

BooM-VVT: Boosting Mask-Free Video Virtual Try-On with Image-Level Pseudo Data

Anonymous Author(s)
Submission Id: 4444



Figure 1: BooM-VVT generates high-fidelity and temporally coherent virtual try-on videos without requiring masks. The left panel shows its robustness in challenging real-world scenarios, including large motion, severe occlusion, and object interaction. The right panel highlights its flexibility in diverse try-on tasks, such as multi-view, cross-category, and layered try-on.

Abstract

Video virtual try-on (VVT) aims to generate realistic videos of a person wearing a target garment. Recent methods leverage a keyframe-driven video generation paradigm to improve in-the-wild performance, yet they still rely on masks to localize try-on regions, making them vulnerable to large motions and severe occlusions. Although mask-free image-based try-on methods have shown promising results by leveraging large-scale pseudo data, extending this paradigm to videos remains difficult, as constructing video-level pseudo data is prohibitively expensive. Furthermore, coarse keyframe sampling and the scarcity of multi-view try-on data limit existing keyframe-driven methods in maintaining garment consistency and handling diverse try-on tasks.

To address these challenges, we propose BooM-VVT, a mask-free VVT framework built upon the keyframe-driven paradigm. To achieve mask-free VVT, we introduce a multi-stage training strategy that leverages image-level pseudo data for mask-free localization learning, substantially reducing the need for costly video-level pseudo data. To improve garment consistency, we propose

Garment-Sensitive Keyframe Sampling, which selects keyframes based on garment-relevant body regions to better capture garment appearance. We further introduce Frame-Shared 3D-RoPE to establish spatiotemporal correspondences between keyframes and target video frames for accurate garment-detail transfer. Finally, we construct OmniView, a large-scale multi-view try-on dataset to support reliable try-on video generation under complex camera viewpoints and diverse try-on tasks. Extensive experiments demonstrate that BooM-VVT achieves superior temporal consistency and garment fidelity over existing methods. Our project page <https://anonymous.4open.science/w/BooMVVT/>.

CCS Concepts

• **Computing methodologies** → **Computer vision problems;** *Computer vision tasks;* *Computer vision representations.*

Keywords

Video Virtual Try-On, Mask-Free Generation, Diffusion Models

1 Introduction

Video virtual try-on (VVT) aims to synthesize realistic videos of a person wearing a target garment, with broad applications in e-commerce and digital content creation. Although recent diffusion-based methods [4, 11, 12, 15, 16, 26] have achieved impressive progress, they remain heavily dependent on scarce paired garment-video data. Due to the limited scale and scene diversity of existing datasets [5, 7], these methods often struggle to preserve garment fidelity and temporal consistency in real-world scenarios.

To address this limitation, recent studies [10, 30] have introduced a keyframe-driven try-on video generation paradigm that decomposes the end-to-end VVT pipeline into two stages: (1) generating try-on images for selected keyframes, and (2) generating the final try-on video conditioned on the resulting keyframe try-on images. Under this paradigm, methods such as DreamVVT [30] can exploit large-scale unpaired video data to train the video generation model, thereby improving generalization in real-world scenarios. Despite these advantages, keyframe-driven methods still face several challenges. **First**, existing VVT methods typically rely on explicit masks to localize try-on regions. However, accurately estimating masks in videos is highly challenging due to large pose variations and severe occlusions, as shown in Fig. 1(a). Inaccurate masks often introduce temporal artifacts such as jittering and flickering. While mask-free image-based try-on methods [8, 9, 28, 29] have shown promising results by leveraging large-scale pseudo data, extending this paradigm to video virtual try-on remains nontrivial, as constructing video-level pseudo data is prohibitively expensive. **Second**, existing methods typically select keyframes based on overall pose changes and full-body visibility. However, this strategy is often influenced by body regions irrelevant to the target garment, such as the lower body when trying on tops. As a result, the selected keyframes may fail to fully capture the garment appearance, thereby compromising garment consistency in the synthesized videos. **Third**, existing multi-view try-on datasets remain limited in scale and coverage, as shown in Table 1. This limitation hinders current methods from generating plausible results under challenging viewpoints (e.g., back views) and handling diverse try-on tasks (see Fig. 1(b)).

To tackle these challenges, we present BooM-VVT, a novel mask-free video virtual try-on framework built upon the keyframe-driven paradigm. To achieve mask-free VVT, we propose a multi-stage training strategy that enables the model to leverage more accessible image-level pseudo data to learn robust mask-free try-on region localization. As a result, the model can adapt to mask-free video virtual try-on with only minimal video-level pseudo data, substantially easing data construction. To improve garment consistency, we first propose a Garment-Sensitive Keyframe Sampling (GSKS) strategy. Specifically, it selects keyframes based on body regions most relevant to the target garment, allowing the selected keyframes to better capture the target garment appearance. We further introduce Frame-Shared 3D-RoPE. By assigning identical positional encodings to keyframe try-on images and their corresponding target video frames, it explicitly establishes spatiotemporal correspondences between them, enabling more accurate garment appearance transfer. Finally, to address the scarcity of multi-view try-on data, we construct OmniView, a large-scale multi-view try-on dataset

Table 1: Comparison with existing multi-view try-on datasets. *Category Diversity* indicates whether the dataset covers diverse garment categories, and *Garment Completeness* indicates whether complete flat garment images are provided.

Dataset	Garment Category Diversity	Garment Completeness	Scale
MVG	✗	✗	1,009
OmniView	✓	✓	6,110

that enables reliable try-on generation across diverse viewpoints and try-on tasks. In summary, our contributions are as follows:

- We propose BooM-VVT, a mask-free video virtual try-on framework that leverages a multi-stage training strategy to learn robust try-on region localization from readily available image-level pseudo data, substantially reducing reliance on costly video-level pseudo data.
- To improve garment consistency, we propose a Garment-Sensitive Keyframe Sampling (GSKS) strategy and Frame-Shared 3D-RoPE. GSKS selects more informative keyframes based on garment-relevant body regions, while Frame-Shared 3D-RoPE facilitates more accurate garment appearance transfer by establishing spatiotemporal correspondences between the keyframes and the target video.
- We construct OmniView, a large-scale multi-view try-on dataset that enables reliable try-on generation across diverse viewpoints and try-on tasks.

2 OmniView Dataset

Overview of OmniView Dataset. To provide comprehensive garment cues for downstream try-on video generation, existing keyframe-driven VVT methods [30] typically need to generate try-on images for keyframes from different viewpoints. However, existing multi-view try-on datasets, such as MVG [21], remain limited in scale, garment category diversity, and garment completeness (see Table 1). To address these limitations, we construct OmniView, a large-scale multi-view try-on dataset containing 6,110 samples. Each sample provides person images from at least two viewpoints, and 88% of the samples include back-view images. As shown in Fig. 2, OmniView offers three key improvements over existing datasets. First, each sample is paired with both front- and back-view flat garment images, providing more complete garment references. Second, in addition to the common upper-body, lower-body, and full-body categories, OmniView also includes an outerwear category. Third, we synthesize pseudo-labeled data for each person image to support diverse try-on tasks, including cross-category and layered try-on. As shown in the third row of Fig. 2(b), the third column illustrates pseudo data for cross-category try-on, while the fourth column shows pseudo data for layered try-on constructed by removing the outerwear.

Data processing. We collect raw data from multiple e-commerce platforms and open-source datasets [18], in strict compliance with their terms of use. Since the raw data often contain noise, such as inconsistent model identities, we first employ Qwen3-VL-32B [1] to



Figure 2: Data samples from the MVG dataset (a) and the proposed OmniView dataset (b). Compared with MVG, OmniView provides higher-quality garment images, more garment categories, and extra pseudo data to support diverse try-on tasks.

annotate both model and garment images. Based on these annotations, we regroup the samples and manually review them to ensure consistent model identity within each group and accurate correspondence between worn garments and garment images. We then leverage Qwen3-VL-32B to generate editing instructions and use Qwen-Image-Edit-2511 [24] to execute them for pseudo-data synthesis. For non-outerwear categories, the instructions support both intra-category and cross-category garment replacement, while for outerwear, we additionally introduce garment removal instructions to support layered try-on. Examples of the synthesized pseudo data are shown in the third row of Fig. 2(b): the first two columns show data generated by intra-category garment replacement, the third column shows data generated by cross-category garment replacement, and the fourth column shows data generated by garment removal. Finally, the synthesized data undergo the same review pipeline to ensure data quality and consistency.

3 Method

3.1 Overview

BooM-VVT is built on the keyframe-driven video generation paradigm and aims to achieve mask-free video virtual try-on, as illustrated in Fig. 3. The framework consists of two components: keyframe try-on (Sec. 3.2) and keyframe-guided try-on video generation (Sec. 3.3). Given an input video V_{in} , we first select informative keyframes using the proposed Garment-Sensitive Keyframe Sampling (GSKS) strategy, and then use a multi-frame try-on model to generate the corresponding try-on results for these keyframes. Benefiting from the proposed OmniView dataset, the multi-frame try-on model can take either single-view or multi-view garment inputs and produce reliable try-on results across diverse viewpoints. The generated keyframe try-on images are subsequently used as appearance conditions for a DiT-based [17] video generation model, which synthesizes the final try-on video. To facilitate accurate garment appearance transfer, we introduce Frame-Shared 3D-RoPE, which explicitly establishes correspondences between keyframe

try-on tokens and their corresponding video-frame tokens by assigning them identical positional encodings. To eliminate reliance on masks, we further devise a multi-stage training strategy for the video generation model, as detailed in Fig. 4.

3.2 Keyframe Try-on

3.2.1 Garment-Sensitive Keyframe Sampling. In keyframe-driven video virtual try-on (VVT), keyframes serve as the primary carrier of target garment information for the final try-on video. An effective keyframe set should satisfy two properties: (i) **high information capacity**: garment-relevant body regions should be sufficiently visible so that fine-grained garment details can be captured; and (ii) **complementary viewpoint coverage**: the selected keyframes should cover diverse viewpoints to provide more complete garment information.

To this end, we propose a Garment-Sensitive Keyframe Sampling (GSKS) strategy. To quantify the information capacity of each frame, we first define a garment-relevant limb set P_c and body region A_c according to the target garment category c . For example, for upper-body garments, P_c includes limbs such as the shoulders and arms, while A_c corresponds to the upper torso. Given an input video V_{in} , GSKS computes an information capacity score $S_{Capacity}^c$ for each frame based on a limb visibility score S_p^c and a region visibility score S_a^c . Specifically, we use DWPose [25] to detect limbs in each frame, and define S_p^c as the ratio of the detected limbs belonging to P_c to the total number of limbs in P_c :

$$S_p^c = \frac{|\hat{P} \cap P_c|}{|P_c|}, \quad (1)$$

where $\hat{P}$ denotes the set of detected limbs in the current frame. Meanwhile, S_a^c denotes the area ratio of the garment-relevant body region A_c to the whole frame, where A_c is estimated using SAM [2]. The final information capacity score is computed as follows:

$$S_{Capacity}^c = \lambda S_p^c + (1 - \lambda) S_a^c. \quad (2)$$

In practice, we set $\lambda = 0.2$.

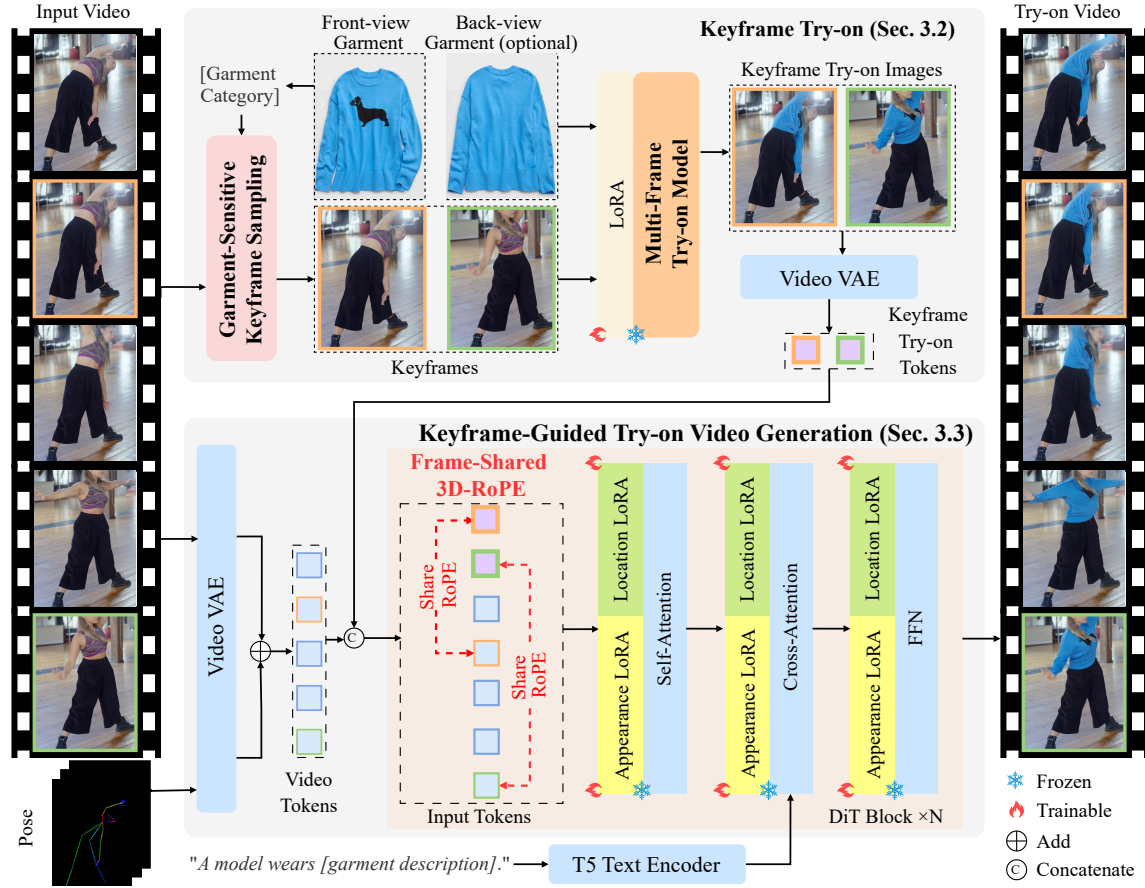


Figure 3: Overview of BooM-VVT. The framework consists of two components: keyframe try-on and keyframe-guided try-on video generation. The first component selects informative keyframes from the input video and generates corresponding try-on images, while the second synthesizes the final try-on video under the guidance of these keyframe try-on results.

After scoring all frames, we select the frame with the highest S_{Capacity}^c as the primary keyframe and add it to the keyframe set $\mathcal{K}$. We then iteratively select the frame that differs the most in viewpoint from the currently selected keyframes and add it to $\mathcal{K}$, so as to ensure complementary viewpoint coverage. We measure viewpoint difference as the average cosine distance between the skeletal joint direction vectors of a candidate frame and those of the selected keyframes. To focus on garment-relevant appearance variation, this computation only involves the limbs in P_c , thus avoiding interference from body regions irrelevant to the target garment. The detailed pseudo-code is provided in the Appendix.

3.2.2 Multi-Frame Try-on Model. Given the selected keyframes, we employ a pretrained image try-on model built upon the MM-DiT architecture [6] to generate the corresponding keyframe try-on images. To improve consistency across frames, we adapt the image try-on model to a multi-frame setting using a simple image concatenation scheme. Specifically, the selected keyframes are concatenated horizontally into a single composite image and fed into the model. The model outputs an image in the same layout, which is then split horizontally to obtain the individual try-on frames.

By processing all keyframes in a single forward pass, the model enables cross-frame information interaction and produces coherent multi-frame try-on results with consistent garment details. When multi-view garment images are available, we apply the same concatenation scheme to organize them as garment inputs. To adapt the model to this multi-frame setting, we fine-tune it on OmniView using lightweight LoRA adapters inserted into the self-attention layers of the MM-DiT blocks. During training, we use both single-view and multi-view garment inputs. As a result, the model can produce plausible try-on results for challenging viewpoints, such as back views, even when only a single-view garment image is available. When multi-view garment references are provided, the visual fidelity can be further improved.

3.3 Keyframe-Guided Try-on Video Generation

Our video generation model builds on Wan-Animate [3], a DiT-based image-to-video framework. As illustrated in Fig. 3, given an input video and its corresponding pose sequence, we encode the video and pose sequence into video tokens and pose tokens, respectively, using a Video VAE. The pose tokens are added to the video tokens to provide structural guidance. In parallel, we encode

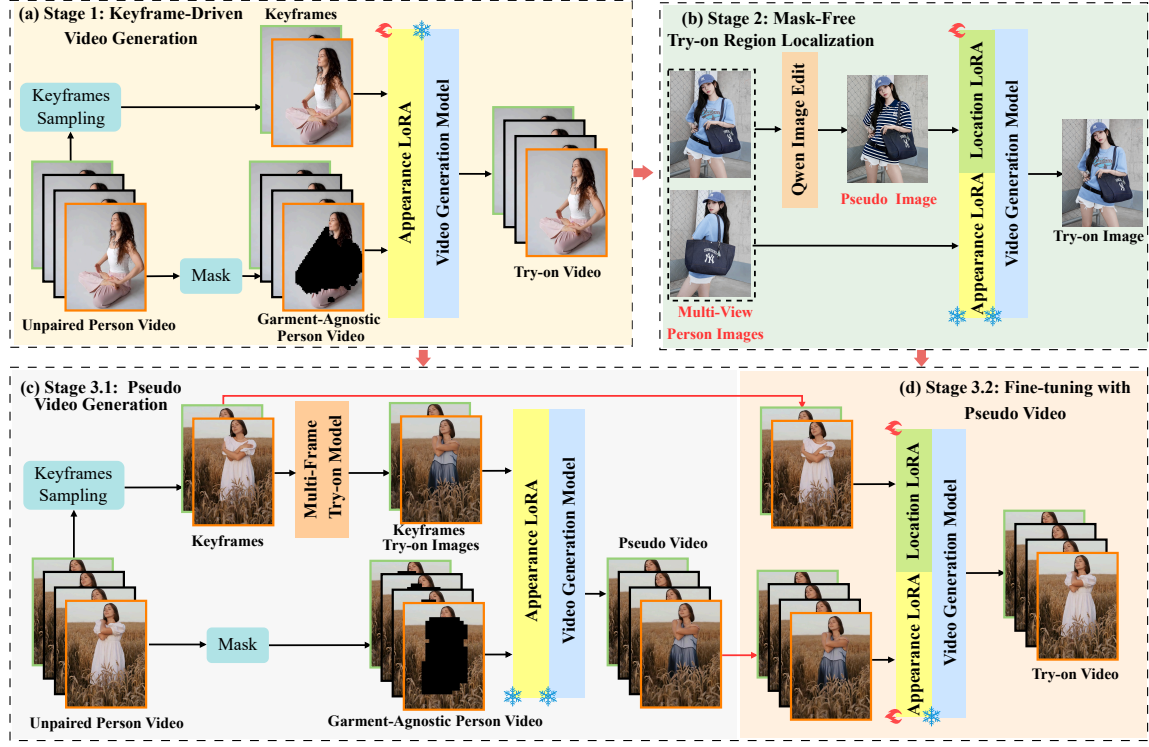


Figure 4: Overview of the proposed multi-stage training strategy for the video generation model in BooM-VVT. (a) Stage 1: The video generation model is trained for keyframe-driven video generation using unpaired videos. (b) Stage 2: The model learns mask-free try-on region localization from image-level pseudo data. (c) Stage 3.1: We use the model trained in Stage 1 to synthesize video-level pseudo data. (d) Stage 3.2: The model is fine-tuned on video-level pseudo data to enable high-quality try-on video generation without masks.

the keyframe try-on images with the same VAE to obtain keyframe tokens. The keyframe tokens are then concatenated with the video tokens along the sequence dimension and fed into the DiT backbone. After iterative denoising, the denoised latents are decoded by the VAE decoder to obtain the final try-on video.

3.3.1 Frame-Shared 3D-RoPE. Following the standard design of image-to-video generation, existing keyframe-driven VVT methods typically assume keyframe try-on images and video frames to be spatiotemporally independent, assigning them different positional encodings. However, this design is suboptimal for keyframe-driven virtual try-on, as it overlooks an important prior: each keyframe try-on image is naturally aligned with its corresponding target frame. Assigning different positional encodings to such aligned representations increases their relative positional distance, thereby weakening the attention between them and hindering accurate garment appearance transfer.

To address this issue, we propose Frame-Shared 3D-RoPE to enable more accurate garment appearance transfer. Our design builds on the distance-aware property of 3D-RoPE [19], where tokens with smaller spatiotemporal distances tend to attend more strongly to each other. Specifically, we assign each keyframe try-on token the same 3D positional encoding as its corresponding target video-frame token. By sharing positional encodings across

aligned keyframe and video-frame tokens, Frame-Shared 3D-RoPE reduces their relative positional distance, thereby strengthening the attention between them and improving garment appearance transfer from the keyframes to the target video.

3.3.2 Multi-Stage Training Strategy. To achieve mask-free video virtual try-on, existing image-based virtual try-on methods typically rely on large-scale pseudo-paired data to learn mask-free try-on region localization. However, directly extending this strategy to the video domain is impractical, since constructing large-scale video-level pseudo data is prohibitively expensive. To address this challenge, we propose a multi-stage training strategy, as illustrated in Fig. 4. Our strategy consists of three stages: we first train the video model for keyframe-guided video generation on large-scale unpaired human-centric videos, then equip it with mask-free try-on region localization using readily available image-level pseudo data, and finally adapt it to mask-free video virtual try-on with minimal video-level pseudo data.

Stage 1: Training for Keyframe-Driven Video Generation. As shown in Fig. 4(a), in the first stage, we follow the standard keyframe-driven training paradigm and train the video generation model on large-scale unpaired human-centric videos. Specifically, for each input video, we first randomly sample several frames as keyframe references. We then use garment masks to remove



Figure 5: Qualitative comparison on the ViViD-S dataset.

the original garments throughout the video, obtaining a garment-agnostic input. Given this garment-agnostic video and the sampled keyframes, the model is trained to reconstruct the original video. Since the try-on regions are explicitly provided by garment masks at this stage, the model can focus on learning accurate garment appearance transfer from the selected keyframes to the full video while preserving temporal coherence.

Stage 2: Training for Mask-Free Try-on Localization. As illustrated in Fig. 4(b), the second stage is designed to equip the video model with mask-free try-on region localization capability. Instead of constructing costly video-level pseudo data, we leverage readily available image-level pseudo pairs for supervision. Specifically, we build a large-scale multi-view person dataset by sampling frames from videos and incorporating existing multi-view person datasets, where each sample contains two images of the same person under different poses. For each sample, we randomly select one image as the target image and use Qwen-Image-Edit-2511 to alter its garment, producing a pseudo image. This pseudo image is fed into the video generation model as a single-frame video, while the other image serves as the keyframe reference. The model is then trained to reconstruct the original target image from the pseudo image, conditioned on the keyframe reference. Since no explicit mask is provided during this process, the model must automatically infer where garment replacement should occur, thereby learning mask-free try-on region localization. To preserve the appearance transfer and video generation priors learned in Stage 1, we freeze the Appearance LoRA and introduce a separate Location LoRA dedicated to mask-free try-on localization.

Stage 3: Training for High-Quality Mask-Free Video Try-on. After the first two stages, the model is already capable of mask-free virtual try-on. However, because Stage 2 relies only on image-level supervision and lacks explicit temporal constraints, the resulting model tends to generate videos with unsatisfactory temporal coherence. To address this issue, we introduce a final fine-tuning stage based on video-level pseudo data. As shown in Fig. 4(c)(d), Stage 3 consists of two sub-stages: Stage 3.1 for pseudo video generation and Stage 3.2 for fine-tuning with pseudo videos. In Stage 3.1, we use the model trained in Stage 1 to construct high-quality video-level pseudo data from unpaired videos. Specifically, we first sample keyframes from unpaired videos and use the multi-frame try-on model to generate the corresponding keyframe try-on images. Meanwhile, we construct garment-agnostic videos by removing the original garments with garment masks. Conditioned on these

keyframe try-on images and garment-agnostic videos, the Stage 1 video generation model synthesizes pseudo videos. In Stage 3.2, we fine-tune the model obtained from Stage 2 using the pseudo videos and their corresponding keyframes as inputs, with the original unpaired videos as supervision. In this stage, explicit video-level temporal supervision allows the model to combine the appearance transfer capability learned in Stage 1 with the mask-free localization capability learned in Stage 2, thereby producing high-quality try-on videos with improved temporal coherence.

4 Experiments

4.1 Experimental Setup

Training Datasets. For the multi-frame try-on model, we use OmniView, MVG [21], and ViViD [7] datasets. From each ViViD video, we sample two frames as multi-view person images to construct multi-view try-on training pairs. In total, we obtain approximately 22K training pairs, each consisting of two person images with different poses and the corresponding garment images. The video generation model uses different training data across stages. For Stage 1, we collect 20K human-centric training videos from public datasets [7, 13, 23] and the Internet. For Stage 2, we sample two frames from each Stage 1 training video to form multi-view person image pairs, and combine them with OmniView to construct 30K training pairs for mask-free try-on region localization. For Stage 3, we select 2,500 videos from the Stage 1 training set and synthesize the corresponding pseudo videos for final fine-tuning.

Evaluation Datasets. For indoor evaluation, we use ViViD-S, which contains 180 samples constructed following the CatV2TON protocol. In addition, we create an in-the-wild benchmark, named WildVVT, consisting of 100 samples that cover diverse human motions, scene conditions, and garment categories.

Implementation Details. For keyframe try-on, we adopt Qwen-Image-Edit-2511 as the Multi-Frame Try-on model and insert a LoRA with rank 64, which is trained for 20K steps. For video generation, we use Wan-Animate as the DiT backbone. In Stage 1, we train an Appearance LoRA ($rank = 128$) for 10K steps. In Stage 2, we freeze the Appearance LoRA, introduce a Location LoRA ($rank = 64$), and train it for 20K steps. In Stage 3, we jointly fine-tune the two LoRA modules for 5K steps. Unless otherwise specified, all models are trained with the AdamW optimizer on 4 NVIDIA RTX Pro 6000 GPUs, using a learning rate of 1×10^{-4} and a batch size of 1. During inference, following DreamVVT, we sample

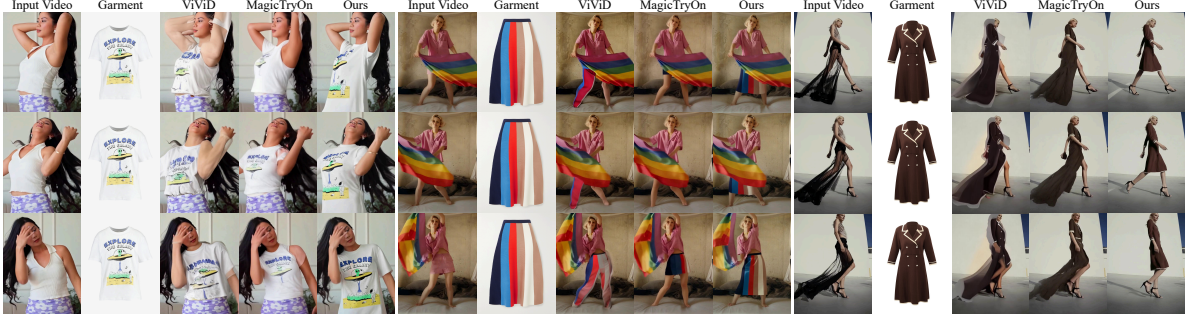


Figure 6: Qualitative comparison on the WildVVT benchmark.

Table 2: Quantitative results on the ViViD-S dataset.

Method	Mask	Unpaired Setting				Paired Setting		
		FVD ^u ↓	ABC ↑	GC ↑	OVQ ↑	FVD ^p ↓	SSIM ↑	LPIPS ↓
ViViD	✓	164.87	3.03	2.26	2.24	139.10	0.842	0.078
CatV2TON	✓	128.53	3.53	2.07	1.98	120.38	0.858	0.077
MagicTryOn	✓	110.66	4.15	2.38	3.12	86.31	0.869	0.068
BooM-VVT (Stage 1)	✓	106.25	4.38	3.36	3.57	88.19	0.872	0.065
BooM-VVT	✗	101.92	4.61	3.95	4.21	79.12	0.891	0.061

two keyframes from each video to balance efficiency and generation quality. For keyframe try-on, our Multi-Frame Try-on model is compatible with Qwen-Image-Edit-Lightning [14], an acceleration LoRA for Qwen-Image-Edit-2511. Thus, we use Qwen-Image-Edit-Lightning for faster inference, with 8 inference steps. We also use the same acceleration LoRA during data construction to speed up pseudo image generation. For video generation, we use 20 inference steps and set the guidance scale to 1.

Evaluation Metrics. Following prior works, we use three widely adopted metrics—FVD [20], SSIM [22], and LPIPS [27]—to evaluate the quality of the generated try-on videos. In addition, to provide a more comprehensive assessment of model performance, we employ Gemini 3.1 Pro as a judge model to rate each result on a scale from 0 (worst) to 5 (best) along three dimensions:

- **Appearance and Background Consistency (ABC):** Measures whether the non-try-on regions, including the person’s identity and the background, remain consistent with the source video.
- **Garment Consistency (GC):** Measures how faithfully the generated garment matches the reference garment in appearance.
- **Overall Video Quality (OVQ):** Measures the overall realism and temporal coherence of the synthesized video.

4.2 Qualitative Comparison

We qualitatively compare BooM-VVT with state-of-the-art video virtual try-on methods, including ViViD [7], CatV2TON [4], and MagicTryOn [12]. As shown in Fig. 5 and Fig. 6, BooM-VVT produces more realistic and temporally coherent try-on videos than previous methods on both ViViD-S and the more challenging WildVVT benchmark. In contrast, CatV2TON and MagicTryOn often exhibit noticeable artifacts due to their reliance on mask-based designs. As illustrated in Fig. 5, when the source and target garments differ substantially in shape, these methods are constrained by the original garment structure and fail to synthesize the correct target garment.

Table 3: Quantitative results on the WildVVT benchmark. The inference time and GPU memory usage are measured on a single NVIDIA A800 when generating a 65-frame video at a resolution of 624×832.

Method	Mask	FVD ↓	ABC ↑	GC ↑	OVQ ↑	GPU Memory	Inference Time
ViViD	✓	301.80	2.13	1.55	1.57	55.8G	142s
CatV2TON	✓	350.72	1.39	1.28	1.23	28.9G	158s
MagicTryOn	✓	250.55	2.12	1.84	1.65	63.5G	342s
BooM-VVT (Stage 1)	✓	241.72	3.96	3.07	3.26	57.9G	281s
BooM-VVT	✗	219.76	4.33	3.72	3.94	57.9G	281s

Table 4: Ablation study on the WildVVT benchmark.

Method	FVD ↓	ABC ↑	GC ↑	OVQ ↑
w/o GSKS	227.88	4.21	3.42	3.85
w/o Frame-Shared 3D-RoPE	224.15	4.34	3.27	3.81
w/o Stage 1	242.61	3.91	3.26	3.66
w/o Stage 2	237.89	4.10	3.37	3.75
Full Model	219.76	4.33	3.72	3.94

Furthermore, as shown in Fig. 6, under highly challenging real-world scenarios involving turning motions and severe occlusions, ViViD and MagicTryOn often fail to preserve plausible garment structures and fine-grained garment details. By comparison, BooM-VVT achieves higher garment fidelity and stronger spatiotemporal consistency across diverse real-world scenarios. More qualitative results are provided in the appendix.

4.3 Quantitative Comparison

Quantitative results are reported in Table 2 and Table 3. Since existing keyframe-driven video virtual try-on methods are not publicly available, a fair direct comparison with them is infeasible. Therefore, we use BooM-VVT (Stage 1) as a baseline for the keyframe-driven setting. Specifically, BooM-VVT (Stage 1) is a variant of our method whose video model is trained only in Stage 1. The results show that the keyframe-driven methods, BooM-VVT (Stage 1) and BooM-VVT, consistently outperform the non-keyframe-driven baselines on both ViViD-S and WildVVT, demonstrating the effectiveness of the keyframe-driven paradigm. However, reliance on masks remains a major limitation of existing methods. When the source and target garments differ significantly in shape, the masks often hinder

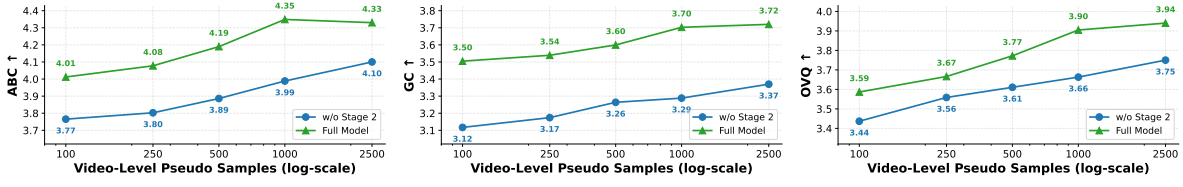


Figure 7: Effectiveness of Stage 2 under different amounts of video-level pseudo data on the WildVVT benchmark.

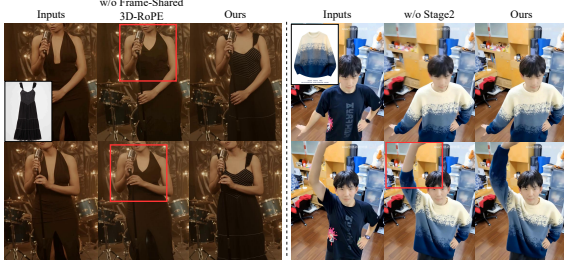


Figure 8: Qualitative results of ablation study on the WildVVT benchmark.

accurate garment transfer, leading to lower GC scores. This is consistent with the qualitative results in Fig. 5. The limitation becomes more pronounced on the more challenging WildVVT benchmark, where large pose variations and severe occlusions make it particularly difficult to obtain accurate masks. As a result, existing methods often fail to maintain appearance consistency and exhibit temporal jittering and flickering, leading to lower ABC and OVQ scores. By contrast, BooM-VVT remains robust in unconstrained real-world scenarios and achieves the best performance across all metrics. Table 3 also reports inference time and GPU memory usage. Compared with MagicTryOn, BooM-VVT delivers clearly better performance without additional inference cost.

4.4 Ablation Study

We conduct ablation studies on the WildVVT benchmark to evaluate the effectiveness of the key components in BooM-VVT, covering both architectural design and training strategy.

On the model architecture. We first examine the two key architectural components of BooM-VVT, namely Garment-Sensitive Keyframe Sampling (GSKS) and Frame-Shared 3D-RoPE. As shown in Table 4, removing either component degrades the overall performance, especially in terms of Garment Consistency (GC). For the *w/o GSKS* variant, we follow DreamVVT and select keyframes mainly based on global pose differences. For the *w/o Frame-Shared 3D-RoPE* variant, keyframe try-on images and their corresponding target video frames are assigned different positional encodings, which weakens the attention between them and hinders accurate garment appearance transfer. Although this variant achieves a slightly higher ABC score, its GC, OVQ, and FVD all deteriorate, suggesting that it tends to preserve non-try-on regions at the expense of garment transfer quality. We further present qualitative comparisons for Frame-Shared 3D-RoPE in Fig. 8, where the *w/o Frame-Shared 3D-RoPE* variant fails to reconstruct the correct shoulder straps and misses the white stripes on the chest.

On the training strategy. We further analyze the proposed multi-stage training strategy, with results reported in Table 4 and Fig. 7. Here, *w/o Stage 1* denotes the variant in which the Appearance LoRA does not undergo Stage 1 training, whereas *w/o Stage 2* denotes the variant in which the Location LoRA does not undergo Stage 2 training. To enable a fair comparison under the same data construction cost, we train the *w/o Stage 2* variant using 400 extra video-level pseudo samples, at a cost comparable to that of the 30K image-level pseudo samples used in Stage 2.

As shown in Table 4, removing Stage 1 causes the largest performance drop, showing that Stage 1 is essential for enabling high-quality keyframe-driven try-on video generation. Removing Stage 2 also leads to clear performance degradation, showing that it is important for accurate mask-free try-on region localization. This is also reflected in the right panel of Fig. 8, where the *w/o Stage 2* variant fails to localize the try-on region correctly, resulting in incorrect garment structures. Fig. 7 further shows that Stage 2 substantially improves fine-tuning efficiency under the same comparable data construction cost. Across all pseudo-video scales, the full model consistently outperforms the *w/o Stage 2* variant. Notably, the full model trained with only 500 video-level pseudo samples already achieves performance comparable to that of the *w/o Stage 2* variant trained with 2,500 video-level pseudo samples. We attribute this advantage to the broad coverage of garment types, human poses, and real-world scenes provided by large-scale image-level pseudo data, which helps the model generalize better to real-world scenarios.

5 Conclusion

In this paper, we present BooM-VVT, a novel mask-free video virtual try-on framework built on the keyframe-driven paradigm. To achieve mask-free video virtual try-on, we introduce a multi-stage training strategy that leverages readily available image-level pseudo data to learn mask-free try-on region localization, substantially reducing the reliance on costly video-level pseudo data. To improve garment consistency, we further introduce Garment-Sensitive Keyframe Sampling (GSKS) and Frame-Shared 3D-RoPE. GSKS selects informative keyframes based on garment-relevant body regions, while Frame-Shared 3D-RoPE improves garment appearance transfer by explicitly establishing spatiotemporal correspondences between keyframes and their corresponding target video frames. In addition, we construct OmniView, a large-scale multi-view try-on dataset that provides more complete garment references and broader garment category coverage, enabling reliable try-on generation under challenging viewpoints and diverse tasks. Extensive experiments demonstrate that BooM-VVT outperforms existing methods in temporal consistency and garment fidelity under unconstrained real-world scenarios.

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631* (2025).
- [2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025).
- [3] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. 2025. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055* (2025).
- [4] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. 2025. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325* (2025).
- [5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. 2019. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1161–1170.
- [6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.
- [7] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. 2024. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794* (2024).
- [8] Yutong Feng, Linlin Zhang, Hengyuan Cao, Yiming Chen, Xiaoduan Feng, Jian Cao, Yuxiong Wu, and Bin Wang. 2025. Omnity: Virtual try-on anything without masks. *arXiv preprint arXiv:2508.13632* (2025).
- [9] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. 2025. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 19085–19096.
- [10] Qingdong He, Xueqin Chen, Yanjie Pan, Peng Tang, Pengcheng Xu, Zhenye Gan, Chengjie Wang, Xiaobin Hu, Jiangning Zhang, and Yabiao Wang. 2025. The devil is in the details: Enhancing Video Virtual Try-On via Keyframe-Driven Details Injection. *arXiv preprint arXiv:2512.20340* (2025).
- [11] Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. 2024. Fashion-vdm: Video diffusion model for virtual try-on. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.
- [12] Guangyuan Li, Siming Zheng, Hao Zhang, Jinwei Chen, Junsheng Luan, Binkai Ou, Lei Zhao, Bo Li, and Peng-Tao Jiang. 2025. MagicTryOn: Harnessing Diffusion Transformer for Garment-Preserving Video Virtual Try-on. *arXiv preprint arXiv:2505.21325* (2025).
- [13] Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. 2025. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7752–7762.
- [14] ModelTC. 2025. Qwen-Image-Lightning: Speed up Qwen-Image model with distillation. Retrieved April 2, 2026 from <https://github.com/ModelTC/Qwen-Image-Lightning> GitHub repository. Includes Qwen-Image-Edit-Lightning models, weights, and workflows..
- [15] Hung Nguyen, Quang Qui-Vinh Nguyen, Khoi Nguyen, and Rang Nguyen. 2025. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6200–6208.
- [16] Yanjie Pan, Qingdong He, Lidong Wang, Bo Peng, and Mingmin Chi. 2025. Once Is Enough: Lightweight DiT-Based Video Virtual Try-On via One-Time Garment Appearance Injection. *arXiv preprint arXiv:2510.07654* (2025).
- [17] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [18] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. 2025. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6795–6804.
- [19] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yinfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- [20] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018).
- [21] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. 2025. Mv-vton: Multi-view virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 7682–7690.
- [22] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [23] Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, et al. 2024. Humanvid: Demystifying training data for camera-controllable human image animation. *Advances in Neural Information Processing Systems* 37 (2024), 20111–20131.
- [24] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).
- [25] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4210–4220.
- [26] Jianhao Zeng, Yancheng Bai, Ruidong Chen, Xuanpu Zhang, Lei Sun, Dongyang Jin, Ryan Xu, Nannan Zhang, Dan Song, and Xiangxiang Chu. 2025. Eevee: Towards Close-up High-resolution Video-based Virtual Try-on. *arXiv preprint arXiv:2511.18957* (2025).
- [27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.
- [28] Wei Zhang, Yeying Jin, Xin Li, Yan Zhang, Xiaofeng Cong, Cong Wang, Fengcai Qiao, and Zhichao Lian. 2026. UniFit: Towards Universal Virtual Try-on with MLLM-Guided Semantic Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 12816–12824.
- [29] Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and An-An Liu. 2025. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26399–26408.
- [30] Tongchun Zuo, Zaiyu Huang, Shuliang Ning, Ente Lin, Chao Liang, Zerong Zheng, Jianwen Jiang, Yuan Zhang, Mingyuan Gao, and Xin Dong. 2025. Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. *arXiv preprint arXiv:2508.02807* (2025).