Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MFKP-VVTO: Mask-Free Video Virtual Try-On via Keyframe Guidance

Xiangyan Fu
Qinghai Normal University
Weihua Pu
Shaanxi Fashion Engineering University
Shengling Geng
2023333341005@stu.qhnu.edu.cn

Qinghai Normal University
Xiaojuan Dang
Shaanxi Fashion Engineering University
Feizhuan Fan
Shaanxi Fashion Engineering University

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11417454/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MFKP-VVTO: Mask-Free Video Virtual Try-On via Keyframe Guidance

Xiangyan Fu[1,2], Weihua Pu[2], Shengling Geng * †[1], XiaoJuan Dang[2], and FeiZhuan Fan[2]

[1]Qinghai Normal University, Xining, 81000, Qinghai, China
[2]Shaanxi Fashion Engineering University, Xi'an, 712000, Shaanxi, China

Abstract

Video virtual try-on aims to accurately transfer a target garment image onto a person in a video, generating try-on results that seamlessly align with human poses and dynamic motion. However, existing methods rely heavily on mask priors, generalize poorly, suffer from unstable temporal semantics, and lack authentic paired training data. Consequently, they yield suboptimal generation performance in challenging scenarios involving crossed limbs, large-scale pose variations, and self-occlusions. To address these limitations, we propose a keyframe-guided mask-free video virtual try-on framework, termed MFKP-VVTO. The proposed method comprises modules for pseudo-pair construction, dual-metric adaptive keyframe sampling (DA-KFS), mask-free keyframe try-on, and keyframe-guided video generation. Specifically, we leverage the pre-trained ViViD model to construct pseudo-paired samples, effectively alleviating the scarcity of real annotated data. DA-KFS integrates weighted joint distances and pose feature similarities to select representative keyframes, capturing global motion through sparse anchor frames. Furthermore, we design a mask-free try-on module built on SD-UNet to fuse dual-modal representations from CLIP and Qwen-VL and characterize garment textures and structures. We also introduce an inter-keyframe garment consistency loss that acts strictly as a regional constraint during training and does not participate in forward inference, thereby satisfying the mask-free inference requirement. Finally, the high-quality keyframe try-on results serve as temporal conditions to steer the video diffusion model, facilitating inter-frame semantic propagation. Extensive experiments on the ViViD, VVT, VITON-HD, and DressCode datasets show that MFKP-VVTO outperforms competitive baselines in occlusion and complex motion scenarios, demonstrating superior generalization and temporal coherence and providing an effective solution for mask-free video virtual try-on under complex dynamic conditions.

Keywords: Diffusion-Based, Video Virtual Try-On, Occlusion Resolution

1 Introduction

As digital technologies advance rapidly, e-commerce has profoundly reshaped consumer shopping behaviors, with the online apparel market showing particular vitality. Nonetheless, conventional static virtual try-on [1–3, 10, 16, 18, 41, 43] remains largely confined to static image generation, producing try-on results from a fixed viewpoint using merely a single person image alongside a target garment image. Such static representations fail to capture clothing's dynamic behavior in realistic scenarios—such as hemlines swaying during walking or cuffs shifting during arm movements. Consequently, consumers cannot fully assess the silhouette, fit, and overall drape of garments, which not only impedes purchasing decisions but also drives high return rates in online shopping. To address these limitations, video virtual try-on technology has emerged [4–8, 37]. By transferring target garments onto persons in dynamic videos, it delivers continuous try-on results across diverse viewpoints and motions, significantly bridging the gap between online shopping and offline fitting experiences. However, when the source video involves substantial body motion or severe occlusions, existing video virtual try-on methods struggle to maintain satisfactory visual fidelity and often produce blurred textures and geometric distortions.

To address these challenges, this paper proposes a keyframe-guided mask-free video virtual try-on framework. Specifically, the framework constructs a large-scale paired pseudo dataset using the pre-trained mask-based ViViD model [8]. Training the core network on these mask-free samples enables accurate garment transfer and robust perception of spatiotemporal dynamics, markedly enhancing reconstruction fidelity in occluded regions. Furthermore, we improve Real-

*Corresponding author: 202333341005@stu.qhnu.edu.cn
†Corresponding Address: Qinghai Normal University

VVT [36]'s keyframe selection strategy by introducing a dual-metric mechanism that combines human skeleton joint distances and pose feature similarities, improving selection precision. Skeleton joint distance serves as the primary metric, while pose feature similarity compensates for inaccuracies in skeletal joint estimation under complex postures, enabling more robust occlusion handling. The overall try-on pipeline has two stages: static try-on for keyframes, followed by keyframe-guided dynamic video try-on. Comprehensive training and empirical evaluations demonstrate that the proposed method effectively mitigates the occlusion bottleneck in video virtual try-on.

Our contributions are summarized as follows:

(1) We propose MFKP-VVTO, a keyframe-guided mask-free video virtual try-on framework. During inference, the network operates without mask inputs, using masks only as regional constraints in the training loss. This design avoids propagating segmentation errors inherent in traditional mask-based paradigms and improves garment reconstruction fidelity in self-occlusion scenarios.

(2) We design a Dual-metric Adaptive Keyframe Sampling (DA-KFS) algorithm, which jointly integrates weighted skeletal joint distances and pose feature similarities for keyframe selection. By leveraging pose feature similarity to compensate for estimation errors from skeletal keypoint detection failures, the algorithm produces a keyframe sequence with comprehensive motion coverage and minimal redundancy, providing reliable temporal anchors for video generation.

(3) We establish a two-stage try-on paradigm: it first achieves high-precision mask-free try-on on keyframes, and subsequently utilizes the latent features of keyframes as temporal anchors to guide the generation of non-keyframes via cross-frame attention. We incorporate multiple consistency losses to regularize inter-frame garment appearance and human poses, effectively suppressing garment semantic drift.

(4) We construct a large-scale paired pseudo-training dataset based on the pre-trained ViViD model. We conduct comprehensive quantitative, qualitative, and ablation experiments on public benchmarks to validate each module's effectiveness and demonstrate the proposed framework's superiority over both mask-based and mask-free baselines.

2 Related work

2.1 Mask-based Video Virtual Try-On

Mask-based video virtual try-on typically frames the try-on task as mask-guided video inpainting, using a human-agnostic mask generated from human parsing to constrain the garment replacement region explicitly and feeding this mask as an input condition during network inference. This category of methods has evolved across three developmental stages: Generative Adversarial Networks (GANs [29]), U-Net-based diffusion models, and temporal generation based on Diffusion Transformers (DiTs). Early approaches were predominantly GAN-based; for example, FW-GAN [7], the pioneering video virtual try-on framework, introduced an optical flow prediction module from Video2Video [27] to synthesize coherent future frames by warping historical frames. Yet, it had limited ability to recover fine-grained textures. Subsequently, researchers introduced diffusion models to the task to replace conventional GAN architectures. Among them, GPD-VVTO [4] was the first to explore diffusion models for video virtual try-on, eliminating the explicit standalone warping module and retaining garment details and bodily alignment through a multi-module collaborative mechanism. Subsequent works, such as ViViD [8], RealVVT [36], and WildVidFit [26], further leveraged diffusion models to improve generation quality and stability, yet remained prone to the adverse effects of garment segmentation errors during the generation process. More recently, efforts like DreamVVT [9], CatV^{2}TON [5], Swifttry [38], Eevee [39], TripVVT [40], LiveVVT [42], and KeyTailor [33] have adopted Diffusion Transformer architectures to bolster temporal modeling capabilities.

Although KeyTailor also introduces a keyframe mechanism, its inference phase still requires an agnostic mask as an input condition, placing it within traditional mask-dependent paradigms and fundamentally distinguishing it from our proposed keyframe-guided paradigm, which dispenses with masks during inference. Overall, in scenarios with large-scale pose variations or limb self-occlusions, mask-based video virtual try-on methods easily produce generative artifacts—such as blurred textures, local structural distortions, and inter-frame flickering—because of boundary inaccuracies or cross-frame inconsistencies in human-parsing masks. This inherent architectural deficiency has propelled in-depth research into mask-free video virtual try-on tailored for complex dynamic scenarios.

2.2 Mask-free Video Virtual Try-On

To overcome the inherent limitations of conventional mask-based paradigms, mask-free video virtual try-on methods have increasingly emerged. In this paper, we define mask-free video virtual try-on as follows: the network receives no segmentation masks as input during inference; masks are used only during training to enforce local-region constraints in the loss function and do not participate in forward propagation. Under this paradigm, models eliminate reliance on segmentation masks at inference time, implicitly perceiving and locating garment-editing regions by leveraging human pose, visual features, or multimodal semantics.

Existing mask-free approaches have explored diverse technical routes: PEMF-VVTO [11] adopts a sparse point-enhanced alignment paradigm, utilizing TAP-Net for cross-frame point tracking and designing point-enhanced attention modules to achieve garment transfer and temporal smoothing. Although it operates without masks during inference by establish-

ing spatial correspondences via sparse feature points, full-frame point tracking incurs substantial computational overhead. It remains susceptible to feature drift under complex motions.BooM-VVT [31] builds on a U-Net video diffusion architecture and operates entirely without human-parsing masks during inference. It builds a garment appearance memory bank to store reference clothing features and reuses them across frames to mitigate semantic drift. However, without an explicit motion-anchoring mechanism, the memory bank is prone to noise accumulation under dramatic pose variations, causing gradual shifts in garment color and style.UniVVT [32] is a multimodal, large-model-guided, mask-free method that leverages vision-language (VL) foundation models to locate garment-editing regions implicitly. Its inference pipeline dispenses with geometric conditions such as masks or pose keypoints, directly processing video sequences, reference garments, and textual instructions to streamline preprocessing. Nonetheless, it relies heavily on the VL model's semantic comprehension, which often leads to garment misalignment in extremely fast dynamic scenes.FlowVVTON [34] introduces optical flow guidance by employing optical flow strictly as a training supervision signal rather than feeding it into forward propagation during inference. It regularizes multi-scale cross-frame feature alignment via a flow-warping latent loss, thereby safeguarding the temporal consistency of generated videos.

In summary, PEMF-VVTO relies on full-frame point tracking, BooM-VVT depends on garment memory banks, UniVVT hinges on the implicit semantics of multimodal large models, and FlowVVTON incorporates optical flow as training supervision. A common deficiency of these methods is the lack of explicit, high-fidelity anchor-frame constraints, leaving long video sequences vulnerable to garment temporal drift. To address this challenge, this paper introduces DA-KFS to generate sparse, high-quality keyframes as temporal anchors. Operating entirely without mask inputs throughout the inference pipeline, our framework guides non-keyframe synthesis using latent features from high-fidelity keyframe try-on results, balancing the robustness of the mask-free paradigm with the temporal stability of dynamic long-video generation.

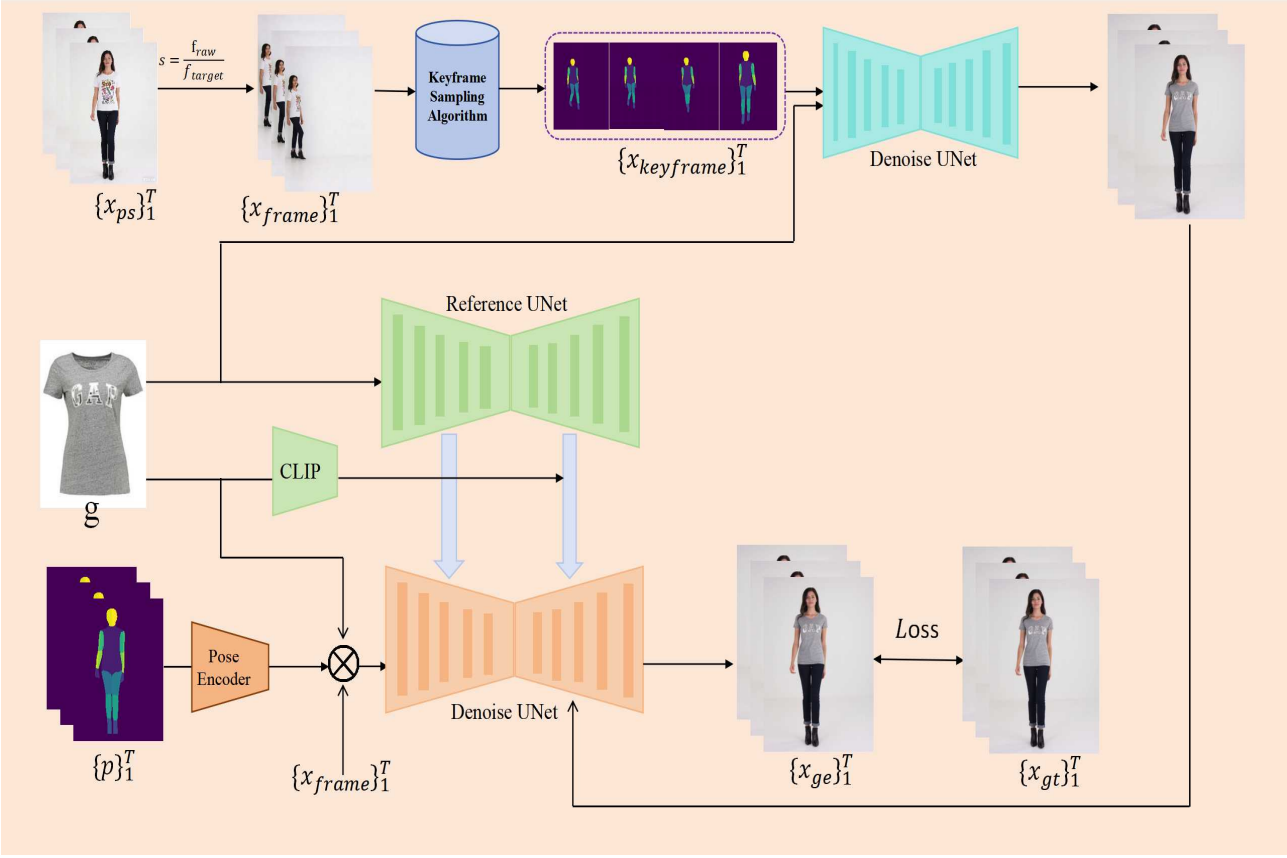


Figure 1: Overview of the overall MFKP-VVTO framework. The proposed method comprises four core modules: pseudo-pair construction, Dual-metric Adaptive Keyframe Sampling (DA-KFS), mask-free keyframe try-on, and keyframe-guided video generation. DA-KFS selects representative keyframes, after which we integrate dual-modal representations from CLIP and Qwen-VL to perform mask-free keyframe try-on, introducing an inter-keyframe garment consistency loss to enforce cross-frame appearance stability. Finally, the keyframe try-on results serve as temporal anchors and, together with skeletal pose information, synthesize the final video virtual try-on results. The subject images displayed in this figure are obtained from the publicly accessible ViViD datasets.

3 Methods

To address the limitations of existing video virtual try-on methods—including heavy reliance on mask priors, limited generalization, unstable temporal semantics, and scarce authentic paired training data—this paper proposes a keyframe-guided, mask-free video virtual try-on framework, MFKP-VVTO (Figure 1). The overall architecture comprises four core components: pseudo-pair construction, dual-metric adaptive keyframe sampling, mask-free keyframe try-on, and keyframe-guided video generation. First, we leverage a pre-trained model to construct video pseudo-paired samples, compensating for the lack of real annotated data and providing rich supervision for mask-free training. Second, we introduce the Dual-metric Adaptive Keyframe Sampling (DA-KFS) algorithm, which integrates weighted joint distances with pose feature similarities. DA-KFS adaptively selects representative keyframes with pronounced pose variations and high dynamic significance, capturing the video's global motion dynamics with sparse keyframes and reducing the complexity of subsequent temporal modeling. Building upon these frames, we construct a mask-free keyframe try-on module that employs SD-UNet as the generative backbone, coupled with a Reference UNet to extract garment features. This module fuses CLIP [22] visual embeddings with Qwen-VL [25] high-level semantic representations to capture garment appearance textures and structural attributes, respectively, realizing cross-modal complementarity and delivering high-fidelity single-frame try-on results without mask inputs. During training, we introduce an inter-keyframe garment consistency loss to regularize the uniformity of garment representations across diverse keyframe poses, effectively mitigating cross-frame texture flickering and semantic drift. Crucially, this loss strictly enforces regional constraints via masks without participating in the network's forward inference, strictly conforming to the mask-free inference paradigm. Finally, the semantically stable, detail-preserving keyframe try-on results serve as temporal conditions to steer the video diffusion model, facilitating inter-frame semantic propagation and temporal feature completion. Consequently, the proposed framework synthesizes video virtual try-on results with superior visual fidelity, robust temporal coherence, and improved generalization.

3.1 Stable Diffusion

Stable Diffusion [14] is currently one of the most widely adopted Latent Diffusion Models (LDMs [21]) in the field. To improve training efficiency while reducing parameter count and computational overhead, this paper builds the generative network on LDMs, which primarily comprise a Variational Autoencoder (VAE [19]) and a denoising U-Net. The VAE maps high-dimensional image data into a low-dimensional latent space. By compressing the representation while faithfully preserving core features—such as garment textures and human poses—it reduces the computational burden of iterative denoising and significantly improves training efficiency. Given an image x and a textual prompt y, the VAE encoder $\mathcal{E}(\cdot)$ first encodes the image into the latent space $z_0 = \mathcal{E}(x)$, and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is subsequently added to the latent representation. Conditioned on the text embeddings encoded by the CLIP text encoder $\tau_\theta(y)$, the denoising U-Net $\epsilon_\theta(\cdot)$ learns to predict the added Gaussian noise, thereby progressively denoising the noisy latent into a clean latent representation. Finally, the VAE decoder $\mathcal{D}(\cdot)$ reconstructs the image representation from the latent space back into the pixel space.

The training loss of the denoising U-Net can be formulated as:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c(y))\|_2^2\right] \quad (1)$$

where z_t denotes the noisy latent embedding at timestep t, and $c(\cdot)$ is the condition-encoding function for the garment feature, which provides precise semantic constraints for the denoising process.

3.2 Pseudo-sample generation

Inspired by [11, 30], we construct pseudo-paired training data utilizing the pre-trained video try-on model ViViD. Given a source person video $\{x_{ps}\}_1^T$ and a garment image g, ViViViD generates the video try-on results $\{x_{ge}\}_1^T$ during inference, along with the corresponding cloth-agnostic video a_1^T, agnostic mask sequence m_1^T, and human pose sequence p_1^T extracted from $\{x_{ps}\}_1^T$. Following ViViD's methodology, we employ a human parsing model [24] to extract the foreground masks a_1^T and m_1^T, and use DensePose [13] to obtain the human pose representations p_1^T. We apply the Segment Anything Model (SAM [23]) to segment the garment mask m_g from the garment image g. After obtaining all necessary inputs, we perform virtual try-on operations by randomly selecting same-category garments from public image and video virtual try-on datasets, ultimately constructing a large-scale, pseudo-paired video training dataset. Because these pseudo samples are synthesized by inferring with the pre-trained mask-based ViViD model, they inevitably inherit the generative biases of the pre-trained network. We further discuss this potential domain shift issue in the ablation study and limitation sections.

3.3 Mask-free Keyframe virtual try-on

Keyframe sampling

Existing keyframe sampling methods for video virtual try-on exhibit pronounced limitations. Rerender-A-Video [35] selects keyframes via uniform sampling and fails to capture dynamic motion variations in the human body adaptively. Although several approaches [9, 34] optimize the sampling strategy using inter-frame similarities—partially mitigating the deficiencies of uniform sampling—they struggle to repre-

sent the dynamic characteristics of human postures effectively. RealVVT [36] screens keyframes based on the L_2 distance of dense pose frames, relying solely on an isolated skeletal metric to measure inter-frame motion disparities. All these methods rely on a single criterion; consequently, in scenarios with complex pose transitions and human occlusions, they fail to quantify inter-frame motion changes accurately. This deficiency leads to unstable sampling, which easily induces garment detail distortions and temporal inconsistencies, falling short of the generation requirements for photo-realistic video virtual try-on. In video virtual try-on tasks, human dynamic poses and inter-frame motion discrepancies are key determinants of synthesis quality. Human self-occlusions and abrupt pose transitions render isolated skeletal evaluation criteria ineffective, triggering garment texture distortions, detail loss, and video flickering that ultimately compromise the realism and stability of the synthesized try-on videos.

To address these challenges, this paper proposes a Dual-metric Adaptive Keyframe Sampling strategy, termed DA-KFS. We first designate a pose-stable frontal frame as the reference anchor, uniformly downsample the source video, and use OpenPose [12] to extract 2D coordinates for 18 skeletal joints per frame to capture human motion dynamics. To address the vulnerability and inaccuracy of isolated skeletal metrics under occlusion, we establish a dual-metric evaluation mechanism that couples weighted geometric joint distances with pose feature similarity. This mechanism uses skeletal geometric variations as the primary criterion and leverages pose semantic features to compensate for keypoint-detection inaccuracies, thereby precisely quantifying inter-frame motion transitions. In the formulation, w_i represents the weight assigned to the i-th human joint. Because limb joints are more critical for characterizing dynamic bodily movements, we assign them higher weights. Conversely, head-related keypoints are more susceptible to occlusion and show less stable detection, so we assign them lower weights. In our implementation, we set the weights as follows: shoulders, elbows, hips, knees, wrists, and ankles are set to 1.0; the neck to 0.6; and the nose, eyes, and ears to 0.2. All weights are subsequently normalized to satisfy $\sum w_i = 1$. Building on this metric, we use an adaptive threshold to identify candidate keyframes with pronounced motion variation, enforce a minimum frame-interval constraint to eliminate temporal redundancy, and incorporate a motion-coverage verification mechanism to replenish missing action frames. As a result, this yields a keyframe sequence with minimal redundancy, strong representational capacity, and comprehensive motion coverage, providing stable, reliable temporal features for subsequent video virtual try-on generation.

The DA-KFS algorithm is described as follows:

Algorithm 1: DA-KFS: Dual-metric Adaptive Key Frame Sampling

Input : N-frame pose sequence $\{J_t\}$, pose features $\{F_t\}$, balance weight $\alpha \in [0.6, 0.8]$, threshold coefficient k, minimum frame interval $L_{\min}$

Output: Final keyframe set Ω_{final}

`// Step 1: Compute inter-frame comprehensive disparity`

for $t \leftarrow 1$ **to** $N-1$ **do**

  $D_{\text{joint}}(t, t+1) = \sqrt{\sum_{i=1}^{18} w_i \cdot [(x_{t,i} - x_{t+1,i})^2 + (y_{t,i} - y_{t+1,i})^2]}$;

  $S_{\text{feat}} = \frac{F_t \cdot F_{t+1}}{\|F_t\|_2 \|F_{t+1}\|_2}$, $D_{\text{feat}} = 1 - S_{\text{feat}}$;

  $D_{\text{total}}(t, t+1) = \alpha D_{\text{joint}} + (1-\alpha) D_{\text{feat}}$;

end for;

`// Step 2: Candidate frame screening via adaptive threshold`

$\mu_D = \text{mean}(D_{\text{total}})$, $\sigma_D = \text{std}(D_{\text{total}})$, $T = \mu_D + k\sigma_D$;

$\Omega_{\text{candidate}} = \{t+1 \mid D_{\text{total}}(t, t+1) > T\}$;

`// Step 3: Redundancy elimination via minimum frame interval`

Sort $\Omega_{\text{candidate}}$ in chronological order; $\Omega_{\text{filtered}} \leftarrow \emptyset$;

for $t_{curr} \in \Omega_{candidate}$ **do**

  if $\Omega_{filtered} \neq \emptyset$ ***and*** $t_{curr} - t_{last} < L_{\min}$ **then**

    Retain the frame with a larger D_{total};

  else

    Add t_{curr} to Ω_{filtered};

end for;

`// Step 4: Motion coverage verification and missing frame supplement`

$\Omega_{\text{supplement}} \leftarrow \emptyset$;

for *uncovered motion category* **do**

  Append the frame with maximum D_{total} in the corresponding sub-sequence to $\Omega_{\text{supplement}}$;

$\Omega_{\text{final}} = \Omega_{\text{filtered}} \cup \Omega_{\text{supplement}}$;

return Ω_{final};

Mask-Free Keyframe Virtual Try-On

For the selected keyframes, our method uses the Stable Diffusion U-Net as the generative backbone for keyframe try-on, while leveraging a Reference U-Net to extract features of the reference garment. This backbone network takes mask-free human keyframe images, the target garment image, CLIP-extracted visual garment features, high-level textual semantic descriptions of the garment generated by Qwen-VL, and 2D skeletal pose sequences of the keyframes as multi-conditional inputs. It synthesizes single-frame virtual try-on images within the latent space, generating high-fidelity keyframe try-on results that serve as temporal con-

ditioning priors for the subsequent video generation stage. Specifically, CLIP is employed to encode the global visual representations of the garment image, capturing appearance information such as garment colors and textures; Qwen-VL extracts high-level semantic textual descriptions regarding garment styles, silhouettes, and cuts from the garment image, serving as auxiliary semantic priors to compensate for CLIP's deficiency in fine-grained garment semantic comprehension. In our framework, we freeze the weights of the Reference U-Net, the CLIP image encoder, and the Qwen-VL model; only the SD-UNet denoising network is fine-tuned. The masks M_i and M_j are used only for regional filtering over garment areas within the loss function and are never fed into the network during forward propagation. Inter-Keyframe Garment Consistency Loss:

$$\mathcal{L}_{\text{consist}} = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} \text{LPIPS}(f_i \odot M_i, f_j \odot M_j) \tag{2}$$

where K denotes the total number of selected keyframes; f_i and f_j represent the try-on output images corresponding to the i-th and j-th keyframes, respectively; M_i and M_j denote the garment region masks of the corresponding keyframes; $\odot$ denotes element-wise pixel multiplication used for mask filtering to preserve the garment region solely within the image; and $\text{LPIPS}(\cdot,\cdot)$ is the perceptual distance metric between two garment region images. This loss regularizes the garment appearance across all keyframe pairs to remain consistent, effectively suppressing abrupt mutations in garment textures and styles between keyframes and ensuring temporal stability in the subsequent video generation. Training Loss for Mask-Free Keyframe Try-On:

$$L_{\text{total}} = l_{\text{LDM}} + \lambda \cdot L_{\text{consist}} \tag{3}$$

where $\mathcal{L}_{LDM}$ denotes the standard latent diffusion loss, and λ is a balancing hyperparameter used to trade off between the noise reconstruction objective and the inter-keyframe garment consistency constraint.

3.4 Keyframe-Guided Mask-Free Video Virtual Try-On

We use supervised learning, using pseudo try-on videos generated by the pre-trained video model ViViD as model input and treating the corresponding try-on results as ground-truth supervision. The model input comprises a full-frame 2D skeletal pose sequence and dual-modal target garment representations, and the backbone network adopts an SD+UNet-based video diffusion architecture. During video generation, latent features from keyframe try-on results sampled via DA-KFS serve as temporal anchors, providing detailed garment references for non-keyframes. We leverage high-level garment semantic representations extracted by Qwen-VL as semantic guidance to synthesize try-on images for non-keyframes. Specifically, we compute the cosine similarity between the skeletal poses of non-keyframes and each keyframe to select the most pose-compatible keyframe as the guiding anchor. During the diffusion denoising of non-keyframes, we fuse cross-frame attention between the latent features of the anchor keyframe and the current frame, driving the non-keyframe garment details to align with the anchor. Concurrently, we encode human poses and garment-fitting semantics from keyframes as temporal conditions and integrate them into each diffusion iteration, constraining garment deformation across adjacent frames to follow human motion dynamics. During inference, users only need to provide the source person video and the reference garment image to obtain the video try-on results. Because the inference phase dispenses with human-agnostic masks, this mask-free model can effectively preserve complex human postures from the source video and achieve natural and plausible garment replacement in the try-on region, outperforming mask-based video virtual try-on methods in scenarios involving complex poses and self-occlusions.For the complete video generation training, the overall loss function is formulated as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LDM}} + \lambda_1 \cdot \mathcal{L}_{\text{align}} + \lambda_2 \cdot \mathcal{L}_{\text{pose}} \tag{4}$$

where $\mathcal{L}_{\text{LDM}}$ denotes the base loss of the latent diffusion model; $\mathcal{L}_{\text{align}}$ represents the cross-frame garment alignment loss that enforces appearance consistency between non-keyframes and anchor keyframes; $\mathcal{L}_{\text{pose}}$ is the pose consistency loss ensuring that the human posture in generated frames aligns faithfully with the original video; and λ_1, λ_2 are the weighting coefficients corresponding to each respective loss term.

4 Experimental Results

ng

4.1 Datasets

In this paper, model training is conducted on public image benchmarks (VITON-HD [2] and DressCode [20]) as well as video datasets (VVT [7] and ViViD [8]); the VVT and ViViD datasets are employed as test sets for performance evaluation to ensure a fair comparison with existing competing methods. To comprehensively validate the effectiveness of the proposed framework, two distinct evaluation protocols are designed: (1) conducting comparative experiments following the input configurations of conventional mask-based baseline methods to demonstrate the performance superiority of our mask-free model; and (2) performing evaluations under the mask-free setting on synthetic pseudo-paired datasets to verify the disruptive artifacts induced by the strong mask prior.

4.2 Evaluation Metrics

For the VITON-HD and DressCode image datasets, we use Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity (SSIM), and Fréchet Inception Distance (FID) as frame-level metrics for image-level evaluation. On the VVT and ViViD video datasets, we use Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity (SSIM), and Video Fréchet Inception Distance (VFID) to assess overall video quality comprehensively. Specifically, VFID jointly evaluates both the spatial fidelity of individual frames and inter-frame temporal coherence, reflecting the overall synthesis quality of the generated videos; a lower VFID indicates higher visual fidelity and smoother temporal continuity across frames.

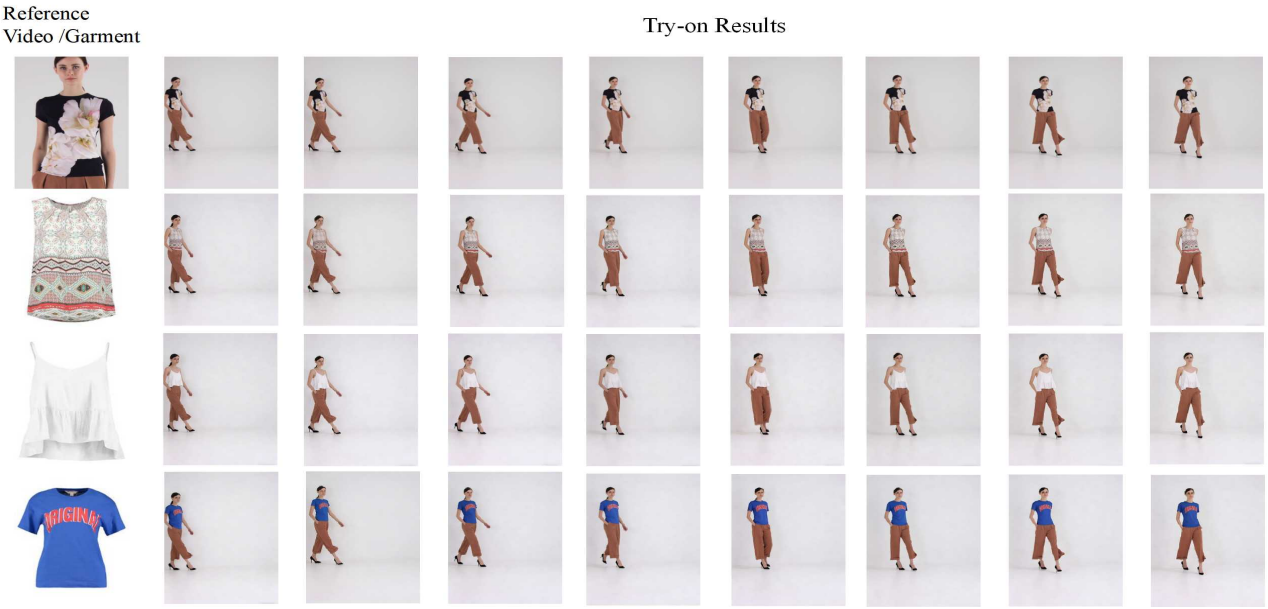


Figure 2: Try-on Results of MFKP-VVTO.The visual training results shown are derived from the publicly accessible ViViD dataset.

4.3 Implementation Details

We initialize the denoising U-Net with pre-trained weights from Stable Diffusion v1.5 [14] and keep the CLIP image encoder consistent with the pre-trained ViViD model. We fine-tune the denoising U-Net with the AdamW optimizer and a fixed learning rate of 5×10^{-5}. We run all experiments on four NVIDIA RTX 4090 GPUs, with 20,000 training iterations in total, and uniformly resize all training data to 512×384. The model uses a three-stage progressive training strategy with varying proportions of image and video samples across phases, setting the video ratio to 0.3, 0.9, and 0.5 for each stage to enable step-by-step acquisition of single-frame try-on and temporal generation capabilities. The hyperparameters of DA-KFS are set to $\alpha = 0.7$, $k = 1.0$, and $L_{\min} = 8$. Loss function hyperparameter settings: $\lambda = 0.1$, $\lambda_1 = 0.5$, $\lambda_2 = 1.0$. Figure 2 illustrates the qualitative training results of the MFKP-VVTO framework on the ViViD video benchmark.

4.4 Quantitative Results

To conduct a comprehensive, fair comparison, this paper presents two distinct experimental evaluation protocols. The first protocol follows the input configurations of conventional mask-based methods to compare our variants—ViViD+MF (mask-free framework only) and ViViD+MF+KPG (mask-free framework with keyframe guidance)—against existing baselines, with results reported in Tables 1 and 2. The second protocol evaluates synthetic pseudo-paired datasets by removing the strong mask prior, providing a more objective reflection of the mask-free paradigm's performance, with the corresponding results in Table 3.

As evidenced by the results in Table 1 and Table 2, even under the evaluation setting where strong mask priors participate in the input, our ViViD+MF (mask-free framework only) achieves noticeable improvements across most metrics compared to the original ViViD baseline; with the subsequent incorporation of the keyframe guidance module (KPG), the full model ViViD+MF+KPG attains the optimal performance. Specifically, on the VITON-HD dataset, compared with the ViViD baseline, SSIM increases from 0.881 to 0.903, LPIPS decreases from 0.089 to 0.056, and FID decreases from 8.67 to 6.07. Notably, in the VVT benchmark evaluation in Table 2, the baseline method ViViD achieves a slightly higher SSIM (0.949) than our proposed ViViD+MF+KPG (0.938). This discrepancy arises because ViViD relies on strong human-agnostic mask priors as input, where non-try-on regions (e.g., extensive plain backgrounds and unaltered limbs) are rigidly preserved and replicated, giving it an unfair mathematical advantage in the pixel-level structural similarity metric (SSIM). However, such local copy-and-paste mechanisms tend to introduce noticeable boundary artifacts and structural tearing in

regions involving limb self-occlusions.

In contrast, our method operates under a mask-free inference paradigm, enabling the network to end-to-end adaptively coordinate the boundaries between the human body and the target garment. Although this slight adaptive boundary resynthesis leads to a marginal drop in SSIM, our approach significantly outperforms ViViD in LPIPS (0.039 vs. 0.068) and VFID (1.04 vs. 3.41), metrics that are far more representative of visual perceptual fidelity and video temporal stability. This convincingly demonstrates that our method achieves superior overall synthesis quality and temporal coherence under complex dynamic motions.These empirical results show that the mask-free framework avoids disruptive interference from mask-parsing errors, while the keyframe anchoring mechanism further improves garment reconstruction fidelity.

Table 3 presents the evaluation results under the mask-free setting, where the test set consists of synthetic pseudo-paired datasets without providing strong mask priors. Compared with the ViViD baseline, ViViD+MF+KPG achieves a 37.1% relative improvement in LPIPS and a 30.0% relative improvement in FID on the VITON-HD dataset, and 21.3% in LPIPS and 20.6% in FID on the ViViD dataset. These results more accurately reflect the capabilities of the mask-free paradigm, showing that even without mask-provided editing-region priors, our mask-free framework, coupled with keyframe guidance, can still achieve high-fidelity garment transfer and improve generation quality under occlusion.

Table 1: Quantitative results on the VITON-HD and DressCode datasets evaluated solely with frame-level metrics (bold indicates the best results; $*$ denotes image-based virtual try-on methods, and + denotes video-based virtual try-on methods).

Method	**VITON-HD**			**DressCode**		
	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	FID (↓)
LaDI-VTON* [17]	0.871	0.0941	13.01	0.915	0.0620	16.71
StableVITON * [15]	0.888	0.073	–	–	–	–
IDM-VTON* [28]	0.870	0.102	6.29	0.920	0.062	8.64
RealVVT$^+$	0.890	0.101	7.844	0.932	0.061	8.881
PEMF-VVTO$^+$	0.894	0.062	6.86	0.927	0.034	3.41
ViViD$^+$	0.881	0.089	8.67	0.907	0.070	8.28
ViViD+MF$^+$	0.893	0.065	7.10	0.926	0.036	3.52
ViViD+MF+KPG (Ours)$^+$	**0.903**	**0.056**	**6.07**	**0.941**	**0.030**	**3.01**

Table 2: Quantitative results on the VVT and ViViD datasets evaluated with video-level metrics (bold indicates the best results; $*$ denotes image-based virtual try-on methods, and + denotes video-based virtual try-on methods).

Method	**VVT**			**ViViD**		
	SSIM (↑)	LPIPS (↓)	VFID (↓)	SSIM (↑)	LPIPS (↓)	VFID (↓)
LaDI-VTON* [17]	0.878	0.190	–	–	–	–
StableVITON* [15]	0.902	0.078	–	–	–	–
RealVVT$^+$	0.936	0.047	2.689	0.878	0.072	0.33
PEMF-VVTO$^+$	0.931	0.044	1.06	0.876	0.068	0.30
ViViD$^+$	**0.949**	0.068	3.41	0.867	0.080	0.34
ViViD+MF$^+$	0.933	0.045	1.07	0.876	0.071	0.34
ViViD+MF+KPG (Ours)$^+$	0.938	**0.039**	**1.04**	**0.879**	**0.063**	**0.27**

4.5 Qualitative Results

To provide an intuitive comparison of visual synthesis effects, qualitative comparisons are conducted between our proposed method and mainstream baselines, including ViViViD and PEMF-VVTO, as illustrated in Figure 3. Red bounding boxes highlight regions where artifacts emerge across baselines, including garment boundary misalignments, texture distortions, and residual foreground artifacts.

From a spatial perspective, ViViD and PEMF-VVTO often produce incomplete textures, silhouette distortions, and boundary misalignments in regions with limb occlusions and large-scale pose variations. In contrast, our MFKP-VVTO faithfully preserves fine-grained details—such as graphic prints and realistic wrinkles—from the reference garments, ensuring that the silhouette adapts naturally to human body postures.

From a temporal perspective, baseline methods of-

ten show abrupt fluctuations in garment colors and textures across consecutive frames. By leveraging temporal anchoring constraints from keyframes, our synthesized results maintain consistent garment appearance across frames, effectively suppressing temporal flickering.

Notably, while our model is trained exclusively on indoor studio datasets of fashion models, it achieves remarkable try-on performance on challenging in-the-wild video datasets, such as StreetVTON and TikTok. This demonstrates that our framework generalizes better across scenarios and adapts robustly to complex poses and varying lighting in realistic open-world environments.

Table 3: Quantitative results on the synthetic pseudo-paired datasets of VITON-HD and ViViD (bold indicates the best results).

Method	**VITON-HD**			**ViViD**		
	SSIM (↑)	LPIPS (↓)	FID (↓)	SSIM (↑)	LPIPS (↓)	VFID (↓)
ViViD	0.827	0.129	29.07	0.840	0.114	2.05
ViViD+MF	0.838	0.084	18.48	0.859	0.079	1.13
ViViD+MF+KPG (Ours)	**0.846**	**0.076**	**16.65**	**0.871**	**0.069**	**0.92**



Figure 3: Visual Comparison Results of Different Models

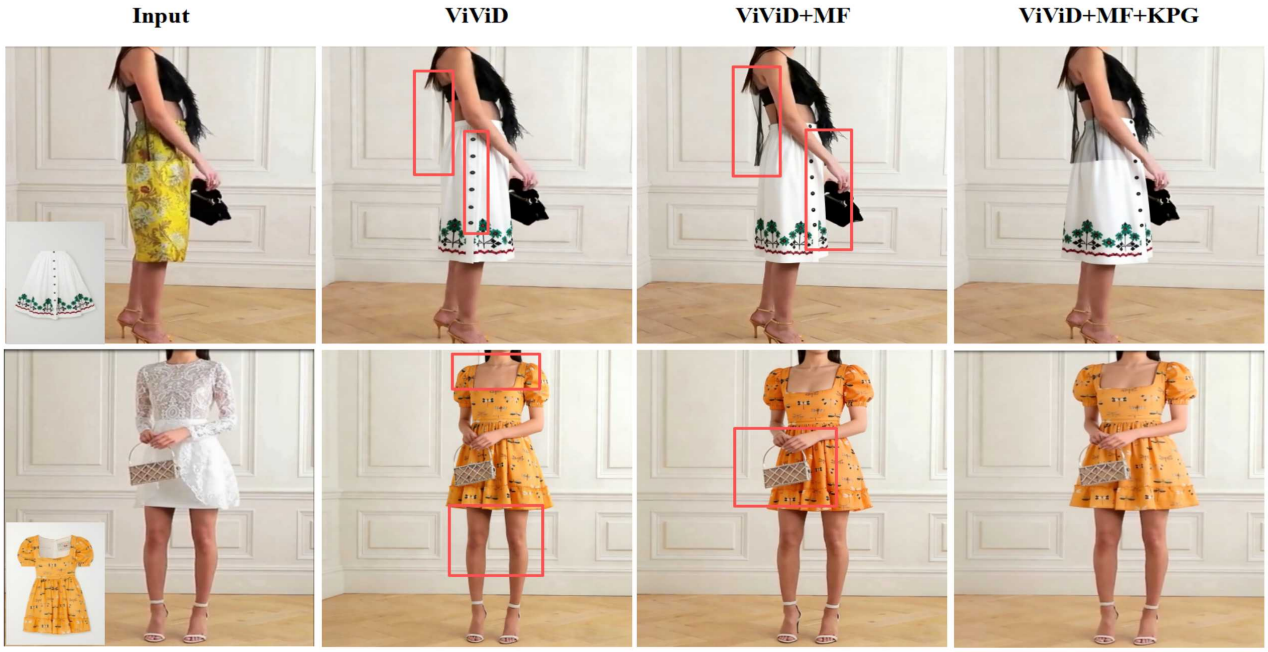


Figure 4: Comparative effects of ablation experiments based on the ViViD base model

4.6 Ablation Study

We conduct ablation studies of the proposed method on the ViViD benchmark. To validate each core component, we run multiple ablation experiments, including overall module ablations, internal dual-metric ablations within DA-KFS, and ablations on the CLIP and Qwen-VL dual-modal representations.

Ablation on Overall Framework

Taking ViViD as the baseline, we sequentially incorporate the mask-free framework (MF) and the keyframe guidance (KPG) module. Table 3 presents the quantitative results, and Figure 4 illustrates the visual ablation comparisons.

(1) Effectiveness of the Mask-Free Framework (MF): Compared with the original ViViD baseline, all metrics of ViViD+MF exhibit marked improvements. Specifically, the LPIPS metric decreases by 34.9% and 30.7% on the VITON-HD and ViViD benchmarks, respectively. This shows that eliminating mask inputs in the mask-free framework avoids the detrimental impact of mask-parsing inaccuracies on video synthesis, promoting natural alignment between garments and human bodies. Moreover, the substantial reduction in the FID metric indicates that the distribution of synthesized samples aligns much closer with authentic try-on videos.

(2) Gains from Keyframe Guidance (KPG): Superimposing the KPG keyframe guidance module upon MF yields further optimizations in both LPIPS and FID metrics. Notably, on the ViViD synthetic pseudo-paired dataset, FID drops from 1.13 to 0.92. Figure 4 shows the visual results. As shown in Figure 4, when using only MF, occluded regions suffer from incomplete garments and boundary distortions; with KPG, garment details in occluded areas are effectively recovered, and inter-frame garment consistency is substantially enhanced. This substantiates the indispensable role of the keyframe anchoring mechanism in handling complex posture scenarios.

Table 4: Ablation Results of the Dual-Metric Mechanism in DA-KFS (bold indicates the best results)

Configuration	**SSIM** (↑)	**LPIPS** (↓)	**FID** (↓)
Joint Distance Only	0.862	0.085	18.21
Pose Feature Similarity Only	0.867	0.081	17.42
DA-KFS (Ours)	**0.871**	**0.076**	**16.65**

Ablation on Dual-Metric Mechanism in DA-KFS

To validate the need for the dual-metric design, which couples weighted skeletal joint distances with pose feature similarities, we conduct comparative ablation experiments across three configurations: using only skeletal joint distances, using only pose feature similarities, and using the full DA-KFS framework. Table 4 presents the ablation results. Relying solely on skeletal joint distances suffers from inaccuracies in keypoint estimation under self-occlusion, which degrades keyframe sampling quality; conversely, relying exclusively on pose feature similarities lacks sensitivity to the human body's spatial motion amplitudes. By combining the strengths of both criteria, the dual-metric DA-KFS adaptively screens for a keyframe sequence with superior motion representativeness and minimal redundancy, thereby improving overall performance.

Table 5: Ablation Results of CLIP Visual Features and Qwen-VL High-Level Semantic Representations (bold indicates the best results)

Configuration	**SSIM** (↑)	**LPIPS** (↓)	**FID** (↓)
CLIP Only	0.863	0.083	17.84
Qwen-VL Only	0.865	0.080	17.10
CLIP + Qwen-VL (Ours)	**0.871**	**0.076**	**16.65**

Ablation on Garment Dual-Modal Features

To validate the complementary roles of CLIP visual features and Qwen-VL high-level semantic representations, we conduct comparative ablation experiments. Table 5 presents the ablation results.

Relying solely on CLIP captures only low-level garment textures and colors, yet falls short of modeling high-level structural semantics such as garment silhouettes, necklines, and sleeve patterns; conversely, relying exclusively on Qwen-VL high-level semantics tends to overlook fine-grained visual textures. Integrating both modalities achieves cross-modal complementarity, thereby further enhancing the overall fidelity of garment reconstruction.

4.7 User Study

Building on the quantitative ablation studies, we further conduct a subjective user evaluation. In this experiment, we recruit 50 female participants with an average age of 35 to perform double-masked scoring of the video virtual try-on results. We use a 5-point Likert scale ranging from 0 to 5, where higher ratings indicate better try-on fidelity. We prepare 100 test sample sets, and each evaluator randomly assesses 50 video sets. We completely randomize the presentation order of all samples and implement a blind evaluation protocol so evaluators are unaware of the underlying algorithmic models corresponding to the videos, mitigating subjective bias. In the PEMF-VVTO group, two questionnaires contained invalid responses and were discarded, leaving a final cohort of 48 valid evaluators. Table 6 summarizes the experimental results.

5 Discussion

Both quantitative comparative evaluations and ablation studies show that our proposed MFKP-VVTO model improves garment transfer precision across spatial and temporal dimensions, markedly enhancing virtual try-on quality in limb-occlusion scenarios. Nevertheless, several limitations remain to be addressed: when confronted with long-duration video sequences and structurally complex apparel, the model is prone

Table 6: User Study Results on the ViViD Dataset. Values Represent Average Scores across Evaluation Metrics (bold indicates the best results)

Method	Participants	Perceptual Evaluation Metrics			
		Style Consistency (↑)	Detail Fidelity (↑)	Visual Realism (↑)	User Satisfaction (↑)
ViViD	50	4.86	4.78	4.85	4.91
RealVVT	50	4.90	4.82	4.88	4.93
PEMF-VVTO	48	4.88	4.92	4.91	4.94
MFKP-VVTO (Ours)	**50**	**4.95**	**4.97**	**4.96**	**4.96**

to synthesis deviations, exhibiting insufficient fine-grained control and editing capability over the try-on region; furthermore, the framework takes only a single frontal garment image as input, lacking multi-view apparel details and thereby precluding multi-angle virtual try-on. In addition, although our approach consistently improves all evaluation metrics over PEMF-VVTO, its inference latency remains relatively high.

In future work, we will first build a more diverse pseudo-paired human training dataset to expand sample variety, then iterate and optimize the model on more challenging in-the-wild videos to further improve algorithmic robustness. Existing dataset cohorts are predominantly composed of individuals with conventional body shapes, demonstrating relatively limited diversity. Moving forward, while adhering to ethical standards and scene generalizability, we will incorporate samples of older people, children, and diverse body somatotypes to enrich body-shape distributions, promote model fairness and universality, and better serve the fitting needs of diverse demographics. Second, for the network architecture, we consider replacing the diffusion U-Net backbone with a Diffusion Transformer (DiT) to enhance global spatiotemporal representation capacity. Finally, we will explore multimodal garment representations to compensate for the information deficiency of single frontal images, enabling multi-view, omnidirectional spatiotemporal video virtual try-on.

6 Conclusions

To address key issues in existing video virtual try-on methods—namely, heavy reliance on mask priors, poor temporal stability, and suboptimal garment fidelity under limb self-occlusions and large-scale pose variations—this paper proposes MFKP-VVTO, a keyframe-guided, mask-free video virtual try-on framework. The overall architecture comprises four core components: pseudo-pair construction, Dual-metric Adaptive Keyframe Sampling (DA-KFS), mask-free keyframe try-on, and keyframe-guided video generation. First, we construct pseudo-paired samples using the pre-trained ViViD model, alleviating the critical scarcity of authentically annotated video try-on data. The DA-KFS algorithm integrates weighted skeletal joint distances with pose-feature similarities to select representative keyframes; it captures the video's global motion dynamics using sparse temporal anchors, effectively mitigating sampling biases from keypoint-detection failures under complex movements. In the mask-free keyframe try-on module, we integrate dual-modal representations from CLIP and Qwen-VL to encode low-level garment appearance textures and high-level silhouette semantics, respectively. We use masks strictly as regional constraints in the loss functions during training and exclude them entirely from the inference pipeline, effectively avoiding error propagation from inaccurate mask parsing. We also introduce an inter-keyframe garment consistency loss and a cross-frame alignment loss to regularize cross-frame appearance stability, using latent features from high-fidelity keyframe try-on results to steer non-keyframe generation and suppress garment semantic drift.

We conduct extensive quantitative, qualitative, and ablation experiments on four public benchmarks: VITON-HD, DressCode, VVT, and ViViD. The results show that, compared with both mask-based and existing mask-free baselines, MFKP-VVTO better preserves garment textures and structural silhouettes under self-occlusion and dramatic pose transitions, striking a favorable balance between single-frame transfer quality and inter-frame temporal coherence. Ablation studies substantiate the necessity of the mask-free framework, the DA-KFS dual-metric sampling strategy, and the CLIP/Qwen-VL dual-modal feature design. Subjective user studies further demonstrate our method's distinct advantages in garment style consistency, detail fidelity, and overall visual realism.

Nevertheless, limitations remain: the proposed method is prone to synthesis deviations under long-duration video sequences and structurally complex apparel, incurs relatively high inference overhead, and currently supports only a single frontal garment image as input. Future work will expand training samples across diverse body somatotypes, optimize the network backbone architecture, and incorporate multi-view garment features to further improve video virtual try-on performance in complex, open-world environments.

7 Acknowledgements

This work is supported by the Key R & D and Transformation Plan of Qinghai Province (2026-SF-147). The Education Department of the Shaanxi Provincial Government supported this research. Project Title: Research on Intelligent Fashion Design Based on Multi-

modal Diffusion Models (No.25JR008).

8 Author contributions

All authors jointly contributed to the experiments, manuscript preparation and submission of this paper. Xiangyan Fu drafted the manuscript. Shengling Geng reviewed the manuscript and provided revision suggestions. Weihua Pu organized the experimental data. Xiaojuan Dang performed experimental analysis. Feizhuan Fan wrote the literature review.

Data Availability

The datasets analyzed during the current study are available in the following public repositories:

- The ViViD dataset is available at: `https://github.com/alibaba-yuanjing-aigclab/ViViD`
- The VVT dataset is available at: `https://github.com/DAIR-Group/FW-GAN`
- The VITON-HD dataset is available at: `https://github.com/shadow2496/VITON-HD`
- The DressCode dataset is available at: `https://github.com/aimagelab/dress-code`

All human participant images shown in Figure 1 and Figure 2 are sourced from the aforementioned publicly accessible academic datasets (ViViD) and were used in accordance with their research licenses.

References

[1] Yu, R., Wang, X., Xie, X. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 10511–10520, 2019.

[2] Choi, S., Park, S., Lee, M., Choo, J. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 14131–14140, 2021.

[3] Minar, M., Tuan, T., Ahn, H., Rosin, P., Lai, Y. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, volume 3, pages 10–14, 2020.

[4] Wang, Y., Dai, W., Chan, L., Zhou, H., Zhang, A., Liu, S. GPD-VVTO: Preserving Garment Details in Video Virtual Try-On. In *Proc. 32nd ACM Int. Conf. Multimedia*, pages 7133–7142, 2024.

[5] Chong, Z., Zhang, W., Zhang, S., Zheng, J., Dong, X., Li, H., Wu, Y., Jiang, D., Liang, X. CatV^{2}TON: Taming Diffusion Transformers for Vision-Based Virtual Try-On with Temporal Concatenation. arXiv preprint arXiv:2501.11325, 2025.

[6] Chen, C., Lo, L., Huang, P., Shuai, H., Cheng, W. Fashionmirror:Co-attention feature-remapping virtual try-on with sequential template poses. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 13809–13818, 2021.

[7] Dong, H., Liang, X., Shen, X., Wu, B., Chen, B., Yin, J. Fw-gan:Flow-navigated warping gan for video virtual try-on. In *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pages 1161–1170, 2019.

[8] 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.

[9] Zuo, T., Huang, Z., Ning, S., Lin, E., Liang, C., Zheng, Z. Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. arXiv preprint arXiv:2508.02807, 2025.

[10] Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., Luo, P. Parser-free virtual try-on via distilling appearance flows. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 8485–8493, 2021.

[11] Chang, T., Wei, X., Chen, Zhang, X., Chen, Q., Luo, W., Yang, X. PEMF-VVTO: Point-Enhanced Video Virtual Try-on via Mask-free Paradigm. arXiv preprint arXiv:2412.03021, 2024.

[12] Cao, Z., Hidalgo, G., Simon, T., Wei, S., Sheikh, Y. Openpose:Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 172–186, 2019.

[13] Güler, R.A., Neverova, N., Kokkinos, I. Densepose: Dense human pose estimation in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 7297–7306, 2018.

[14] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R. Sdxl:Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952, 2023.

[15] Kim, J., Gu, G., Park, M., Park, S., Choo, J. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 8176–8185, 2024.

[16] Lin, E., Zhang, X., Zhao, F., Luo, Y., Dong, X., Zeng, L., Liang, X. Dreamfit: Garment-centric human generation via a lightweight anything-dressing encoder. In *Proc. AAAI Conf. Artif. Intell.*, pages 5218–5226, 2025.

[17] Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proc. 31st ACM Int. Conf. Multimedia*, pages 8580–8589, 2023.

[18] Xu, Y., Gu, T., Chen, W., Chen, A. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proc. AAAI Conf. Artif. Intell.*, pages 8996–9004, 2025.

[19] Kingma, D.P., Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[20] Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R. Dress code: High-resolution multi-category virtual try-on. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 2231–2235, 2022.

[21] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 10684–10695, 2022.

[22] Radford, A., Wook, J., Aditya, H., Gabriel, R., Sandhini, G., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., et al. CLIP: Learning Transferable Visual Models from Natural Language Supervision. arXiv preprint, 2019.

[23] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R. Segment Anything. arXiv:2304.02643, 2023.

[24] Li, P.K., Xu, Y.Q., Wei, Y.C., Yang, Y. Self-Correction for Human Parsing. arXiv preprint arXiv:1910.09777, 2019.

[25] Wang, P., Bai, S., Tan, S., Wang, S.J., Fan, Z., Bai, J.Z., Chen, K.Q., Liu, X.J., Wang, J.L., Ge, W.B. Qwen-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[26] He, Z., Chen, P., Wang, G., Li, G., Torr, P., Lin, L. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *Eur. Conf. Comput. Vis.*, pages 123–139, 2024.

[27] Wei, X., Zhu, J., Feng, S., Su, H. Video-to-video translation with global temporal consistency. In *ACM Int. Conf. Multimedia*, pages 18–25, 2018.

[28] Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J. Improving diffusion models for authentic virtual try-on in the wild. In *Eur. Conf. Comput. Vis.*, pages 206–235, 2024.

[29] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. Generative adversarial networks. Communications of the ACM, 63(11):139-144, 2020.

[30] Zhang, X., Song, D., Zhan, P., Chang, T., Zeng, J., Chen, Q., Luo, W., Liu, A.A. BooW-VTON: Boosting In-the-Wild Virtual Try-On via Mask-Free Pseudo Data Training. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1–11, 2025.

[31] Zhang, W., Li, X., Shi, P., Gao, J., Peng, X., Lian, Z., Jin, Y. BooM-VVT: Boosting Mask-Free Video Virtual Try-On with Image-Level Pseudo Data. arXiv preprint arXiv:2609.04120, 2026.

[32] Cao, Y., Feng, S., Shen, F., Peng, H., Xia, J., Zhu, Y., Yu, C. UniVVT: A Unified End-to-End Framework for High-Fidelity Video Virtual Try-on. arXiv preprint arXiv:2608.05745, 2026.

[33] He, Q., Chen, X., Pan, Y., Tang, P., Xu, P., Gan, Z., Wang, Y. The devil is in the details: Enhancing Video Virtual Try-On via Keyframe-Driven Details Injection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9182–9191, 2026.

[34] Chen, S., Sun, X., Zhang, L., Zhang, J. FlowVVTON: Flow-Guided Mask-Free Video Virtual Try-On. arXiv preprint arXiv:2608.30450, 2026.

[35] Yang, S., Zhou, Y., Liu, Z., Loy, C. C. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia*, pages 1–11, 2023.

[36] Li, S., Jiang, Z., Zhou, J., Liu, Z., Chi, X., Wang, H. Realvvt: Towards photorealistic video virtual try-on via spatio-temporal consistency. arXiv preprint arXiv:2501.08682, 2025.

[37] Jiang, J., Wang, T., Yan, H., Liu, J. Clothformer: Taming video virtual try-on in all module. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10789–10798, 2022.

[38] Nguyen, H., Nguyen, Q. Q. V., Nguyen, K., Nguyen, R. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *Proc. AAAI Conf. Artif. Intell.*, volume 39, number 6, pages 6200–6208, 2025.

[39] Zeng, J., Bai, Y., Chen, R., Zhang, X., Sun, L., Jin, D., Chu, X. Eevee: Towards close-up high-resolution video-based virtual try-on. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4614–4624, 2026.

[40] Shao, D., Wu, S., Wang, S., Wang, Y., Tang, Z., Liu, F., Yi, Z. TripVVT: A Large-Scale Triplet Dataset and a Coarse-Mask Baseline for In-the-Wild Video Virtual Try-On. In *Eur. Conf. Comput. Vis.*, pages 210–228, 2026.

[41] Fu, X., Geng, S., Pu, W., Dang, X. A review of deep learning-based virtual try-on research. *Journal on Image and Video Processing*, 2026(1):5, 2026.

[42] Cao, Y., Feng, S., Duan, R., Wang, L., Shi, D., Yu, C., Xing, J. LiveVVT: High-Fidelity Video Virtual Try-On in Real Time. arXiv preprint arXiv:2608.26714, 2026.

[43] X. Fu, M. Zhou, W. Pu, et al. CAF-VTON: cross-attention layered fusion based latent diffusion virtual try-on, *Signal, Image and Video Processing*, vol. 19, no. 12, p. 975, 2025.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- figure1.jpg
- figure2.jpg
- figure3.jpg
- figure4.jpg