VTON 360: High-Fidelity Virtual Try-On from Any Viewing Direction

Zijian He[1] Yuwei Ning[1] Yipeng Qin[2] Wangrun Wang[1] Sibei Yang[3] Liang Lin[1,4,5] Guanbin Li[1,4,5*]
[1]Sun Yat-sen University [2]Cardiff University [3]ShanghaiTech University
[4]Guangdong Key Laboratory of Big Data Analysis and Processing [5]Peng Cheng Laboratory
hezj39@mail2.sysu.edu.cn, yuwei_ning@hust.edu.cn,{qinyipeng1991,wanggrun}@gmail.com,
yangsb@shanghaitech.edu.cn linliang@ieee.org, liguanbin@mail.sysu.edu.cn

Figure 1. **Results of VTON 360.** Our VTON 360 enables high-fidelity 3D Virtual Try-On (VTON) by seamlessly adapting E-commerce garments onto a clothed 3D human model, supporting full 360° view rendering. The highlighted bounding boxes (dashed line) demonstrate our method's ability to preserve intricate clothing details and patterns (*e.g.*, collar accessories, horizontal line patterns, logos, texts, numbers) across diverse garment types.

Abstract

Virtual Try-On (VTON) is a transformative technology in e-commerce and fashion design, enabling realistic digital visualization of clothing on individuals. In this work, we propose VTON 360, a novel 3D VTON method that addresses the open challenge of achieving high-fidelity VTON that supports any-view rendering. Specifically, we leverage the equivalence between a 3D model and its rendered multi-view 2D images, and reformulate 3D VTON as an extension of 2D VTON that ensures 3D consistent results across multiple views. To achieve this, we extend 2D VTON models to include multi-view garments and clothing-agnostic human body images as input, and propose several novel techniques to enhance them, including: i) a pseudo-3D pose representation using normal maps derived from the SMPL-X 3D human model, ii) a multi-view spatial attention mechanism that models the correlations between features from different viewing angles, and iii) a multi-view CLIP embedding that enhances the garment CLIP features used in 2D VTON with camera information. Extensive experiments on large-scale real datasets and clothing images from e-commerce platforms demonstrate the effectiveness of our approach. Project page: https://scnuhealthy.github.io/VTON360.

1. Introduction

Virtual Try-On (VTON) enables realistic digital visualization of clothing on individuals and has emerged as a transformative technology in e-commerce and fashion design. While significant research efforts have been made on 2D VTON solutions [8, 12, 19, 33, 39], these approaches are inherently limited in their representation of view-related features. To overcome this limitation and enable high-fidelity any-view rendering, 3D VTON methods were introduced.

3D VTON requires accurate garment transfer onto a 3D human body while ensuring realistic garment fitting, texture preservation, and 3D consistency. The two primary aims of 3D VTON are i) achieving *high-fidelity* and ii) supporting *any-view rendering*. Leveraging the inherent capability of 3D models for *any-view rendering*, early 3D VTON methods [13, 15, 28] make clothing simulation on synthetic human bodies. Specifically, these methods utilized 3D scanners to capture clothing meshes, followed by the development of specialized dressing algorithms. Although effec-

*Corresponding author is Guanbin Li.

tive, these methods rely on costly 3D scanning equipment and the physical presence of the human body/clothing (*i.e.*, not fully virtual), restricting their practicality in real-world applications. As a byproduct, most early methods focused on developing geometrically correct dressing algorithms using standard templates of human body and clothing models. Addressing this limitation, researchers extended 3D VTON by introducing algorithms that reconstruct 3D clothing models from input images, enabling the use of image-based clothing inputs [3, 32, 41, 42]. However, since input clothing images (usually frontal) are inherently 2D and lack multi-view information, this approach struggles to reconstruct high-fidelity clothing models that can be rendered well from all viewing directions.

To complement this missing information, DreamVTON [50] introduces a novel approach that leverages Text-to-Image (T2I) diffusion models to reconstruct both the human body and clothing from input images. Its key insight is that T2I models learned view-agnostic "concepts" of both bodies and garments during their training, and that the corresponding concepts for the input body and clothing images can be obtained using LoRA [21]. By utilizing Score Distillation Sampling (SDS) [37], DreamVTON can generate visual-pleasing 3D VTON results by ensuring consistency between renderings from arbitrary viewpoints and the concepts. Nonetheless, DreamVTON's high flexibility comes at the cost of low fidelity. This limitation stems from the fact that the concepts learned by T2I models are semantic in nature, thus lacking 3D geometric consistency and pixel-level accuracy with respect to the input body and clothing images. Recently, a concurrent work, namely GaussianVTON [6], partially addressed this limitation by formulating 3D VTON as a 3D scene editing task, where a given 3D human model is edited using multi-view images generated by 2D VTON methods. While it significantly enhances the fidelity of the human body, the fidelity and 3D consistency of clothing remain problematic, as there are no 2D VTON methods that can generate multi-view images with 3D consistency. Therefore, to the best of our knowledge, achieving high-fidelity 3D VTON that supports any-view rendering remains an open challenge.

In this work, we address the above-mentioned challenge via proposing VTON 360, a novel 3D VTON method that achieves high-fidelity VTON from arbitrary viewing directions. Similar to GaussianVTON [6], our method edits a given 3D human model by inpainting the rendered images using a latent diffusion model. However, we set ourselves apart through our novel garment fidelity preservation strategy that can generate high-fidelity on-body garments in all viewing directions. Specifically, we first extend both the garment and clothing-agnostic human body inputs to typical 2D VTON models to leverage multi-view information, including paired front and back view garment images as well as a set of multi-view clothing-agnostic human body images sampled from random azimuth angles. Then, we propose several novel enhancements to bridge the gap between typical 2D VTON methods and our multi-view 3D consistency requirements: i) We propose a pseudo-3D pose representation using normal maps derived from the SMPL-X 3D human model, which captures fine-grained surface orientation details and provides more consistent geometry across views compared to the 2D pose representations (semantic segmentation maps) used in 2D VTON models. ii) We design a Multi-view Spatial Attention mechanism that models the correlations between features from different viewing angles, featuring a novel "correlation" matrix modeling the relationships among different input views. iii) We propose a multi-view CLIP embedding that enhances the garment CLIP embedding used in 2D VTON methods with camera information, thereby facilitating network learning of features relevant to a particular view. Together, these innovations enable our 2D VTON model to generate high-quality, multi-view and 3D-consistent virtual try-on results. Extensive experiments on Thuman2.0 [55] and MVHumanNet [51] datasets demonstrate that our method achieves high fidelity 3D VTON which supports any-view rendering. In addition, we show the effectiveness and generalizability of our methodology by testing it using garments from e-commerce platforms. Our conclusions include:

- We propose a novel 3D Virtual Try-On (VTON) method, namely *VTON 360*, which achieves high-fidelity VTON from arbitrary viewing directions.
- Leveraging the *equivalence* between a 3D model and its rendered multi-view 2D images, we reformulate 3D VTON as an extension of 2D VTON that ensures consistent results across multiple views. Specifically, we introduce several novel techniques, including: (i) pseudo-3D pose representation; (ii) multi-view spatial attention; and (iii) multi-view CLIP embedding. These innovations enhance traditional 2D VTON models to generate multi-view and 3D-consistent results.
- Extensive experimental results on two large real datasets as well as real clothing images from e-commerce platforms demonstrate the effectiveness of our approach.

2. Related Work

2D Virtual Try-On. 2D Virtual Try-On (VTON) aims to transfer a desired garment to the corresponding region of a target human image while preserving the human pose and identity. Early methods [2, 8, 10, 11, 16, 18, 29, 31, 38, 54, 56] use Generative Adversarial Networks (GANs) to deform the garments to match the target body shape, which a critical step for achieving realistic VTON. However, accurately adapting to diverse real-world conditions remains a significant challenge. Addressing this issue, recent methods [12, 19, 33, 60] reframe 2D VTON as a conditioned in-

painting task, leveraging the strong priors provided by diffusion models [20, 43, 45] to achieve promising results. This strategy is further improved by [9, 26, 53], which introduce a ReferenceNet to extract hierarchical garment features and apply attention mechanisms to condition the Main UNet.

3D Virutal Try-On. For 3D Virtual Try-On (VTON), traditional methods [4, 13, 15, 28, 36] rely on 3D scanning or cloth simulation to generate highly precise body and garment geometry. These methods were then extended by learning-based methods [3, 32] that employ differentiable rendering to dress the SMPL [30] model with a desired garment mesh. Despite their effectiveness, such methods rely on costly 3D scanning and the physical presence of human body/clothing, limiting their application in the real world. Addressing this limitation, M3D-VTON [59] proposes a depth-based 3D VTON framework to reconstruct 3D clothed human models from 2D human and garment images, but the results often suffer from explicit warping artifacts. To improve 3D VTON results, recent methods [23, 24, 50, 62] resort to text-to-image (T2I) diffusion models and employ the Score Distillation Sampling (SDS) loss [40] to ensure consistency among different viewing directions. Specifically, TeCH [24] adapts the generative priors of T2I diffusion model to the specific person and clothes by training descriptive text prompts with DreamBooth [40]. DreamWaltz [23] leverages Pose ControlNet [57] to attain clothed human body models. DreamVTON [50] introduces a multi-concept LoRA [21] to personalize the T2I diffusion model, and uses a template-based optimization mechanism that combines with SDS loss to better preserve patterns on the garment. Although effective, these methods often produce results lacking in fidelity, as the concepts learned by T2I models are semantic rather than at the pixel level. Concurrent to our work, GaussianVTON [6] proposes an alternative approach by combining Gaussian Splatting [25] with pre-trained 2D VTON models and formulate it as an editing task. However, since there are no 2D VTON methods that can generate multi-view images with 3D consistency, the fidelity and 3D consistency of the clothing generated remain problematic.

Radiance Field-based 3D Human or Scene Editing. Recently, radiance field-based editing has attracted significant interest due to its efficient differentiable rendering capabilities, sparking substantial advancements in text-driven 3D editing. For example, InstructN2N [17] employ an image-based diffusion model InstructP2P [5] to modify the rendered image by the user's text description with a variant of the score distillation sampling (SDS) [37] loss. GaussianEditor [7] applies Gaussian Splatting [25] as 3D representation instead of NeRF, adopting Gaussian semantic tracking to track target Gaussian values, significantly improving editing speed and controllability. To enable accurate location and appearance control, subsequent works [47, 61] specify the editing region using the attention score or with a segmentation model. TIP-Editor [63] proposes a content personalization step dedicated to the reference image based on LoRA, achieving the editing following a hybrid text-image prompt. GaussCtrl [48] leverage depth conditions and attention-based latent code alignment to achieve 3D-aware multi-view consistent editing instead of iteratively editing single views using SDS loss. However, these works primarily focus on global appearance modifications within a text-driven pipeline, while our approach emphasizes preserving fine textural details from different viewing directions throughout the editing process.

3. Preliminary

Latent Diffusion Model. Latent Diffusion Model [39] is a variant of diffusion models that performs denoising within the latent space of a Variational Autoencoder (VAE) [27]. During training, given a fixed encoder $\mathcal{E}$, an input image x is transformed into its latent representation $z_0 = \mathcal{E}(x)$. A conditional diffusion model $\hat{\epsilon}_\theta$, typically implemented with a UNet architecture, is then trained using a weighted denoising score matching objective:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{z},\mathbf{c},\boldsymbol{\epsilon},t}[\boldsymbol{\epsilon} - \|\hat{\epsilon}_\theta(\mathbf{z_t}; \mathbf{c}, t)\|_2^2] \tag{1}$$

where $\mathbf{z}_t := \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ denotes the forward diffusion process at timestep t; α_t, σ_t are time-dependent functions defined by the diffusion model formulation; $\mathbf{c}$ denotes the conditional input and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ is Gaussian noise. During inference, data samples are generated by initiating from Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$ and iteratively refining it using a DDIM [44] sampler.

4. Method

Our method leverages the *equivalence* between a 3D model and its rendered multi-view 2D images to achieve high-fidelity, any-view 3D VTON. Specifically, as Fig. 2 shows, given an input 3D human model and a garment image, our method 1) renders the 3D model into multi-view 2D images and 2) formulates 3D VTON as a consistent, unified 2D VTON process across these rendered views; 3) By reconstructing the edited images into a 3D model using existing 3D reconstruction methods, we ensure visual coherence and precise garment alignment from any viewing angle. Among them, the second step is crucial as existing 2D VTON methods [9, 26, 53] lack 3D knowledge, preventing them from generating multi-view images with 3D consistency.

To address this challenge, we propose several novel techniques (Sec. 4.2) that equip a typical 2D VTON network (Sec. 4.1), which is built on a latent diffusion model [39], with the capability to generate 3D-consistent results. We use Gaussian Splatting [25] as our 3D representation.

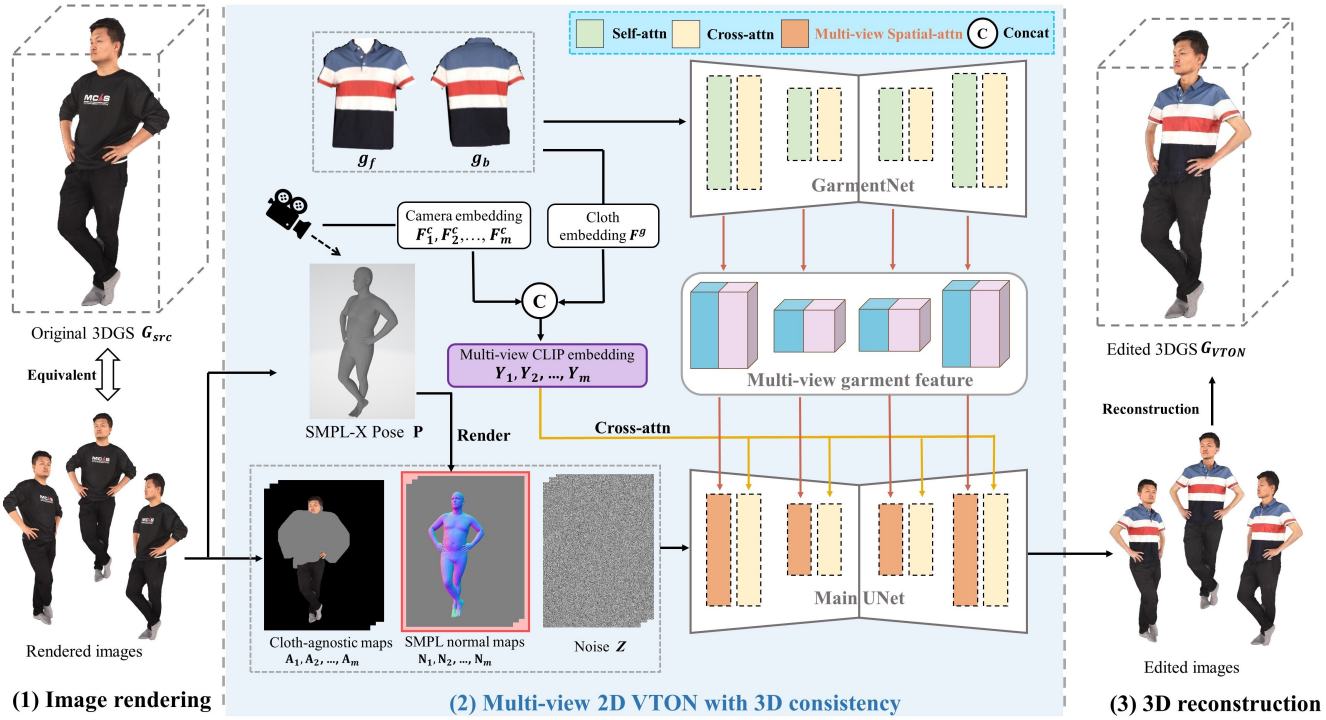

Figure 2. **Overview of VTON 360.** Given an input 3D human model $\mathbf{G}_{\text{src}}$ and a pair of garment images (g_f, g_b), our method 1) renders $\mathbf{G}_{\text{src}}$ into multi-view 2D images (left) and 2) edits the rendered multi-view 2D images (middle); 3) reconstructs the edited images into a 3D model $\mathbf{G}_{\text{VTON}}$ (right). In the crucial step 2), we propose three novel techniques to equip a typical 2D VTON network with the capability to generate 3D-consistent results: 1) Pseudo-3D Pose Input, 2) Multi-view Spatial Attention, and 3) Multi-view CLIP Embedding.

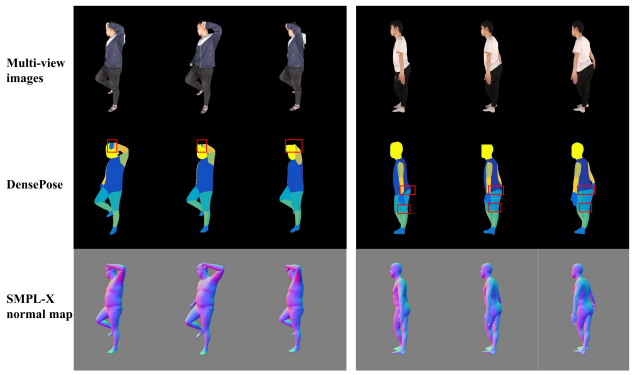

Figure 3. **DensePose (2D) vs. SMPL-X normal map (pseudo-3D) representations.** DensePose applies uniform labels per body part, lacking 3D consistency across views and causing artifacts and temporal inconsistencies (highlighted with red boxes). In contrast, SMPL-X normal maps capture fine surface details, ensuring geometric coherence and stable, realistic shading across views.

4.1. Recap of 2D VTON Framework

Following previous works [12, 26, 53], we formulate 2D VTON as an exemplar-based inpainting problem, aiming to fill an input clothing-agnostic image $\mathbf{A}$ with a given garment image g, where $\mathbf{A}$ is obtained following the method used in [53]. As illustrated in Fig. 2 (middle), the network architecture is based on the latent diffusion model [39] with an encoder $\mathcal{E}$ and comprises two components:

- A GarmentNet [9, 53] that extracts features from $\mathcal{E}(g)$.
- A Main UNet that inpaints $\mathbf{A}$ according to i) detailed garment features extracted by the GarmentNet; ii) the 2D pose of $\mathbf{A}$ represented by semantic labels using DensePose [14]; iii) CLIP embeddings of input garment g. Among them, i) and ii) together with noise are input to the self-attention layers, while iii) is input to the cross-attention layers of the Main UNet.

Both the GarmentNet and the Main UNet share the same network architecture.

4.2. Multi-view 2D VTON with 3D Consistency

To enable the aforementioned 2D VTON model to generate multi-view and 3D-consistent results, we propose the following novel enhancements to its design:

Multi-view Inputs. We extend both inputs to the model:

- *Multi-view Garment Inputs:* We extend the input garment representation from a single image g to paired front and back view images g_f, g_b, providing comprehensive garment information across all viewing angles. Accordingly, we use the encoder $\mathcal{E}$ to encode g_f, g_b into their latent representations $\mathcal{E}(g_f), \mathcal{E}(g_b)$ and feed them into GarmentNet to obtain their multi-layer features F_f^l and F_b^l at layer l, respectively.
- *Multi-view Clothing-agnostic Image Inputs:* Based on

the *equivalence* between a 3D human model and its rendered multi-view 2D images, we extend the input human body representation from a single, clothing-agnostic image, $\mathbf{A}$, to a set of m multi-view images, denoted as $\mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_m}$. These images are sampled from random azimuth angles, allowing the 2D VTON model to access comprehensive, multi-view information from the input 3D human model.

Pseudo-3D Pose Input. As shown in Fig. 3, the 2D DensePose representations [14] commonly used in state-of-the-art 2D VTON methods [9, 26] assign a uniform semantic label to all pixels within each body part (*e.g.*, thigh), inherently lack 3D geometric consistency across multiple views, and often introduce artifacts and temporal inconsistencies. To address these limitations, we propose a novel pseudo-3D pose representation: the normal maps $\mathbf{N}$ derived from the SMPL-X [35] model of the input body. These normal maps capture fine-grained surface orientation details, providing a more consistent representation across views by preserving geometric structure in the 3D space. Furthermore, they enable smoother, temporally stable transitions and realistic shading effects across multi-view scenarios. In practice, we employ a lightweight PoseEncoder $\mathcal{E}'$ [22] and feed $\mathcal{E}'(\mathbf{N})$ into the Main UNet. We obtain the SMPL-X model from the multi-view images of the input body using EasyMoCap [1].

Accordingly, we concatenate three components as the enhanced input to the Main UNet: i) a noise latent z_t; ii) the encoded pseudo-3D poses $\mathcal{E}'(\mathbf{N_1}), \mathcal{E}'(\mathbf{N_2}), ..., \mathcal{E}'(\mathbf{N_m})$; and iii) the encoded multi-view clothing-agnostic images $\mathcal{E}(\mathbf{A_1}), \mathcal{E}(\mathbf{A_2}), ..., \mathcal{E}(\mathbf{A_m})$. Let $F_1^l, F_2^l, ..., F_m^l$ be the feature representations at layer l of the Main UNet, and recall the garment features F_f^l and F_b^l defined above, we enhance the self-attention layers of the Main UNet as:

Multi-view Spatial Attention. To cope with the aforementioned multi-view input features and ensure their consistency, we draw insights from the *temporal* attention layer commonly used in video generation and editing [49, 58] and extend it to our multi-view *spatial* attention layer, denoted as $\mathrm{MVAttention}$. The key distinction of our $\mathrm{MVAttention}$ is that its input multi-view features $F_1^l, F_2^l, ..., F_m^l$ are from images captured from non-uniform spatial intervals, with the viewing angles varying randomly. Consequently, features from similar views exhibit a higher correlation, while those from distinct views are largely independent. To model this relationship, we construct a "correlation" matrix C based on the angular disparity obtained from camera rotation matrices of the input multi-view images, and define our $\mathrm{MVAttention}$ as follows:

$$\begin{aligned} &\mathbf{F^l} = [F_1^l \oplus F_2^l \ldots \oplus F_m^l],\ \hat{\mathbf{F}}^\mathbf{l} = [\mathbf{F^l} \oplus F_f^l \oplus F_b^l] \\ &Q = W^Q\ \mathbf{F^l},\ K = W^K\ \hat{\mathbf{F}}^\mathbf{l},\ V = W^V\ \hat{\mathbf{F}}^\mathbf{l} \\ &A_i = \mathrm{softmax}(\frac{Q_i \times (C_i \cdot K^T)}{\sqrt{d}}),\ \mathbf{H_i^l} = A_i \times V_i \end{aligned} \tag{2}$$

where $i \in \{1, 2, ..., m\}$ denotes i-th view; the Query Q comes directly from $\mathbf{F^l}$ and the concatenation of $[\mathbf{F^l}, F_f^l, F_b^l]$ serves as the key K and the value V; $\oplus$ indicates matrix concatenation along the token axis; d denotes the dimension; W^Q, W^K, W^V represent the linear transformation matrices; we omitted the l of the attention matrices and parameters for simplicity; $C \in \mathbb{R}^{m \times m}$, C_i represents i-th row in C, and its "correlation" value between i-th and j-th features is C_{ij}:

$$C_{ij} = ((\mathrm{trace}(R_i^T R_j) - 1)/2 + 1)/2 \tag{3}$$

where R_i and R_j are the extrinsic rotation matrices of the corresponding camera views, $(\mathrm{trace}(R_i^T R_j) - 1)/2$ is the cosine value of the angle between these camera views.

Multi-view CLIP Embedding. Camera viewpoints can serve as an effective condition signal to enhance 3D consistency in video content generation [52]. Building on this insight, we incorporate camera condition within our try-on network by encoding camera parameters as an additional token, enabling the generation of more consistent multi-view images. Specifically, we define a world coordinate system in which the camera faces the subject directly. For each input image (view) $\mathbf{A_i}$, $1 \leq i \leq m$, we extract the rotation matrix from the camera's corresponding extrinsic matrix. This rotation matrix is then reshaped into a 9-dimensional tensor $\mathbf{r_i}$, which undergoes positional encoding to effectively integrate the camera parameters into the feature representation F_i^c.

$$\begin{aligned} F_i^c = (&\sin(2^0 \pi \mathbf{r_i}), \cos(2^0 \pi \mathbf{r_i}), ..., \\ &\sin(2^{L-1} \pi \mathbf{r_i}), \cos(2^{L-1} \pi \mathbf{r_i})) \end{aligned} \tag{4}$$

where L is the length of positional embedding. We then project F_i^c to match the dimensionality of the garment CLIP image embedding F^g via an MLP and concatenate them along the token axis to form Y_i. This combined representation, Y_i, is subsequently used in the key K_x and value V_x of the cross-attention layers of the Main UNet:

$$\begin{aligned} Y_i &= F^g \oplus \mathrm{MLP}(F_i^c) \\ Q_\mathrm{x} = W_\mathrm{x}^Q\ \mathbf{H_i^l},\ K_\mathrm{x} &= W_\mathrm{x}^K\ Y_i,\ V_\mathrm{x} = W_\mathrm{x}^V\ Y_i \\ F_i^{(l+1)} &= \mathrm{softmax}(\frac{Q_\mathrm{x} K_\mathrm{x}^T}{\sqrt{d_\mathrm{x}}}) V_\mathrm{x}, \end{aligned} \tag{5}$$

where $\mathbf{H^l}$ is the output of the $\mathrm{MVAttention}$ of the l-th layer; we omitted the l of the cross attention matrices and parameters for simplicity.

Training. Our enhanced multi-view 2D VTON network can be trained by minimizing the following latent diffusion model loss function:

$$\mathcal{L}_\mathrm{ldm} = \mathbb{E}_{z_t, \eta, \psi, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \eta, \psi, \zeta))\|_2^2\right], \tag{6}$$

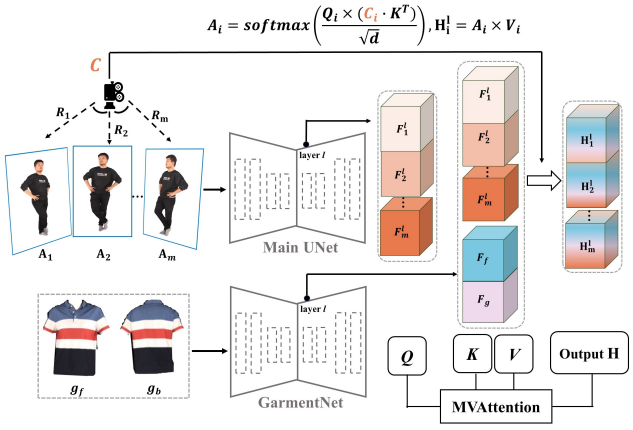

Figure 4. **Illustration of the proposed Multi-view Spatial Attention.** Query (Q): multi-view features $\mathbf{F^l}$; Key (K) and Value (V): concatenation of $\mathbf{F^l}$ and garment features F_f^l and F_b^l. The attention score between viewpoints i and j is modulated by a weight C_{ij}, determined by the cosine of the angle between them.

where $\eta = [\mathcal{E}(g_f); \mathcal{E}(g_b); \mathcal{E}(\mathbf{N_i})_{i=1}^m]$ represents the input latent garment images and latent normal maps; $\zeta = [\mathcal{E}'(\mathbf{A_i})_{i=1}^m]$ denotes the input latent clothing-agnoistic images; $\psi = \mathbf{Y}$ is the proposed multi-view CLIP embedding.

5. Experiments

5.1. Experimental Setup

Datasets. We conduct experiments on two public datasets: Thuman2.0 [55] and MVHumanNet [51]. Thuman2.0 comprises 526 reconstructed clothed human scans, from which we render multi-view input images. Of these samples, 426 are used for training, while the remaining 100 are set aside for testing. To further evaluate the effectiveness and robustness of our method, we also perform experiments on MVHumanNet, a large-scale dataset of multi-view human images that encompasses a diverse range of subjects, daily outfits and motion sequences. The images in MVHumanNet are captured using a multi-view system with either 48 or 24 cameras. We use 4,990 subjects from this dataset, allocating 4,790 to training and 200 for tests. For each subject, we randomly select two frames of multi-view images from its entire motion sequence. While MVHumanNet provides multi-view images directly for editing and reconstruction, we render uniformly distributed views around each human subject in Thuman2.0 to ensure consistent input.

Baselines. We primarily compare our method with three existing methods: DreamWaltz [23], GaussCtrl [48], and TIP-Editor [63]. DreamWaltz is a method designed for directly generating 3D human bodies based on textual descriptions, while GaussCtrl and TIP-Editor are two radiance-based editing methods. GaussCtrl is based on Stable Diffusion, using a description-like prompt to edit the scene. TIP-Editor accepts both text and image prompts. We configure it by specifying the human body as the editing region and the desired garment as the image prompt. We use ChatGPT to generate the text prompts corresponding to the clothing images.

Evaluation Metrics. For quantitative evaluation, we assess garment-to-person alignment between the edited person and the reference image. Following [63], we calculate the average DINO similarity [34] between the reference image and the rendered multi-view images of the edited 3D scene. Additionally, to evaluate multi-view consistency, we compute the CLIP Directional Consistency Score as outlined in [17]. Given the large scale of experiments (repeated 3DGS reconstruction), we selected a subset of examples from the dataset for metric evaluation. Specifically, from the test sets of Thuman and MVHumanNet, we randomly sampled 10 human scans each, performing virtual try-on with 6 randomly chosen garments per human scan.

We further conducted a user study involving 50 participants who rated the results of our method and three baseline methods based on two criteria: overall "Quality" and "Alignment" with the reference image. Each evaluation consisted of two questions: (1) Which method produces the highest quality of the edited 3D human? and (2) Which method achieves the most consistent alignment with the target clothing? Participants viewed the VTON results as rotating randomized video sequences.

Implementation Details. During pre-processing, we crop the multi-view images to the bounding box around the person and resize them to a resolution of 768×576. The front view and the back view of garment images are obtained from the corresponding clothed human images. After editing, we pad the images back to their original size. The data processing pipeline is the same for both Thuman2.0 and MVHumanNet datasets.

The Main UNet and the GarmentNet are initialized by the Stable Diffusion V1.5 [39]. The training process is divided into two stages. In the first stage, each view is trained independently, during which we establish the feature extraction capabilities of both the PoseEncoder and GarmentNet, as well as the generative capability of the Main UNet. The second stage involves multi-view training, where we randomly select M views for each human subject. This stage is focused on training the proposed MVAttention module to enhance the network's ability to maintain consistency across views. Due to memory constraints, we set $M = 8$ for the training phase. During the testing phase, we uniformly sampled 32 views from a 360-degree rotation around the subject. The editing of these 32 views is conducted in two batches, with each batch processing $M = 16$ views.

5.2. Comparisons with State-of-the-Art Methods

Qualitative Evaluation. Fig. 5 shows visual comparisons between our method and the baselines. DreamWaltz [23]

Method	Thuman2.0 [55]				MVHumanNet [51]			
	$\text{CLIP}_{cons}\uparrow$	$\text{DINO}_{sim}\uparrow$	$\text{Vote}_{quality}$	Vote_{align}	$\text{CLIP}_{cons}\uparrow$	$\text{DINO}_{sim}\uparrow$	$\text{Vote}_{quality}$	Vote_{align}
DreamWaltz [23]	0.887	0.556	0.46%	1.54%	0.935	0.495	0.46%	0.46%
TIP-Editor [63]	**0.939**	0.569	0.92%	0.62%	**0.948**	0.512	2.15%	1.38%
GaussCtrl [48]	0.931	0.577	1.08%	1.38%	0.938	0.521	1.69%	1.23%
Ours	0.923	**0.633**	**97.54%**	**96.46%**	0.933	**0.623**	**95.69%**	**96.92%**

Table 1. **Quantitative comparisons.** CLIP_{cons} denotes the CLIP Direction Consistency Score. DINO_{sim} is the DINO similarity.

Figure 5. **Qualitative comparison.** The first two rows show the results on Thuman2.0 dataset while the last two rows show the results on MVHumaNet dataset. Our method achieves good texture preservation (highlighted by the blue boxes), while three baseline methods mostly fail.

regenerates 3D clothed humans from text prompts but struggles to accurately retain both body and clothing characteristics. GaussCtrl [48], lacking support for image prompts, fails to maintain detailed clothing textures. While Tip-Editor [63] leverages Lora [21] for personalization, it encounters difficulties in consistently mapping clothing inputs from two views into the 3D human because the personalized concept are semantic in 2D space. In contrast, our method effectively preserves intricate clothing details, such as text, stripes, and logos.

Quantitative Evaluation. Tab. 1 shows the results for the CLIP Directional Consistency Score and DINO similarity on Thuman2.0 and MVHumanNet datasets. Our approach surpasses other methodes on DINO_{sim}, clearly illustrating the superiority of our method in terms of garment texture preservation. While our results on CLIP_{cons} are comparable to those of other methods, it is important to note that these methods incorporate the SDS loss, which to some extent smooths the representation of humans in 3D space. Additionally, the "flatter" textures of other methods could also result in artificially higher consistency scores. Furthermore, user studies have shown that our method significantly exceeds baselines in terms of edited 3D human quality and the alignment of clothing details.

5.3. Visual Results using E-commerce Garment

Fig. 6 showcases VTON results using garments from the MVG dataset [46], whose images are from e-commerce

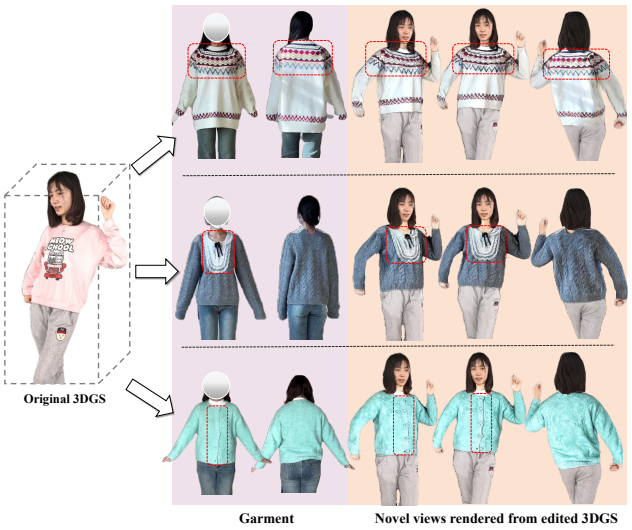

Figure 6. **Generalization to e-commerce garments (the MVG dataset).** Our method, trained on the THuman2.0 dataset, shows strong generalizability when applied to e-commerce garments. For clarity in visualization, we display garment images on human models; however, in the actual VTON process, the garments are segmented from the models using parse maps.

platforms like YOOX NET-A-PORTER, Taobao, and TikTok*, and a model trained on the Thuman2.0 dataset [55]. The results demonstrate that our method effectively preserves intricate garment details and textures. For instance, it accurately maintains the stripe patterns in the first row, the cute tie in the second row, and the buttons in the third row, highlighting the robustness of our approach in handling diverse and realistic clothing items.

5.4. Ablation Study

We conduct an ablation study on Thuman2.0 dataset in Tab. 2 and Fig. 7 to evaluate the impact of our three proposed modules in enhancing a typical 2D VTON network with 3D-consistent generation capabilities. Starting with the 2D VTON baseline [53] using DensePose, we progressively replace DensePose with our pseudo-3D pose, incorporate multi-view CLIP embeddings, and ultimately integrate MVAttention in the final configuration. Results in Tab. 2 indicate that each module contributes to metric improvements. Fig. 7 visualizes an example of multi-view image editing. The incorporation of pseudo-3D pose substantially improves limb generation compared to the 2D VTON baseline. Comparing rows 4 and 5, prior to the integration of multi-view CLIP embedding, the model captures limited spatial information, resulting in detail loss at specific angles (columns 3, 4, and 6). Finally, the proposed MVAttention achieves a more coherent generation across views.

*https://net-a-porter.com, www.taobao.com, www.douyin.com

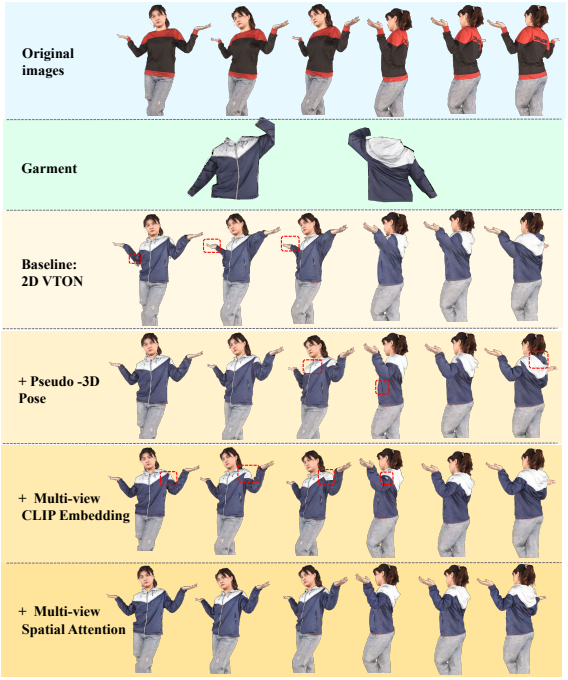

Figure 7. **Visualization of the impact of the three proposed techniques on multi-view consistent editing.** The red boxes highlight the artifacts. Starting from the 2D VTON baseline, the pseudo-3D pose improves limb generation quality, multi-view CLIP embedding enhances detail across different viewing directions, and finally, MVAttention further strengthens consistency in the generated images.

Methods	$\text{CLIP}_{cons}\uparrow$	$\text{DINO}_{sim}\uparrow$
2D-VTON	0.892	0.609
+ Pseudo-3D Pose	0.910	0.626
+ Multi-view CLIP Embedding	0.913	0.631
+ Multi-view Spatial Attention	**0.923**	**0.633**

Table 2. **Ablation studies.** We ablate the impact of the three proposed techniques on Thuman2.0 dataset.

6. Conclusions

In this work, we proposed VTON 360, a novel 3D Virtual Try-On (VTON) method that achieves high-fidelity VTON with the ability to render clothing from arbitrary viewing directions. Our method features a novel formulation of 3D VTON as an extension of 2D VTON that ensures 3D consistent results across multiple views. To bridge the gap between 2D VTON models and 3D consistency requirements, we introduce several key innovations, including multi-view inputs, pseudo-3D pose representation, multi-view spatial attention, and multi-view CLIP embedding. Extensive experiments demonstrate the effectiveness of our approach, significantly outperforming prior 3D VTON techniques in both fidelity and any-view rendering.

Acknowledgement

This work is supported in part by the National Key R&D Program of China under Grant No.2024YFB3908503, in part by the National Natural Science Foundation of China under Grant NO. 62322608 and in part by the CCF-Kuaishou Large Model Explorer Fund (NO. CCF-KuaiShou 2024007).

References

[1] Easymocap - make human motion capture easier. Github, 2021. 5

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision*, pages 409–425. Springer, 2022. 2

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430, 2019. 2, 3

[4] Robert Bridson, Ronald Fedkiw, and John Anderson. Robust treatment of collisions, contact and friction for cloth animation. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 594–603, 2002. 3

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 3

[6] Haodong Chen, Yongle Huang, Haojian Huang, Xiangsheng Ge, and Dian Shao. Gaussianvton: 3d human virtual try-on via multi-stage gaussian splatting editing with image prompting. *arXiv preprint arXiv:2405.07472*, 2024. 2, 3, 1

[7] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21476–21485, 2024. 3

[8] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2021. 1, 2

[9] Yisol Choi, Sangkyung Kwak, Kyungmin Lee, Hyungwon Choi, and Jinwoo Shin. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*, 2024. 3, 4, 5

[10] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Daniel K Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3480–3489, 2022. 2

[11] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8485–8493, 2021. 2

[12] 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. *arXiv preprint arXiv:2308.06101*, 2023. 1, 2, 4

[13] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012. 1, 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 4, 5

[15] Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert W Sumner, Forrester Cole, Mark Meyer, Tony DeRose, and Markus Gross. Subspace clothing simulation using adaptive bases. *ACM Transactions on Graphics (TOG)*, 33(4):1–9, 2014. 1, 3

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[17] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 3, 6, 1

[18] 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)*, pages 3470–3479, 2022. 2

[19] Zijian He, Peixin Chen, Guangrun Wang, Guanbin Li, Philip HS Torr, and Liang Lin. Wildvidfit: Video virtual try-on in the wild via image-based controlled diffusion models. In *European Conference on Computer Vision*, pages 123–139. Springer, 2024. 1, 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 7

[22] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023. 5

[23] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 6, 7

[24] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. Tech: Text-guided

reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, pages 1531–1542. IEEE, 2024. 3

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 3, 1

[26] Jeongho Kim, Gyojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. *arXiv preprint arXiv:2312.01725*, 2023. 3, 4, 5

[27] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[28] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pages 667–684, 2018. 1, 3

[29] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 2

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[31] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5084–5093, 2020. 2

[32] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034, 2020. 2, 3

[33] Davide Morelli, Alberto Baldrati, Giuseppe Cartella, Marcella Cornia, Marco Bertini, and Rita Cucchiara. Ladi-vton: Latent diffusion textual-inversion enhanced virtual try-on. *arXiv preprint arXiv:2305.13501*, 2023. 1, 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[35] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 5

[36] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15, 2017. 3

[37] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[38] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13535–13544, 2022. 2

[39] 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*, pages 10684–10695, 2022. 1, 3, 4, 6

[40] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 3

[41] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11763–11773, 2021. 2

[42] Igor Santesteban, Miguel Otaduy, Nils Thuerey, and Dan Casas. Ulnef: Untangled layered neural fields for mix-and-match virtual try-on. *Advances in Neural Information Processing Systems*, 35:12110–12125, 2022. 2

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[45] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[46] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. Mv-vton: Multi-view virtual try-on with diffusion models. *arXiv preprint arXiv:2404.17364*, 2024. 7

[47] Junjie Wang, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, and Qi Tian. Gaussianeditor: Editing 3d gaussians delicately with text instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20902–20911, 2024. 3

[48] Jing Wu, Jia-Wang Bian, Xinghui Li, Guangrun Wang, Ian Reid, Philip Torr, and Victor Adrian Prisacariu. Gaussctrl: multi-view consistent text-driven 3d gaussian splatting editing. *arXiv preprint arXiv:2403.08733*, 2024. 3, 6, 7

[49] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 5

[50] Zhenyu Xie, Haoye Dong, Yufei Gao, Zehua Ma, and Xiaodan Liang. Dreamvton: Customizing 3d virtual try-on with personalized diffusion models. *arXiv preprint arXiv:2407.16511*, 2024. 2, 3

[51] 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*, pages 19801–19811, 2024. 2, 6, 7

[52] Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 5

[53] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779*, 2024. 3, 4, 8

[54] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3460–3469, 2022. 2

[55] 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*, pages 5746–5756, 2021. 2, 6, 7, 8

[56] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7982–7990, 2021. 2

[57] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 3

[58] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. 5

[59] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021. 3

[60] 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*, pages 4606–4615, 2023. 2

[61] Jingyu Zhuang, Chen Wang, Liang Lin, Lingjie Liu, and Guanbin Li. Dreameditor: Text-driven 3d scene editing with neural fields. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[62] Jingyu Zhuang, Di Kang, Linchao Bao, Liang Lin, and Guanbin Li. Dagsm: Disentangled avatar generation with gs-enhanced mesh. *arXiv preprint arXiv:2411.15205*, 2024. 3

[63] Jingyu Zhuang, Di Kang, Yan-Pei Cao, Guanbin Li, Liang Lin, and Ying Shan. Tip-editor: An accurate 3d editor following both text-prompts and image-prompts. *arXiv preprint arXiv:2401.14828*, 2024. 3, 6, 7, 1

VTON 360: High-Fidelity Virtual Try-On from Any Viewing Direction

Supplementary Material

Appendix A introduces the preliminaries of 3DGS. The detailed formulations of the two quantitative metrics are presented in Appendix B. Additionally, Appendix C outlines the post-processing techniques applied to ensure the preservation of human characteristics in image editing. Appendix D elaborates on the failure cases and proposes a mitigation strategy to address it. Finally, Appendix E showcases additional VTON results, including those from a real 3D scene used in GaussianVTON [6].

A. 3D Representation: Gaussian Splatting

3D Gaussian Splatting (3DGS) [25] has emerged as a prominent technique in 3D reconstruction due to its ability to render high-quality scenes in real-time. Unlike traditional point cloud based methods, which directly represent scenes as discrete points, 3DGS models each point as a continuous Gaussian function g_i:

$$g_i(x;\mu_i,\Sigma_i) = e^{-\frac{1}{2}(x-\mu_i)^\top \Sigma_i (x-\mu_i)}, \tag{7}$$

where x is the position vector of g_i, $\mu_i \in \mathbb{R}^3$ and $\Sigma_i \in \mathbb{R}^{3\times3}$ are g_i's mean and covariance matrix, respectively. Then, g_i is projected onto a 2D image plane to facilitate rendering. This projection yields a new mean vector ${\mu_i}' \in \mathbb{R}^2$ and an updated covariance matrix $\Sigma_i' \in \mathbb{R}^{2\times2}$ defined as:

$${\mu_i}' = KT[{\mu_i}^\top, 1]^\top, \Sigma_i' = JT\Sigma_i T^\top J^\top, \tag{8}$$

where J is the Jacobian matrix derived from the affine approximation of the perspective projection, T and K denote the extrinsic and intrinsic matrices, respectively. Given the color c_i and opacity α_i at the Gaussian center point, the rendered color at a 2D pixel p is calculated as follows:

$$\begin{aligned} C_p &= \sum_{i=1}^{N} \alpha_i c_i T_i g_i(p;\mu_i',\Sigma_i') \\ T_i &= \prod_{j=1}^{i-1} (1-\alpha_j g_j(p;\mu_j',\Sigma_j')), \end{aligned} \tag{9}$$

where T_i denotes the cumulative transmission along the ray.

B. Metrics

In the quantitative evaluation, we employ two metrics:

- Average DINO Similarity [63], which measures the alignment between the garment image and the edited 3D human.
- CLIP Directional Consistency Score [17], which evaluates multi-view consistency.

Specifically, given an edited 3D human (after VTON), 120 views are uniformly projected around its central axis. These views are divided into three categories based on orientation: S_f, S_b, and S_s, corresponding to 40 front views, 40 back views, and 40 side views, respectively. Let $D(\cdot)$ represent the normalized DINO embedding and $C(\cdot)$ denote the normalized CLIP embedding. Using these, we formally define the two metrics as follows:

$$\begin{aligned} \text{DINO}_{sim} &= \frac{1}{80}\Big(\sum_{i\in S_f} D(g_f)\cdot D(e_i) + \sum_{i\in S_b} D(g_b)\cdot D(e_i)\Big) \\ \text{CLIP}_{cons} &= \frac{1}{120}\sum_{i}(C(e_i)-C(o_i))\cdot(C(e_{i+1})-C(o_{i+1})) \end{aligned} \tag{10}$$

where e_i, e_{i+1} and o_i, o_{i+1} denotes the two consecutive novel views from the edited 3DGS and the original 3DGS, respectively.

C. Post-processing

The clothing-agnostic maps $\mathbf{A}$ often mask parts of the face and hair, particularly for females. Due to the inherent properties of the diffusion model, it is unable to fully restore the intricate details of these masked regions. To ensure high-fidelity preservation of human characteristics, we apply a post-processing step where, after editing the rendered views, we "copy" the face and hair from the original image o onto the edited image e. Specifically, let m represent the region corresponding to the face and hair, which can be extracted from the parsed map during pre-processing, we implement post-processing as:

$$e = (1-m)\cdot e + m\cdot o \tag{11}$$

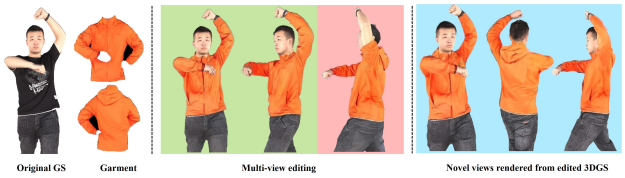

Figure 8. Our multi-view editing may fail in certain views with complex poses (red box in pink background) but these views can be automatically discarded to mitigate their impact on 3D VTON (blue background).

D. Limitations

As shown in Fig. 8, our method may fail in certain views with complex postures. To address this, we use Z-Score

Normalization to automatically identify and discard problematic views based on the view reconstruction loss during the process of lifting multiple views to 3D space, mitigating their adverse impact.

E. Additional Visualization Results

Fig. 9 illustrates additional VTON results. The first two rows showcase results from the THuman2.0 dataset; the middle two rows showcase results from the MVHumanNet dataset. To further demonstrate the effectiveness of our method, we apply it on a real 3D scene used in GaussianVTON [6]. The last two rows in Fig. 9 illustrate these VTON results with the model trained on Thuman2.0 dataset. Despite the data gap, including w/wo background and unseen camera poses, our method exhibits robust performance and preserves the details of the clothing well.

Figure 9. **Additional visualization results.** The first, middle, and last two rows show results on Thuman2.0, MVHumanNet, and a real 3D scene used in GaussianVTON, respectively.