

# TripVVT: A Large-Scale Triplet Dataset and a Coarse-Mask Baseline for In-the-Wild Video Virtual Try-On

Dingbao Shao<sup>1,2\*</sup>, Song Wu<sup>2\*</sup>, Shenyi Wang<sup>1</sup>, Ye Wang<sup>3†</sup>, Ziheng Tang<sup>1</sup>, Fei Liu<sup>4</sup>, Jiang Lin<sup>1</sup>, Xinyu Chen<sup>1</sup>, Qian Wang<sup>2</sup>, Ying Tai<sup>1</sup>, Jian Yang<sup>1</sup>, and Zili Yi<sup>1✉</sup>

<sup>1</sup>Nanjing University, China

<sup>2</sup>China Mobile Jiutian Artificial Intelligence Technology (Beijing) Co., Ltd., China

<sup>3</sup>Jilin University, China <sup>4</sup>ByteDance Inc., China

522024710012@smail.nju.edu.cn yi@nju.edu.cn



**Fig. 1:** Our method leverages the large-scale, diverse TripVVT-10K dataset and a coarse human-mask paradigm to produce realistic and temporally coherent try-on results under challenging real-world conditions. The top row illustrates dynamic motion in low-light scenes, the middle row shows seated poses that differ from the common standing cases, and the bottom row demonstrates performance in crowded public environments such as subway interiors, with complex backgrounds and multiple people in view. To support the research community, we have open-sourced the TripVVT-10K dataset and the associated benchmark.

\*Equal contribution. †Project lead. ✉Corresponding author.

**Abstract.** Due to the scarcity of large-scale in-the-wild triplet data and the improper use of masks, the performance of video virtual try-on models remains limited. In this paper, we first introduce **TripVVT-10K**, the largest and most diverse in-the-wild triplet dataset to date, providing explicit video-level cross-garment supervision that existing video datasets lack. Built upon this resource, we develop **TripVVT**, a Diffusion Transformer-based framework that replaces fragile garment masks with a simple, stable human-mask prior, enabling reliable background preservation while remaining robust to real-world motion, occlusion, and cluttered scenes. To support comprehensive evaluation, we further establish **TripVVT-Bench**, a 100-case benchmark covering diverse garments, complex environments, and multi-person scenarios, with metrics spanning video quality, try-on fidelity, background consistency, and temporal coherence. Compared to state-of-the-art academic and commercial systems, TripVVT achieves superior video quality and garment fidelity while markedly improving generalization to challenging in-the-wild videos. We publicly release the dataset and benchmark at <https://huggingface.co/datasets/TripVVT/TripVVT-10K>, which we believe provide a solid foundation for advancing controllable, realistic, and temporally stable video virtual try-on.

**Keywords:** Video Virtual Try-on · Dataset · Benchmark

## 1 Introduction

Video Virtual Try-on (VVT) [9, 12, 24] focuses on seamlessly replacing a person’s clothing in video sequences with a given garment reference, while preserving key attributes such as the individual’s identity, pose, motion, and background. It has gained significant attention due to its creative potential across various domains, including e-commerce, digital human creation, and the metaverse.

Current research in video virtual try-on can be broadly categorized into two paradigms: mask-based and mask-free approaches. Mask-based methods [9, 11, 12, 18, 24, 52], due to the lack of suitable paired data, primarily focus on reconstructing region-level content from the original video, referencing garments directly from the input video itself. However, this reliance on high-quality masks limits their ability to generalize effectively to diverse garment references. To overcome the dependency on masks, mask-free methods [5, 36, 48] aim to directly generate the try-on results conditioned on the input video sequence and the reference garment, without requiring masks to explicitly specify the editing regions. However, such methods often inadvertently alter other regions, making it challenging to preserve the integrity of the background. Moreover, most existing methods address virtual try-on in indoor video scenarios, but perform poorly on real, diverse, and complex in-the-wild video scenarios. These challenges significantly hinder the development and practical applications of video virtual try-on.

To address these challenges, we first introduce **TripVVT-10K**, the first large-scale and high-quality in-the-wild **Triplet** video dataset for **Virtual Video**



**Try-on. TripVVT-10K** comprises 10,031 triplets of ⟨original video, garment reference, try-on video⟩, including 30 types of clothing categories from upper body, lower body, and full body, and covering diverse, real, and complex outdoor scenarios. We construct the dataset from an initial collection of web-crawled videos and filter out low-quality samples. The resulting high-quality subset forms the original video in triplet. Subsequently, we build an automated and efficient data synthesis pipeline based on Wan-Animate [1]. Specifically, we first use Nano Banana [34] to create a high-quality swapped initial frame. We get the final try-on video via Wan-Animate under the guidance of the swapped frame and its original pose sequence. We then employ Nano Banana to extract the garment reference, which, together with original video and try-on video, constitutes the final triplet.

Leveraging TripVVT-10K, we introduce an end-to-end virtual try-on framework, TripVVT, based on a Diffusion Transformer (DiT) [30]. A novel hybrid paradigm is proposed to resolve the core trade-off between mask-based and mask-free methods. Its key design lies in using a simple human mask as the condition. This offers a stable spatial prior that confines edits to the human body, thereby robustly avoiding the over-constraints of garment masks and preserving the background. The model is trained using the try-on videos as input, with the original videos serving as the ground truth in a reverse design that facilitates robust supervision.

In addition, to support comprehensive evaluation and foster advancements, we introduce **TripVVT-Bench**, a dedicated benchmark tailored for in-the-wild scenarios. TripVVT-Bench incorporates a multi-dimensional evaluation framework across video quality, clothing fidelity, background preservation, and temporal consistency. We hope that TripVVT-Bench will serve as a standard benchmark for evaluating in-the-wild video virtual try-on and drive progress in this field.

**Our main contributions are as follows:**

- We construct **TripVVT-10K**, the first large-scale, high-resolution in-the-wild *video triplet* dataset for virtual try-on, enabling fully supervised training and evaluation under real-world conditions.
- We propose **TripVVT**, a Diffusion Transformer-based framework that harmoniously reconciles the mask-based and mask-free paradigms. By utilizing a coarse human mask as a spatial prior, it achieves state-of-the-art performance on both studio-style and challenging in-the-wild videos.
- We introduce **TripVVT-Bench**, the first comprehensive benchmark for video virtual try-on in real-world scenarios, providing a standardized and multi-dimensional evaluation platform covering video quality, try-on fidelity, background preservation, and temporal consistency.

**Table 1:** Comparison of virtual try-on datasets. “Pair” denotes model–garment paired data; “Pseudo” denotes unpaired/pseudo paired image data; “Triplet” denotes video triplets of the same subject across garments.

| Dataset                   | Modality | Resolution | Scene  |         | Person |          | Pairing | Number        |
|---------------------------|----------|------------|--------|---------|--------|----------|---------|---------------|
|                           |          |            | Studio | Outdoor | Single | Multiple |         |               |
| VITON-HD [6]              | Image    | 768×1024   | ✓      |         | ✓      |          | Pair    | 11,647        |
| DressCode [28]            | Image    | 768×1024   | ✓      |         | ✓      |          | Pair    | 53,792        |
| StreetTryOn [10]          | Image    | Var.       |        | ✓       | ✓      |          | Pseudo  | 14,453        |
| LAION-G. [15]             | Image    | Var.       | ✓      |         | ✓      |          | Triplet | 60,000        |
| VVT [11]                  | Video    | 192×256    | ✓      |         | ✓      |          | Pair    | 791           |
| ViViD [12]                | Video    | 624×832    | ✓      |         | ✓      |          | Pair    | 9,700         |
| <b>TripVVT-10K (ours)</b> | Video    | 720×1280   |        | ✓       | ✓      | ✓        | Triplet | <b>10,031</b> |

## 2 Related Work

### 2.1 Image-based Virtual Try-on

Early image-based virtual try-on typically followed a “warp–then–blend” pipeline [27, 41, 44, 47, 53]. With diffusion models, recent methods [7, 17, 21, 37, 43, 49] rely on human parsing or rule-based masks to preserve background and identity. While effective for single images, these fine-grained masks are sensitive to errors—tight masks truncate garment details, while loose masks overwrite background regions. Newer methods explore looser or optional spatial priors [13, 15, 19, 22], confirming that pixel-accurate garment masks are not essential. However, these approaches do not address the video-specific challenge of maintaining consistent spatial priors over time.

### 2.2 Video Virtual Try-on

Video try-on additionally requires temporal coherence and consistent motion following. Existing diffusion-based systems—such as ViViD [12], CatV2TON [9], and MagicTryOn [24]—extend image pipelines by incorporating temporal modeling. Similarly, Fashion-VDM [20] leverages advanced video diffusion architectures to significantly enhance temporal smoothness and garment details. However, these methods all fundamentally depend on precise garment masks or human parsing to explicitly define editing regions. These priors often become unreliable in real-world videos due to pose changes, occlusions, and lighting variations, leading to accumulated artifacts or background distortion. PEMF-VTO [5] further shows the difficulty of fully mask-free generation under limited pseudo-paired supervision. Collectively, prior works expose a fundamental tension: precise but fragile garment masks vs. flexible but unstable mask-free generation. This motivates our use of a coarse yet robust human-mask prior to balance spatial control and temporal stability.



**Fig. 2:** Overview of the TripVVT-10K Dataset. (a) Distribution of Clothing Categories: A sunburst chart visualizes the hierarchical composition of our dataset’s wardrobe, spanning major categories like upper-body, lower-body, and full-body, each broken down into numerous fine-grained styles. (b) Representative Examples: A gallery of sample triplets showcases the dataset’s diversity in capture conditions. These examples feature varied scenes, lighting, human dynamics, and the corresponding high-quality try-on results.

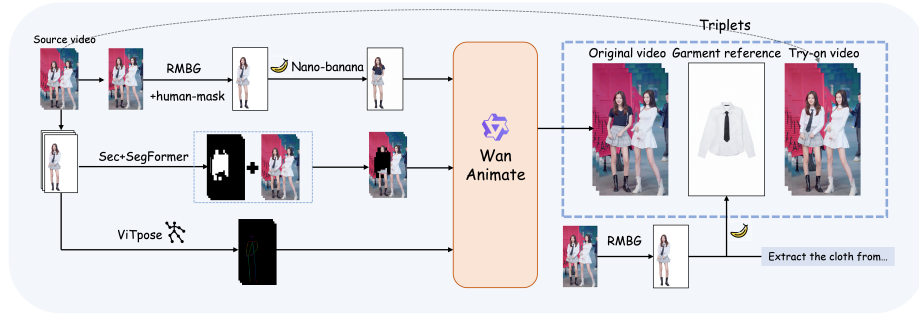
### 2.3 Datasets and Benchmarks

Existing datasets for virtual try-on are predominantly image-based. VITON-HD [6] and DressCode [28] provide paired person–garment images in studio environments, while Street TryOn [10] offers in-the-wild but unpaired single images. More recently, Any2AnyTryOn [15] introduces the LAION-Garment dataset, which supplies image-level triplets of a person in one outfit, a garment reference, and the same person in another outfit, enabling supervised cross-garment transfer at the image level. On top of these datasets, VTBench [16] organizes a comprehensive benchmark for image-based try-on. However, all these resources are static and do not provide temporal information for video modeling.

For videos, public resources such as VVT [11] and the video portion of ViViD [12] are captured indoors with limited scene diversity and only pairwise supervision. Prior works have repeatedly highlighted the need for larger, more diverse, and temporally reliable video datasets [5, 9, 25], yet, to our knowledge, there is still no publicly available in-the-wild video dataset with high-resolution triplets and an accompanying benchmark jointly assessing fidelity, background preservation, and temporal consistency. As comprehensively compared in Table 1, TripVVT-10K and TripVVT-Bench are designed to fill this critical gap.

## 3 Method

In this section, we first present an overview of the TripVVT-10K dataset in Sec. 3.1. We then detail the data synthesis pipeline in Sec. 3.2. Finally, we introduce the TripVVT in Sec. 3.3.



**Fig. 3:** Overview of the TripVVT-10K data construction pipeline. We synthesize training triplets ⟨original video, garment reference, try-on video⟩ from in-the-wild source videos. (a) Original Video Synthesis: A garment-swapped anchor frame is first generated (via [11]); conditioned on this frame and extracted pose sequences, the source video is re-rendered through masked inpainting to create a synthesized "original" version. (b) Garment Reference: A canonical product image is generated by performing image-to-garment synthesis on the initial person crop. (c) Try-on Video: The raw source video serves as the ground-truth target. Our pipeline ensures spatial-temporal consistency across the generated triplets.

### 3.1 Dataset Overview

TripVVT-10K is a large-scale video dataset comprising 10,031 triplets. It bridges the gap between controlled studio data and complex real-world environments. As shown in Fig. 2, the dataset demonstrates unprecedented diversity in both garment variety and capture conditions. This dual diversity is essential for developing robust models that can handle diverse clothing types while maintaining consistency in unconstrained settings. By systematically incorporating these variations, TripVVT-10K offers a valuable resource to advance virtual try-on technology toward practical real-world application.

### 3.2 Dataset Curation

To support supervised learning for in-the-wild video virtual try-on, we construct **TripVVT-10K**, a large-scale triplet dataset consisting of ⟨original video, garment reference, try-on video⟩. Each triplet is derived from a real source video through an automated synthesis pipeline. Following a reverse design, we treat the synthesized garment-swapped video as the model input (original video), while the authentic source video is used as the ground-truth try-on video. This reverse-training paradigm ensures that the supervision signal always comes from real videos, avoiding the quality limitations of synthetic targets.

*Stage 1: Source Video Collection and Preprocessing.* We first crawl approximately 20K in-the-wild human videos from the web, covering diverse scenes, viewpoints, and garments. We normalize all videos to a fixed resolution, filter

out low-quality or heavily compressed clips, and uniformly subsample each remaining video to 121 frames. The resulting high-quality subset serves as our *source videos*.

To robustly track the target person, we combine Gemini-2.5-Flash [34] and SAM2 [32] to obtain a temporally consistent *human-mask video* for each source clip. These masks both provide clean person regions for downstream garment reconstruction and act as a key conditioning signal in our TripVVT model, serving as a simple yet effective spatial prior. Conditioning on human masks instead of fine-grained garment masks avoids the over-constrained behavior of traditional mask-based methods while still confining edits to the human body and preserving the background.

*Stage 2: Triplet and Condition Generation.* After preprocessing the source videos, we construct the three core components of each triplet (original video, garment reference, try-on video) together with auxiliary modalities required by our model. Following the reverse-training paradigm, we first synthesize the *original video* (i.e., the model input), use the unmodified source video as the *try-on video*, and then reconstruct the *garment reference* from the source video.

To facilitate video synthesis, we additionally construct a *garment-mask video* that is used only during generation. Concretely, we segment the garment region on the first frame using SegFormer [39] applied to the human-cropped person, and then propagate this garment mask across the entire sequence via SEC [50].

*First-frame editing.* We begin by generating a swapped initial frame where the subject wears a new target garment as shown in Fig. 3. A garment is randomly sampled from DressCode [28], and Gemini-2.5-Flash [34] is used to produce a textual description tightly bound to this specific garment instance. The subject portrait extracted from the first source frame is combined with this garment-specific description and fed into Nano Banana [34] to synthesize a high-fidelity swapped first frame. By grounding each description on a concrete sampled garment and reusing it in Nano Banana, we ensure that the text is consistently matched to a specific garment, mitigating hallucination and repetition issues that arise when directly asking a language model to imagine garments.

*Video synthesis via guided inpainting.* Next, we synthesize a full garment-swapped video. We extract a pose sequence from the source video using ViT-Pose [42] and prepare an inpainting-ready version of the video by masking out the original garment region with the propagated garment masks. Conditioned on the swapped first frame, the pose sequence, and the inpainting-ready source video, Wan-Animate performs guided inpainting to produce a temporally consistent garment-swapped sequence that preserves the original motion and background. This synthesized sequence is stored as the *original video* in TripVVT-10K, while the unedited source video is directly used as the *try-on video*.

*Garment reference reconstruction.* Finally, we construct the garment reference image corresponding to the clothing worn in the try-on video. We extract the subject from the first frame of the source video using the human mask and RMBG-2.0 [51] to remove complex background clutter and other people. The resulting clean foreground crop is then fed into Nano Banana [34] to reconstruct



a canonical product-style garment image, representing the clothing originally worn by the subject in a background-free, front-view style.

*Stage 3: Quality Filtering.* We apply a two-stage filtering procedure to ensure dataset quality. First, an automatic check based on Gemini-2.5-Flash evaluates the fidelity of the reconstructed garment reference and the consistency between the synthetic original video and the source video, considering attributes such as garment color, texture, and structure. Triplets with low garment fidelity or severe artifacts are discarded. Second, the remaining candidates are manually reviewed to remove cases with identity shifts, noticeable temporal flickering, or visible editing artifacts.

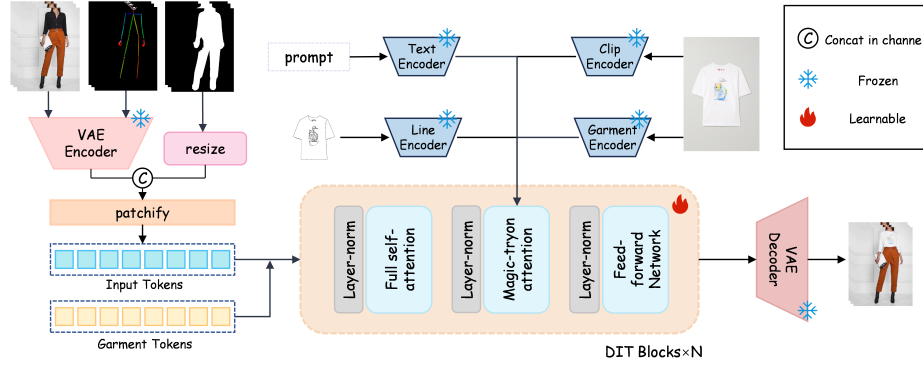
*Augmentation.* To further enrich the dataset beyond in-the-wild videos for training, we additionally incorporate supplemental data generated from existing public datasets. For image-based sources, we use **CatVTON** [8] to convert image garment-try-on pairs into triplets following the same  $\langle$ original video, garment reference, try-on video $\rangle$  structure. For video-based sources, we employ **MagicTryOn** [24] to synthesize video triplets in the same format. Since these synthetic image/video sources differ from TripVVT-10K in both content and generation characteristics, we design dedicated filtering strategies tailored to CatVTON and MagicTryOn outputs rather than reusing the TripVVT-10K pipeline. The detailed conversion and filtering procedures for these augmented samples are provided in the Supplementary Material. These samples are used solely as additional training data and are not included in the TripVVT-10K.

*Auxiliary Modalities.* Beyond the core triplet  $\langle$ original video, garment reference, try-on video $\rangle$ , each TripVVT-10K sample is accompanied by several auxiliary conditions that are explicitly used by our TripVVT framework during training: (1) a *human-mask video*, serving as the primary spatial prior and replacing traditional garment masks as the conditioning signal; (2) a *pose video* estimated by DWPose [45] from the original video, providing motion guidance; (3) a *garment line map*, extracted from the garment reference using the AniLines tool [29] to capture garment structure; and (4) a *text prompt* describing the overall scene and the target person’s outfit over the entire video, generated by Qwen-VL-Max [3]. These auxiliary modalities are consistently generated for all triplets and tightly match the model design described in Sec. 3.3.

### 3.3 Model Architecture

To effectively leverage the rich supervision provided by our TripVVT-10K dataset, we propose TripVVT, an end-to-end virtual try-on framework. As illustrated in Fig. 4, the model’s architecture is specifically designed to capitalize on the dataset’s structure. It builds upon the design of MagicTryOn [24] and is tailored to process the three synchronized inputs available in our dataset: the original video, the corresponding human-mask video, and the pose video of the target person. We introduce these specific inputs to enhance control and robustness. The

human mask serves two key functions: first, it preserves background consistency by strictly constraining the editing area to the subject, and second, it aids in localizing the person of interest. Compared to garment masks, human masks are also more robust and easier to obtain at high quality. The pose video, extracted via DWPose [45], is provided as an auxiliary condition to ensure the generated motion faithfully preserves the subject’s original movements. The core process begins with a VAE encoder extracting spatio-temporal features from the original video, which are then concatenated with the resized human-mask features. In parallel, the reference garment image is passed through a garment encoder to produce garment tokens. These fused features are fed into a DiT-based backbone. Finally, we condition the network on a global text prompt describing the entire scene (person, action, and environment) to serve as the textual control signal.



**Fig. 4:** Architecture of TripVVT. The model takes the original video, pose and human mask as inputs, reusing the DiT-based backbone of MagicTryOn [24], while garment and line encoders inject garment appearance and structural cues into the DiT blocks.

## 4 Benchmark and Evaluation

### 4.1 TripVVT-Benchmark

To address the limitations of existing benchmarks, which often rely on studio-only data and narrow evaluation metrics, we introduce TripVVT-Bench. Built upon the TripVVT-10K dataset, it consists of 100 high-quality and challenging triplets (original video, garment reference, try-on video) for in-the-wild evaluation, fully held out from the training set. The test set covers a diverse range of scenarios, including various garment types (upper, lower, and full-body), scenes with complex backgrounds, and both single- and multi-person videos. To ensure broad compatibility with different methods, we also provide auxiliary inputs such as garment masks and DensePose [14] information, generated using the

methodology from CatV2TON [9]. Our evaluation protocol for TripVVT-Bench is multi-faceted, assessing performance across four key dimensions. For video quality, we compute VFID (using both I3D [4] and ResNeXt [40] backbones) and supplement it with the reference-based metrics SSIM [38] and LPIPS [46], applicable due to our ground-truth videos. For try-on fidelity, we use CLIP-I [31] for semantic similarity and leverage Gemini-2.5-Flash [34](Gemini-SR) to automatically score the try-on success rate. Background consistency is evaluated via both pixel-level (BG-L1-Err) and perceptual (BG-DINO-Err [33]) differences in non-garment regions. Finally, for temporal consistency, we use CLIP-F [31] to measure feature similarity between adjacent frames. Further details on these metric calculations are provided in the Supplementary Material.

## 4.2 Implementation Details

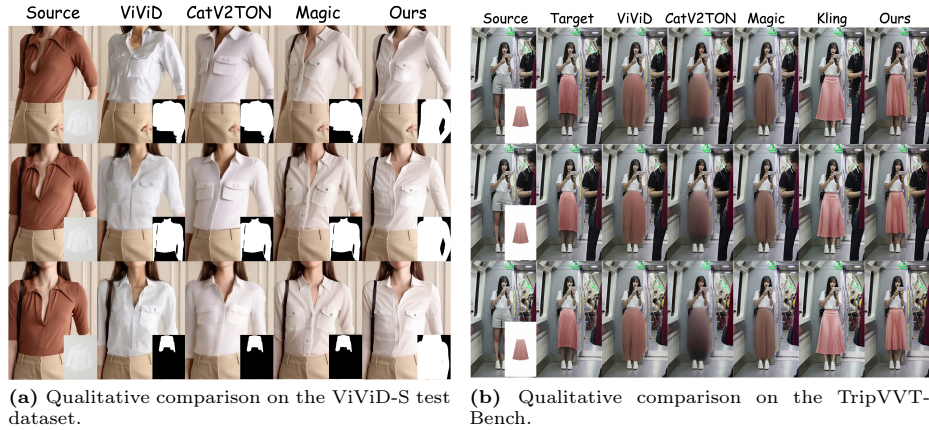
We adopt wan2.1-Fun-14B-control [2] (fine-tuned from Wan2.1-I2V-14B [35]) as our base model and train it with a three-stage progressive strategy that gradually increases data diversity and resolution. Stage 1 uses the TripVVT supplementary set for 25k steps at  $512 \times 384$ . Stage 2 adds TripVVT-10K and trains another 25k steps at mixed resolutions ( $512 \times 384$  and  $592 \times 334$ ). Stage 3 ups the resolution to  $832 \times 624$  and  $960 \times 544$  for 5.5k steps. All stages use 49-frame clips, batch size 2, AdamW [26] with a constant learning rate of  $1 \times 10^{-5}$ . Training runs on 8 H100 GPUs.

## 4.3 Experimental Settings

We conduct a comprehensive evaluation on two distinct test sets to assess performance in both controlled and in-the-wild environments. First, on the established ViViD-S test dataset [9], we compare TripVVT against state-of-the-art academic methods, including ViViD [12], CatV2TON [9], and MagicTryOn [24]. We also include DreamVVT [54] in this comparison for publicly reported metrics. However, as DreamVVT is not open-source, we could not run our custom evaluation scripts on its outputs, so it is excluded from comparisons involving our proposed metrics on this test set. Second, on our newly proposed TripVVT-Bench, which features challenging real-world scenarios and relies entirely on our comprehensive evaluation protocol, we benchmark TripVVT against the open-source academic models (ViViD, CatV2TON, MagicTryOn) and the commercial video editing tool Kling 1.6 [23].

## 4.4 Quantitative Experiments

*Results on ViViD-S test.* As summarized in Tab. 2, TripVVT delivers competitive or superior results on most metrics on ViViD-S, with clear improvements in perceptual video quality and temporal consistency compared to prior methods. Thanks to the strong supervision of TripVVT-10K and its supplementary training sets, our model achieves strong try-on fidelity, while being slightly inferior to



**Fig. 5:** Overall qualitative comparisons. Left: results on ViViD-S; Right: results on our TripVVT-Bench.

mask-based methods in background-related metrics and CLIP-based similarity. We attribute this gap in background consistency to the use of a coarse human mask instead of a fine-grained garment mask, which prioritizes robustness in real-world scenes over strict pixel-level alignment.

**Table 2:** Quantitative comparison with other methods on ViViD-S [9] test dataset. The best and second-best results are demonstrated in bold and underlined, respectively.

| Method          | VFID <sub>I</sub> ↓ | VFID <sub>R</sub> ↓ | CLIP-I ↑      | CLIP-F ↑      | BG-L1 ↓       | BG-DINO ↓     |
|-----------------|---------------------|---------------------|---------------|---------------|---------------|---------------|
| ViViD [12]      | 21.8032             | 0.8212              | <b>0.7270</b> | 0.9574        | 0.0571        | <b>0.0055</b> |
| CatV2TON [9]    | 19.5131             | 0.5283              | 0.7109        | 0.9304        | <u>0.0507</u> | 0.0071        |
| MagicTryOn [24] | 17.5710             | 0.5073              | 0.7072        | <u>0.9647</u> | <b>0.0493</b> | <u>0.0056</u> |
| DreamVVT [54]   | <u>16.9468</u>      | <u>0.4285</u>       | -             | -             | -             | -             |
| Ours            | <b>16.4398</b>      | <b>0.3315</b>       | <u>0.7267</u> | <b>0.9670</b> | 0.0624        | 0.0059        |

*Results on TripVVT-Bench.* On the more challenging TripVVT-Bench (Tab. 3), TripVVT clearly outperforms both academic and commercial competitors in video quality and temporal stability, and achieves competitive try-on fidelity, while also narrowing the gap to mask-based methods in background preservation. This highlights the advantage of training on diverse in-the-wild data: our model generalizes well where studio-trained baselines struggle. On reference-based metrics, TripVVT is slightly worse than MagicTryOn in SSIM, likely because the full-body human mask forces the model to reconstruct the entire person during try-on, making exact pixel-level alignment more difficult.

**Table 3:** Quantitative comparison with other methods on TripVVT-Bench dataset. The best and second-best results are demonstrated in bold and underlined, respectively.

| Method          | VFID <sub>I</sub> ↓ | VFID <sub>R</sub> ↓ | SSIM ↑        | LPIPS ↓       | CLIP-I ↑      | CLIP-F ↑      | BG-L1 ↓       | BG-DINO ↓     | Gemini-SR ↑ |
|-----------------|---------------------|---------------------|---------------|---------------|---------------|---------------|---------------|---------------|-------------|
| ViViD [12]      | 26.7620             | 0.7083              | 0.8393        | 0.1343        | <b>0.7249</b> | 0.9349        | <u>0.0691</u> | <u>0.0053</u> | 9%          |
| CatV2TON [9]    | 34.2275             | 3.3266              | 0.7845        | 0.2130        | 0.6394        | 0.9212        | 0.1032        | 0.0109        | 4%          |
| MagicTryOn [24] | 22.9238             | <u>0.5502</u>       | <b>0.8641</b> | <u>0.1150</u> | 0.6934        | 0.9495        | <b>0.0515</b> | <b>0.0036</b> | 43%         |
| Kling [23]      | <u>22.4242</u>      | 0.6289              | 0.6662        | 0.1353        | <u>0.7148</u> | <u>0.9504</u> | 0.1662        | 0.0064        | <u>78%</u>  |
| Ours            | <b>20.7245</b>      | <b>0.3163</b>       | <u>0.8538</u> | <b>0.1053</b> | 0.7110        | <b>0.9606</b> | 0.0852        | 0.0059        | <b>91%</b>  |

**Table 4:** Ablation study results on TripVVT-Bench.

| Method          | VFID <sub>I</sub> ↓ | VFID <sub>R</sub> ↓ | SSIM ↑        | LPIPS ↓       | CLIP-I ↑      | CLIP-F ↑      | BG-L1 ↓       | BG-DINO ↓     |
|-----------------|---------------------|---------------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Base            | <b>20.7245</b>      | <b>0.3163</b>       | <b>0.8538</b> | <b>0.1053</b> | <u>0.7110</u> | <b>0.9606</b> | <b>0.0852</b> | <b>0.0059</b> |
| w/o TripVVT-10K | <u>22.5105</u>      | <u>0.4569</u>       | <u>0.7605</u> | <u>0.1182</u> | <b>0.7133</b> | 0.9483        | <u>0.1552</u> | <b>0.0059</b> |
| w/o Pose        | 33.6635             | 0.9949              | 0.5762        | 0.2744        | 0.6989        | 0.9540        | 0.3244        | 0.0112        |
| w/o Human mask  | 33.2486             | 5.9749              | 0.6087        | 0.2570        | 0.6839        | <u>0.9566</u> | 0.2942        | 0.0118        |
| w Garment mask  | 28.5826             | 0.6946              | 0.5956        | 0.2451        | 0.7046        | 0.9466        | 0.3007        | 0.0102        |

## 4.5 Qualitative Experiments

We provide qualitative comparisons in Fig. 5a and Fig. 5b to visually demonstrate the advantages of TripVVT.

*Results on ViViD-S test.* Fig. 5a shows a single video sequence from the ViViD-S test dataset with three frames: the first frame has a poor garment mask, while the subsequent frames have relatively stable masks. In the first frame, where the garment mask is inaccurate, mask-based methods fail to reproduce the reference shirt’s sleeve length and introduce noticeable artifacts. Thanks to the robust human-mask prior, TripVVT is the only method that correctly recovers the sleeve length while maintaining coherent edits. In the later frames with stable masks, our method still produces the most convincing results, with sharper details, more realistic fabric texture, and better alignment with the body.

*Results on TripVVT-Bench.* Fig. 5b presents qualitative results on our challenging TripVVT-Bench. Since previous methods are mainly trained on indoor or studio-style data, they struggle to produce satisfactory try-on results in these in-the-wild scenarios, often showing severe distortions and inconsistent clothing appearance. The commercial video editing model Kling can successfully replace the outfit, but the generated garment texture and style differ noticeably from the reference image. In contrast, TripVVT better preserves human structure and produces more natural and realistic try-on effects, outperforming both open-source and commercial baselines. More qualitative results are provided in the Supplementary Material.

**Table 5:** User study results comparing different methods on TripVVT-Bench.

| Method | ViViD [12] | CatV2TON [9] | MagicTryOn [24] | Kling [23] | Ours  |
|--------|------------|--------------|-----------------|------------|-------|
| Rank 1 | 2.6%       | 0.2%         | 12.2%           | 17.2%      | 67.6% |

#### 4.6 User Study

We conducted a user study on TripVVT-Bench to assess perceptual quality. For each triplet (original video, garment reference, try-on video), participants viewed anonymized outputs from all methods in random order and selected their top three preferences. They were asked to holistically balance video quality, try-on fidelity, background consistency, and temporal coherence; videos could be replayed and zoomed to reduce bias. We collected 2,500 votes from 50 participants. Following prior work, we report the proportion of first-place (Rank-1) votes as the primary metric. As shown in Table 5, our method is consistently preferred, achieving the highest user preference.

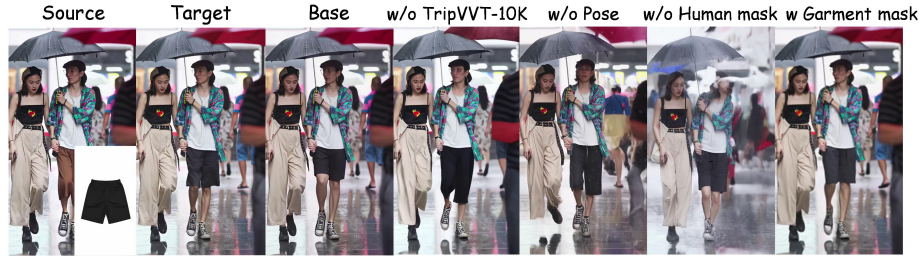
#### 4.7 Ablation Studies

We conduct a series of ablation studies to validate the effectiveness of the key components in our approach as shown in Tab. 4 and Fig. 6, respectively.

*Effectiveness of the TripVVT-10K.* To demonstrate the crucial role of TripVVT-10K in enhancing generalization, we train a model variant exclusively on our supplementary dataset. As shown in Tab. 4, this variant exhibits a significant performance degradation across all quantitative metrics compared to our full model. The qualitative comparison in Fig. 6 further reveals that without the rich, in-the-wild data from TripVVT-10K, the model fails to handle complex outdoor scenes, confirming the value of our curated dataset.

*Effectiveness of the Human Mask.* To validate our human-mask prior, we evaluate two variants: removing the mask entirely, and replacing it with a garment mask (derived via CatV2TON [9]). As shown in Tab. 4 and Fig. 6, both variants underperform our baseline. The mask-free variant suffers the most; without spatial constraints, the model struggles to localize edits, leading to severe boundary inconsistencies. Conversely, while the garment-mask variant improves upon the mask-free setting, it still introduces noticeable visual artifacts compared to our human-mask baseline. We attribute this performance gap to the inherent fragility and inaccuracy of garment masks in complex real-world scenarios.

*Effectiveness of Pose Guidance.* Removing pose guidance has a minor impact on quantitative scores (Tab. 4), but qualitatively it leads to distorted poses and unnatural motion (Fig. 6), confirming its importance for maintaining motion fidelity.



**Fig. 6:** Qualitative results of ablation studies on TripVVT-Bench.



**Fig. 7:** A limitation example where the try-on model slightly alters non-target regions: the backpack disappears and minor changes occur on the skirt and boots during garment transfer.

## 5 Limitations

Our current pipeline relies solely on a human mask to localize the person to be edited, leaving the model to automatically infer which regions should undergo garment replacement. As a result, in some cases the model slightly alters clothing regions that are not intended for editing, as illustrated in Fig. 7. This limitation suggests that relying on coarse spatial priors may be insufficient for strictly controlled garment transfer. Future work may explore new model architectures or training strategies that incorporate stronger structural guidance—such as explicit garment segmentation, region-level constraints, or attention supervision—to achieve more precise and disentangled try-on editing.

## 6 Conclusion

We introduced TripVVT-10K, the first large-scale, high-resolution in-the-wild video triplet dataset, providing the explicit cross-garment supervision for training reliable video virtual try-on models. Built on this foundation, we developed TripVVT, a DiT-based framework that leverages a simple yet stable human-mask prior to balance spatial control with robustness in real-world scenes. To support standardized evaluation, we further proposed TripVVT-Bench, a 100-case

benchmark for in-the-wild video virtual try-on. Experiments on both ViViD-S and TripVVT-Bench demonstrate that TripVVT achieves strong video quality, garment fidelity, and temporal consistency, with clear advantages in challenging in-the-wild conditions where previous systems struggle. We hope that the dataset, benchmark, and method presented in this work provide a unified foundation for advancing controllable, realistic, and temporally stable video virtual try-on.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62406134), Jiangsu Provincial Science & Technology Major Project (Grant No. BG2024042), the Suzhou Key Technologies Project (Grant No. SYG2024136) and the Nanjing University-China Mobile Communications Group Co. Ltd. Joint Institute.

## References

1. Alibaba, H.T.T.L.: Wan-animate: Unified character animation and replacement with holistic replication. (2025), <https://arxiv.org/abs/2509.14055>
2. Alibaba PAI: Wan2.1-fun-14b-control. <https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-Control> (2025), accessed: 2025-02-28
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 4724–4733. IEEE Computer Society (2017). <https://doi.org/10.1109/CVPR.2017.502>, <https://doi.org/10.1109/CVPR.2017.502>
5. Chang, T., Chen, X., Wei, Z., Zhang, X., Chen, Q.G., Luo, W., Song, P., Yang, X.: Pemf-vto: Point-enhanced video virtual try-on via mask-free paradigm. arXiv preprint arXiv:2412.03021 (2024)
6. Choi, S., Park, S., Lee, M., Choo, J.: VITON-HD: high-resolution virtual try-on via misalignment-aware normalization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 14131–14140. Computer Vision Foundation / IEEE (2021). <https://doi.org/10.1109/CVPR46437.2021.01391>, [https://openaccess.thecvf.com/content/CVPR2021/html/Choi\\_VITON-HD\\_High-Resolution\\_Virtual\\_Try-On\\_via\\_Misalignment-Aware\\_Normalization\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Choi_VITON-HD_High-Resolution_Virtual_Try-On_via_Misalignment-Aware_Normalization_CVPR_2021_paper.html)
7. Choi, Y., Kwak, S., Lee, K., Choi, H., Shin, J.: Improving diffusion models for authentic virtual try-on in the wild. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVI. Lecture Notes in Computer Science, vol. 15144, pp. 206–235. Springer (2024). [https://doi.org/10.1007/978-3-031-73016-0\\_13](https://doi.org/10.1007/978-3-031-73016-0_13), [https://doi.org/10.1007/978-3-031-73016-0\\_13](https://doi.org/10.1007/978-3-031-73016-0_13)



8. Chong, Z., Dong, X., Li, H., Zhang, S., Zhang, W., Zhang, X., Zhao, H., Liang, X.: Catvton: Concatenation is all you need for virtual try-on with diffusion models (2024), <https://arxiv.org/abs/2407.15886>
9. Chong, Z., Zhang, W., Zhang, S., Zheng, J., Dong, X., Li, H., Wu, Y., Jiang, D., Liang, X.: Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. arXiv preprint **arXiv:2501.11325** (2025). <https://doi.org/10.48550/ARXIV.2501.11325>, <https://doi.org/10.48550/arXiv.2501.11325>
10. Cui, A., Mahajan, J., Shah, V., Gomathinayagam, P., Liu, C., Lazebnik, S.: Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2025, Tucson, AZ, USA, February 26 - March 6, 2025. pp. 1414–1423. IEEE (2025). <https://doi.org/10.1109/WACV61041.2025.00145>, <https://doi.org/10.1109/WACV61041.2025.00145>
11. Dong, H., Liang, X., Shen, X., Wu, B., Chen, B., Yin, J.: FW-GAN: flow-navigated warping GAN for video virtual try-on. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 1161–1170. IEEE (2019). <https://doi.org/10.1109/ICCV.2019.00125>, <https://doi.org/10.1109/ICCV.2019.00125>
12. Fang, Z., Zhai, W., Su, A., Song, H., Zhu, K., Wang, M., Chen, Y., Liu, Z., Cao, Y., Zha, Z.: Vivid: Video virtual try-on using diffusion models. arXiv preprint **arXiv:2405.11794** (2024). <https://doi.org/10.48550/ARXIV.2405.11794>, <https://doi.org/10.48550/arXiv.2405.11794>
13. Feng, Y., Zhang, L., Cao, H., Chen, Y., Feng, X., Cao, J., Wu, Y., Wang, B.: Omnistry: Virtual try-on anything without masks. arXiv preprint **arXiv:2508.13632** (2025). <https://doi.org/10.48550/ARXIV.2508.13632>, <https://doi.org/10.48550/arXiv.2508.13632>
14. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. pp. 7297–7306. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00762>, [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Guler\\_DensePose\\_Dense\\_Human\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Guler_DensePose_Dense_Human_CVPR_2018_paper.html)
15. Guo, H., Zeng, B., Song, Y., Zhang, W., Liu, J., Zhang, C.: Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19085–19096 (2025)
16. Hu, X., Yujie, L., Luo, D., Xu, P., Zhang, J., Zhu, J., Wang, C., Fu, Y.: Vtbench: Comprehensive benchmark suite towards real-world virtual try-on models. arXiv preprint **arXiv:2505.19571** (2025). <https://doi.org/10.48550/ARXIV.2505.19571>, <https://doi.org/10.48550/arXiv.2505.19571>
17. Jiang, B., Hu, X., Luo, D., He, Q., Xu, C., Peng, J., Zhang, J., Wang, C., Wu, Y., Fu, Y.: Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on. arXiv preprint **arXiv:2411.10499** (2024). <https://doi.org/10.48550/ARXIV.2411.10499>, <https://doi.org/10.48550/arXiv.2411.10499>
18. Jiang, J., Wang, T., Yan, H., Liu, J.: Clothformer: Taming video virtual try-on in all module. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022. pp. 10789–10798. IEEE (2022). <https://doi.org/10.1109/CVPR52688.2022.01053>, <https://doi.org/10.1109/CVPR52688.2022.01053>

19. Jo, Y., Park, M., Kang, D.o.: Up-vton: A unified virtual try-on framework supporting mask, mask-free, and prompt-driven guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6971–6979 (2025)
20. Karras, J., Li, Y., Liu, N., Zhu, L., Yoo, I., Lugmayr, A., Lee, C., Kemelmacher-Shlizerman, I.: Fashion-vdm: Video diffusion model for virtual try-on. In: SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024. pp. 93:1–93:11. ACM (2024). <https://doi.org/10.1145/3680528.3687623>, <https://doi.org/10.1145/3680528.3687623>
21. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. arXiv preprint **arXiv:2312.01725** (2023). <https://doi.org/10.48550/ARXIV.2312.01725>, <https://doi.org/10.48550/arXiv.2312.01725>
22. Kim, J., Jin, H., Park, S., Choo, J.: Promptdresser: Improving the quality and controllability of virtual try-on via generative textual prompt and prompt-aware mask. arXiv preprint **arXiv:2412.16978** (2024). <https://doi.org/10.48550/ARXIV.2412.16978>, <https://doi.org/10.48550/arXiv.2412.16978>
23. Kuaishou Technology: Kling ai model. <https://app.klingai.com/cn/> (2025), accessed: 2025-11-08
24. Li, G., Zheng, S., Zhang, H., Chen, J., Luan, J., Ou, B., Zhao, L., Li, B., Jiang, P.: Magictyron: Harnessing diffusion transformer for garment-preserving video virtual try-on. arXiv preprint **arXiv:2505.21325** (2025). <https://doi.org/10.48550/ARXIV.2505.21325>, <https://doi.org/10.48550/arXiv.2505.21325>
25. 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). <https://doi.org/10.48550/ARXIV.2501.08682>, <https://doi.org/10.48550/arXiv.2501.08682>
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017), <https://api.semanticscholar.org/CorpusID:53592270>
27. Morelli, D., Baldrati, A., Cartella, G., Cornia, M., Bertini, M., Cucchiara, R.: Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In: Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023– 3 November 2023. pp. 8580–8589. ACM (2023). <https://doi.org/10.1145/3581783.3612137>, <https://doi.org/10.1145/3581783.3612137>
28. Morelli, D., Fincato, M., Cornia, M., Landi, F., Cesari, F., Cucchiara, R.: Dress code: High-resolution multi-category virtual try-on. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022. pp. 2230–2234. IEEE (2022). <https://doi.org/10.1109/CVPRW56347.2022.00243>, <https://doi.org/10.1109/CVPRW56347.2022.00243>
29. Pan, Z.: Anilines - anime lineart extractor. <https://github.com/zhenglinpan/AniLines-Anime-Lineart-Extractor> (2025)
30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023. pp. 4172–4182. IEEE (2023). <https://doi.org/10.1109/ICCV51070.2023.00387>, <https://doi.org/10.1109/ICCV51070.2023.00387>
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual

- Event. *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763. PMLR (2021), <http://proceedings.mlr.press/v139/radford21a.html>
32. Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net (2025), <https://openreview.net/forum?id=Ha6RTeWMd0>
  33. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S.E., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3. arXiv preprint **arXiv:2508.10104** (2025). <https://doi.org/10.48550/ARXIV.2508.10104>, <https://doi.org/10.48550/arXiv.2508.10104>
  34. Team, G.: Gemini 2.5: Pushing the frontier with advanced reasoning, multi-modality, long context, and next generation agentic capabilities. arXiv preprint **arXiv:2507.06261** (2025). <https://doi.org/10.48550/ARXIV.2507.06261>, <https://doi.org/10.48550/arXiv.2507.06261>
  35. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
  36. Wan, Z., Xu, Y., Hu, D., Cheng, W., Chen, T., Wang, Z., Liu, F., Liu, T., Gong, M.: MF-VITON: high-fidelity mask-free virtual try-on with minimal input. arXiv preprint **arXiv:2503.08650** (2025). <https://doi.org/10.48550/ARXIV.2503.08650>, <https://doi.org/10.48550/arXiv.2503.08650>
  37. Wang, C., Chen, T., Chen, Z., Huang, Z., Jiang, T., Wang, Q., Shan, H.: FLDM-VTON: faithful latent diffusion model for virtual try-on. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. pp. 1362–1370. ijcai.org (2024), <https://www.ijcai.org/proceedings/2024/151>
  38. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004). <https://doi.org/10.1109/TIP.2003.819861>, <https://doi.org/10.1109/TIP.2003.819861>
  39. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. *CoRR abs/2105.15203* (2021), <https://arxiv.org/abs/2105.15203>
  40. Xie, S., Girshick, R.B., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. pp. 5987–5995. IEEE Computer Society (2017). <https://doi.org/10.1109/CVPR.2017.634>, <https://doi.org/10.1109/CVPR.2017.634>
  41. Xie, Z., Huang, Z., Dong, X., Zhao, F., Dong, H., Zhang, X., Zhu, F., Liang, X.: GP-VTON: towards general purpose virtual try-on via collaborative local-flow

- global-parsing learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 23550–23559. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.02255>, <https://doi.org/10.1109/CVPR52729.2023.02255>
42. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (2022)
  43. Xu, Y., Gu, T., Chen, W., Chen, A.: Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 8996–9004. AAAI Press (2025). <https://doi.org/10.1609/AAAI.V39I9.32973>, <https://doi.org/10.1609/aaai.v39i9.32973>
  44. Yang, Z., Jiang, Z., Li, X., Zhou, H., Dong, J., Zhang, H., Du, Y.: D<sup>4</sup>-vton: Dynamic semantics disentangling for differential diffusion based virtual try-on. arXiv preprint **arXiv:2407.15111** (2024). <https://doi.org/10.48550/ARXIV.2407.15111>, <https://doi.org/10.48550/arXiv.2407.15111>
  45. Yang, Z., Zeng, A., Yuan, C., Li, Y.: Effective whole-body pose estimation with two-stages distillation. In: IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023. pp. 4212–4222. IEEE (2023). <https://doi.org/10.1109/ICCVW60793.2023.00455>, <https://doi.org/10.1109/ICCVW60793.2023.00455>
  46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 586–595. Computer Vision Foundation / IEEE Computer Society (2018). <https://doi.org/10.1109/CVPR.2018.00068>, [http://openaccess.thecvf.com/content\\_cvpr\\_2018/html/Zhang\\_The\\_Unreasonable\\_Effectiveness\\_CVPR\\_2018\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html)
  47. Zhang, S., Han, X., Zhang, W., Lan, X., Yao, H., Huang, Q.: Limb-aware virtual try-on network with progressive clothing warping. IEEE Trans. Multim. **26**, 1731–1746 (2024). <https://doi.org/10.1109/TMM.2023.3286278>, <https://doi.org/10.1109/TMM.2023.3286278>
  48. Zhang, X., Song, D., Zhan, P., Chang, T., Zeng, J., Chen, Q., Luo, W., Liu, A.: Boov-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 26399–26408. Computer Vision Foundation / IEEE (2025). <https://doi.org/10.1109/CVPR52734.2025.02458>, [https://openaccess.thecvf.com/content/CVPR2025/html/Zhang\\_BooW-VTON\\_Boosting\\_In-the-Wild\\_Virtual\\_Try-On\\_via\\_Mask-Free\\_Pseudo\\_Data\\_Training\\_CVPR\\_2025\\_paper.html](https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_BooW-VTON_Boosting_In-the-Wild_Virtual_Try-On_via_Mask-Free_Pseudo_Data_Training_CVPR_2025_paper.html)
  49. Zhang, X., Li, X., Kampffmeyer, M., Dong, X., Xie, Z., Zhu, F., Dong, H., Liang, X.: Warpdiffusion: Efficient diffusion model for high-fidelity virtual try-on. arXiv preprint **arXiv:2312.03667** (2023). <https://doi.org/10.48550/ARXIV.2312.03667>, <https://doi.org/10.48550/arXiv.2312.03667>
  50. Zhang, Z., Ding, S., Dong, X., He, S., Lin, J., Tang, J., Zang, Y., Cao, Y., Lin, D., Wang, J.: Sec: Advancing complex video object segmentation via progressive concept construction. arXiv preprint arXiv:2507.15852 (2025)
  51. Zheng, P., Gao, D., Fan, D., Liu, L., Laaksonen, J., Ouyang, W., Sebe, N.: Bilateral reference for high-resolution dichotomous image segmentation. arXiv preprint

- arXiv:2401.03407** (2024). <https://doi.org/10.48550/ARXIV.2401.03407>, <https://doi.org/10.48550/arXiv.2401.03407>
52. Zhong, X., Wu, Z., Tan, T., Lin, G., Wu, Q.: MV-TON: memory-based video virtual try-on network. In: MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021. pp. 908–916. ACM (2021). <https://doi.org/10.1145/3474085.3475269>, <https://doi.org/10.1145/3474085.3475269>
  53. Zhu, L., Yang, D., Zhu, T., Reda, F., Chan, W., Saharia, C., Norouzi, M., Kemelmacher-Shlizerman, I.: Tryondiffusion: A tale of two unets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 4606–4615. IEEE (2023). <https://doi.org/10.1109/CVPR52729.2023.00447>, <https://doi.org/10.1109/CVPR52729.2023.00447>
  54. Zuo, T., Huang, Z., Ning, S., Lin, E., Liang, C., Zheng, Z., Jiang, J., Zhang, Y., Gao, M., Dong, X.: Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. arXiv preprint **arXiv:2508.02807** (2025). <https://doi.org/10.48550/ARXIV.2508.02807>, <https://doi.org/10.48550/arXiv.2508.02807>