

Eevee: Towards Close-up High-resolution Video-based Virtual Try-on

Jianhao Zeng^{1*} Yancheng Bai^{1*†} Ruidong Chen^{1,2} Xuanpu Zhang² Lei Sun¹
 Dongyang Jin¹ Ryan Xu¹ Nannan Zhang^{3‡} Dan Song² Xiangxiang Chu¹
¹Amap, Alibaba Group ²Tianjin University
³Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences



Figure 1. An illustrative sample from our proposed virtual try-on dataset. Compared to existing datasets, ours provides a richer collection of data, including both full-shot and close-up videos of the person, detailed garment images, and rich textual descriptions. Furthermore, all images and videos are provided at a higher resolution. More examples can be found in the Appendix.

Abstract

Video virtual try-on technology provides a cost-effective solution for creating marketing videos in fashion e-commerce. However, its practical adoption is hindered by two critical limitations. First, the reliance on a single garment image as input in current virtual try-on datasets limits the accurate capture of realistic texture details. Second, most existing methods focus solely on generating full-

shot virtual try-on videos, neglecting the business’s demand for videos that also provide detailed close-ups. To address these challenges, we introduce a high-resolution dataset for video-based virtual try-on. This dataset offers two key features. First, it provides more detailed information on the garments, which includes high-fidelity images with detailed close-ups and textual descriptions; Second, it uniquely includes full-shot and close-up try-on videos of real human models. Furthermore, accurately assessing consistency becomes significantly more critical for the close-up videos, which demand high-fidelity preservation of garment details. To facilitate such fine-grained evaluation, we propose a new garment consistency metric VGID (Video Garment Inception Distance) that quantifies the preservation of both texture and structure. Our experiments validate these contributions. We demonstrate that by utilizing the detailed im-

*Equal contribution.

†Project Lead.

‡Corresponding author: Nannan Zhang (nn.zhang2@siat.ac.cn).

§The research is supported in part by the National Natural Science Foundation of China (62571369). The code is available at <https://github.com/AMAP-ML/Eevee>.

ages from our dataset, existing video generation models can extract and incorporate texture features, significantly enhancing the realism and detail fidelity of virtual try-on results. Furthermore, we conduct a comprehensive benchmark of recent models. The benchmark effectively identifies the texture and structural preservation problems among current methods.

1. Introduction

Driven by the rapid growth of e-commerce, the demand for apparel marketing videos is surging, making virtual try-on technology [50, 51, 64, 65] a crucial content generation tool. A foundational task in this domain is video-based virtual try-on [10, 16, 26, 31, 32], where the goal is to synthesize a video from a target garment image and a person’s source video. In the resulting video, the individual is realistically dressed in the target garment while maintaining their original movements and identity.

However, due to dataset limitations, existing video-based virtual try-on models face several common technical hurdles in practical applications. First, existing datasets typically only provide a single full-shot video as the ground truth. This results in videos that are more like a general, low-detail preview, failing to capture how the garment truly behaves and failing to meet the business demand for detailed close-up videos. Consequently, the practical value of this technology is significantly limited. Moreover, constrained by the low resolution of training datasets like ViViD [16] (832×624) and VVT [15] (256×192), existing models [30, 31, 68, 71] produce videos with insufficient clarity. This coarse visual output fails to provide users with adequate detail to assess the garment’s true material and fit. Additionally, the majority of these models rely on a single, flat image of the garment as input, making it challenging to generate lifelike fabric folds and intricate details. In practice, however, vendors can readily supply multi-angle detail shots of their garments, yet this valuable information remains largely underutilized.

To address the aforementioned issues, we introduce Eevee, a high-resolution dataset designed to support a wide range of virtual try-on tasks. Compared to previous datasets, Eevee’s main distinctions are as follows: First, in addition to providing a full-shot video of a person, it also offers a close-up video. Second, the dataset includes detailed images and a detailed textual description of the garment to assist the model in generating fine details and textures. Finally, Eevee features the highest resolution currently available for this task. The practical benefit of this is evident in our comparative analysis: as illustrated in Fig. 2, a magnified, side-by-side comparison with the ViViD dataset shows that Eevee provides substantially clearer details and superior texture fidelity, overcoming the blur common in lower-



Figure 2. A comparison of garment image and video frame, along with their zoomed-in details, from the ViViD dataset shown on top and the Eevee dataset shown on bottom.

resolution benchmarks. Fig. 1 illustrates a sample from our dataset. The left panel displays the raw collected data; the right panel shows its corresponding parsing annotation. To enhance usability, we applied further data processing beyond this annotation step.

Close-up virtual try-on videos, where the garment occupies the majority of the frame, necessitate a more precise evaluation of garment fidelity. Existing metrics such as VFID [53] and VBench [24] primarily assess overall video quality and realism, but they are insufficient for gauging the consistency of the generated apparel with the input garment image. To bridge this gap, we propose a new metric, VGID (Video Garment Inception Distance), specifically designed to measure garment consistency. This metric operates by calculating the cosine similarity between deep features extracted from the generated garment region in the video and those from the input garment image.

Leveraging our Eevee dataset, we conduct a comprehensive benchmark of recent state-of-the-art models [16, 27, 31] using our proposed VGID metric alongside established metrics such as VFID, SSIM [59], LPIPS [66], and VBench. This evaluation reveals significant disparities in texture and structural preservation among current methods, thereby establishing a strong baseline for future research in high-fidelity virtual try-on.

In summary, we introduce Eevee, a high-resolution virtual try-on dataset featuring close-up videos and detailed garment data for fine texture generation. To evaluate close-up garment consistency where metrics like VFID or VBench fall short, we propose a novel metric, VGID. Furthermore, we establish a strong baseline for future research by comprehensively benchmarking state-of-the-art models using our proposed dataset and metric.

Dataset	Garment Category	Detailed Image	Close-up Video	Resolution		Total number			
				Image	Video	Train pairs	Test pairs	Images	Frames
VITON [20]	Upper-body	✗	✗	256 × 192	–	14,221	2,032	32,506	–
MPV [14]	Upper-body	✗	✗	256 × 192	–	52,336	10,544	49,211	–
VITON-HD [8]	Upper-body	✗	✗	1024 × 768	–	11,647	2,032	27,358	–
DressCode [40]	Multi	✗	✗	1024 × 768	–	48,392	5,400	161,376	–
DeepFashion [37]	✗	✗	✗	1101 × 750	–	52,712	–	52,712	–
VVT [15]	Upper-body	✗	✗	256 × 192	256 × 192	661	130	1,582	205,675
TikTokDress [42]	Upper-body	✗	✗	720 × 540	720 × 540	693	124	817	272,548
ViViD[16]	Multi	✗	✗	1380 × 920	832 × 624	7,759	1,941	9,700	1,213,694
Eevee (Ours)	Multi	✓	✓	2400 × 1800	1088 × 816	8,364	1,000	28,092	1,905,973

Table 1. Comparison between our dataset and the most widely used datasets for virtual try-on and other related tasks.

2. Related Work

2.1. Video-based Virtual Try-on

With the rapid development of Diffusion models [6, 7, 22, 29, 35, 46, 47, 52, 56–58, 67] and Transformers [1, 2, 11–13, 33, 34, 39, 54, 62], video generation [63, 69] and video editing [19, 27, 70] models have achieved significant success. This progress is now impacting virtual try-on technology, which aims to synthesize a target garment onto a human model to generate try-on results. While significant progress has been made in image-based virtual try-on [9, 17, 28, 41, 50, 61, 64], research on the more complex video-based virtual try-on is still in its early stages. In this emerging field, researchers have begun their initial explorations. For example, RealVVT [32] employs a dual U-Net [25] architecture to generate precise virtual try-on videos, enhancing both spatial and temporal consistency. CatV2TON [10] is a Diffusion Transformer (DiT) [54] framework for both image and video virtual try-on, which implements a DiT [45] architecture by temporally concatenating garment and person inputs. Similarly, Magic-Tryon [31] builds upon the DiT backbone from Wan2.1 [55] and fully leverages its self-attention mechanism to achieve unified modeling of spatiotemporal coherence. Due to dataset limitations, these methods often only generate full-shot, low-resolution virtual try-on videos that also lack detailed garment textures, which restricts their practical application.

2.2. Virtual Try-on Datasets

The availability of high-quality datasets is indispensable for the development of virtual try-on. Early developments in this field were primarily centered on static, image-based collections. Seminal datasets like VITON [20] first established the task by providing paired images of a person and a target in-shop garment. This was later superseded by VITON-HD [8], which significantly increased the image resolution to 1024×768 , enabling the generation of more realistic and detailed results. Following this, Dress-

Code [40] further expanded the scale and complexity, offering a large-scale repository with more diverse poses and challenging garment types. While effective for static try-on, these image-based datasets lack the temporal information necessary for video-based applications. Efforts to bridge this gap have led to the creation of video datasets, most notably VVT [15] and ViViD [16]. However, the VVT dataset is of limited utility due to its low video resolution and small scale. Although ViViD represents a significant advancement in terms of resolution and data volume, it possesses critical limitations. First, it provides only a single, full-shot video, failing to meet the commercial demand for detailed close-up videos. Furthermore, it only provides a single, flat-lay image for each garment, which is insufficient for capturing the intricate textures, material properties, and deformation details required for high-fidelity virtual try-on.

3. Eevee

3.1. Dataset

Existing public video-based virtual try-on datasets suffer from three limitations. First, existing public datasets for video-based virtual try-on typically only contain video from a single global perspective. However, practical applications often require more diverse perspectives, such as close-ups on specific areas, which these datasets lack. Additionally, these datasets typically represent a garment with only a single, flat in-shop image. This overlooks a key opportunity in practical applications: merchants can readily provide detailed close-up shots, and leveraging such information could enhance the fine-grained details of the generated results. Finally, the majority of these public datasets suffer from low resolution, which places a hard ceiling on the visual quality of the virtual try-on, often resulting in blurry outputs where essential garment details are lost.

To facilitate the development of video-based virtual try-on, we introduce Eevee. It is a high-resolution dataset designed to overcome prior limitations. It provides both full-shot videos and corresponding close-up videos, address-

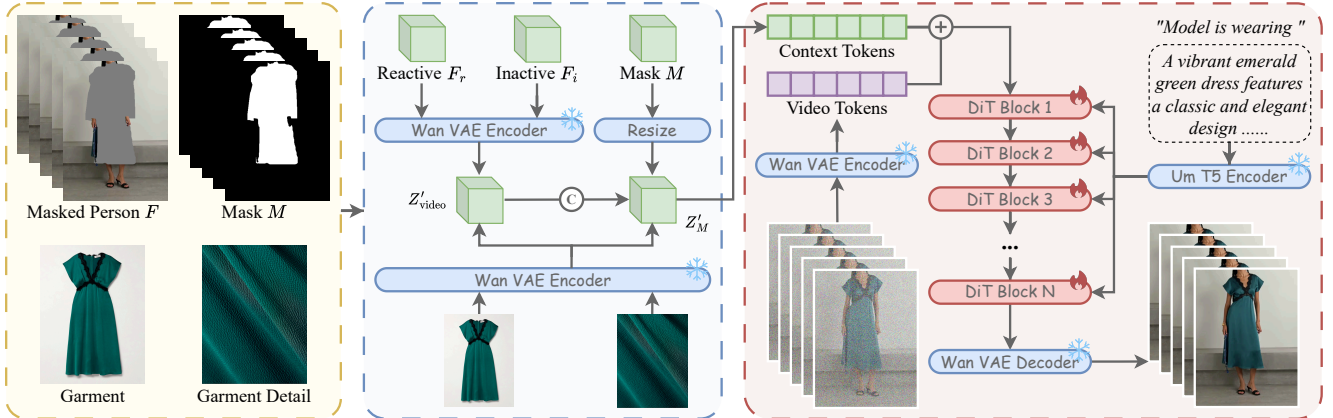


Figure 3. The overall fine-tuning pipeline of the VACE on our dataset. Compared to other video-based virtual try-on models, our VACE fine-tuning process includes an additional step where detailed garment images are fed into the network through the Wan VAE Encoder.

ing the need for detailed dynamic views. Additionally, it moves beyond the single in-shop photo by providing detailed close-up images for each garment. This rich data is designed to support a wide range of virtual try-on tasks. The Eevee dataset features 9,364 high-resolution, video-based virtual try-on pairs, establishing it as the highest-resolution collection available with images up to 2400×1800 and videos up to 1088×816 . The dataset is categorized into three main garment types, split into training and testing sets: upper-body (4492 training / 500 test pairs), lower-body (2308 training / 250 test pairs), and dresses (1564 training / 250 test pairs). The total number of images in our dataset, excluding parsing annotations, is 28,092, and the total number of video frames is 1,905,973.

We first collected a variety of images and related video data from e-commerce websites. The corresponding raw data encompasses in-shop garment images, close-up details of the garment, images of the person, as well as full-shot and close-up videos of the person. To make the dataset easier to use, we have added a series of annotations. (1) We utilize the large language model Qwen-VL-Max [3] to generate detailed textual descriptions and classify garments from images. (2) For full-shot virtual try-on videos and person images, we use OpenPose [4] to obtain human parsing results and then generate the mask. For close-up virtual try-on videos, where obtaining human parsing results is often difficult, we instead use Grounded SAM-2 [36, 48, 49] to perform semantic segmentation on the garment region to extract the mask. We then apply a 31-pixel padding around the mask to mitigate boundary inaccuracies and prevent the leakage of the original garment’s shape information. We also conduct manual random sampling to identify and remove severe segmentation failures. (3) To adapt to other virtual try-on models, we use AniLines [44] to extract garment contour maps and Detectron2 [60] to obtain DensePose UV coordinates for the human body. Fig. 1 illustrates

a sample from our dataset.

Tab. 1 summarizes the key characteristics of our dataset in comparison with existing datasets for virtual try-on and fashion-related tasks. Notably, the Eevee dataset is the first to provide both detailed images of garments and close-up virtual try-on videos. Furthermore, it boasts the highest image and video resolutions currently available.

3.2. Method

To validate our dataset, especially the garment detail images, we fine-tune the pre-trained VACE [27] model. VACE, a diffusion transformer-based framework for video generation and editing built on Wan2.1 [55], is ideal for video-based virtual try-on. This task functions as a masked video-to-video editing process. We chose VACE as our base model for two key reasons. First, it requires only video masks, avoiding the reliance on extra parsing data (like DensePose [18]) seen in other methods such as MagicTryon [31]. This minimizes interference from annotation inaccuracies. Second, VACE is highly extensible, making it a suitable platform for future work involving more complex multimodal inputs.

The overall fine-tuning pipeline is shown in Fig. 3. Given an input masked person video F and a mask video M , VACE first decouples F using the provided segmentation mask. This decoupling, based on M , yields two distinct frame sequences identical in shape:

$$\begin{aligned} F_r &= F \odot M \\ F_i &= F \odot (1 - M), \end{aligned} \quad (1)$$

where F_r , called reactive frames, contains all the pixels to be changed (such as the gray pixels), while all the pixels to be kept are stored in F_i , the inactive frames.

After decoupling, we process these conditional inputs. As shown in the diagram, F_r and F_i are processed inde-

pendently by the Wan VAE Encoder to produce latent representations, which we denote Z_r and Z_i respectively. The mask M is processed by first being rearranged into patches to create a 64-channel representation, which is then down-sampled via interpolation to match the latent dimensions, yielding the mask latent Z_M . These processed representations are then used to form the Context Tokens. Specifically, the inactive latents Z_i and reactive latents Z_r are first concatenated along the channel dimension to create a unified video latent representation Z_{video} . Next, the garment and garment detail images (treated as f individual reference frames) are encoded by the VAE to produce a set of reference latents $Z_{\text{ref},j}$. To match the channel structure of Z_{video} , reference latent $Z_{\text{ref},j}$ is concatenated with a zero tensor of the same shape along the channel dimension. These prepared reference latents are then prepended to the main video latent Z_{video} along the temporal dimension. Concurrently, the interpolated mask latent Z_M is prepended with a block of zero-latents along the temporal dimension to create Z'_M . Finally, Z'_{video} is concatenated with Z'_M along the channel dimension. This resulting tensor forms the Context Tokens, which are then combined with the Video Tokens and fed into the DiT model.

3.3. Metric

Unlike video-based virtual try-on which typically focuses on full-shot video generation, close-up video-based virtual try-on presents a distinct set of challenges. In these zoomed-in perspectives, the garment itself occupies a significantly larger portion of the frame. This shift means that the accurate reconstruction of intricate garment textures and patterns has become the most critical measure of quality. However, existing metrics are ill-equipped for this specific task. For instance, VFID [53] evaluates the overall video quality. Furthermore, while VBench’s Subject Consistency [24] is useful, it primarily measures temporal stability, ensuring that the garment remains consistent within the video. However, it does not assess fidelity to the input garment image. To address this limitation, we propose VGID (Video Garment Inception Distance), a new metric designed to precisely quantify the fidelity between the generated garment texture and the source garment image.

VGID’s core idea is to measure the consistency between the generated video’s garment and the input garment image. To do this while avoiding interference from the background, a naive approach would be to use segmentation masks. However, to obviate the critical dependency on these unreliable or unavailable masks, VGID leverages the inherent properties of vision transformers. We leverage the DINO-V2 [43] model, which is renowned for its powerful emergent segmentation capabilities developed during self-supervised training; the model learns to segment objects without explicit supervision. We harness this property by

extracting and processing the [CLS]-to-patch self-attention maps from the final Transformer block. This map is then upsampled and normalized to create a soft mask M , which assigns high values (near 1.0) to the salient garment and low values (near 0.0) to the background. However, we observe that for categories such as ‘lower body’ and ‘upper body’, the soft masks generated by DINO-V2 frequently fail to isolate the target item, often incorporating parts of adjacent garment pieces. This cross-contamination makes the mask unreliable for evaluating multi-garment outfits. Given that dresses typically function as a single, holistic garment and are less susceptible to this ambiguity, we report this consistency metric exclusively for the ‘dresses’ category to ensure a more accurate and meaningful evaluation.

VGID provides a score, computed as the cosine similarity between the masked, spatially-averaged features of the garment image I_s and a generated frame I_v . To calculate this, we first extract feature maps $F_s = \Phi(I_s)$ and $F_v = \Phi(I_v)$ using a feature extractor Φ . The corresponding soft masks, \hat{M}_s and \hat{M}_v , are resized to match the spatial dimensions of the feature maps and then applied via element-wise multiplication ($F' = F \odot \hat{M}$). This step effectively zeros out background features. The resulting masked features are subjected to global average-pooling (GAP), and their consistency is quantified using cosine similarity, as defined by the equation:

$$\text{VGID}(I_s, I_v) = \frac{\text{GAP}(F'_s) \cdot \text{GAP}(F'_v)}{\|\text{GAP}(F'_s)\| \|\text{GAP}(F'_v)\|}. \quad (2)$$

This calculation forms our semantic fidelity metric, denoted as VGID. It leverages deep, semantically-aware features from DINO-V2 to measure the integrity of high-level patterns—for example, ensuring that a plaid pattern remains recognizably plaid. For a full video V with T frames, the final score is computed as the average across all frames, providing a robust assessment of semantic consistency throughout the video.

4. Experiments

4.1. Implementation Details

We conduct all experiments on the Eevee dataset. To facilitate a direct comparison, we performed the experiments at resolutions of 1088×816 and 832×624 . Although the videos in the dataset are all longer than 49 frames, we use the first 49 frames of each video as input and generate 49 output frames. Our experiments are designed in two configurations: paired and unpaired. The paired configuration uses the person’s own garment as input to test the model’s fidelity in preserving person-specific features. The unpaired configuration, which simulates a realistic use case, challenges the model to dress the person in a different target garment. Besides, we performed virtual try-on on the full-shot videos and the close-up videos, respectively.

Resolution	Type	Method	Unpaired			Paired				
			VFID _R ^u ↓	VFID _I ^u ↓	VGID ^u ↑	VFID _R ^p ↓	VFID _I ^p ↓	VGID ^p ↑	SSIM ↑	LPIPS ↓
1088 × 816	Full-shot	ViViD	0.565	12.859	0.514	0.487	10.306	0.521	0.874	0.119
		MagicTryon	0.187	9.783	0.512	0.137	6.168	0.524	0.906	0.092
		VACE	0.454	11.260	0.522	0.282	7.290	0.525	0.899	0.083
		Ours	0.220	10.065	0.524	0.076	6.555	0.534	0.910	0.078
832 × 624	Full-shot	ViViD	0.389	12.194	0.506	0.285	9.211	0.508	0.871	0.111
		MagicTryon	0.161	9.865	0.520	0.106	6.167	0.514	0.908	0.074
		VACE	0.587	11.647	0.524	0.343	7.624	0.528	0.888	0.090
		Ours	0.219	9.777	0.528	0.076	6.520	0.527	0.905	0.071
1088 × 816	Close-up	ViViD	1.253	12.665	0.533	0.976	10.900	0.509	0.785	0.185
		MagicTryon	0.752	11.285	0.538	0.607	8.714	0.519	0.807	0.163
		VACE	0.913	12.844	0.533	0.429	9.331	0.515	0.778	0.174
		Ours	0.535	11.621	0.540	0.274	8.429	0.524	0.803	0.154
832 × 624	Close-up	ViViD	0.936	12.198	0.533	0.663	10.116	0.509	0.775	0.176
		MagicTryon	0.595	11.262	0.534	0.479	8.524	0.513	0.791	0.152
		VACE	1.039	12.783	0.545	0.593	9.572	0.519	0.755	0.178
		Ours	0.530	11.475	0.548	0.273	8.349	0.526	0.786	0.148

Table 2. This table compares the performance of different models on full-shot and close-up videos, evaluated at different video resolutions using various metrics. The subscripts ‘u’ and ‘p’ respectively represent the unpaired setting and paired setting. The best results are in bold.

For implementation, we initialize our model with the pre-trained weights from VACE and fine-tune it using LoRA [23]. The AdamW [38] optimizer is used with a fixed learning rate of 1e-4. All training are conducted on 8 NVIDIA H20 (96GB) GPUs. The number of inference steps during testing is set to 50. Furthermore, to ensure a fair comparison, all model variants in the ablation study are trained under identical hyperparameter configurations.

For evaluation, we employ SSIM [59] and LPIPS [66]. To assess video quality, we use VFID [53] to evaluate both spatial quality and temporal consistency. We report results using two different backbones: I3D [5] (VFID_I) and ResNeXt [21] (VFID_R). In addition, we utilize select metrics from VBench [24] to comprehensively evaluate video quality and consistency. Finally, to evaluate the consistency between the generated video and the input garment, we report VGID. Due to space limitations, we report the weighted average results across our test categories (dresses, upper body, lower body). Detailed per-category results are in the Appendix.

4.2. Quantitative Comparison

All of our quantitative comparisons are performed on the Eevee dataset, as our fine-tuned VACE model requires additional garment detail images as input. To ensure a comprehensive and fair evaluation, we designed our experiments across several distinct settings.

As shown in Tab. 2, both MagicTryon [31] and our fine-tuned VACE model [27] achieved the best results. While MagicTryon demonstrated superior performance on the VFID_I and LPIPS [66] metric, our fine-tuned VACE model excelled on the other metrics, particularly in the

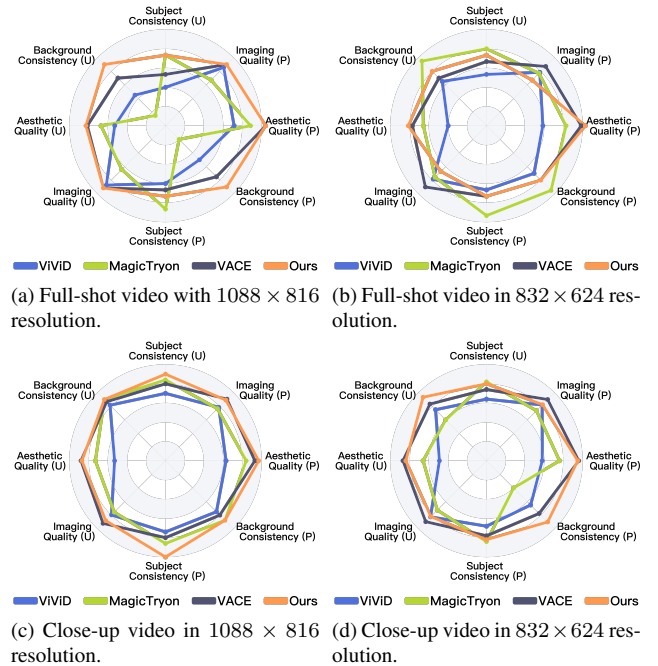


Figure 4. VBench evaluation results of video-based virtual try-on models. We visualize the evaluation results of four models in 4 VBench dimensions under both paired and unpaired settings. We normalize the results per dimension for clearer comparisons. For comprehensive numerical results, please refer to the Appendix.

paired setting. We attribute this advantage to the fine-tuning process and detailed garment image. This stage specifically enhanced the model’s ability to learn the intricate relationship between the paired inputs, allowing it to better preserve



Figure 5. Qualitative comparison of close-up video virtual try-on results on the Eevee dataset. Our fine-tuned VACE model not only accurately reproduces the patterns of the input garment, but the generated video is also more realistic overall. Please zoom in for better visualization.

garment details while realistically warping it onto the target pose. Additionally, the performance of all models decreased to some extent, which is attributable to the increased difficulty of high-resolution and close-up virtual try-on video generation tasks.

Additionally, we utilize the VBench [24] benchmark to comprehensively evaluate the overall quality and consistency of the generated videos. As shown in Fig. 4, our fine-tuned VACE model secured only a marginal lead. We observed that the metrics across most models showed no significant differentiation. This suggests that the VBench metrics are more heavily influenced by the inherent nature of the task itself rather than the specific capabilities of the models. Consequently, VBench appears limited in its ability to effectively measure or distinguish the true performance differences among various virtual try-on models.

4.3. Qualitative Comparison

In Fig. 5, we present the visual results of different methods on the Eevee dataset. We observe that our fine-tuned VACE [27] model achieves outstanding performance in generating try-on videos. We attribute this enhanced fidelity directly to the detailed garment images utilized as input during the fine-tuning process, which provide the model with the rich visual information necessary to preserve these specific attributes. Furthermore, this is consistent with the previously mentioned performance on the Eevee. ViViD [16] and MagicTryon [31] often struggle to accurately preserve the patterns of the input garment. The original VACE tends to directly paste the garment onto the target area rather than performing a genuine try-on. As shown in Fig. 5, the tag on

the garment’s collar in the VACE result should be hidden behind the model. Additionally, since VACE is not optimized for video-based virtual try-on, it occasionally experiences generation failures, such as producing solid white clothing or failing to generate human body parts entirely. This leads to poor quantitative metrics for VACE.

4.4. Ablation Study

To better understand the impact of detailed garment images and the fine-tuning process, we conducted a series of comparative experiments based on the VACE model. We compared four VACE [27] model configurations: the original model, the model using detailed garment images as additional input, the model fine-tuned on the Eevee dataset with LoRA [23], and the model both fine-tuned with LoRA and using the additional detailed garment images.

As shown in Tab. 3, our experimental data reveals that the impact of directly inputting detailed garment images alone is negligible, with the model showing performance comparable to the original VACE baseline. This suggests the model struggles to leverage this extra information without specific adaptation. Conversely, the value of these detailed images becomes apparent during fine-tuning. Specifically, the model fine-tuned with LoRA while using the additional detailed garment images achieves superior results compared to the model fine-tuned with LoRA without them. This strongly indicates that the fine-tuning process is crucial for enabling the model to effectively exploit the high-fidelity information present in the detailed images.

This conclusion is further corroborated by the qualitative visual comparisons presented in Fig. 6. The original

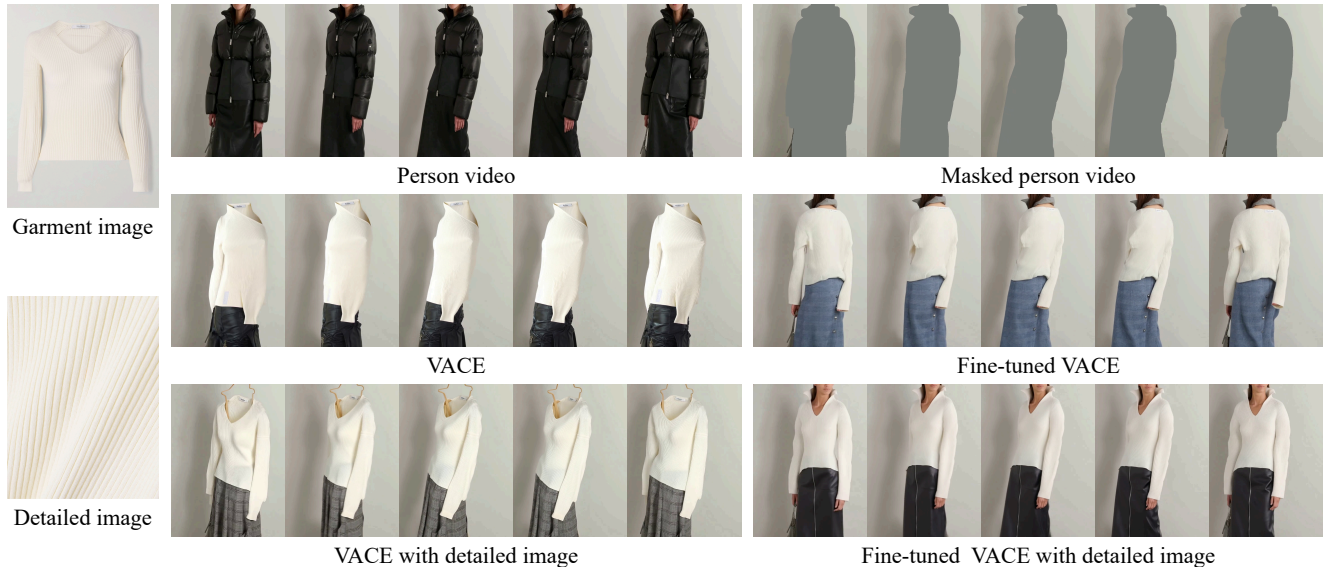


Figure 6. Qualitative comparison of the ablation study. The model combining LoRA fine-tuning and detailed images generates textures with significantly enhanced clarity and fidelity compared to other models. Please zoom in for better visualization

Resolution	Type	LoRA Finetune	Detailed Image	Unpaired			Paired				
				VFID _R ^u ↓	VFID _I ^u ↓	VGID ^u ↑	VFID _R ^p ↓	VFID _I ^p ↓	VGID ^p ↑	SSIM↑	LPIS↓
1088 × 816	Full-shot	×	×	0.454	11.260	0.522	0.282	7.290	0.525	0.899	0.083
		×	✓	0.661	11.290	0.523	0.400	7.380	0.531	0.896	0.084
		✓	×	0.225	10.219	0.520	0.093	6.558	0.529	0.909	0.079
		✓	✓	0.220	10.065	0.524	0.076	6.555	0.534	0.910	0.078
1088 × 816	Close-up	×	×	0.913	12.844	0.533	0.429	9.331	0.515	0.778	0.174
		×	✓	1.218	12.666	0.525	0.628	9.162	0.520	0.770	0.174
		✓	×	0.608	11.757	0.529	0.220	8.441	0.517	0.802	0.156
		✓	✓	0.535	11.621	0.540	0.274	8.429	0.524	0.803	0.154

Table 3. Ablation study of LoRA fine-tuning and detailed images across full-shot and close-up videos. The subscripts ‘u’ and ‘p’ respectively represent the unpaired setting and paired setting. The best results are in bold.

VACE model produced incorrect results, generating videos that featured only the garment with no person. While providing additional detailed images of the garment made the output more natural, the human body was still missing. In contrast, the VACE model, after being fine-tuned on the Eevee dataset, successfully generated a human. Furthermore, the fine-tuned VACE model, when combined with the detailed garment image, not only produced a human body but also achieved a more natural-looking result overall.

Finally, recognizing that distribution metrics can sometimes be inconsistent with human perception, we conducted a user study (N=53) focusing on semantic alignment. Raters preferred our configuration with detailed guidance over the baseline without it in preserving texture (79.25%), pattern (83.02%), and overall quality (84.91%). This subjective evaluation confirms that the detailed image is critical for accurately preserving specific garment identity.

5. Conclusion

In this paper, we addressed the critical limitations of existing video-based virtual try-on: low resolution, only full-shot videos, and the failure to capture fine-grained garment textures from single input garment images. Our primary contribution is the introduction of Eevee, a high-resolution dataset designed to support a wide range of virtual try-on tasks. It is the first dataset to provide both full-shot and close-up videos, and corresponding detailed close-up images for the garments. Recognizing that this task requires better evaluation, we also proposed a new metric VGID to specifically measure garment consistency. Leveraging this dataset and metric, we conducted a comprehensive benchmark of recent state-of-the-art models, establishing a baseline for future research in virtual try-on. We hope our work inspires further research into virtual try-on.

References

- [1] Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024. 3
- [2] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025. 3
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 4
- [4] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 4
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6
- [6] Ruidong Chen, Lanjun Wang, Weizhi Nie, Yongdong Zhang, and An-An Liu. Anyscene: Customized image synthesis with composited foreground. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8724–8733, 2024. 3
- [7] Ruidong Chen, Yancheng Bai, Xuanpu Zhang, Jianhao Zeng, Lanjun Wang, Dan Song, Lei Sun, Xiangxiang Chu, and Anan Liu. Layer-wise instance binding for regional and occlusion control in text-to-image diffusion transformers. *arXiv preprint arXiv:2603.05769*, 2026. 3
- [8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14131–14140, 2021. 3
- [9] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for authentic virtual try-on in the wild. In *European Conference on Computer Vision*, pages 206–235. Springer, 2024. 3
- [10] Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025. 2, 3
- [11] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in neural information processing systems*, 34:9355–9366, 2021. 3
- [12] Xiangxiang Chu, Jianlin Su, Bo Zhang, and Chunhua Shen. Visionllama: A unified llama backbone for vision tasks. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.
- [13] Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. *ICCV*, 2025. 3
- [14] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9026–9035, 2019. 3
- [15] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1161–1170, 2019. 2, 3
- [16] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. 2, 3, 7
- [17] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607, 2023. 3
- [18] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 4
- [19] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3
- [20] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 3
- [21] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 6
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3
- [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6, 7
- [24] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 2, 5, 6, 7
- [25] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on*

- computer-based medical systems (CBMS)*, pages 558–564. IEEE, 2020. 3
- [26] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10799–10808, 2022. 2
- [27] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2, 3, 4, 6, 7
- [28] Jeongho Kim, Guojung Gu, Minhoo Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8176–8185, 2024. 3
- [29] Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025. 3
- [30] Dong Li, Wenqi Zhong, Wei Yu, Yingwei Pan, Dingwen Zhang, Ting Yao, Junwei Han, and Tao Mei. Pursuing temporal-consistent video virtual try-on via dynamic pose interaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22648–22657, 2025. 2
- [31] Guangyuan Li, Siming Zheng, Hao Zhang, Jinwei Chen, Junsheng Luan, Binkai Ou, Lei Zhao, Bo Li, and Peng-Tao Jiang. Magictryon: Harnessing diffusion transformer for garment-preserving video virtual try-on, 2025. 2, 3, 4, 6, 7
- [32] Siqi Li, Zhengkai Jiang, Jiawei Zhou, Zhihong Liu, Xiaowei Chi, and Haoqian Wang. Realvvt: Towards photorealistic video virtual try-on via spatio-temporal consistency. *arXiv preprint arXiv:2501.08682*, 2025. 2, 3
- [33] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024. 3
- [34] Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos, 2025. 3
- [35] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3
- [36] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4
- [37] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 3
- [38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [39] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer, 2024. 3
- [40] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress code: High-resolution multi-category virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2231–2235, 2022. 3
- [41] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. In *Proceedings of the 31st ACM international conference on multimedia*, pages 8580–8589, 2023. 3
- [42] Hung Nguyen, Quang Qui-Vinh Nguyen, Khoi Nguyen, and Rang Nguyen. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6200–6208, 2025. 3
- [43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [44] Zhenglin Pan. Anilines - anime lineart extractor. <https://github.com/zhenglinpan/AniLines-Anime-Lineart-Extractor>, 2025. 4
- [45] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 3
- [46] Long Peng, Wenbo Li, Renjing Pei, Jingjing Ren, Jiaqi Xu, Yang Wang, Yang Cao, and Zheng-Jun Zha. Towards realistic data generation for real-world super-resolution. *arXiv preprint arXiv:2406.07255*, 2024. 3
- [47] Long Peng, Anran Wu, Wenbo Li, Peizhe Xia, Xueyuan Dai, Xinjie Zhang, Xin Di, Haoze Sun, Renjing Pei, Yang Wang, et al. Pixel to gaussian: Ultra-fast continuous super-resolution with 2d gaussian modeling. *arXiv preprint arXiv:2503.06617*, 2025. 3
- [48] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 4
- [49] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kun-chang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. 4
- [50] Dan Song, Jian-Hao Zeng, Min Liu, Xuan-Ya Li, and An-An Liu. Fashion customization: Image generation based on editing clue. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4434–4444, 2023. 2, 3

- [51] Dan Song, Juan Zhou, Jianhao Zeng, Hongshuo Tian, Bolun Zheng, Rongbao Kang, and An-An Liu. Mef-gd: Multimodal enhancement and fusion network for garment designer. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3
- [53] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2, 5, 6
- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [55] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 4
- [56] Jiyuan Wang, Chunyu Lin, Cheng Guan, Lang Nie, Jing He, Haodong Li, Kang Liao, and Yao Zhao. Jasmine: Harnessing diffusion prior for self-supervised depth estimation. *arXiv preprint arXiv:2503.15905*, 2025. 3
- [57] JiYuan Wang, Chunyu Lin, Lei Sun, Rongying Liu, Lang Nie, Mingxing Li, Kang Liao, Xiangxiang Chu, and Yao Zhao. From editor to dense geometry estimator. *arXiv preprint arXiv:2509.04338*, 2025.
- [58] Jiyuan Wang, Chunyu Lin, Lei Sun, Zhi Cao, Yuyang Yin, Lang Nie, Zhenlong Yuan, Xiangxiang Chu, Yunchao Wei, Kang Liao, et al. Geometry-guided reinforcement learning for multi-view consistent 3d scene editing. *arXiv preprint arXiv:2603.03143*, 2026. 3
- [59] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 6
- [60] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019. 4
- [61] Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, and Xiaodan Liang. Gp-vton: Towards general purpose virtual try-on via collaborative local-flow global-parsing learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23550–23559, 2023. 3
- [62] Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. *arXiv preprint arXiv:2507.19946*, 2025. 3
- [63] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3
- [64] Jianhao Zeng, Dan Song, Weizhi Nie, Hongshuo Tian, Tongtong Wang, and An-An Liu. Cat-dm: Controllable accelerated virtual try-on with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8372–8382, 2024. 2, 3
- [65] Nannan Zhang, Yijiang Li, Dong Du, Zheng Chong, Zhengwentai Sun, Jianhao Zeng, Yusheng Dai, Zhengyu Xie, Hairui Zhu, and Xiaoguang Han. Robust-mvton: Learning cross-pose feature alignment and fusion for robust multi-view virtual try-on. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16029–16039, 2025. 2
- [66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 2, 6
- [67] Xuanpu Zhang, Xuesong Niu, Ruidong Chen, Dan Song, Jianhao Zeng, Penghui Du, Haoxiang Cao, Kai Wu, and An-an Liu. Group relative attention guidance for image editing. *arXiv preprint arXiv:2510.24657*, 2025. 3
- [68] Jun Zheng, Fuwei Zhao, Youjiang Xu, Xin Dong, and Xiaodan Liang. Viton-dit: Learning in-the-wild video try-on from human dance videos via diffusion transformers. *arXiv preprint arXiv:2405.18326*, 2024. 2
- [69] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 3
- [70] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10477–10486, 2023. 3
- [71] Tongchun Zuo, Zaiyu Huang, Shuliang Ning, Ente Lin, Chao Liang, Zerong Zheng, Jianwen Jiang, Yuan Zhang, Mingyuan Gao, and Xin Dong. Dreamvvt: Mastering realistic video virtual try-on in the wild via a stage-wise diffusion transformer framework. *arXiv preprint arXiv:2508.02807*, 2025. 2