existing baselines in preserving structural consistency and garment identity across complex camera maneuvers.

Keywords: Video Virtual Try-on · Camera Control · 4D Reconstruction

1 Introduction

Video Virtual Try-on (VVT) aims to overlay target garments onto dynamic human subjects videos while preserving spatio-temporal consistency. As a critical engine for immersive "try-before-you-buy" experiences, VVT has become a cornerstone of digital fashion and e-commerce, increasingly serving as a focal point for research. Early studies predominantly rely on image-based warping or frame-wise priors [16, 40, 47]. To further enhance generative stability in intricate environments, recent works [20, 53] integrate powerful DiT-based architectures, pushing the boundaries of VVT toward more diverse and complex scenarios.

Despite this progress, existing VVT research shares a fundamental limitation: **they are confined to the fixed camera trajectory of the input monocular video**. Such limitation conflicts with real try-on needs, where users expect to freely rotate and inspect garments from novel viewpoints to assess details such as side profiles and back appearance. In light of that, this paper introduce a new frontier: Camera-controllable Video Virtual Try-On (**CaM-VVT**). By decoupling synthesis from the input trajectory, CaM-VVT enables interactive, 3D-aware virtual fitting beyond passive video replay. It requires view-consistent garment synthesis from novel viewpoints while maintaining strict geometric alignment with the background under arbitrary camera motion.

A straightforward approach for CaM-VVT is a two-stage pipeline: first perform video virtual try-on, then apply a video-to-video (V2V) camera control

model. However, such sequential design remains suboptimal for digital fashion due to three fundamental hurdles: (i) *Cascaded error accumulation.* Texture inconsistencies from the initial video virtual try-on stage are often amplified by the subsequent V2V camera control models. This is because the control models may not handle the out-of-distribution try-on results. (ii) *Limited garment modeling.* Existing V2V camera control methods lack explicit modeling of human geometry and garment deformation. As a result, directly applying them to try-on videos often fails to preserve the spatio-temporal coherence of garment structure. (iii) *Heavy computational burden.* Both current video virtual try-on and V2V camera control models are heavy and slow at inference. Naively cascading them makes end-to-end inference impractical and can nearly double the runtime.

To overcome these hurdles, we present **TryOnCrafter**, a unified and end-to-end Diffusion Transformer (DiT) framework for CaM-VVT. In detail, TryOnCrafter consists of two core components: (i) **4D Try-on Proxy.** Unlike prior camera-control works [23, 43, 44, 48, 49] that rely on and often fragmented dynamic point cloud to represent the entire 4D scene, our proxy explicitly decouples the human subject from the environment. By distilling high-quality 2D image try-on priors into a clothed 3DGS-based avatar and synchronizing it with SMPL-X motion sequences, we achieve a complete reconstruction of both the human geometry and its articulated dynamics. Compared to sparse and unstable point cloud, this 4D try-on proxy provides significantly more robust and comprehensive garment textures and motion cues, ensuring structural integrity even under radical viewpoint transitions. (ii) **Proxy-Anchored Try-on Video Diffusion Transformer.** Building upon this foundation, we introduce the Proxy-Anchored Video Diffusion Transformer. This model leverages the rendered proxy sequence as a primary structural anchor to synthesize photorealistic try-on videos. By grounding the generative process in explicit 4D guidance, TryOnCrafter ensures that the synthesized results are not only visually stunning but also strictly constrained by physically plausible deformations and target camera trajectories.

To further support our task, we establish **CaM-VVTBench**, a comprehensive evaluation protocol for controllable try-on. Extensive experiments show TryOnCrafter significantly outperforms existing baselines in preserving structural consistency and garment identity. Moreover, TryOnCrafter supports downstream applications, including human relocalization, "bullet time" effects and 360° orbital viewing, paving a new way for interactive digital fashion experiences.

In summary, the contributions of this paper are threefold:

- We propose TryOnCrafter, the first framework to extend video virtual try-on to the Camera-controllable Video Virtual Try-on (CaM-VVT) task, decoupling synthesis from the constraints of the input camera trajectory.
- We introduce the Renderable 4D Try-on Proxy combined with a Proxy-Anchored Video DiT, establishing a pioneering re-rendering-based paradigm that provides robust structural grounding and motion integrity.
- We establish CaM-VVTBench, a comprehensive evaluation protocol for controllable try-on. Extensive experiments demonstrate that TryOnCrafter sig-

nificantly outperforms existing baselines in preserving structural consistency and garment identity under unconstrained camera movements.

2 Related Work

Video Virtual Try-On. Recent advances in video diffusion models have substantially advanced video virtual try-on [8, 10, 16, 25, 32, 40, 47], thereby meeting the e-commerce demand for dynamic visualization. This task aims to transfer a target garment onto a person in a source video while preserving spatio-temporal consistency. Early research primarily relied on image diffusion models to generate videos. For example, ViViD [6] enhances temporal coherence by embedding temporal attention mechanisms and additionally constructs a large-scale VVT dataset to facilitate benchmarking. With the advancement of diffusion transformer architectures, subsequent studies aim at more scalable and expressive generation. CatV^{2}TON [5] directly concatenates garment and video latents for multimodal attention interaction, eliminating the reliance on extra auxiliary networks. Magic-Tryon [20] develops a refined feature injection strategy to maintain detailed garment textures and logos. More recently, DreamVVT [53] extends video try-on task from simple indoor scenes to complex outdoor environments. Despite these advancements, existing methods lack explicit camera modeling, restricting the output to fixed viewpoints. This inherently falls short of the growing virtual fitting demand for dynamic and multi-view try-on experiences. To fill this gap, this work pioneers re-rendering-based paradigm, unifying video try-on and camera-controllable video generation into a single framework.

Camera Control in Video-to-Video Generation. Driven by industry demand, camera-controllable video generation has become a focal point in controllable diffusion models [7, 12, 19, 21, 22, 30, 33–35, 51, 52]. Pioneering methods for camera-controllable video generation [2, 9, 41] typically rely on specific external conditions. To improve efficiency, subsequent approaches [11, 24, 31] have transitioned toward simulating camera motion through a tuning-free paradigm. Recently, some works [1, 23, 43, 44, 48, 49] have integrated camera control into video-to-video translation to achieve perspective redirection, which is most relevant with our task. ReCamMaster [1] injects extrinsic embeddings to enable perspective translation, but the absence of explicit geometry often leads to physical inconsistencies. While TrajectoryCrafter [48] and CAT4D [43] explicitly reconstruct dynamic 3D scenes using point clouds to ensure physical consistency, it fails to decouple camera motion from human motion. In contrast, our TryOnCrafter is the first to unify virtual try-on, human motion, and camera trajectory within a single video generation framework, which is unexplored in prior works.

3 Method

The objective of TryOnCrafter is to synthesize high-fidelity virtual try-on videos across arbitrary camera trajectories, ensuring the preservation of intricate garment textures and robust temporal coherence. As illustrated in Fig. 2, our frame-

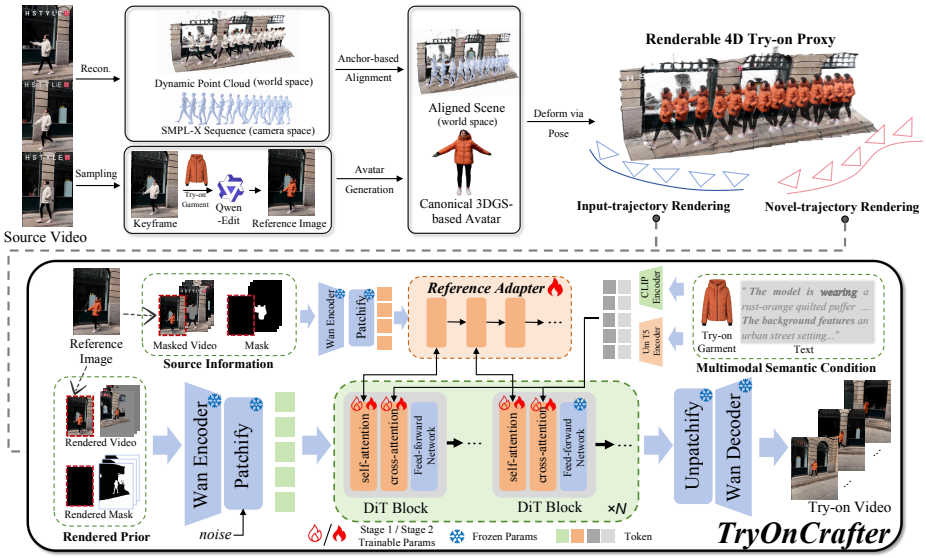


Fig. 2: Overview of TryOnCrafter. *Top*: 4D Try-on Proxy Construction. A metric-aligned scene in world space is established via anchor-based alignment of dynamic point clouds and SMPL-X sequences. Integrating a canonical 3DGS-based avatar distilled from reference image, the resulting 4D Try-on Proxy supports interactive editing and rendering across arbitrary camera trajectories. *Bottom*: Proxy-Anchored Try-on Video Generation. The try-on video is synthesized by the DiT under the integrated guidance of rendered priors, cross-view source features, and multimodal semantic conditions.

work comprises two primary stages: 4D Try-on Proxy Construction (Sec. 3.1) and Proxy-Anchored Video Generation (Sec. 3.2).

3.1 4D Try-on Proxy Construction.

Scene Reconstruction and Anchor-based Alignment. Given the source video $\mathcal{V}^s$, we first reconstruct the dynamic scene from $\mathcal{V}^s$ by estimating dense depth $\mathcal{D}_t$, confidence $\mathcal{W}_t$, and camera parameters $\{\mathbf{K}_t, \mathbf{R}_t, \mathbf{t}_t\}$ via feed-forward MVS frameworks [37–39, 50]. The resulting back-projected global point cloud $\mathcal{P}$ is partitioned into a human component $\mathcal{P}^h$ and a background component $\mathcal{P}^b$ using SAM 2 [28]. Specifically, $\mathcal{P}^h$ captures the human geometry clothed in the source garment, it is subsequently replaced by our high-fidelity, 3DGS-based avatar dressed in the target try-on apparel. Concurrently, $\mathcal{P}^b$ is preserved as the background context to enable seamless rendering within a unified, metric-aligned world space $\mathcal{S}^w$.

Then, we estimate a temporally consistent SMPL-X sequence $\{\Theta_t\}_{t=1}^T$ using GVHMR [29] to decouple the human motion from $\mathcal{V}^s$. A fundamental challenge in our pipeline is the coordinate discrepancy between the parametric SMPL-X sequence $\{\Theta_t\}_{t=1}^T$, estimated in the local camera space $\mathcal{S}^c$, and the dynamic point cloud $\{\mathcal{P}_t\}_{t=1}^T$, reconstructed in the global world space $\mathcal{S}^w$. While these

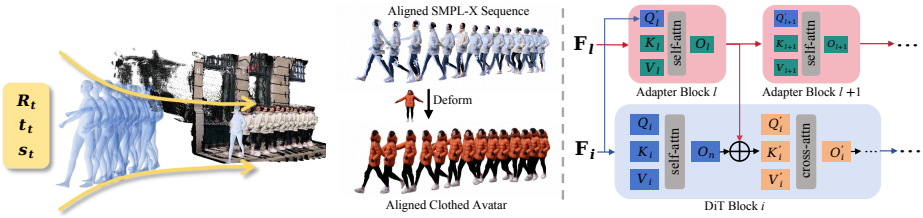
two representations are inherently consistent in their 2D projections, they reside in different metric scales and 3D reference frames. To bridge this gap, as illustrated in Fig. 3(a), we utilize $\{\mathcal{P}_t^h\}$ as a geometric anchor and formulate a robust similarity transformation $\mathcal{T}(\mathbf{v}) = s_t R_t \mathbf{v} + t_t$ to map the SMPL-X vertices $V_t^{smpl} \subset \mathcal{S}^c$ into the world space $\mathcal{S}^w$.

To account for the inherent noise and incompleteness of the monocularly reconstructed $\{\mathcal{P}_t^h\}$, we propose a confidence-aware point-to-surface alignment objective. We minimize the weighted residual between the world-space observation $\mathbf{p}_i$ and the transformed parametric model:

$$\min_{s_t, R_t, t_t} \sum_{i \in \mathcal{P}_t^n} w_i \cdot \left(\min_{v \in V_t^{smpl}} \| s_t R_t \mathbf{v} + t_t - \mathbf{p}_i \|_2 \right) \tag{1}$$

where w_i is a multi-modal confidence weight derived from the MVS depth confidence $\mathcal{W}_t$ and the semantic boundary proximity, ensuring that the optimization prioritizes reliable geometric regions (e.g., the torso) over noisy peripherals.

This process effectively resolves the scale ambiguity inherent in monocular reconstruction and ensures that the subsequent 3DGS avatar is precisely localized within the 3D scene environment, maintaining strict temporal and spatial coherence.



(a) Anchor-based Alignment (b) Deform via SMPL-X (c) Cross-view Reference Adapter (CRA)

Fig. 3: Details of 4D Try-on Proxy Construction and CRA Module. (a) A similarity transformation $\{R_t, t_t, s_t\}$ maps the SMPL-X sequence from camera space to the world space, utilizing the human point cloud as a geometric anchor. (b) The canonical 3DGS-based avatar is dynamically warped via LBS driven by the aligned SMPL-X sequence, preserving structural integrity across the motion. (c) The CRA module facilitates interaction between DiT features $\mathbf{F}_i$ and reference features $\mathbf{F}_l$. By sharing K-V projections and utilizing independent Q'-O' weights, it integrates a reference residual $\mathbf{O}'_l$ into the backbone to ensure cross-view identity and background consistency.

Canonical 3DGS-based Avatar Generation. To construct a high-fidelity representation of the human in the target garment, we employ image try-on [14, 42] and image-to-avatar foundation models [15, 26, 27] to generate a canonical 3DGS-based avatar. The process initiates with the identification of an optimal

keyframe I_{t^*} that maximizes frontal visibility and minimizes self-occlusion, ensuring the extraction of complete texture information. Sampling strategy is in the Appendix **??**.

Following the selection of I_{t^*}, we leverage image try-on frameworks to synthesize a high-fidelity reference image $\hat{I}_{t^*}$ featuring the target garment. By feeding $\hat{I}_{t^*}$ and its corresponding SMPL-X pose $\{\Theta_{t^*}\}$ into the avatar foundation model, we distill a canonical 3DGS avatar $\mathcal{G}_c$ that captures both the personalized identity and the intricate geometry of the new apparel. We define the Gaussian primitives as $\mathcal{G}_c = \{\mathbf{x}_i, \mathbf{\Sigma}_i, \sigma_i, \mathbf{c}_i\}$, where each primitive is anchored to the upsampled surface of the SMPL-X model in a canonical A-pose $\{\Theta_c\}$. Here, $\mathbf{x}_i$, $\mathbf{s}_i$, and $\mathbf{r}_i$ denote the 3D position, anisotropic scale, and quaternion rotation, respectively, while σ_i and $\mathbf{c}_i$ capture opacity and view-dependent appearance. By distilling the 2D try-on result into an explicit avatar, we establish a geometrically-grounded reference that inherently maintains structural integrity across varying views.

Deformation and Rendering. To animate the canonical avatar $\mathcal{G}_c$, we employ Linear Blend Skinning (LBS) driven by the aligned SMPL-X sequence $\{\Theta_t\}$. As illustrated in Fig. 3(b), the canonical clothed avatar is dynamically warped to match the target poses prescribed by the motion sequence. For each Gaussian primitive, skinning weights w are assigned by querying the nearest vertices of the upsampled SMPL-X mesh in the canonical space. To ensure the avatar is correctly situated within the environment, each primitive's position $\mathbf{x}$ and covariance $\mathbf{\Sigma}$ are first transformed to the posed space and subsequently mapped to the world space $\mathcal{S}^w$ via the similarity transformation $\mathcal{T}$:

$$\mathbf{x} = \mathcal{T}\left(\sum_{j=1}^{J} w_j \mathbf{T}_j \mathbf{x}_c\right), \quad \mathbf{\Sigma} = R\left(\sum_{j=1}^{J} w_j \mathbf{R'}_j \mathbf{\Sigma}_c \mathbf{R'}_j^{\top}\right) R^{\top}, \tag{2}$$

where $\mathbf{T}_j$ and $\mathbf{R'}_j$ denote the rigid bone transformations, and R represents the rotation component of $\mathcal{T}$. This deformation pipeline yields a posed avatar $\mathcal{G}_p$ that preserves fine-grained garment details while maintaining strict structural alignment with the background context $\mathcal{P}^b$.

The reconstructed 4D Try-on Proxy serves as a unified foundation for video synthesis under both input and novel camera trajectories. To synthesize the structural guidance, we adopt a hybrid rendering strategy: the dressed human avatar $\mathcal{G}_p$ is rendered via a differentiable 3DGS rasterizer [17], while the background point cloud $\mathcal{P}^b$ is rendered using the PyTorch3D rasterizer. To handle mutual occlusions, the final pixel color $C(p)$ is determined by a depth-buffer fusion scheme:

$$C(p) = \begin{cases} C_b(p), & \text{if } D_b(p) < D_h(p), \\ \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), & \text{if } D_b(p) \geq D_h(p), \end{cases} \tag{3}$$

where D_b and D_h represent the rendered depth of the background and human, respectively. This integrated rendering provides a geometrically consistent and temporally coherent video proxy, serving as a high-quality spatial-temporal guidance for the subsequent generative refinement stage.

In summary, by integrating high-fidelity 3DGS-based avatars with persistent background point cloud, our approach yields a geometrically and temporally coherent 4D Try-on Proxy. This representation not only enables high-fidelity rendering across arbitrary camera trajectories but also supports explicit content manipulation within the dynamic 4D scene. By providing versatile spatio-temporal guidance, the proxy effectively bridges the gap between 4D geometric modeling and the subsequent Proxy-Anchored Video Generation stage.

3.2 Proxy-Anchored Try-on Video Generation

To synthesize photorealistic try-on videos from the 4D proxy, we present a Video Diffusion Transformer (DiT) anchored by rendered priors. While generic I2V pre-train models often suffer from structural distortion and identity drift under radical motion, our framework circumvents these pitfalls by leveraging pixel-aligned structural metadata to rigorously constrain the generative process. This design ensures physically plausible garment deformations that strictly adhere to the target trajectory $\mathcal{T}$. As shown in Fig. 2, TryOnCrafter employs a three-tiered conditioning hierarchy: the Rendered Priors provide foundational spatio-temporal grounding, while the Cross-view Reference Adapter (CRA) and Multi-Modal Semantic Conditioning supply auxiliary appearance and semantic cues to ensure identity fidelity and scene harmony.

Rendered Prior & Reference Frame Injection. To enforce strict adherence to the target camera trajectory $\mathcal{T}$, we leverage the rendered video $\mathcal{V}_{render}$ derived from our 4D try-on proxy as a fundamental structural prior. Unlike implicit motion modeling, $\mathcal{V}_{render}$ serves as a pixel-aligned spatio-temporal constraint that explicitly encodes global geometric structures, articulated body poses, and high-fidelity appearance, all precisely synchronized with the prescribed camera trajectory $\mathcal{T}$. By integrating this content-rich rendered prior, the generative model gains a dense structural and texture anchor, ensuring that synthesized garment deformations and character orientations remain geometrically consistent and texture-persistent within the underlying 4D space across all frames. Additionally, to stabilize the global appearance and prevent silhouette drift during the early denoising stages, we inject the high-fidelity reference frame $\hat{I}_{t^*}$ as the first frame of rendered video.

Technically, the augmented rendered video is processed through the Wan Encoder [36] and integrated with a downsampled binary mask to highlight valid geometric regions. During the patchify stage, these encoded features are concatenated along the channel dimension with the latent noise.

Cross-view Reference Adapter (CRA). While the rendered prior provides global spatio-temporal guidance, the Cross-view Reference Adapter (CRA) is

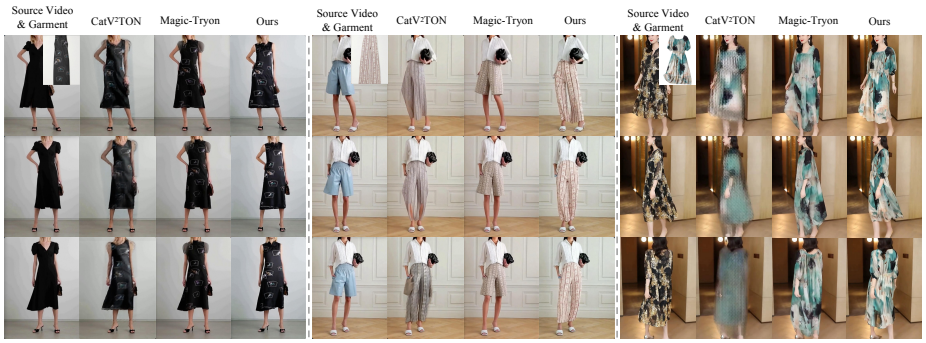


Fig. 4: Qualitative comparison on the video try-on benchmark. The left two columns show results on ViViD-S [6], and the right column shows results on the in-the-wild test set.

designed to recover fine-grained identity and background details from the source information domain. As illustrated in Fig. 3(c), CRA facilitates a synergistic interaction between the main DiT blocks and the reference branch. For a given DiT feature $\mathbf{F}_i$ and its corresponding reference feature $\mathbf{F}_l$, CRA employs a weight-sharing attention mechanism to bridge the two domains.

To ensure feature alignment while preserving the pre-trained generative prior, CRA shares the key (W_K) and value (W_V) projection weights with the DiT backbone, but utilizes independent query (W'_Q) and output (W'_O) weights. The reference residual $\mathbf{O}'_l$ is computed as:

$$\mathbf{O}'_l = \text{Softmax}\left(\frac{(\text{sg}(\mathbf{F}_i)W'_Q)(\mathbf{F}_l W_K)^\top}{\sqrt{d}}\right)(\mathbf{F}_l W_V)\cdot W'_O. \tag{4}$$

A stop-gradient operator $\text{sg}(\cdot)$ is applied to $\mathbf{F}_i$ to maintain training stability. This reference signal is then integrated into the main branch via an additive residual connection:

$$\mathbf{F}'_i = \text{Attn}(\mathbf{F}_i, Q_i, K_i, V_i, O_i) + \mathbf{O}'_l, \tag{5}$$

where $\text{Attn}(\cdot)$ denote the standard attention operation. By complementing the structural guidance of the 4D proxy with these fine-grained source features, the CRA module ensures high-fidelity appearance preservation across unobserved viewpoints with minimal computational overhead.

Multi-Modal Semantic Conditioning. To complement explicit geometric priors, we incorporate a multi-modal semantic mechanism to ensure global consistency, particularly in occluded or sparsely reconstructed regions. Specifically, we leverage a CLIP image encoder to extract garment features $\mathbf{f}_{gar}$ (texture and material cues) and a UmT5 text encoder for high-level attributes $\mathbf{f}_{text}$. These tokens are concatenated and injected into DiT blocks via cross-attention.

Fig. 5: Qualitative comparison results on CaM-VVTBench. **Left** trajectory: *orbit right* and *zoom in*, **Right** trajectory: *orbit left*. † denotes restricted to input trajectories.

This semantic grounding effectively prevents drift in unobserved viewpoints and harmonizes the synthesized apparel with the scene context during long-horizon generation.

4 Experiments

4.1 Implementation Details

Our TryonCrafter is built upon the pretrained Wan2.1-I2V-14B [36] foundation, featuring a progressive two-stage training strategy. All experiments are conducted on 16 NVIDIA A100 (80GB) GPUs. For inference, we employ 20 denoising steps and a CFG scale of 6.0. Detailed training strategies and parameter settings are provided in the Appendix **??**.

4.2 Dataset and Metrics

Video Virtual Try-On Benchmark. For VVT evaluation, we adopt the ViViD [6] benchmark, using its official training split (7,759 samples) and evaluating on the ViViD-S test set (180 samples). Following prior works [6, 20, 53], we adopt VFID with I3D [3] (VFID_I) and ResNext [45] (VFID_R) under both paired and unpaired settings. For paired evaluation, we additionally report SSIM and LPIPS to measure structural and perceptual similarity to ground truth.

Fig. 6: Versatile Applications of TryOnCrafter. Leveraging the decoupled 4D proxy, our model enables controllable synthesis: (a) Human Relocalization: Translating the motion sequence within $\mathcal{S}^w$ while maintaining scene-geometry consistency. (b) 360-degree Orbital Viewing: Synthesizing full orbital trajectories with robust structural integrity in unobserved viewpoints. (c) Bullet Time: Rendering a frozen temporal moment from a moving camera by anchoring the generation to a single pose.

Camera-controllable Video Virtual Try-On Benchmark. For CaM-VVT task, we establish a new benchmark, termed CaM-VVTBench. The training set comprises approximately 60K video clips collected from online sources, spanning diverse subjects, garment categories, and camera trajectories to ensure broad coverage of real-world scenarios. The test set contains 96 carefully curated samples with six predefined camera motion patterns, including *tilt up*, *tilt down*, *zoom in*, *zoom out*, *orbit left*, and *orbit right*. These standardized motion primitives enable controlled and reproducible evaluation, forming a dedicated benchmark. For the evaluation metrics, we use VBench [13], which includes Subject Consistency, Imaging Quality, Background Consistency, Aesthetic Quality, Motion Smoothness, Temporal Flickering, Overall Consistency scores and Overall Score.

4.3 Evaluation on Video Virtual Try-On Benchmark

Baselines. We compare our method with state-of-the-art image/video virtual try-on approaches, including StableVITON [18], OOTDiffusion [46], IDM-VTON [4], ViViD [6], CatV2TON [5], Magic-Tryon [20], and DreamVVT [53]. These methods cover diverse paradigms, including GAN-based frameworks, diffusion-based try-on models, and video-aware architectures with temporal constraints.

Quantitative Comparison. As reported in Table 1, our method establishes a new SOTA across both paired and unpaired settings, significantly improving video-level fidelity metrics VFID_I and VFID_R. In the challenging unpaired scenario, our approach substantially outperforms all baselines. Similarly, top performance in paired VFID_I^p and VFID_R^p reflects enhanced spatiotemporal consistency and garment-aware alignment. While some methods yield marginally higher SSIM in paired cases, our framework delivers superior perceptual quality (LPIPS) while maintaining competitive structural similarity. Overall, these

Table 1: Quantitative comparison with state-of-the-art methods on the ViViD [6] dataset. The bold represent the best results. p and u denote the paired setting and unpaired setting.

Method	$\text{VFID}_I^p \downarrow$	$\text{VFID}_R^p \downarrow$	$\text{VFID}_I^u \downarrow$	$\text{VFID}_R^u \downarrow$	SSIM↑	LPIPS↓
StableVITON	34.2446	0.7735	36.8985	0.9064	0.8019	0.1338
OOTDiffusion	29.5253	3.9372	35.3170	5.7078	0.8087	0.1232
IDM-VTON	20.0812	0.3674	25.4972	0.7167	0.8227	0.1163
ViViD	17.2924	0.6209	21.8032	0.8212	0.8029	0.1221
CatV2TON	13.5962	0.2963	19.5131	0.5283	0.8727	0.0639
Magic-Tryon	12.1988	0.2346	17.5710	0.5073	**0.8841**	0.0815
DreamVVT	11.0180	0.2549	16.9468	0.4285	0.8737	0.0619
Ours	**9.6085**	**0.1817**	**10.7563**	**0.2088**	0.8675	**0.0567**

Table 2: Ablation study of different try-on components in TryOnCrafter on the ViViD dataset. The bold and underlined entries represent the best and second-best results, respectively.

Method	$\text{VFID}_I^p \downarrow$	$\text{VFID}_R^p \downarrow$	$\text{VFID}_I^u \downarrow$	$\text{VFID}_R^u \downarrow$	SSIM↑	LPIPS↓
w/o $\hat{I}_{t^*}$	11.8614	0.2382	<u>18.1947</u>	<u>0.2996</u>	0.8526	0.0669
w/o mask	9.9393	<u>0.1927</u>	18.5710	0.4212	0.8633	0.0582
w/o $\mathbf{f}_{text}$	18.7144	0.3500	18.5158	0.3707	0.8189	0.1319
w/o $\mathbf{f}_{par}$	**9.1740**	0.2257	18.4331	0.3487	**0.8680**	**0.0562**
w/o $\mathcal{V}_{render}$	16.4608	0.3937	20.1280	0.4159	0.8346	0.0925
full (Ours)	<u>9.6085</u>	**0.1817**	**10.7563**	**0.2088**	<u>0.8675</u>	<u>0.0567</u>

Table 3: Quantitative comparison with two-stage methods on the CaM-VVTBench. The bold represent the best results. † denotes restricted to input trajectories.

Method	Overall Score↑	Subject Consis.↑	Imaging Quality↑	Background Consis.↑	Aesthetic Quality↑	Motion Smooth.↑	Temporal Flicker.↑	Overall Consis.↑
Magic-Tryon+TrajectoryCrafter	71.64	87.76	68.27	90.76	49.89	97.88	94.53	5.15
Magic-Tryon+ReCaMaster	61.92	87.93	66.69	91.82	46.76	98.39	95.50	8.29
TryOnCrafter†+TrajectoryCrafter	71.95	88.52	68.93	91.18	51.29	97.93	94.60	7.53
TryOnCrafter†+ReCaMaster	74.32	88.90	67.70	91.72	47.64	**98.45**	**95.54**	8.35
TryOnCrafter (Ours)	**75.47**	**92.71**	**74.28**	**92.05**	**51.44**	98.32	94.90	**9.61**

results substantiate the effectiveness of our framework in enhancing video-level realism and coherence across diverse protocols without compromising visual fidelity.

Qualitative Comparison. Fig. 4 demonstrates TryOnCrafter's superiority over SOTA frameworks across diverse scenarios. Our method exhibits higher fidelity in texture transfer and color preservation compared to CatV^{2}TON and Magic-Tryon, precisely maintaining intricate patterns and original vibrancy. Notably, by leveraging geometry-aware rendered priors, TryOnCrafter accurately reconstructs complex silhouettes (e.g., full-length trousers), whereas Magic-Tryon fails to resolve structural geometry, erroneously collapsing long garments into shorts. Furthermore, in unconstrained "in-the-wild" environments, our 4D try-on proxy enables the synthesis of realistic dynamics and natural folds for challenging apparel like loose dresses. Conversely, baselines suffer from significant structural distortion and lack the spatiotemporal coherence required for physically plausible deformations in complex scenes.

4.4 Evaluation on CaM-VVTBench

Baselines. Given the absence of end-to-end models for the CaM-VVT task, we establish rigorous benchmarks by coupling open-source state-of-the-art (SOTA) VVT methods with representative camera control frameworks. Specifically, we employ Magic-Tryon [20] as the VVT baseline. For camera steering, we integrate two distinct paradigms: TrajectoryCrafter [48], which leverages explicit 3D point cloud reconstruction and geometric priors, and ReCaMaster [1], which

utilizes implicit parameter injection. Furthermore, to ensure a comprehensive and fair comparison, we evaluate combinations of these control methods with TryOnCrafter† († denotes **restricted to input trajectories**).

Qualitative Comparison. As illustrated in Fig. 5, while two-stage baselines achieve plausible initial synthesis by leveraging the try-on priors of TryOnCrafter†, their performance degrades significantly as camera movements escalate in intensity and perspective complexity. Specifically, TrajectoryCrafter frequently exhibits structural breakdown and limb distortions (e.g., subjects' arms) during aggressive rotations due to the lack of explicit human priors and its reliance on fragmented 3D point clouds. In contrast, our unified framework provide the human avatar in Renderable 4D Try-on Proxy as a continuous geometric foundation, significantly outperforming the point-cloud baseline in preserving limb integrity and background coherence. Furthermore, ReCaMaster suffers from a profound deficit in structural comprehension; the absence of explicit geometric constraints yields severe warping across both the human subject and the complex background manifold (e.g., distorted bushes and bench). Most crucially, while baselines fail to maintain texture consistency or generate physically plausible clothing motion under violent maneuvers, our TryOnCrafter effectively achieves viewpoint-agnostic synthesis of fine-grained textures from unobserved angles. By synthesizing realistic, trajectory-aligned cloth deformations, our Proxy-Anchored Video DiT demonstrates superior structural robustness and high-fidelity appearance across unconstrained trajectories.

Quantitative Comparison. As shown in Table 3, TryOnCrafter establishes a new SOTA on CaM-VVTBench, significantly outperforming two-stage baselines. Our model leads substantially in Overall Score, Subject Consistency, and Imaging Quality, underscoring the efficacy of the Proxy-Anchored DiT in mitigating error accumulation and preserving fine-grained textures. While ReCaMaster-based variants show marginal gains in Motion Smoothness, qualitative results reveal this often stems from excessive blurring and identity loss during aggressive maneuvers. Conversely, TryOnCrafter achieves a superior equilibrium between temporal stability and high-fidelity synthesis, evidenced by top-tier Aesthetic Quality and Background Consistency. These metrics validate the Renderable 4D Proxy as a robust structural foundation for physically plausible, 3D-aware virtual try-on.

4.5 Ablation Study

Ablation on 4D Try-on Proxy. We systematically evaluate the core components of the 4D Try-on Proxy—the clothed avatar $\mathcal{G}_p$, the background point cloud $\mathcal{P}^b$, and the Anchor-based Alignment—using the CaM-VVTBench. As illustrated in Fig. 7, omitting $\mathcal{G}_p$ deprives the model of essential pose and texture priors, leading to stochastic motion and severe structural artifacts. The absence of $\mathcal{P}^b$ restricts the generative scope solely to the human subject, failing to capture the environmental context and ambient lighting. Furthermore, removing the Anchor-based Alignment disrupts spatio-temporal synchronization, resulting in misaligned poses and discontinuous trajectories.

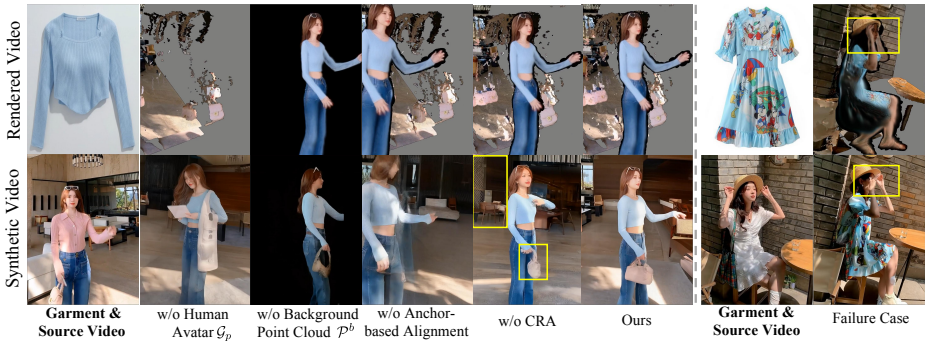


Fig. 7: Left: Ablation studies of main component of TryOnCrafter. **Right**: Failure case caused by perspective-induced parallax and inaccuracies of SMPL-X estimation.

Ablation on DiT Conditioning Techniques. We further investigate the conditioning mechanisms in Proxy-Anchored Video DiT, comprising rendered priors, the CRA module, and multimodal semantic cues. Quantitative results in Table 2 confirm the necessity of this integration; notably, the significant performance decline in the w/o Rendered Video setting underscores that explicit geometric anchoring is indispensable for maintaining structural integrity and visual fidelity. Regarding specific components, while removing garment semantics $\mathbf{f}_{gar}$ yields marginal gains in simple paired settings, it causes severe degradation in unpaired scenarios, highlighting the importance of semantic priors for generalization. Furthermore, omitting the CRA module—responsible for source identity and articulated cues—drastically reduces cross-view geometric consistency and ID preservation (Fig. 7). These findings substantiate that the synergy between our explicit 4D proxy and multi-modal conditioning is vital for high-fidelity, camera-controllable try-on.

5 Applications

The explicit 4D Try-on Proxy enables the decoupling of human subjects, background, and camera trajectories within a unified world space $\mathcal{S}^w$. Beyond standard synthesis, we explore three generative applications that demonstrate the framework's versatility.

Human Relocalization. By integrating the subject and background into $\mathcal{S}^w$, we can perform arbitrary spatial edits on the motion sequence. As shown in Fig. 6(a), our geometry-grounded approach facilitates Human Relocalization, ensuring consistent occlusion and metric-scale harmony between the edited subject trajectory and the environment—a task where traditional 2D-based methods often suffer from perspective distortion.

360-degree Orbital Viewing. While existing frameworks [1,48] are often limited to narrow-angle synthesis (e.g., $\pm 60°$), TryOnCrafter achieves high-fidelity 360-degree orbital viewing generation. By leveraging the 4D try-on proxy as a

persistent geometric anchor, our model maintains structural stability and texture consistency even in unobserved viewpoints (Fig. 6(b)), effectively resolving the challenges of extreme disocclusion.

Bullet Time Virtual Try-on. We extend our framework to Bullet Time effects, where a moving camera renders a frozen temporal instant. This is achieved by anchoring the generation to a specific SMPL-X pose Θ_{t^*} and its corresponding canonical frame I_{t^*} across the sequence. As illustrated in Fig. 6(c), the resulting videos exhibit smooth, view-consistent transitions around a single pose, offering a novel immersive try-on experience.

6 Limitation

Despite its strong performance, TryOnCrafter possesses certain limitations. Specifically, our framework relies on a parametric 4D Try-on Proxy for geometric guidance; however, extreme viewpoint transitions often introduce significant parallax and ambiguity, challenging the structural alignment between the subject and the background. As shown in Fig. 7, such spatial shifts can exacerbate estimation inaccuracies in the SMPL-X model, occasionally manifesting as misaligned hand poses or subtle geometric inconsistencies during radical articulated motions. Furthermore, the iterative denoising process of our DiT-based framework incurs high inference costs, hindering real-time interaction for trajectory edits.

7 Conclusion

We pioneer Camera-controllable Video Virtual Try-on (CaM-VVT) to overcome the trajectory dependency of existing paradigms. Our framework, TryOnCrafter, integrates two synergistic components: a Renderable 4D Try-on Proxy for explicit subject-environment decoupling, and a Proxy-anchored DiT that leverages a multi-tiered conditioning hierarchy for high-fidelity synthesis. By anchoring the generative process to the 4D proxy, TryOnCrafter ensures physically plausible garment deformations and temporal coherence across unconstrained camera movements. Extensive evaluations on CaM-VVTBench demonstrate that our method significantly outperforms SOTA baselines. Furthermore, the 4D representation's versatility enables diverse applications like human relocalization and 360-degree orbital viewing, paving a new way for interactive digital fashion.

Acknowledgement

This work was supported by the Institute Innovation Project of the Institute of Computing Technology, Chinese Academy of Sciences under Grant No. E561090. This work was supported by Alibaba Group through Alibaba Research Intern Program.

References

1. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
2. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–12 (2025)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: European Conference on Computer Vision. pp. 206–235. Springer (2024)
5. Chong, Z., Zhang, W., Zhang, S., Zheng, J., Dong, X., Li, H., Wu, Y., Jiang, D., Liang, X.: Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. arXiv preprint arXiv:2501.11325 (2025)
6. Fang, Z., Zhai, W., Su, A., Song, H., Zhu, K., Wang, M., Chen, Y., Liu, Z., Cao, Y., Zha, Z.J.: Vivid: Video virtual try-on using diffusion models. arXiv preprint arXiv:2405.11794 (2024)
7. Gong, Y., Song, Y., Li, Y., Li, C., Zhang, Y.: Relationadapter: Learning and transferring visual relation with diffusion transformers. arXiv preprint arXiv:2506.02528 (2025)
8. Guo, H., Zeng, B., Song, Y., Zhang, W., Zhang, C., Liu, J.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. arXiv preprint arXiv:2501.15891 (2025)
9. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)
10. He, Z., Chen, P., Wang, G., Li, G., Torr, P.H., Lin, L.: Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In: European Conference on Computer Vision. pp. 123–139. Springer (2024)
11. Hou, C., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
12. Huang, S., Song, Y., Zhang, Y., Guo, H., Wang, X., Liu, J.: Arteditor: Learning customized instructional image editor from few-shot examples. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17651–17662 (2025)
13. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
14. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
15. Jiang, T., Ho, H.I., Kaufmann, M., Song, J.: Prioravatar: Efficient and robust avatar creation from monocular video using learned priors. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)

16. Karras, J., Li, Y., Liu, N., Zhu, L., Yoo, I., Lugmayr, A., Lee, C., Kemelmacher-Shlizerman, I.: Fashion-vdm: Video diffusion model for virtual try-on. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
17. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
18. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: 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. pp. 8176–8185 (2024)
19. Li, C., Zhang, C., Xu, W., Lin, J., Xie, J., Feng, W., Peng, B., Chen, C., Xing, W.: Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. arXiv preprint arXiv:2412.09262 (2024)
20. Li, G., Zheng, S., Zhang, H., Chen, J., Luan, J., Ou, B., Zhao, L., Li, B., Jiang, P.T.: Magictryon: Harnessing diffusion transformer for garment-preserving video virtual try-on. arXiv preprint arXiv:2505.21325 (2025)
21. Li, X., Lai, Z., Xu, L., Qu, Y., Cao, L., Zhang, S., Dai, B., Ji, R.: Director3d: Real-world camera trajectory and 3d scene generation from text. Advances in neural information processing systems **37**, 75125–75151 (2024)
22. Lin, J., Wu, Y., Wang, Z., Liu, X., Guo, Y.: Pair-id: A dual modal framework for identity preserving image generation. IEEE Signal Processing Letters (2024)
23. Liu, K., Shao, L., Lu, S.: Novel view extrapolation with video diffusion priors. arXiv preprint arXiv:2411.14208 (2024)
24. Meng, Y., Zhu, Z., Hui, L., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In: The Thirteenth International Conference on Learning Representations (2024)
25. Nguyen, H., Nguyen, Q.Q.V., Nguyen, K., Nguyen, R.: Swifttry: Fast and consistent video virtual try-on with diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6200–6208 (2025)
26. Qiu, L., Gu, X., Li, P., Zuo, Q., Shen, W., Zhang, J., Qiu, K., Yuan, W., Chen, G., Dong, Z., et al.: Lhm: Large animatable human reconstruction model from a single image in seconds. arXiv preprint arXiv:2503.10625 (2025)
27. Qiu, L., Zhu, S., Zuo, Q., Gu, X., Dong, Y., Zhang, J., Xu, C., Li, Z., Yuan, W., Bo, L., et al.: Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21148–21158 (2025)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
29. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
30. Song, Q., Lin, M., Zhan, W., Yan, S., Cao, L., Ji, R.: Univst: A unified framework for training-free localized video style transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
31. Song, Q., Lin, Z., Zeng, Z., Zhang, Z., Cao, L., Ji, R.: Lightmotion: A light and tuning-free method for simulating camera motion in video generation. arXiv preprint arXiv:2503.06508 (2025)
32. Song, Q., Shen, Y., Chen, M., Sun, H., Lan, J., Zhu, X., Zheng, B., Cao, L.: Fashionchameleon: Towards real-time and interactive human-garment video customization. arXiv preprint arXiv:2605.15824 (2026)

33. Song, Q., Song, Y., Peng, K., Gao, Y., Shou, M.Z.: Worldwander: Bridging egocentric and exocentric worlds in video generation. arXiv preprint arXiv:2511.22098 (2025)
34. Song, Q., Zhou, D., Lin, J., Shen, F., Wang, J., Hu, X., Chen, C., Heng, P.A.: Scenedecorator: Towards scene-oriented story generation with scene planning and scene consistency. arXiv preprint arXiv:2510.22994 (2025)
35. Song, Y., Liu, C., Shou, M.Z.: Makeanything: Harnessing diffusion transformers for multi-domain procedural sequence generation. arXiv preprint arXiv:2502.01572 (2025)
36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
38. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20697–20709 (2024)
39. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025)
40. Wang, Y., Dai, W., Chan, L., Zhou, H., Zhang, A., Liu, S.: Gpd-vvto: Preserving garment details in video virtual try-on. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7133–7142 (2024)
41. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
42. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
43. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26057–26068 (2025)
44. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. arXiv preprint arXiv:2411.19324 (2024)
45. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
46. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 8996–9004 (2025)
47. Xu, Z., Chen, M., Wang, Z., Xing, L., Zhai, Z., Sang, N., Lan, J., Xiao, S., Gao, C.: 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. pp. 3199–3208 (2024)
48. Yu, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 100–111 (2025)

49. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
50. Zhang, J., Herrmann, C., Hur, J., Jampani, V., Darrell, T., Cole, F., Sun, D., Yang, M.H.: Monst3r: A simple approach for estimating geometry in the presence of motion. arXiv preprint arXiv:2410.03825 (2024)
51. Zhang, Y., Yuan, Y., Song, Y., Wang, H., Liu, J.: Easycontrol: Adding efficient and flexible control for diffusion transformer. arXiv preprint arXiv:2503.07027 (2025)
52. Zhang, Y., Zhang, Q., Song, Y., Zhang, J., Tang, H., Liu, J.: Stable-hair: Real-world hair transfer via diffusion model. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 10348–10356 (2025)
53. Zuo, T., Huang, Z., Ning, S., Lin, E., Liang, C., Zheng, Z., Jiang, J., Zhang, Y., Gao, M., Dong, X.: Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. arXiv preprint arXiv:2508.02807 (2025)