This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# Pose-Diverse Multi-View Virtual Try-on from a Single Frontal Image via Diffusion Transformer

Seonghee Han* Minchang Chung* Gyeongsu Cho Kyungdon Joo Taehwan Kim†
Ulsan National Institute of Science and Technology (UNIST), Republic of Korea
{seonghee, minchang, threedv, kyungdon, taehwankim}@unist.ac.kr

## Abstract

*We study multi-view virtual try-on framework from a single reference image of themselves and a single frontal image of a garment. While most existing approaches focus on single-view synthesis, their reliance on a single, fixed viewpoint limits their application in immersive environments that require diverse poses and viewpoints. The ability to generate a multi-view virtual try-on image is crucial for a comprehensive user experience, as if the user to inspect the garment from multiple views, including the back and sides, providing a similar experience to a real fitting room. In this paper, we propose a novel framework for pose-controllable, multi-view virtual try-on from a single image. Our method incorporates the proposed cross attention injection for high-quality synthesis and allows for fine-grained pose control. Our framework consists of two stages: the first stage generates a clean frontal try-on result using an off-the-shelf model, and the second stage creates diverse viewpoints using a diffusion transformer. Additionally, an attention injection mechanism ensures consistent identity and garment preservation across synthesized views. Unlike conventional methods that require multiple images of the user or the garment from various angles, our model eliminates these constraints by synthesizing multi-view results from a single input image pair. Our method not only generates realistic images but also enables users to virtually inspect the fit and arrangement of the garment from multiple angles without the need for additional data. Our extensive experiments demonstrate that our framework outperforms existing 3D-based multi-view virtual try-on methods in terms of image quality and pose diversity.*

## 1. Introduction

Virtual try-on (VTON) is a task that aims to synthesize a photo-realistic image of a reference person wearing a target garment. The demand for this technology has surged in recent years, propelled by the rapid growth of the online fashion market. This technology enhances the shopping experience by allowing consumers to visualize how the garments look on them while also helping retailers reduce the costs associated with returns. VTON helps reduce the fundamental uncertainty consumers face when buying apparel online, making it a critical tool for modern e-commerce[13]. However, existing image-based VTON methods [3, 7, 9, 11, 13, 14, 22, 26, 39, 40, 43] still face limitations. They are restricted to single-view synthesis, and a fixed frontal viewpoint does not provide a comprehensive and realistic visualization of the fit, texture, and drape of a garment. As a result, users are unable to observe the back or sides of the clothing, and a static input pose makes it difficult to fully grasp the true style of the garment. This limitation fails to meet the requirements of modern immersive applications, which require multi-angle views of clothing. We aim to bridge the gap between existing single-view try-on and immersive 3D experiences by enabling view-consistent synthesis without 3D supervision.

Although recent methods such as VTON360 [16] and MV-VTON [36] have attempted to address multi-view try-on, they require multiple human images from different views as well as both front and back images of the garment. These demanding input requirements hinder their deployment in real-world scenarios where usually only a single reference image is available. To overcome these limitations, we propose a novel framework that synthesizes high-fidelity, pose-controllable, multi-view try-on results from a single input image-garment pair, as shown in Figure 1. Our framework consists of a two-stage approach that decouples frontal view virtual try-on from multi-view synthesis, leaving the frontal generation to an off-the-shelf model. In the first stage, we leverage Gemini 2.5 Flash Image [4], serving as an off-the-shelf VTON model, to perform a frontal try-on. In the second stage, our proposed Multi-view Synthesis Diffusion Transformer takes this frontal result and synthesizes realistic novel views under arbitrary poses. Crucially, our model supports pose control, allowing users to observe

* Equal contribution.
† Corresponding author.

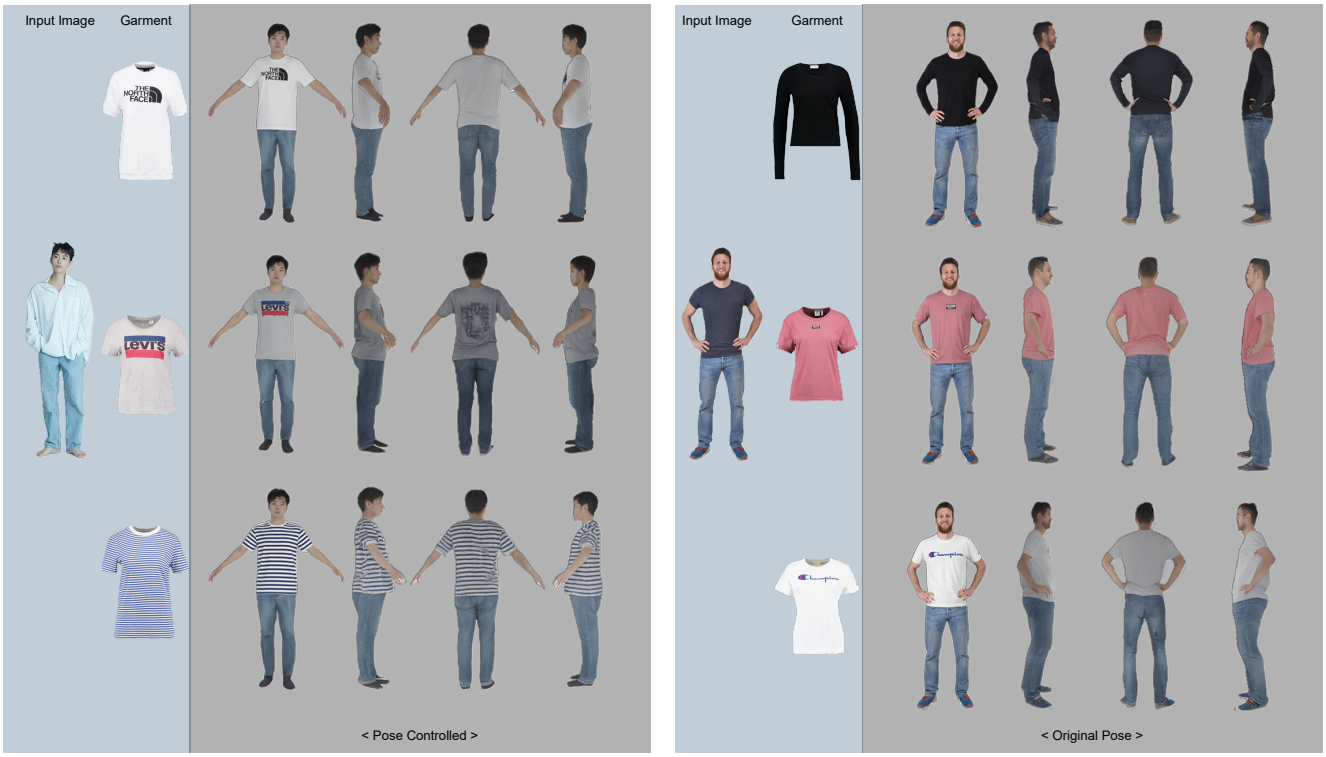



Figure 1. In-the-wild generation results using our method. The left column shows pose-controlled try-on outputs, where the garment is rendered in a new target pose. The right column shows results with the same pose as the input person's image.

how the garment fits across different angles and body poses. This capability significantly enhances the practical utility of VTON systems and bridges the gap between online shopping and real-world fitting experiences.

In addition, we introduce a novel attention manipulation strategy within our Diffusion Transformer, which explicitly injects information from the source image into the cross-attention layers. Our mechanism better preserves facial identity and garment fidelity across views, preventing appearance drift.

Our contributions are summarized as follows:

- We propose a novel two-stage framework for multi-view virtual try-on from a single input image, enabling realistic garment visualization from a single frontal image and arbitrary poses.
- Our method restructures the try-on pipeline by first employing an off-the-shelf frontal try-on model to obtain a single view and then introducing a dedicated Diffusion Transformer that expands this frontal result into high-quality, view-consistent novel poses and viewpoints.
- Unlike previous methods that require multi-view human and garment images, our approach works with a single front-view human image and a garment image, significantly improving practicality.
- We enable fine-grained pose control, allowing users to explore how garments fit across diverse body postures and viewing angles, thus providing a realistic try-on experience akin to a real fitting room.

## 2. Related Work

### 2.1. Diffusion Models

Recent advances in text/image-to-image generation have been largely driven by diffusion models such as Stable Diffusion [30], which can generate high-quality images efficiently. To further steer the generation process, frameworks like ControlNet [42] introduce auxiliary conditioning inputs (e.g., pose or depth maps), allowing for precise structural control over the outputs. More recently, the Diffusion Transformer(DiT) [28] has been introduced as an alternative backbone for diffusion models, demonstrating enhanced generation quality. For instance, Stable Diffusion 3 [6] exemplifies this trend with a multi-modal DiT (mmDiT) architecture integrated with a rectified flow model, employing a unified attention mechanism that processes both text and image modalities. This scalable architecture is particularly advantageous for complex generation tasks requiring global context modeling, such as ensuring consistency across multiple viewpoints. Our method is built upon a DiT backbone and extends it for multi-view human generation,

where maintaining view-consistent identity and garment appearance is particularly challenging. This design improves text-image alignment while supporting stable and efficient training for high-resolution generation.

### 2.2. Multi-view Diffusion Models

A pioneer work on multi-view diffusion model is Zero-1-to-3 [24], which first proposed using camera viewpoints as conditions to achieve novel view synthesis from a single image. However, its view-by-view generation process often led to inconsistencies due to the independent sampling and stochastic nature of diffusion models. To address this, methods such as Zero123++ [31] improve the stability and consistency of the generation per view. Later works, such as MVDream [32], adopt an all-view-at-once approach with shared multi-view attention across viewpoints, which ImageDream [37] later adapted for image-conditioned tasks. Wonder3D [25] further enhanced geometric consistency by incorporating normal maps and cross-domain attention. A common limitation of these methods is their training on generic object datasets; they struggle with facial or garment details, often causing identity loss or blurred textures.

To optimize the computation of multi-view attention, recent studies have proposed more efficient mechanisms. EpiDiff [17] employs epipolar attention to constrain matching candidates along epipolar lines, while Era3D [23] uses row-wise and geometry-aware attention under orthographic projection assumptions. We utilize the power of multi-view diffusion for single image, single garment to VTON.

### 2.3. Image-based Virtual Try-on

Image-based Virtual Try-On (VTON) aims to transfer a target garment onto a person's image while preserving their identity and pose [19]. Early approaches are predominantly based on Generative Adversarial Networks (GANs) [8] and typically follow a two-stage process. The first stage involves deforming the garment to align with the target's body shape and pose. Seminal works like VITON [13] and CP-VTON [35] utilize Thin Plate Spline transformations for this warping process, while others [14, 15] employ optical flow estimation to handle garment deformation. The second stage then fuses the warped garment with the person's image using a GAN-based generator. However, these warping-based methods often struggle with cross-sample alignment, and their effectiveness diminishes when dealing with intricate clothing details or complex poses. Furthermore, the GAN-based generators are prone to training instability, mode collapse, and often produce artifacts.

To overcome the limitations of GANs, recent works have reframed VTON as a conditioned inpainting task, leveraging the powerful generative priors of large-scale diffusion models. This shift lead to more natural-looking results. Initial methods [3, 34] explore various ways to condition the diffusion process. For example, TryOnDiffusion [44] introduces a parallel U-Net architecture to process the person and garment features separately, while LaDI-VTON [26] encodes the garment using textual inversion to serve as a condition. DCI-VTON [9] proposes a hybrid framework that combines a pre-trained warping network with a diffusion model.

Despite these advances, early diffusion-based VTON models often have difficulty preserving the high-frequency details of the original clothing. Subsequent research focuses on enhancing detail fidelity through more sophisticated conditioning mechanisms. Methods such as StableVITON [22], OOTDiffusion [39], and FitDiT [20] introduce specialized encoders and attention mechanisms to better integrate garment features. A notable trend is the use of a ReferenceNet to extract hierarchical garment features, which are then injected into the main generation network via attention. While high-fidelity results are achieved, many of these models still rely on intensive preprocessing steps like human parsing, masking, and pose estimation. To address this, the latest works such as Boow-VTON [43] and Any2AnyTryon [11] are pioneering fully mask-free approaches, aiming to make virtual try-on more applicable to in-the-wild images without dependencies on precise annotations. Nevertheless, these approaches remain inherently single-view, which lack an explicit mechanism for multi-angle or pose-controllable generation. Moreover, they cannot propagate a single frontal view to unseen viewpoints without auxiliary views or 3D assets.

In contrast, our framework combines identity-preserving virtual try-on with multi-view, pose-controllable synthesis, which produces view-consistent results across diverse poses by coupling a Diffusion transformer with an attention-injection mechanism and a pose-conditioned ControlNet.

## 3. Method

In this section, we present our framework for multi-view virtual try-on (VTON) from a single pair of human image and garment image. Our approach is designed to generate high-quality VTON results across diverse viewpoints while maintaining both garment fidelity and human identity. Instead of performing VTON for each target view, we first generate a frontal view try-on image, which then serves as a strong condition for our multi-view diffusion model to synthesize various viewpoints as visualized in Figure 1. Our pipeline is composed of two main stages. In the first stage, we put a garment on a frontal view image as illustrated in Figure 2. Optionally, we transform the original input image into an A-pose to ensure geometric consistency and maximize body visibility. This optional branch results A-posed frontal image for the 2D VTON, where the off-the-shelf image model applies the garment to the transformed image. In the second stage which is the multi-view synthesis stage, as

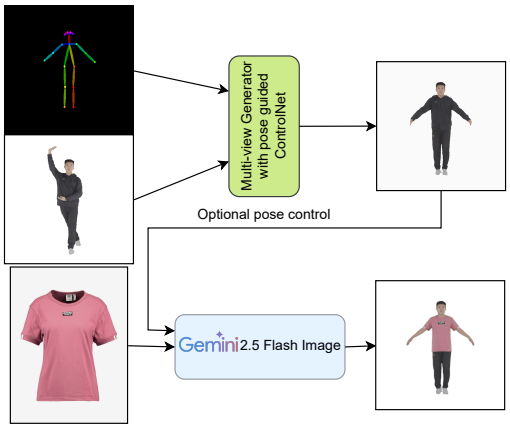



Figure 2. Generated pose control result and corresponding virtual try-on result. The Pose-guided Multi-view Generator synthesizes views based on the target pose, which are then processed by Gemini 2.5 Flash to produce the final frontal try-on image.

shown in Figure 3, a single high-quality try-on image is fed into our multi-view generator to produce a set of consistent multi-view outputs.

### 3.1. First stage: Image Virtual Try-on

The frontal image is first passed to an off-the-shelf image-based VTON module (Gemini 2.5 Flash Image), together with the target garment image as illustrated in Figure 2. This module leverages its extensive generative prior, enabling the seamless transfer of the garment onto the subject. The output is a single high-fidelity frontal image of the person wearing the new garment in a desired pose. Optionally, we use pose control to change the pose of the input image before we perform VTON on frontal image.

### 3.2. Second stage: Multi-view Generation

Our multi-view generation framework is built upon the Stable Diffusion 3.5 Medium [33]. While Stable Diffusion 3.5 demonstrates high fidelity to text prompts, our approach is conditioned on a source image rather than text. Consequently, we adapt the model by replacing the text conditioning mechanism to vision-based mechanisms. This mechanism integrates information from two distinct encoders. First, a global image embedding is extracted from the input image using an OpenCLIP-bigG [18] vision encoder. This embedding is projected into the conditioning space of the diffusion transformer. Second, to capture fine-grained facial details crucial for identity preservation, we use an Arc2Face [27] encoder to process ArcFace-based [5] facial features. By combining these two, our model can faithfully reproduce both the overall look and the specific identity of the person. The final conditioning representation is a fusion of these two embeddings. As illustrated in Figure 3, the global image embedding and the facial identity embedding are concatenated to form a unified conditioning. This multi-modal vector serves as the input to the transformer blocks, enabling the model to generate robust multi-view outputs that preserve both the overall appearance and specific facial features of the source image.

### 3.3. Image conditioning via Attention Injection

To further ensure high fidelity, we employ a dual conditioning strategy. Beyond the fused embedding described above, we introduce an explicit attention injection strategy within the cross-attention layers, specifically designed to leverage our high-quality source images.

Given that our conditioning image is a high-quality VTON output, we move beyond simple feature concatenation. As depicted in Figure 4, our process begins by caching the hidden states from the source image's forward pass. Then, during the multi-view generation step, our attention injection directly substitutes a portion of the noisy hidden states at the current timestep with these clean, cached states. This operation is performed in the latent patch space, where the patch region corresponding to the front view is replaced entirely. Subsequently, the full set of cached reference states is also concatenated into these modified hidden states before the cross-attention computation, providing additional global context. After the attention operation, the concatenated dimension is removed to maintain architectural consistency. This hybrid replace-and-concatenate approach ensures that the high-fidelity appearance of the source image is robustly anchored, leading to superior preservation of identity and apparel.

### 3.4. Pose Guidance

A key advantage of our approach is the ability to control the output pose, a capability that is often limited in existing multi-view virtual try-on methods. As highlighted in the Section 1, this enables the garment to be displayed effectively on the user, highlighting how the garment fits better. Furthermore, this capability is crucial for addressing a core challenge in virtual try-on: placing garments when the input image is not in a canonical pose, as not all body parts are fully visible. For this, we introduce pose control mechanism by training ControlNet [42] with the following loss:

$$L_{RF} = \mathbb{E}_{t,\epsilon\sim\mathcal{N}(0,I)}\left[\|v_\theta(\boldsymbol{z}, t, \boldsymbol{c}_{img}, \boldsymbol{c}_{pose}) - (\epsilon - \boldsymbol{x}_0)\|_2^2\right]$$

where $\boldsymbol{z}$ represents the noisy feature at time $t$, $\boldsymbol{c}_{img}$ and $\boldsymbol{c}_{pose}$ are the input conditions (frontal image and target pose), $v_\theta$ denotes the velocity field, $\boldsymbol{x}_0$ refers to the ground truth target image feature, and $\epsilon$ is the sampled noise.

This transformation to a desired pose (e.g. A-pose) is also technically crucial for overall pipeline. This is because attempting multi-view VTON directly on an image with a complex pose may lead to noticeable artifacts. By transforming the person into a desired pose, we ensure that the

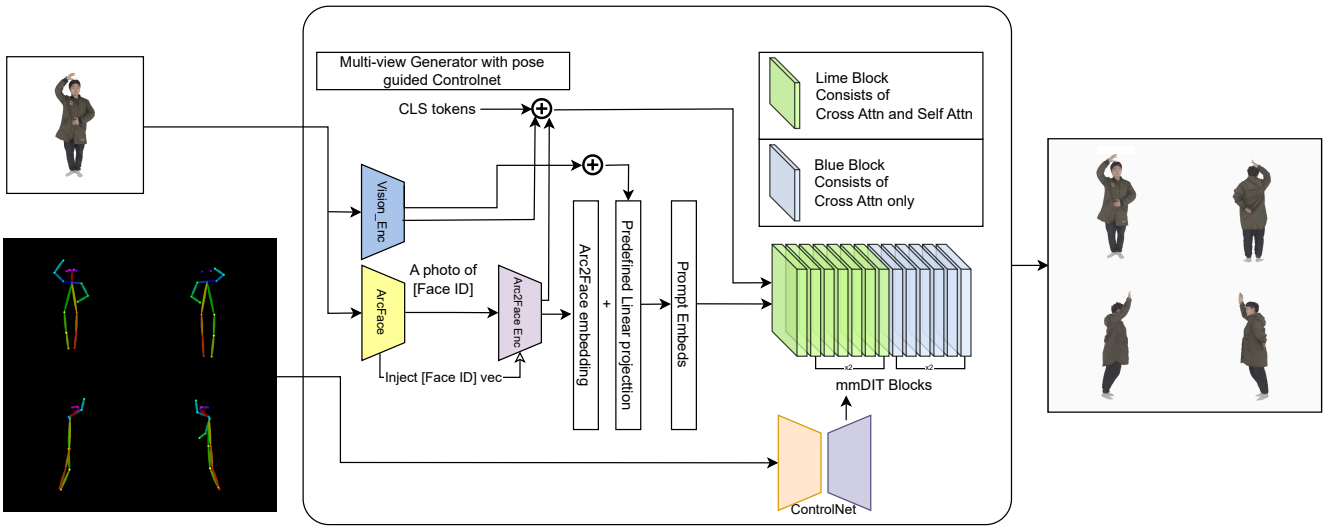



Figure 3. The architecture of our proposed second stage multi-view synthesis model. The symbol + denotes concatenation. The predefined linear projection is a fixed mapping that projects the pooled output of the vision encoder into the prompt embedding space.

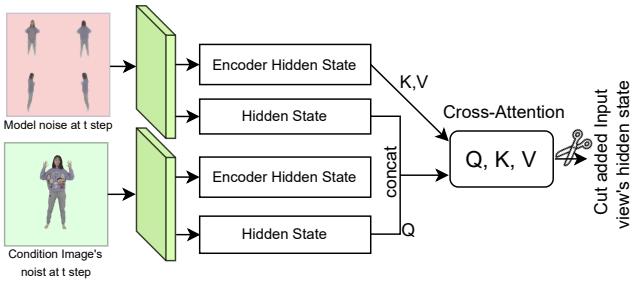



Figure 4. Our proposed cross attention injection.

subsequent virtual try-on stage operates under a consistent and stable condition, free from such complexity. This facilitates a high-quality frontal try-on result, which serves as a robust foundation for the final multi-view synthesis. To achieve this, we generate multi-view images under specific poses where key body regions are clearly visible. We use OpenPose [1] to generate the control sequence.

# 4. Experiment

## 4.1. Implementation Details

We trained our model in two phases on four 80GB Nvidia A100 GPUs. The base multi-view generation model was first trained for roughly 560,000 iterations. To enable pose control, we then enhanced this model with a pose-conditioned ControlNet and performed further training for an additional 280,000 iterations. For both training stages, we utilized the AdamW optimizer with a fixed learning rate of $3 \times 10^{-5}$, a batch size of 1, and a CosineAnnealingWarmRestarts learning rate schedule. A mixed-precision strategy was consistently applied, using bfloat16 for the Transformer and CLIP, while the VAE operated in float32.

## 4.2. Training Dataset

Our model is trained on an extensive and diverse corpus of 3D human data, aggregating high-fidelity scans with real-world captures. The foundation of our dataset consists of 4,497 high-quality 3D scans, which are sourced from 2K2K (2,050) [12], THuman2.0 (500) [41], its extension THuman2.1 (1,464), and Renderpeople (478) [29]. This collection primarily features subjects in simple clothing. To enhance variety in attire and posture, we augment this set with 10,135 multi-view captures of 3,172 individuals from MVHumanNet [38]. We transformed these captures into 3D representations by reconstructing Gaussian splats via the SplatFacto [21] algorithm. In total, our composite training set contains 14,632 distinct 3D human samples. Details on how we process MVHumanNet in our work are provided in the supplementary material.

To unify this diverse collection, all assets undergo a standardized rendering pipeline. Each 3D model is projected into four canonical views using cameras positioned at 90-degree intervals. These views, initially rendered at 4096×4096, are resized to 512×512 for training. Furthermore, we extract facial features from every sample using ArcFace [5] to capture identity-specific details.

To train Pose Control component, we use a combination of the MVHumanNet dataset and the Renderpeople dataset. Our Renderpeople subset contains rigged human models originally in canonical poses, which we retarget to cover a wide range of diverse body poses. MVHumanNet provides multi-frame sequences of the same subject with varying poses. We train ControlNet by using the front-view image at a given timestep as input and conditioning on a target

pose from another timestep, enabling the model to generate multi-view images aligned with the desired pose.

## 4.3. Evaluation

Evaluating multi-view virtual try-on remains a challenging task primarily due to the lack of publicly available benchmark datasets that provide paired multi-view images with corresponding try-on results. This absence makes it difficult to perform objective and quantitative assessments of model performance. To overcome this limitation, we propose a self-referential evaluation pipeline that enables comprehensive quantitative analysis without relying on ground-truth multi-view try-on pairs.

First, we transform each input human image to an A-pose using our mmDiT-based multi-view diffusion model and render multi-view images. We treat these images as the pseudo ground truth (without applying try-on). Second, we perform virtual try-on with the target garment on the input frontal image, and then generate multi-view images using our multi-view generator. To enable pose- and apparel-aligned comparison, we subsequently apply an "original-garment" virtual try-on to the frontal view obtained in the previous step and generate multi-view images again. Finally, we compare these round-trip outputs against the pseudo ground truth. This pipeline ensures that both pseudo GT and predicted results are aligned in pose, viewpoint, and apparel, enabling fair quantitative evaluation.

This evaluation examines whether the model retains identity and garment information when undergoing a complete forward-and-reverse cycle of try-on synthesis, effectively testing the preservation of latent consistency across poses and apparel changes. Unlike direct one-shot metrics that measure fidelity for a single generation, this cyclic framework reveals whether the DiT maintains a stable latent representation of body geometry and visual identity throughout multi-stage processes. This self-referential formulation thus serves not only as a quantitative metric but also as a diagnostic tool, revealing how well the model preserves identity-relevant cues through multi-step passes.

### 4.3.1. Quantitative Evaluation

We evaluate our model's performance using the Renderpeople and THuman datasets. For quantitative comparison, we utilize three standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR) for pixel-level accuracy, Structural Similarity Index (SSIM) for perceived structural similarity, and Learned Perceptual Image Patch Similarity (LPIPS) to measure perceptual differences.

### 4.3.2. Validation of Multi-View and Pose Control

To validate our proposed evaluation pipeline, we first demonstrate the robustness of our multi-view diffusion model's core components to establish the validity of our evaluation pipeline.

**Evaluation on Pose control on multi-view images** To evaluate pose-conditioned synthesis accuracy, we use the Renderpeople dataset, which provides paired renders of each subject in an A-pose and in arbitrary poses. Given the arbitrary-pose input, our model generates the corresponding A-pose, which is directly compared to the ground-truth. The evaluation yields strong quantitative results (PSNR 39.21, SSIM 0.94, LPIPS 0.059), as reported in Table 1, demonstrating the model's ability to effectively disentangle pose from appearance.

**Evaluation on Multi-view capability** In a complementary validation, the model's multi-view generation fidelity is evaluated on the THuman 2.1 dataset under original, unconstrained poses. The task involves generating a complete four-view set from a single frontal image. Achieving a PSNR of 38.21, SSIM of 0.92, and LPIPS of 0.077, the model demonstrates its capability to generate geometrically consistent and high-fidelity multi-view outputs. The demonstrated reliability in these fundamental tasks substantiates the use of our end-to-end evaluation pipeline.

| Evaluation Task | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| ***Validation of Multi-View and Pose Control Capabilities*** | | | |
| Pose control (Renderpeople) | 39.21 | 0.94 | 0.059 |
| Multi-view synthesis (THuman) | 38.21 | 0.92 | 0.077 |
| ***Self-Referential Pipeline*** | | | |
| End-to-End (Renderpeople) | 39.62 | 0.95 | 0.054 |
| End-to-End (THuman) | 36.44 | 0.86 | 0.146 |

Table 1. Quantitative evaluation results. We validate the model's core capabilities (Pose control and Multi-view synthesis) and then present the end-to-end results using our self-referential pipeline.

### 4.3.3. Self-Referential Evaluation

We define the Pseudo-Ground Truth (Pseudo-GT) based on our pose control diffusion model. Given an input image that may contain arbitrary poses, we first normalize it into an A-pose using our pose transfer module. The resulting A-pose image serves as the Pseudo-GT, as it reflects the reliable pose alignment capability of our model. Subsequently, we perform a virtual try-on with different garments and then restore the original garment on top of these try-on results. The restored outputs, referred to as the Predictions, are compared against the Pseudo-GT to assess both pose consistency and garment fidelity. The quantitative results of this evaluation across both datasets are presented in Table 1.

A PSNR of 39.62 on the Renderpeople dataset demonstrates the model's robust capability to preserve appearance and identity consistency across multi-stage generation. Furthermore, the high SSIM and low LPIPS values specifically indicate that the model successfully preserves intricate garment details and human features across viewpoints. No-

tably, all garments used in our evaluation are sourced from the VITON-HD [2] dataset, which includes diverse textures, fabric types, and color patterns.

#### 4.3.4. Comparison with Baseline

To evaluate the effectiveness of our method, we compare our method against VTON360 [16], a recent and relevant baseline for multi-view virtual try-on. VTON360 follows a 3D-centric, per-view try-on pipeline. It first renders multiple 2D views from a 3D human model and then applies virtual try-on independently to each view. This provides explicit multi-view supervision, but is computationally expensive and requires full 3D geometry, making it unsuitable for typical in-the-wild settings due to absence of 3D model.

In contrast, our approach operates directly from a single frontal human image. We first generate a complete multi-view set via our multi-view diffusion model, and only a single frontal view garment is required for virtual try-on. The garment information is then preserved and propagated across all views by our multi-view generation model. Consequently, our method removes the need for 3D assets or repeated per-view try-on for both human image and garment, offering both greater efficiency and higher consistency across viewpoints.

| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|---|---|---|---|
| VTON360 (3D based) | 34.59 | 0.84 | 0.14 |
| Ours (w/o Pose control) | **37.89** | **0.91** | **0.08** |

Table 2. Baseline comparison with VTON360 on 4-views without using the pose controllability of our model, evaluated on THuman2.0. Our method substantially outperforms VTON360 in all metrics, demonstrating superior fidelity and consistency without requiring 3D geometry. Our model takes a frontal view human image and a garment, and VTON360 takes multi-view human images rendered from a mesh and front and back views of the garment.

Table 2 compares the quantitative performance of VTON360 and our method. As noted earlier, VTON360 requires front and back of the garment and multi-view human images, while ours requires a single front garment and a frontal human image as input. VTON360 achieves 34.59 PSNR, 0.84 SSIM, and 0.14 LPIPS, while our model substantially outperforms it with 37.89 PSNR, 0.91 SSIM, and 0.08 LPIPS on the Renderpeople dataset. The significant performance gap underscores the advantages of our design: by decoupling try-on from view synthesis, our method maintains garment fidelity and identity consistency while also applying to in-the-wild single-image inputs.

#### 4.3.5. Qualitative Evaluation

Figure 5 shows the qualitative comparison between our method and VTON360 on the THuman2.0 dataset. The garments for this experiment are from VITON-HD [2]. As shown in figure, our framework requires only a single frontal garment image and a frontal view of the target person to generate multi-view try-on results. In contrast, VTON360 requires both front and back views of the garment, and a 360° capture of the person, which must be preprocessed by rendering a 3D human mesh into multiple 2D views before training and inference.

When comparing the generated results, several artifacts of VTON360 become apparent. In the left example, VTON360 fails to apply the red garment correctly on side views, leaving noticeable black regions around the arms where the try-on process breaks down. In both examples, the frontal view results from VTON360 exhibit unnatural artifacts around the neckline, producing empty or misaligned regions where the garment fabric should appear. By comparison, our method produces natural and complete garments across all views, with consistent preservation of neckline structure and arm coverage.

#### 4.3.6. Ablation studies

To analyze the contributions of our components, we investigate the impact of two mechanisms: our attention injection mechanism and a frontal-view replacement technique.

| Model Variant | Renderpeople | | | THuman | | |
|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
| **Full Model** | **39.62** | **0.95** | **0.054** | **36.44** | **0.86** | **0.146** |
| w/o attn injection | 38.43 | 0.93 | 0.072 | 36.14 | 0.85 | 0.149 |
| w/o frontal-view replacement | 38.98 | 0.94 | 0.059 | 36.42 | 0.86 | 0.147 |

Table 3. Ablation on Renderpeople and THuman.

As shown in Table 3, both components contribute to the final performance. On the Renderpeople dataset, removing our attention processor causes the PSNR to drop from 39.62 to 38.43, SSIM to decrease 0.95 to 0.93, and LPIPS to increase from 0.054 to 0.072. This result underscores the critical role of the attention mechanism in preserving high-frequency details. The direct injection of clean, cached hidden states from the source VTON image into the noisy multi-view generation process ensures garment textures, facial features, and overall identity are robustly maintained across different viewpoints.

Similarly, removing the frontal-view replacement also degrades performance across all metrics, though to a lesser extent, with PSNR falling to 38.98 and LPIPS increasing to 0.059. These components ensure that the frontal view of the generated multi-view set is identical to high-quality VTON input. Despite the marginal performance drop, our analysis indicates that without guidance of the strong frontal view, inconsistencies arise in the side and back views, compromising the overall geometric and textural coherence. This suggests that it plays an important role in the model.

A similar trend is observed on the more challenging THuman2.0 dataset. Without our proposed attention pro-




Figure 5. Qualitative comparison on the THuman2.0 dataset. The first row shows the original four-view images, the second row presents our results, and the third row shows outputs from VTON360. The back view of the garment is generated by Gemini 2.5 Flash Image, which is for VTON360. The corresponding inputs for each method are shown on the left side of each row. More samples with complex garments and inference times comparisons are provided in the supplementary material.

cessor, PSNR drops from 36.44 to 36.14, SSIM from 0.86 to 0.85, and LPIPS increases from 0.146 to 0.149. The removal of the frontal-view replacement shows a marginal but impact on this dataset. These results clearly demonstrate that both mechanisms are beneficial for performance.

## 5. Discussion

**Advantages over Existing Methods** Most existing virtual try-on methods require extensive preprocessing, such as DensePose [10], region masks for garment placement, OpenPose, and human parsing. While our pipeline still relies on pose estimation, it does not require such extensive annotations, substantially reducing the input requirements.

**Use of Off-the-Shelf Model** Our first-stage virtual try-on leverages an off-the-shelf model, Gemini 2.5 Flash Image, which is conditioned on both text and images. However, the level of control it provides is limited, and it occasionally fails to generate plausible images. Since it is a text-and-image conditioned model, precise pose and view control are challenging when applied to multi-view VTON tasks. Failed samples are provided in the supplementary material.

**Limitations of our approach** While our pipeline demonstrates strong performance, pose control in the multi-view diffusion model is sensitive to the given pose. We observed that scaling the target A-pose to match the subject's height improves results. During evaluation, we align poses using the subject's head and foot positions. Our multi-view diffusion model takes a frontal image as input, and for the best performance, the subject's face should be clearly visible and directly facing the camera.

## 6. Conclusion

We proposed a novel framework for multi-view virtual try-on from a single image, addressing the limitations of existing methods that require multiple human and garment views. By decoupling frontal try-on and multi-view synthesis, and introducing a Multi-view Synthesis Diffusion Transformer, our method enables realistic results from arbitrary poses and viewpoints. While the system assumes a frontal-view input and may perform better for subjects closer to the distribution of the training dataset, it substantially reduces input requirements and improves practicality for real-world applications.

**Acknowledgment** This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-II220612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI, No.RS-2022-II220608/2022-0-00608, Artificial intelligence research about multimodal interactions for empathetic conversations with humans, No.IITP-2025-RS-2024-00360227, Leading Generative AI Human Resources Development, No.RS-2025-25442824, AI Star Fellowship Program(Ulsan National Institute of Science and Technology, & No.RS-2020-II201336, Artificial Intelligence graduate school support(UNIST)) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00219959).

## References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 5

[2] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2021. 7

[3] 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 (ECCV)*, 2024. 1, 3

[4] Google DeepMind. Introducing gemini 2.5 flash image (aka "nano banana"). Google Developers Blog / Model Card, 2025. Accessed: 2025-08-27. 1

[5] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4, 5

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. 2

[7] 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, 2024. 1

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. 3

[9] 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 (ACMMM)*, 2023. 1, 3

[10] 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 (CVPR)*, 2018. 8

[11] Hailong Guo, Bohan Zeng, Yiren Song, Wentao Zhang, Jiaming Liu, and Chuang Zhang. Any2anytryon: Leveraging adaptive position embeddings for versatile virtual clothing tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1, 3

[12] Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, and Hae-Gon Jeon. High-fidelity 3d human digitization from single 2k resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[13] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 3

[14] Xintong Han, Weilin Huang, Xiaojun Hu, and Matthew Scott. Clothflow: A flow-based model for clothed person generation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[15] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 3

[16] Zijian He, Yuwei Ning, Yipeng Qin, Wangrun Wang, Sibei Yang, Liang Lin, and Guanbin Li. Vton 360: High-fidelity virtual try-on from any viewing direction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 7

[17] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu Qiao, Bo Dai, and Lu Sheng. EpiDiff: Enhancing Multi-View Synthesis via Localized Epipolar-Constrained Diffusion . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 4

[19] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017. 3

[20] Boyuan Jiang, Xiaobin Hu, Donghao Luo, Qingdong He, Chengming Xu, Jinlong Peng, Jiangning Zhang, Chengjie Wang, Yunsheng Wu, and Yanwei Fu. Fitdit: Advancing the authentic garment details for high-fidelity virtual try-on, 2024. 3

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 5

[22] Jeongho Kim, Guojung Gu, Minho 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 (CVPR)*, 2024. 1, 3

[23] Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wenhan Luo, Ping Tan, Wenping Wang, Qifeng Liu, and Yike Guo. Era3d: High-resolution multiview diffusion using efficient row-wise attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023. 3

[25] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[26] 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 ACM International Conference on Multimedia (ACMMM)*, 2023. 1, 3

[27] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 4

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2

[29] Renderpeople. Renderpeople 3d human model dataset, 2024. Accessed: 2025-05-20. 5

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022. 2

[31] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023. 3

[32] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. MVDream: Multi-view Diffusion for 3D Generation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[33] Stability AI. Introducing stable diffusion 3.5, 2024. Accessed: 2025-05-22. 4

[34] Siqi Wan, Yehao Li, Jingwen Chen, Yingwei Pan, Ting Yao, Yang Cao, and Tao Mei. Improving virtual try-on with garment-focused diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[35] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[36] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. Mv-vton: multi-view virtual try-on with diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025. 1

[37] Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023. 3

[38] Zhangyang Xiong, Chenghong Li, Kenkun Liu, Hongjie Liao, Jianqiao Hu, Junyi Zhu, Shuliang Ning, Lingteng Qiu, Chongjie Wang, Shijie Wang, et al. Mvhumannet: A large-scale dataset of multi-view daily dressing human captures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[39] Yuhao Xu, Tao Gu, Weifeng Chen, and Arlene Chen. Ootd-iffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025. 1, 3

[40] Zhaotong Yang, Yuhui Li, Shengfeng He, Xinzhe Li, Yangyang Xu, Junyu Dong, and Yong Du. Omnivton: Training-free universal virtual try-on. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 1

[41] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[42] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023. 2, 4

[43] Xuanpu Zhang, Dan Song, Pengxin Zhan, Tianyu Chang, Jianhao Zeng, Qingguo Chen, Weihua Luo, and Anan Liu. Boow-vton: Boosting in-the-wild virtual try-on via mask-free pseudo data training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 3

[44] Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, and Ira Kemelmacher-Shlizerman. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3