Vanast: Virtual Try-On with Human Image Animation via Synthetic Triplet Supervision

Hyunsoo Cha Wonjung Woo Byungjun Kim Hanbyul Joo
Seoul National University
{243stephen, agnes2327, byungjun.kim, hbjoo}@snu.ac.kr
https://hyunsoocha.github.io/vanast/

Figure 1. **Vanast.** Given a human image and one or more garment images, our method generates virtual try-on with human image animation conditioned on a pose video while preserving identity.

Abstract

We present Vanast, a unified framework that generates garment-transferred human animation videos directly from a single human image, garment images, and a pose guidance video. Conventional two-stage pipelines treat image-based virtual try-on and pose-driven animation as separate processes, which often results in identity drift, garment distortion, and front–back inconsistency. Our model addresses these issues by performing the entire process in a single unified step to achieve coherent synthesis. To enable this setting, we construct large-scale triplet supervision. Our data generation pipeline includes generating identity-preserving human images in alternative outfits that differ from garment catalog images, capturing full upper and lower garment triplets to overcome the single-garment–posed video pair limitation, and assembling diverse in-the-wild triplets without requiring garment catalog images. We further introduce a Dual Module architecture for video diffusion transformers to stabilize training, preserve pretrained generative quality, and improve garment accuracy, pose adherence, and identity preservation while supporting zero-shot garment interpolation. Together, these contributions allow Vanast to produce high-fidelity, identity-consistent animation across a wide range of garment types.

1. Introduction

A fundamental question arises when aiming to generate a garment-transferred human animation video: *how can one synthesize a realistic video of a person wearing a target garment given only a single human image, one or more garment images, and a pose guidance video?* A straightforward approach to building such a system with existing methods is to first apply an image-based virtual try-on model to a human image and garment images [5, 8, 9, 16], then animate the synthesized result using a pose-driven video generation model [13, 18, 27, 40]. While such a two-stage pipeline can produce plausible videos, it suffers from several inherent limitations. First, discrepancies between the training

distributions of image try-on and video animation models cause identity drift, garment distortion, and compounding artifacts at inference time. Second, the decomposition into two separate models is computationally inefficient. Third, garments have distinct front–back geometry, yet standard video animation models operate from a single static image and therefore lose information required to synthesize consistent appearances across diverse viewpoints. These issues indicate that high-quality virtual try-on animation requires a unified, single-stage generation framework. Motivated by these challenges, we introduce Vanast, an end-to-end system that directly synthesizes garment-transferred human animation videos from a human image, one or more garment images, and a pose guidance video.

Building such a model requires triplet supervision consisting of a human image, garment images, and a RGB video of the same person moving while wearing the target garment. However, no existing public dataset provides this structure. Garment images and fashion videos can be collected from online retail sources, but the human image cannot simply be a frame sampled from the video. If the human image shows the same clothing as the video, the model learns to prioritize motion reenactment rather than garment transfer. The human image must therefore depict the person wearing different clothing from the video garment, a configuration that current datasets do not offer [6, 7, 21]. To address this gap, we present a method to generate human images in alternative outfits while preserving identity. Our method transforms a given human image into a photorealistic version wearing garments different from those in the video, enabling the construction of accurate triplet pairs. To overcome the limitation of online shopping videos, where each clip typically features only one garment category, we additionally capture a dataset containing full upper and lower garment triplets.

Yet two challenges remain. First, transferring garments from real in-the-wild images is difficult because online catalog-style garments differ greatly from casually worn garments in unconstrained environments. Second, captured triplets suffer from limited scene diversity, leading to degradation when models are deployed outside their narrow capture conditions. To overcome these limitations, we introduce a scalable pipeline for constructing triplets from in-the-wild videos that lack garment images. This pipeline discovers garments, filters motion segments, and enforces identity consistency without requiring paired catalog images, expanding both visual diversity and garment variability.

A straightforward approach to train on such triplets is to modify a video diffusion transformer [28] by feeding the human image, garment images, and pose video as additional conditioning signals. One option is to concatenate or fuse these inputs at the token level, but this severely hinders fast convergence and destabilizes optimization [15]. An alternative is to introduce a dedicated context module and fine-tune it to encode all conditions, yet this approach often fails to propagate the full set of constraints uniformly into the generated video, causing certain conditions to be underrepresented or ignored. To address these challenges, we propose a **Dual Module** architecture that preserves the pretrained text-to-video backbone while introducing dedicated pathways for garment transfer and pose guidance while preserving identity. This design converges quickly, maintains the generative fidelity of the original model, and significantly improves garment accuracy, pose adherence, and identity preservation. It further enables zero-shot garment interpolation through its modular conditioning structure, allowing smooth transitions between garment styles without additional finetuning while preserving identity.

In summary, our contributions are as follows:

- We introduce the first unified framework, Vanast, that directly synthesizes human image animation videos with virtual try-on from a single human image, garment images, and a pose guidance video, eliminating the limitations of two-stage virtual try-on pipelines.
- We construct scalable triplet supervision for training by (1) generating identity-preserving human images wearing alternative garments, (2) capturing full upper and lower garment triplets to overcome single-garment constraints of online sources, and (3) building diverse in-the-wild triplets without requiring garment images.
- We propose a Dual Module architecture for video DiT that enables fast and stable convergence, preserves pretrained generative quality, and enhances garment transfer accuracy, pose adherence, and identity preservation, while also supporting zero-shot garment interpolation.

2. Related Work

Virtual Try-On. Virtual try-on has evolved from geometry-driven clothing warping to correspondence-based synthesis that models fine-grained human–garment interactions. Early approaches rely on human and clothing parsing combined with geometric alignment and appearance blending [2, 16, 35]. While effective under well-aligned conditions, their dependence on explicit 2D warping often limits robustness in the presence of large pose variation, occlusions, or non-rigid garment deformation. Recent diffusion-based methods have reshaped the landscape of image-based virtual try-on by replacing hand-designed warping modules with learned correspondence priors. Mask-conditioned dual-UNet architectures [3–5, 33] employ segmentation masks to enforce explicit spatial control and enhance compositional editing. Meanwhile, transformer-based diffusion models [8, 9] leverage global self-attention to implicitly infer garment–body correspondences without relying on mask supervision, demonstrating improved generalization across diverse poses and body shapes. Despite these advances, existing approaches remain fundamentally image-centric. When

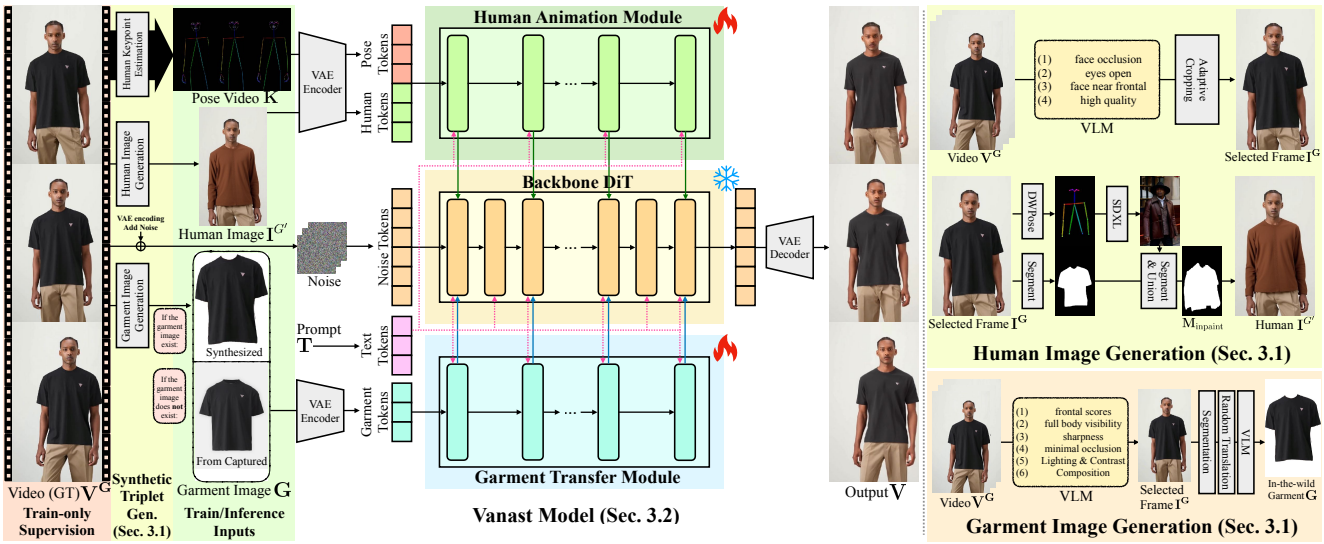

Figure 2. **Overview of Vanast Pipeline.** Our Vanast framework generates virtual try-on human animation videos from a human image, garment images, and a pose video. By incorporating scalable human-image and garment-image generation pipelines, our method avoids dataset-specific constraints and trains effectively at scale. The Dual Modules architecture ensures that the three conditioning signals, human image $\mathbf{I}^{\mathbf{G}'}$, garment images $\mathbf{G}$, and pose video $\mathbf{K}$, are faithfully reflected in the resulting video.

applied to videos, they exhibit temporal flickering and identity drift due to the absence of mechanisms for maintaining consistency across frames. In this work, we revisit VTON from an image-to-animation perspective and introduce a video-native diffusion framework with pose-conditioned generation, enabling unified and temporally coherent virtual try-on in dynamic settings.

Diffusion-based Human Animation. Recent advances in human and portrait animation have been driven by diffusion models that focus on generating coherent motion from a single reference image. Methods such as Animate Anyone and Champ [13, 40] extend 2D UNet architectures [25] with temporal attention layers [10], enabling them to leverage the strong prior knowledge learned from large-scale 2D image models while generating temporally consistent animation videos from only a single image input. Contemporary state-of-the-art avatar animation systems [27, 29, 36] further employ video diffusion frameworks to achieve robust pose transfer and strong identity preservation, even under complex motion patterns. DisPose [18] incorporates ControlNet-based keypoint conditioning to provide accurate, pose-driven control for single-image animation. Existing animation models, however, lack mechanisms for garment transfer and thus are unable to generate video-based virtual try-on results from a human image paired with a separate garment input. Our pipeline addresses this gap by integrating VTON with pose-conditioned video animation to produce motion-consistent videos with garment fidelity.

Subject-driven Image and Video Generation. Subject-driven generation methods condition on identity while composing additional attributes. Diffusion Transformer–based image synthesis models [19, 26, 31] place strong emphasis on identity preservation and compositional control, and they can be adapted to virtual try-on through localized inpainting. However, similar to image-based VTON systems, these models still rely on a separate animation stage to achieve temporal coherence. Recently, diffusion-based models emerge that generate videos directly from a subject [15, 20, 34, 38]. These models take a subject image and a text prompt as input and control the subject's actions and background according to the prompt. VACE [15] builds upon a video diffusion transformer [11, 28] and shows that diverse tasks such as video to video editing and reference to video generation can be unified through a single auxiliary module. Through this process, VACE performs reference to video generation followed by pose conditioned video to video synthesis, thereby enabling pose controlled virtual try on. However, when pose, garment, and human image are jointly conditioned through a single auxiliary module, the model often struggles to preserve fine garment details or to synthesize pose motion accurately. Our approach addresses this limitation by separating pose and garment conditioning into independent network modules and training them jointly. This design improves pose motion synthesis and garment detail accuracy while simultaneously enhancing identity preservation.

3. Method

Our goal is to generate a human animation video with garment transfer in a single stage framework, given a target human image, one or more target garment images, and a motion guidance video, as shown in Fig. 1. Our framework takes, as input, the target garment images $\mathbf{G}$, a target human

image $\mathbf{I}^{\mathbf{G}'}$ showing the person wearing an arbitrary garment $\mathbf{G}' \neq \mathbf{G}$, a motion guidance video $\mathbf{K}$, and a text prompt $\mathbf{T}$ describing human actions and background context. As output, our model generates an F-frame animation sequence denoted as $\mathbf{V} = \{\mathbf{I}_t^{\mathbf{G}}\}_{t=1}^{F}$, where each frame $\mathbf{I}_t^{\mathbf{G}}$ represents a temporally coherent synthesis of the person wearing the target garment items $\mathbf{G}$. Our model supports multiple garment or accessory items, expressed as $\mathbf{G} = \{\mathbf{G}_1, ..., \mathbf{G}_n\}$, encompassing upper and lower garments and accessories such as hats. We define our model as follows:

$$\mathbf{V} = \text{Vanast}(\mathbf{G}, \mathbf{I}^{\mathbf{G}'}, \mathbf{K}, \mathbf{T}). \tag{1}$$

To train Vanast, we construct a triplet dataset consisting of $\mathbf{I}^{\mathbf{G}'}$, $\mathbf{G}$, and a ground truth video $\mathbf{V}^{\mathbf{G}}$ in which the person is in motion while wearing the target garment $\mathbf{G}$. The input motion sequence $\mathbf{K} = \{\mathbf{k}_t\}_{t=1}^{F}$ is obtained by applying an off-the-shelf 2D keypoint estimator DWPose [37] to $\mathbf{V}^{\mathbf{G}}$. Examples from our triplet dataset are shown in Fig. 3.

3.1. Synthetic Triplet Dataset Generation

Constructing the required triplet dataset ($\mathbf{I}^{\mathbf{G}'}$, $\mathbf{G}$, $\mathbf{V}^{\mathbf{G}}$) is challenging, and no publicly available dataset provides the necessary quality or structure for our task. As a key contribution, we introduce a pipeline to generate large quantities of high-quality triplet datasets from multiple data sources. We begin by collecting garment-video pairs ($\mathbf{G}$, $\mathbf{V}^{\mathbf{G}}$) from online shopping platforms, where videos feature diverse identities and backgrounds. However, constructing $\mathbf{I}^{\mathbf{G}'}$, the target person image wearing an arbitrary garment different from $\mathbf{G}'$, is non-trivial, because such platforms rarely include images of the same person wearing multiple outfits. A naive strategy would be to select a frame from $\mathbf{V}^{\mathbf{G}}$, resulting in $\mathbf{I}^{\mathbf{G}'}$ with the identical garment to $\mathbf{G}$. As shown in our ablation study, this causes the model to overfit to animating the human appearance following motion cues $\mathbf{K}$, rather than learning garment transfer.

To address this, we present a synthesis pipeline that generates $\mathbf{I}^{\mathbf{G}'}$ using pre-trained image diffusion models, producing realistic images of the same person wearing different garments. In addition, we introduce a complementary strategy to consgtructs the entire triplet directly from in-the-wild videos $\mathbf{V}^{\mathbf{G}}$, by automatically extracting both $\mathbf{G}$ and $\mathbf{I}^{\mathbf{G}'}$. This enables us to incorporate more diverse garment appearances that go beyond the clean, studio-style images typically found in online shopping environments. Finally, we collect our own studio-captured, high-quality dataset to support scenarios involving multiple garments, where $\mathbf{G} = \{\mathbf{G}_1, ..., \mathbf{G}_n\}$. Examples of our newly collected multi-garment dataset is shown in Fig. 3.

Synthesizing $\mathbf{I}^{\mathbf{G}'}$ from ($\mathbf{G}$, $\mathbf{V}^{\mathbf{G}}$). We leverage a pre-trained diffusion inpainting model, FLUX [17] to modify only the garment or object regions of $\mathbf{I}^{\mathbf{G}}$, where $\mathbf{I}^{\mathbf{G}}$ is a selected frame from $\mathbf{V}^{\mathbf{G}}$. For high quality synthesis, our pipeline

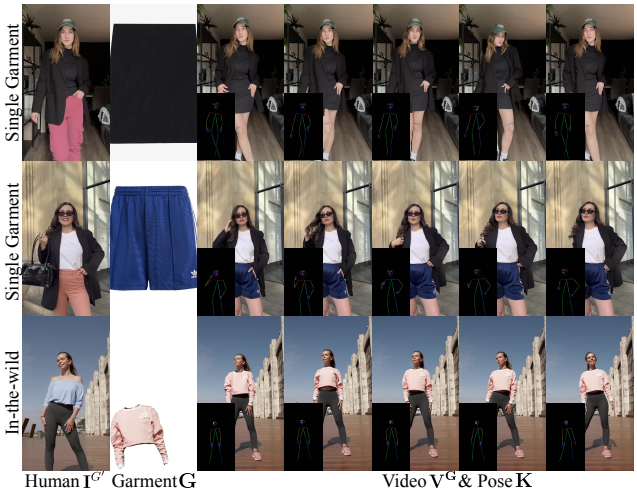

Figure 3. **Samples of Synthetic Triplet Datasets.** We show samples of the datasets used for generation and training. The triplet construction contributes to enabling the model to preserve identity while accurately transferring garments and producing animation videos that follow the target pose.

consists of three stages: (1) selecting suitable candidate frames $\mathbf{I}^{\mathbf{G}}$, (2) constructing proper and effective inpainting mask $\mathbf{M}_{\text{inpaint}}$ to ensure realistic and plausible garment synthesis, and (3) diversifying the synthesized garment type in $\mathbf{I}^{\mathbf{G}}$ using LLM-based prompting.

First, we describe how to select candidate images $\mathbf{I}^{\mathbf{G}}$ from $\mathbf{V}^{\mathbf{G}}$. We randomly sample n frames from each video and select a representative frame using the Vision-Language Model (VLM), Qwen2.5-VL [1]. The selected frame satisfies the following criteria: (1) the face is not occluded by hands or objects (sunglasses or masks are allowed); (2) both eyes are open; (3) the face is near-frontal; and (4) the quality score reflecting focus, noise, and exposure is at least $95/100$. If no frame meets these conditions, the first frame of the video is selected. After frame selection, we perform adaptive cropping to support a wide range of human image scales. After frame selection, the face and body detection model [24] detects the largest face and full-body bounding boxes. Each bounding box is expanded according to a predefined scale, and an interpolated bounding box is generated through random linear interpolation between them. A rectangular crop region with a $9 : 16$ aspect ratio that fully contains the interpolated box is then computed and adjusted to fit within the image boundaries.

For each selected candidate image $\mathbf{I}^{\mathbf{G}}$, we construct a target mask region $\mathbf{M}_{\text{inpaint}}$ for the inpainting. A key requirement is that the mask must not simply follow the silhouette of the original garment $\mathbf{G}$, as the inpainting model tends to preserve the existing garment shape instead of generating a diverse garment for $\mathbf{I}^{\mathbf{G}'}$. Thus, the mask should reflect the expected garment region rather than the observed one. Inspired by PERSE [4], we leverage a text-to-image model [23] to

first synthesize an auxiliary images that maintains the same pose as $\mathbf{I^G}$ but features an arbitrary garment and identity. Then, we extract a garment mask from this synthesized image using an off-the-shelf segmentation model [32].

Finally, using the chosen $\mathbf{I^G}$ and constructed mask $\mathbf{M}_{\text{inpaint}}$, we generate $\mathbf{I^{G'}}$ as an image of the same person wearing a different garment from the same category as $\mathbf{G}$. The text prompts for inpainting are randomly composed from a predefined pool of garment types and colors using ChatGPT [22]. To ensure gender-consistent garment descriptions, we employ the VLM [1] to classify each image as male or female and incorporate the result into the text prompt. Finally, diffusion inpainting model [17] modifies only the garment or object regions based on the generated masks $\mathbf{M}_{\text{inpaint}}$ and prompt, producing high-quality human fashion images $\mathbf{I^{G'}}$. We show an overview of the human image generation process in Fig. 2.

Synthesizing $(\mathbf{I^{G'}}, \mathbf{G})$ from $\mathbf{V^G}$. To further increase pose and background diversity, we present a pipeline to collect the desired triplet data from an in-the-wild video dataset [30]. For constructing $\mathbf{I^{G'}}$, we use the same aforementioned pipeline. However, because in such videos the corresponding garment image $\mathbf{G}$ is not available, we therefore introduce a method to generate $\mathbf{G}$ directly from $\mathbf{V^G}$, enabling the construction of a complete triplet dataset. We design a process select sutable candidate frame to synthesize garment images $\mathbf{G}$ from in-the-wild videos $\mathbf{V^G}$. From each video, n frames are randomly sampled, and frontal scores are obtained using the VLM [1]. Among the top k frames with the highest frontal scores, the best frame is selected according to the following priority criteria evaluated by the VLM: (1) full-body visibility (from head to toe), (2) sharpness, (3) minimal occlusion by hands, arms, or objects, (4) lighting and contrast, and (5) composition. The most suitable frame is then chosen. From the selected frame, we extract an upper-clothing mask using the segmentation model [32]. A garment-highlighted image is generated by filling the background outside the garment mask with white. To prevent the model from being biased by garment position, random translation is applied based on the mask bounding box. Finally, the VLM determines whether each segmented garment or object is valid as the target garment, filtering out unstable segmentations from the segmentation model. This synthetic garment image generation process enables the creation of synthetic triplet datasets from in-the-wild videos, contributing to dataset scale-up and enhancing model robustness. We describe the garment image generation process in Fig. 2.

3.2. Model Architecture

Dual Modules. We introduce a training strategy for our video diffusion model using the constructed synthetic triplet dataset. A common approach is to tokenize all conditions and either concatenate or fuse them, but this often leads to slow convergence during training [15]. Moreover, fine-tuning with a single context module makes it difficult to balance the control of the three conditions [14, 15].

To address these limitations, we propose a **Dual Module** architecture. We adopt a distributed and cascaded structure from the backbone T2V (text-to-video) DiT model [28] and divide it into two specialized modules. Inspired by VACE [15], we design a Human Animation Module (HAM) that focuses on generating human animation using human and pose images, and a Garment Transfer Module (GTM) that handles garment transfer using garment images. HAM and GTM share part of the block architecture with the backbone DiT. This distributed and cascaded design allows the model to progressively integrate contextual information across multiple levels of representation space, leading to richer conditioning effects compared to single-point injection. The overall formulation is defined as follows:

$$h_{l+1} = \begin{cases} \mathrm{B}_l^{\mathrm{T2V}}(h_l)\,, & \text{if } l \neq 2k, \\ \mathrm{B}_l^{\mathrm{T2V}}(h_l) + \alpha \cdot \mathrm{B}_l^{\mathrm{HAM}}(h_l) \\ \qquad\qquad + \beta \cdot \mathrm{B}_l^{\mathrm{GTM}}(h_l)\,, & \text{if } l = 2k. \end{cases} \tag{2}$$

where B denotes each transformer block in the DiT backbone, l is the index of a transformer block, k is a non-negative integer ranging from 0 to 14, and h denotes the hidden state which is the input or output of a block. The scalar $\alpha = 0.5, \beta = 0.5$ controls the relative strength between the HAM and GTM, determining the balance of their contributions during feature integration. We freeze the backbone DiT during training and optimize only the HAM and GTM modules. An overview of our model architecture is presented in Fig. 2. For detailed model architecture and implementation, refer to supp.mat. Sec. B.

Tokenization. To provide tokenized inputs for the Dual Modules, we convert each component of the synthetic triplet dataset into latent representations using the pretrained VAE encoder $\mathcal{E}_{\text{VAE}}$ [28]. Let z_{H}, z_{G}, and z_{P} denote the encoded latents of $\mathbf{I^{G'}}$, $\mathbf{G}$, and $\mathbf{K}$, respectively. For the HAM module, we construct a motion-conditioned appearance context by performing frame-wise concatenation of z_{H} and z_{P} along the temporal dimension, following previous approach [15]. For the GTM module, we use z_{G} alone as input; to match its temporal dimension with HAM, a zero tensor is appended as a placeholder before concatenation. Finally, a 3D convolutional projection layer maps each concatenated latent volume into token embeddings suitable for downstream processing.

Garment interpolation. Our model is capable of transferring interpolated garments, constructed from two garments $\mathbf{G_A}, \mathbf{G_B}$ belonging to the same category, to the human animation in a zero-shot manner without any additional optimization [4, 39]. To achieve this, we obtain the outputs of the GTM transformer blocks for each garment image and compute an γ-weighted summation of the two representations as

follows:

$$h_{l+1} = \mathrm{B}_l^{\mathrm{T2V}}(h_l) + \alpha \cdot \mathrm{B}_l^{\mathrm{HAM}}(h_l) + \gamma \cdot \mathrm{B}_l^{\mathrm{GTM}}(h_l; \mathbf{G_A}) + (1-\gamma) \cdot \mathrm{B}_l^{\mathrm{GTM}}(h_l; \mathbf{G_B}), \tag{3}$$

where $\gamma \in [0, 1]$ denotes the interpolation ratio. This allows the model to produce smooth and semantically coherent interpolations between garments.

4. Experiments

Datasets. We train our model on a total of $9,135$ videos, each ranging from 3 to 10 seconds in length. These videos are sourced from public internet shopping-mall sites, our captured dataset, and an in-the-wild video dataset [30]. We construct two evaluation datasets to compare our model with the baselines. The first dataset, referred to as the Internet dataset, is sourced from publicly available shopping-mall websites and consists of videos and standalone garment product images. The second dataset is built using the official test split of the ViViD dataset [7]. Since the ViViD videos do not contain visible faces, we generate $\mathbf{I}^{\mathbf{G}'}$ using an image outpainting model [17] before leveraging our synthetic triplet dataset generation pipeline. Although the alignment requirements for human images differ across VTON models, they generally support full-body inputs. Therefore, for evaluation, we use only full-body or near–full-body images as $\mathbf{I}^{\mathbf{G}'}$. In total, we randomly sample 80 samples from the Internet dataset and 50 samples from the ViViD dataset. Note that the garment images, human images, and pose videos used for evaluation are entirely disjoint from those in the training set. Further details are provided in the supp.mat.

Metrics. We evaluate each method with standard metrics for fidelity and perceptual quality. In particular, we report six metrics: L_1, PSNR, SSIM, LPIPS, FID [12], and VFID [7]. We apply L1 distance, PSNR, SSIM, and LPIPS frame-wise between generated video and ground-truth. Together, these metrics quantify both the accuracy of the garment transfer and the degree of identity preservation. We also leverage the Fréchet Inception Distance (FID) and VFID to compute perceptual realism and temporal consistency.

4.1. Comparisons

Baselines. Since no current method directly produces a virtual try-on video from a human image, a garment image, and a keypoint video, we construct a two-stage pipeline for quantitative comparison. Stage 1 synthesizes a single image of the person wearing the target garment, and Stage 2 animates the image generated in stage 1 using the target motion. We again divide stage 1 into two types, image virtual try-on generation and subject-to-image generation, resulting in 16 model combinations in total.

The image virtual try-on models we use for garment transfer in Stage 1 are as follows: OOTDiffusion [33], a mask-conditioned dual-Unet diffusion model for reference-guided garment transfer; CatVTON [5], a diffusion-based try-on model that supports mask-conditioned synthesis; OmniTry [8] and Any2AnyTryon [9], diffusion transformer models that operate without explicit masks for reference-guided try-on. Regarding CatVTON, the model produces cropped outputs. To match the aspect ratio and resolution across methods, we composite CatVTON results back onto the original image. This may favor pixel-wise metrics by reducing framing discrepancies. For subject-to-image generation models we use: VisualCloze [19], MOSAIC [26], and UNO [31], diffusion-transformer-based models for subject-conditioned image synthesis including virtual try-on. For human image animation in Stage 2, we employ two models: StableAnimator [27], video diffusion-based human image animation models; DisPose [18], a video diffusion human animation model with ControlNet module.

We additionally evaluate VACE [15], a diffusion transformer-based model for subject-driven video generation. While VACE does not explicitly support a single-stage try-on pipeline, its unified architecture can accept multiple conditioning signals simultaneously. Therefore, we also evaluate VACE in a single-stage inference setting, where the model receives a pose video, a human image, and a garment image as joint inputs to directly produce the final video. For completeness, we additionally assess a two-step variant in which we first generate a video of the person wearing the garment from the human and garment images, and then feed the first

Figure 4. **Qualitative Comparisons (Subject-to-Image-based).** We compare our results with baselines constructed by combining subject-to-image generation and animation models. Our method produces the most accurate pose following and garment transfer while preserving identity with high fidelity.

Figure 5. **Qualitative Comparisons (Virtual Try-On-based).** We compare our results with baselines formed by combining image virtual try-on models with animation models. Our method achieves the most accurate pose following and garment transfer while preserving identity with the highest fidelity.

Table 1. **Quantitative Comparison with the Combination of Subject-to-Image and Animation Models.** We compare our model with a baseline that combines a subject-to-image model and an animation model. Our model achieves the best performance across all metrics. **Bold text** indicates the best score in each column.

		Internet Dataset							ViViD Dataset						
Img.Gen.	Animation	L_1 ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	$VFID_{I3D}$ ↓	$VFID_{ResNeXt}$ ↓	L_1 ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	$VFID_{I3D}$ ↓	$VFID_{ResNeXt}$ ↓
VisualCloze	StableAnimator	0.2266	11.08	0.7210	0.4676	210.37	34.48	1.69	0.2835	8.80	0.6355	0.6061	190.83	49.04	7.21
	DisPose	0.2590	9.49	0.6527	0.5884	205.68	40.49	1.93	0.3043	8.25	0.5880	0.6871	214.00	55.82	11.71
MOSAIC	StableAnimator	0.1875	12.20	0.7128	0.4400	158.76	32.36	2.34	0.2382	10.02	0.6617	0.5583	158.47	44.46	2.55
	DisPose	0.1714	12.36	0.7133	0.4641	155.18	35.24	1.54	0.2008	10.67	0.6686	0.5619	155.98	50.90	4.91
UNO	StableAnimator	0.2025	11.79	0.7148	0.4381	162.15	31.73	3.12	0.2556	9.63	0.6622	0.5577	158.31	45.39	2.95
	DisPose	0.1774	12.06	0.7071	0.4734	154.27	34.08	1.65	0.2125	10.41	0.6610	0.5609	140.26	45.55	3.03
VACE (2-stage)		0.1708	12.44	0.6618	0.4507	141.89	49.87	5.44	0.1994	11.03	0.6053	0.5678	143.00	46.22	4.59
VACE (1-stage)		0.1453	13.09	0.6894	0.4052	115.40	47.31	5.86	0.1733	11.62	0.6245	0.5363	134.96	43.97	3.20
Ours		**0.0719**	**17.95**	**0.7550**	**0.2370**	**91.05**	**22.52**	**0.39**	**0.1077**	**14.67**	**0.6686**	**0.3649**	**105.89**	**35.72**	**1.30**

Table 2. **Quantitative Comparison with the Combination of Image Virtual Try-On and Animation Models.** We compare our model with a baseline that combines a image virtual try-on model and an animation model. Our model achieves the best performance across all metrics. **Bold text** indicates the best score in each column.

		Internet Dataset							ViViD Dataset						
Img.Gen.	Animation	L_1 ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	$VFID_{I3D}$ ↓	$VFID_{ResNeXt}$ ↓	L_1 ↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	$VFID_{I3D}$ ↓	$VFID_{ResNeXt}$ ↓
OOTDiffusion	StableAnimator	0.1305	13.91	0.7412	0.3428	174.79	33.17	2.55	0.2431	9.99	0.6566	0.5528	168.46	43.39	2.66
	DisPose	0.1143	14.86	0.7551	0.3325	172.94	31.48	1.68	0.2101	10.72	0.6421	0.5927	169.92	45.13	8.32
CatVTON	StableAnimator	0.1242	14.56	0.7649	0.3273	132.09	26.43	0.90	0.2415	10.04	**0.6772**	0.5441	167.27	43.12	4.62
	DisPose	0.0987	15.87	**0.7779**	0.3238	120.69	25.00	1.05	0.2083	10.66	0.6430	0.6018	155.30	43.82	9.62
OmniTry	StableAnimator	0.1227	14.53	0.7671	0.3178	121.04	25.66	1.18	0.2372	10.10	0.6726	0.5425	161.33	43.36	4.21
	DisPose	0.0968	15.46	0.7775	0.3151	111.08	24.60	0.97	0.2027	10.75	0.6496	0.5867	149.84	43.76	8.45
Any2AnyTryon	StableAnimator	0.1250	14.37	0.7596	0.3271	126.15	26.28	1.36	0.2460	9.78	0.6685	0.5518	165.04	45.13	4.26
	DisPose	0.0950	15.53	0.7778	0.3160	116.39	25.39	0.78	0.2083	10.58	0.6426	0.5969	154.38	43.69	8.07
Ours		**0.0719**	**17.95**	0.7550	**0.2370**	**91.05**	**22.52**	**0.39**	**0.1077**	**14.67**	0.6686	**0.3649**	**105.89**	**35.72**	**1.30**

frame of that video together with the pose video to obtain the final animation.

Results. As shown in Tab. 1, our model achieves the best performance across all metrics when compared with combinations of subject-to-image generation models and animation models. Qualitative results in Fig. 4 further confirm that our approach produces the most accurate pose following and garment transfer, while preserving identity more faithfully than all subject-to-image–based baselines.

As presented in Tab. 2, our model also outperforms all combinations of virtual try-on and animation models on every metric except SSIM, for which our score remains comparable to the best-performing method. Qualitative comparisons in Fig. 5 demonstrate that our results most closely resemble the ground truth among all image virtual try-on–based baselines. Additional results are provided in the supplemental materials.

Figure 6. **Result of Single Garment Transfer.** We present virtual try-on with human image animation results generated from a single garment image.

4.2. Ablation Study

We evaluate the contributions of our key architectural components as well as the effectiveness of the synthetic triplet dataset. Quantitative results are reported in Tab. 3, and qualitative comparisons are shown in Fig. 7. **"Single Module"** refers to a baseline in which the T2V backbone is frozen and a single trainable module is used. All conditions from the triplet dataset are concatenated and fed into this module for training. **"Backbone-LoRA"** denotes a model that directly concatenates all input conditions into the T2V backbone without introducing any additional modules, and fine-tunes the model using LoRA layers, applied to every DiT block. This setup enables faster convergence while preserving the generative capabilities of T2V. **"w/o SynthHuman"** shares the same architecture as our full model but is trained without using $I^{G'}$, relying solely on I^{G}.

As shown in Tab. 3, our model achieves the highest performance across all metrics. The qualitative results in Fig. 7 further demonstrate that our outputs most closely resemble the ground truth. The "Single Module" baseline fails to accu-

Table 3. **Ablation Study.** We conduct ablation study for each component of our model and dataset configuration. **Bold text** indicates the best score in each column.

Method	$L_1 \downarrow$	PSNR↑	SSIM↑	LPIPS↓	FID↓	$\text{VFID}_{\text{I3D}} \downarrow$	$\text{VFID}_{\text{ResNeXt}} \downarrow$
Single Module	0.1162	14.28	0.6609	0.3974	108.84	39.64	1.76
Backbone-LoRA	0.1359	13.17	0.6314	0.4502	120.97	42.47	1.87
w/o SynthHuman	0.1163	14.62	0.6653	0.3943	110.76	38.89	1.93
Ours	**0.1069**	**14.74**	**0.6657**	**0.3673**	**104.59**	**35.60**	**1.21**

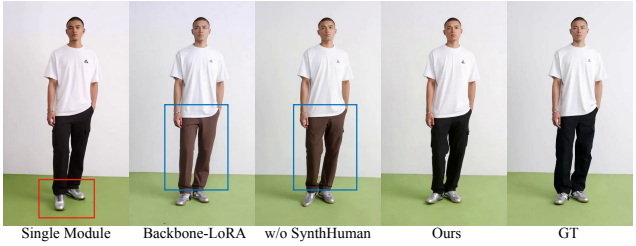

Figure 7. **Ablation Study.** We present the ablation study results for the lower garment transfer. The red box in the "Single Module" result demonstrates vulnerability to pose conditions. Both "Backbone-LoRA" and "w/o SynthHuman" fail to achieve accurate garment transfer, as indicated in blue box. In contrast, our full model produces results most similar to the ground truth.

rately control pose conditions, while the "Backbone-LoRA" and "w/o SynthHuman" variants struggle to perform correct garment transfer.

4.3. Application

Garment interpolation. As shown in Fig. 10, our model is capable of generating human animation videos in a single stage without any additional training, while interpolating and transferring garments between two input garments according to the interpolation weight α. Both upper and lower garments exhibit smooth and natural interpolation.

Multiple garment transfer. Our model supports single-garment transfer, as illustrated in Fig. 6, and further enables multiple-garment transfer through the GTM trained on the synthetic triplet dataset. As shown in Fig. 8, our approach can transfer both upper and lower garments simultaneously and generate human image animations without any additional training. The results demonstrate that garment details for both regions are well preserved while maintaining strong identity consistency.

In-the-wild garment transfer. Benefiting from the combination of the in-the-wild dataset and the synthetic triplet dataset, our model is able to perform garment transfer from in-the-wild garment images. As shown in Fig. 9, the model successfully transfers garments despite the mismatch between the garment image pose and the human pose during animation, while maintaining strong temporal consistency throughout the generated sequence.

Figure 8. **Result of Multiple Garment Transfer.** We present zero-shot garment transfer results where both upper and lower garments are transferred simultaneously. The logos and fine details of the garments are well preserved and accurately reflected in the generated animation videos.

Figure 9. **Result of In-the-wild Garment Transfer.** We present garment transfer results using in-the-wild garment images provided by the TikTokDress [21] dataset.

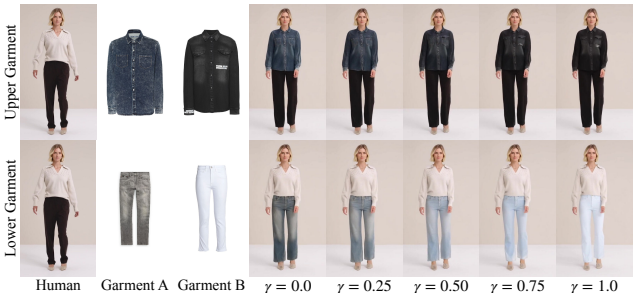

Figure 10. **Result of Garment Interpolation.** Without requiring any additional finetuning, our Vanast model performs zero-shot transfer of interpolated garments by GTM. γ denotes a scalar interpolation weight.

5. Conclusion

We introduce Vanast, a unified framework that synthesizes garment-transferred human animation videos directly from a single human image, garment images, and a pose guidance video. By constructing scalable triplet supervision from in-the-wild videos and complementary upper–lower garment captures, our pipeline addresses the structural limitations of existing online data and enables identity-preserving training with diverse garment variations. The proposed Dual Module architecture significantly improves garment fidelity, pose adherence, and identity preservation while supporting zero-shot garment interpolation and multi-garment transfer without finetuning. Extensive experiments show that Vanast consistently surpasses two-stage pipelines combining state-of-the-art virtual try-on and subject-to-image with human animation models, and generalizes effectively to in-the-wild garment images with strong temporal coherence.

Acknowledgments. We constructed part of the training dataset for the Vanast model using the Pixophilia-SNU dataset. This work was supported by Samsung Electronics MX Division, NRF grant funded by the Korean government (MSIT) (No. RS-2022-NR070498), and IITP grant funded by the Korea government (MSIT) [No. RS-2024-00439854, No. 2022-0-00156, No. RS-2025-25441838, and No. RS-2021-II211343]. H. Joo is the corresponding author.

References

[1] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4, 5

[2] H. Cha, B. Kim, and H. Joo. Pegasus: Personalized generative 3d avatars with composable attributes. In *CVPR*, 2024. 2

[3] H. Cha, B. Kim, and H. Joo. Durian: Dual reference image-guided portrait animation with attribute transfer. *arXiv preprint arXiv:2509.04434*, 2025. 2

[4] H. Cha, I. Lee, and H. Joo. Perse: Personalized 3d generative avatars from a single portrait. In *CVPR*, 2025. 4, 5

[5] Z. Chong, X. Dong, H. Li, S. Zhang, W. Zhang, X. Zhang, H. Zhao, D. Jiang, and X. Liang. Catvton: Concatenation is all you need for virtual try-on with diffusion models. *arXiv preprint arXiv:2407.15886*, 2024. 1, 2, 6

[6] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *ICCV*, 2019. 2

[7] Z. Fang, W. Zhai, A. Su, H. Song, K. Zhu, M. Wang, Y. Chen, Z. Liu, Y. Cao, and Z.-J. Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024. 2, 6

[8] Y. Feng, L. Zhang, H. Cao, Y. Chen, X. Feng, J. Cao, Y. Wu, and B. Wang. Omnitry: Virtual try-on anything without masks. *arXiv preprint arXiv:2508.13632*, 2025. 1, 2, 6

[9] H. Guo, B. Zeng, Y. Song, W. Zhang, C. Zhang, and J. Liu. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. *arXiv preprint arXiv:2501.15891*, 2025. 1, 2, 6

[10] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 3

[11] Y. HaCohen, N. Chiprut, B. Brazowski, D. Shalem, D. Moshe, E. Richardson, E. Levin, G. Shiran, N. Zabari, O. Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 6

[13] L. Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 1, 3

[14] Z. Jiang, C. Mao, Z. Huang, A. Ma, Y. Lv, Y. Shen, D. Zhao, and J. Zhou. Res-tuning: A flexible and efficient tuning paradigm via unbinding tuner from backbone. *NeurIPS*, 2023. 5

[15] Z. Jiang, Z. Han, C. Mao, J. Zhang, Y. Pan, and Y. Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025. 2, 3, 5, 6

[16] J. Kim, G. Gu, M. Park, S. Park, and J. Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024. 1, 2

[17] B. F. Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 4, 5, 6

[18] H. Li, Y. Li, Y. Yang, J. Cao, Z. Zhu, X. Cheng, and L. Chen. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*, 2024. 1, 3, 6

[19] Z.-Y. Li, R. Du, J. Yan, L. Zhuo, Z. Li, P. Gao, Z. Ma, and M.-M. Cheng. Visualcloze: A universal image generation framework via visual in-context learning. *arXiv preprint arXiv:2504.07960*, 2025. 3, 6

[20] L. Liu, T. Ma, B. Li, Z. Chen, J. Liu, G. Li, S. Zhou, Q. He, and X. Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 3

[21] H. Nguyen, Q. Q.-V. Nguyen, K. Nguyen, and R. Nguyen. Swifttry: Fast and consistent video virtual try-on with diffusion models. In *AAAI*, 2025. 2, 8

[22] OpenAI. Chatgpt (gpt-5). https://chat.openai.com, 2025. 5

[23] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 4

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 4

[25] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[26] D. She, S. Fu, M. Liu, Q. Jin, H. Wang, M. Liu, and J. Jiang. Mosaic: Multi-subject personalized generation via correspondence-aware alignment and disentanglement. *arXiv preprint arXiv:2509.01977*, 2025. 3, 6

[27] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu. Stableanimator: High-quality identity-preserving human image animation. In *CVPR*, 2025. 1, 3, 6

[28] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3, 5

[29] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 2025. 3

[30] Z. Wang, Y. Li, Y. Zeng, Y. Fang, Y. Guo, W. Liu, J. Tan, K. Chen, T. Xue, B. Dai, et al. Humanvid: Demystifying training data for camera-controllable human image animation. *NeurIPS*, 2024. 5, 6

[31] S. Wu, M. Huang, W. Wu, Y. Cheng, F. Ding, and Q. He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 3, 6

[32] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 5

[33] Y. Xu, T. Gu, W. Chen, and A. Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *AAAI*, 2025. 2, 6

[34] Z. Xu, Z. Huang, J. Cao, Y. Zhang, X. Cun, Q. Shuai, Y. Wang, L. Bao, J. Li, and F. Tang. Anchorcrafter: Animate cyber-anchors saling your products via human-object interacting video generation. *arXiv preprint arXiv:2411.17383*, 2024. 3

[35] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020. 2

[36] S. Yang, H. Li, J. Wu, M. Jing, L. Li, R. Ji, J. Liang, H. Fan, and J. Wang. Megactor-σ: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer, 2025. 3

[37] Z. Yang, A. Zeng, C. Yuan, and Y. Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023. 4

[38] S. Yuan, J. Huang, X. He, Y. Ge, Y. Shi, L. Chen, J. Luo, and L. Yuan. Identity-preserving text-to-video generation by frequency decomposition. In *CVPR*, 2025. 3

[39] K. Zhang, Y. Zhou, X. Xu, B. Dai, and X. Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *CVPR*, 2024. 5

[40] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, 2024. 1, 3