IPVTON: Image-based 3D Virtual Try-on with Image Prompt Adapter

Xiaojing Zhong[1,2], Zhonghua Wu[3], Xiaofeng Yang[2], Guosheng Lin [2*], Qingyao Wu [1,4*]

[1]School of Software Engineering, South China University of Technology, China
[2]Nanyang Technological University, Singapore
[3]SenseTime Research, Singapore
[4]Peng Cheng Laboratory, China
vzxj12@gmail.com, gslin@ntu.edu.sg, qyw@scut.edu.cn

Abstract

Given a pair of images depicting a person and a garment separately, image-based 3D virtual try-on methods aim to reconstruct a 3D human model that realistically portrays the person wearing the desired garment. In this paper, we present IPVTON, a novel image-based 3D virtual try-on framework. IPVTON employs score distillation sampling with image prompts to optimize a hybrid 3D human representation, integrating target garment features into diffusion priors through an image prompt adapter. To avoid interference with non-target areas, we leverage mask-guided image prompt embeddings to focus the image features on the try-on regions. Moreover, we impose geometric constraints on the 3D model with a pseudo silhouette generated by ControlNet, ensuring that the clothed 3D human model retains the shape of the source identity while accurately wearing the target garments. Extensive qualitative and quantitative experiments demonstrate that IPVTON outperforms previous methods in image-based 3D virtual try-on tasks, excelling in both geometry and texture.

Introduction

Human generation has been a prominent task in the AIGC field, with virtual try-on attracting widespread attention due to its significant commercial and entertainment value. Image-based 2D virtual try-on technology, which generates a realistic photo of a person wearing a desired garment by combining the person's image with the garment's image, is valued for its user-friendliness and resource efficiency. However, this method is limited by its reliance on a fixed viewpoint (see Fig. 1 (a)), which poses challenges in real-world applications where users often need to assess the garment from multiple angles. On the other hand, the traditional 3D virtual try-on method provides the advantage of multi-angle views but requires complex processes such as garment-body registration and physics simulations (see Fig. 1 (b)), making it labor-intensive. The challenge of reconstructing accurate 3D models from 2D images, an inherently ill-posed problem, further complicates efforts to integrate image-based and 3D-based virtual try-on techniques.

Owing to the remarkable progress in diffusion models for Text-to-Image (T2I) (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019), the field of

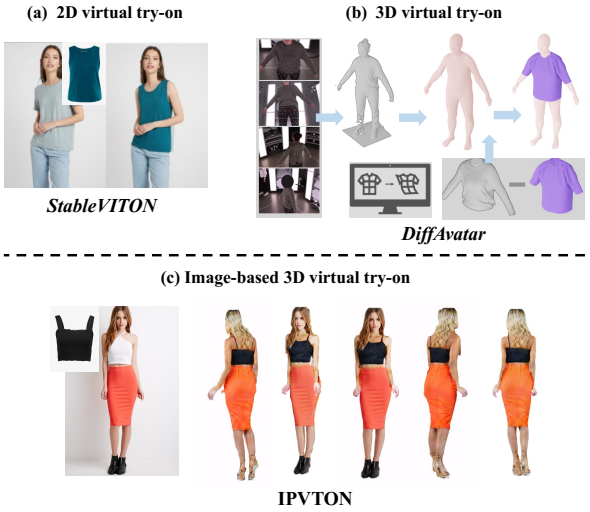

Figure 1: Compared to 2D virtual try-on (Kim et al. 2024) with its fixed viewpoint and 3D virtual try-on (Li et al. 2024) that require complex processes, IPVTON can generate 3D try-on results from just a human image and a garment image.

3D content generation has seen significant advancements. Recent works (Chen et al. 2023; Wang et al. 2024; Qian et al. 2023; Zhong et al. 2025) leverage 2D generative priors from pre-trained T2I models (e.g., StableDiffusion (SD)) combined with the Score Distillation Sampling (SDS) loss (Poole et al. 2022) to optimize 3D representations, resulting in high-quality 3D objects. Despite the success in synthesizing images with specific concepts (Ruiz et al. 2023; Kumari et al. 2023), extending these techniques to customized 3D object generation remains challenging. For instance, incorporating personalized modules such as LoRAs (Hu et al. 2021) into the SD model diminishes its ability to generate consistent multi-view images (Xie et al. 2024). Additionally, fine-tuning with only a few images struggles to capture the complex features of garments necessary for 3D virtual try-on.

Image prompt adapter (IP-Adapter) (Ye et al. 2023) introduces a cross-attention layer for image prompts in diffusion models, enabling controllable generation based on provided

*Corresponding Authors

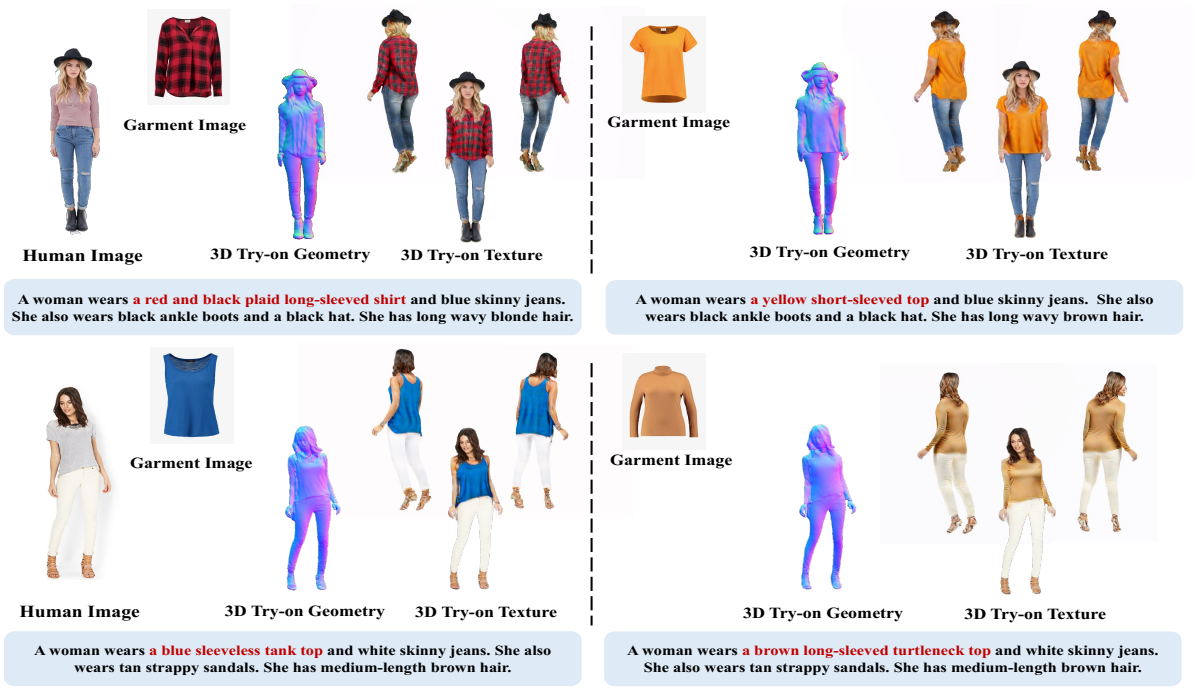

Figure 2: **3D Try-on results.** Given a human image, a garment image and a text prompt, IPVTON can generate realistic 3D human models with the desired garment shapes and textures while preserving the source identity.

images. In this paper, we propose IPVTON, a data-efficient image-based 3D virtual try-on framework that integrates an image prompt adapter with customized diffusion models to optimize a hybrid 3D human model using SDS loss. Since IP-Adapter is compatible with existing diffusion models, it eliminates the need for additional parameter fine-tuning with limited-viewpoint images, preserving multi-view generative priors for consistent 3D model generation. Moreover, combining textual and visual prompts effectively encapsulates the high-level semantics of garments. Specifically, we adopt a two-stage 3D generation framework that independently optimizes the geometry and texture of a hybrid 3D human model initialized with SMPL-X (Pavlakos et al. 2019), using an image prompt encoder to extract features from the target garment image and its corresponding normal map to guide the respective optimizations. Unlike using a reference image to influence the entire output (Zeng et al. 2023; Ran et al. 2024), virtual try-on requires preserving the non-try-on regions of the source human image during optimization. To address this problem, we employ mask-guided image prompt embeddings to focus the image prompt features on the targeted region, reducing unintended effects on surrounding areas. Furthermore, while mask guidance mitigates interference, it may limit the effectiveness of image prompts in guiding geometry generation. To overcome this problem, we introduce a Pseudo Silhouette Loss (PSL) to ensure the generated 3D human conforms to the desired garment shapes. Fig. 2 illustrates the 3D try-on results generated from a human image and an in-shop garment image. Overall, our contributions are summarized as follows:

- We design a data-efficient image-based 3D virtual try-on framework that generates 3D human models seamlessly wearing the desired garments, which can be observed from any viewpoint.
- We combine score distillation sampling with image prompts to optimize a hybrid 3D human representation, using an image prompt adapter to integrate garment features into the diffusion prior. We leverage mask-guided image prompt embeddings to focus the image features on the masked region, preserving the source identity in non-try-on areas.
- To ensure the generated model accurately reflects the desired garment shapes, we propose a pseudo silhouette loss to optimize the 3D human geometry.

Related Work

2D and 3D Virtual Try-on. Image-based 2D virtual try-on aims to fit an in-shop garment onto a clothed human in an image. Traditional methods primarily rely on Generative Adversarial Networks (GANs) (Goodfellow et al. 2020; Zhong et al. 2023a; Wu et al. 2020; Shi et al. 2021), where the garment is first deformed to align with the person's pose, followed by a generator that blends the deformed garment with the person's image (Zhong et al. 2021; Wu et al. 2019; Choi et al. 2021; Ge et al. 2021). Building on the advancements of diffusion models in image editing, virtual try-on research has increasingly focused on their application, leveraging pre-trained diffusion models to blend garments with human appearances (Kim et al. 2024; Choi et al. 2024; Zhu et al. 2023). Despite the success of 2D virtual

try-on methods, they struggle to generate multi-view try-on results, which are crucial for real-world applications.

With the increasing demand for 3D virtual try-on, (Bhatnagar et al. 2019; Mir, Alldieck, and Pons-Moll 2020; Patel, Liao, and Pons-Moll 2020; Zhong et al. 2023b) represent garments layered over the SMPL model (Loper et al. 2023; Pang et al. 2024). M3D-VTON (Zhao et al. 2021) constructs a 3D clothed human by predicting dual depth maps for a person's image and applies these depth values to the results of 2D virtual try-on. To leverage the powerful generative prior of diffusion models, DreamVTON (Xie et al. 2024) combines SDS loss with LoRAs (Hu et al. 2021) to generate 3D humans with customized identities and clothing. However, the need to fine-tune the LoRA layers for each pair of samples incurs a time cost. Efficiently integrating desired garment features into a diffusion model remains a challenge.

Text-guided 3D Human Generation. Avatar-CLIP (Hong et al. 2022) initializes the geometry of 3D human using a shape VAE network and refines geometry and texture with CLIP loss (Radford et al. 2021). Dreamwaltz (Huang et al. 2024b) improves SDS loss by incorporating 3D-aware skeleton conditioning, while Humannorm (Huang et al. 2024a) and AvatarVerse (Zhang et al. 2024) utilize the hybrid 3D representation DMTet (Shen et al. 2021) combined with structural condition maps to achieve more detailed and realistic geometry. TADA (Liao et al. 2024) enhances the upsampled SMPL-X model by adding a displacement layer and texture map. TeCH (Huang et al. 2024c) combines SDS loss with DreamBooth. However, while the geometry and texture of the generated 3D human can be altered by modifying the text prompt, the results often deviate from the provided image.

Customizing Diffusion Models. DreamBooth (Ruiz et al. 2023) fine-tunes the network on a small set of subject-specific images, enabling the customization of diffusion models to closely match the style or subject of the provided images. LoRA (Hu et al. 2021) reduces trainable parameters by learning rank-decomposition matrices, enabling efficient fine-tuning of pre-trained diffusion models with specific concepts. Custom Diffusion (Kim et al. 2024) fine-tunes a small subset of weights in the cross-attention layers, focusing on the key and value mappings from text to latent features. IP-Adapter (Ye et al. 2023) proposes decomposing the cross-attention layers for text and image features, allowing an image prompt adapter to incorporate additional image styles. (Choi et al. 2024) first customize a diffusion model with IP-Adapter for 2D virtual try-on. IPDreamer (Zeng et al. 2023) combines SDS loss with IP-Adapter, enabling the customization of 3D models. However, since it is designed for general 3D objects, applying it directly to humans yields coarse results due to the complexity of human topology.

Preliminaries

Latent Diffusion Model (LDM) performs diffusion in a lower-dimensional latent space for decreasing computing cost. Specifically, LDM employs an autoencoder to encode an input image x into a latent code $z = \mathcal{E}(x)$ and decode z to $x = \mathcal{D}(z)$. During the forward stage, the initial latent code z_0 is gradually perturbed by adding Gaussian noise ϵ over the time step t to match the Gaussian distribution: $z_t \sim \mathcal{N}(0, I)$. In the reverse stage, a noise predictor ϵ_ϕ based on a U-Net structure (Ronneberger, Fischer, and Brox 2015) and parameterized by ϕ is trained to predict the noise added at each corresponding step of the forward process. The training uses the following loss function:

$$\min_{\phi} \mathbb{E}_{\mathbf{z}_0, \epsilon \sim \mathcal{N}(0,I), t} \left\| \epsilon_\phi \left(z_t; y, t \right) - \epsilon \right\|_2^2, \tag{1}$$

where y represents a conditional text prompt and ϵ denotes the added random noise.

Score Distillation Sampling (SDS) is proposed to optimize a 3D representation parameterized by η using differentiable rendering, ensuring that the rendered 2D images conform to the diffusion prior. Given a random camera pose, the differentiable rendering function $\mathbf{g}$ generates the rendered image I via $I = \mathbf{g}(\eta)$. η is optimized for 3D consistency by computing the gradient of $\mathcal{L}_{\text{SDS}}$ with respect to z, which is encoded from the rendered image I:

$$\nabla_\eta \mathcal{L}_{\text{SDS}}(\phi, z) = \mathbb{E}_{t,\epsilon} \left[w(t) \left(\hat{\epsilon}_\phi \left(z_t; y, t \right) - \epsilon \right) \frac{\partial z}{\partial \eta} \right], \tag{2}$$

where $w(t)$ is a time-dependent weighting function that varies with t and z_t is the noised latent vector. Compared to ϵ_ϕ, $\hat{\epsilon}_\phi$ incorporates classifier-free guidance (Ho and Salimans 2022) to align the diffusion process with the target prompt.

Method

We first introduce an efficient 3D hybrid representation, initialized with the SMPL-X human body prior (Loper et al. 2023), to model the source identity's body shape and pose. Building on this model, we adopt a two-stage, text-guided 3D generation framework that independently optimizes the geometry and texture using SDS loss with mask-guided image prompt embeddings. To ensure the generated 3D human conforms to the desired garment shape, we employ a pseudo silhouette loss to constrain the geometry generation.

3D Hybrid Human Representation

We utilize DMtet (Shen et al. 2021) as our 3D representation because it combines explicit and implicit forms to efficiently model the 3D clothed human and can be easily converted into meshes. Inspired by (Huang et al. 2024c), we create an outer shell M_{shell} of SMPL-X (Feng et al. 2021) to form an outer shell tetrahedral grid (V_{shell}, T_{shell}), with V_{shell} representing the set of vertices and T_{shell} representing the set of tetrahedrons in the grid. For each vertex $v_i \in V_{shell}$, we train an MLP-based neural network Ω_g, parameterized by ϕ_g, to predict its Signed Distance Field (SDF) value: $\Omega_g(v_i) = s(v_i; \phi_g)$. We initialize Ω_g as follows:

$$\mathcal{L}_{SDF}(\Omega_g) = \sum_{x \in \mathbf{P}} \left\| s\left(x; \phi_g\right) - \text{SDF}\left(x\right) \right\|_2^2, \tag{3}$$

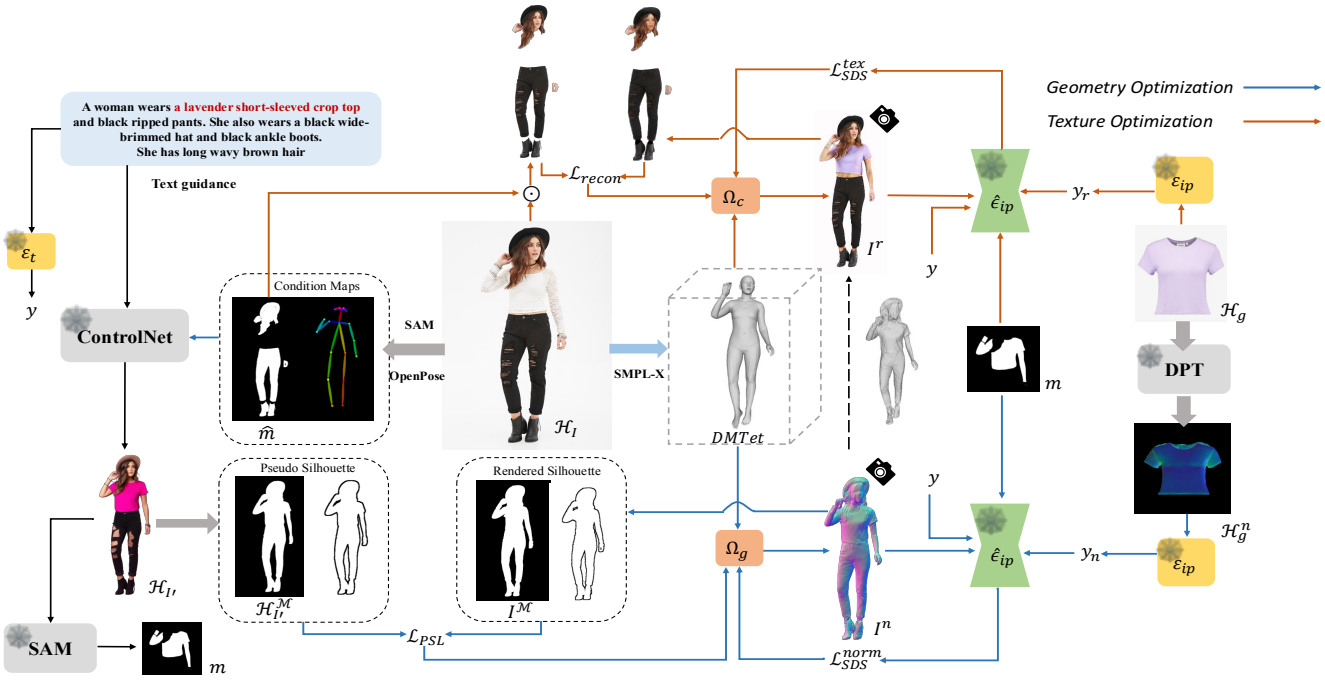

Figure 3: **Overview of IPVTON.** Given a human image $\mathcal{H}_I$, we first construct a DMTet-based 3d representation initialized with SMPL-X to model the human, with its geometry and texture generated through Ω_g and Ω_c, respectively. During geometry optimization, the rendered human normal map I^n is encoded into the diffusion model $\hat{\epsilon}_{ip}$ and, along with y_n and m, is used to compute $\mathcal{L}_{SDS}^{norm}$. y_n is the normal image prompt embedding encoded from $\mathcal{H}_g^n$ via $\mathcal{E}_{ip}$, and m is a mask covering the try-on region, derived from $\mathcal{H}_I'$. During texture optimization, the rendered human image I^r is encoded into $\hat{\epsilon}_{ip}$ and along with y_r, y and m, is used to compute $\mathcal{L}_{SDS}^{tex}$. y is the text prompt embedding encoded from the target texts via $\mathcal{E}_t$, and y_r is the image prompt embedding encoded from $\mathcal{H}_g$ via $\mathcal{E}_{ip}$. $\odot$ denotes pixel-wise multiplication.

where $\mathbf{P}$ is the set of random sampling points near M_{shell}.

Next, we employ the Marching Tetrahedra (MT) algorithm (Doi and Koide 1991) for iso-surface extraction, resulting in triangular meshes. Additionally, we train another MLP-based neural network Ω_c, parameterized by ϕ_c, to generate its albedo map. Given a sampled camera pose, we can utilize differentiable rasterization (Laine et al. 2020) to render the human mesh's normal map I^n, color map I^r, and mask $I^{\mathcal{M}}$.

Geometry Optimization Stage

To optimize geometry guided by text prompts, (Chen et al. 2023; Wang et al. 2024; Huang et al. 2024a) encode the rendered normal map, with the resulting encoding serving as input to the diffusion model for calculating the normal SDS loss. However, representing garments with complex shapes remains challenging without prompt engineering, as text prompts typically describe limited garment dimensions (e.g., length, width). To capture the high-level semantics of garments, we utilize IP-Adapter (Ye et al. 2023) to combine textual and image prompts with a decoupled cross-attention mechanism. Specifically, an image prompt adapter $\mathcal{E}_{ip}$ is used to project the image into a sequence of features that are combined with the textual embedding. As shown in Fig. 3, we extract the image prompt feature of the target garment's normal map, denoted as $y_n = \mathcal{E}_{ip}(\mathcal{H}_g^n)$, where the normal image $\mathcal{H}_g^n$ is obtained from the target garment image $\mathcal{H}_g$ via normal map estimation DPT (Ranftl, Bochkovskiy, and Koltun 2021). The calculation of normal SDS loss is as follows:

$$\nabla_{\phi_g}\mathcal{L}_{SDS}^{norm}(\phi', z^n) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_{ip}\left(z_t^n; y, y_n, t\right) - \epsilon\right)\frac{\partial z^n}{\partial \phi_g}\right], \tag{4}$$

where $\hat{\epsilon}_{ip}$ refers to the diffusion model employed in IP-Adapter, with ϕ' denoting its parameters, and z^n represents the latent codes encoded from I^n. However, using global image prompt embeddings can unintentionally affect the entire body, including areas that should remain unchanged during geometry optimization. To address this problem, we employ mask-guided image prompt embeddings to focus the image prompt features on the masked region. Given the query features $\mathbf{Z}$, the output of the cross-attention module for the image prompt, $\mathbf{Z}'$, is computed as follows:

$$\mathbf{Z}' = m\,\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{ip}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{ip}, \tag{5}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \mathbf{K}_{ip} = y_n\mathbf{W}'_k$ and $\mathbf{V}_{ip} = y_n\mathbf{W}'_v$ are the query, key, and value matrices for the normal image

prompt features y_n. $\mathbf{W}_q, \mathbf{W'}_k, \mathbf{W'}_v$ are the projection matrices used for linear transformation. m denotes the partial mask of the pseudo mask $\mathcal{H}_{I'}^{\mathcal{M}}$ generated by SAM, covering the try-on regions, which will be described later.

Although mask guidance reduces interference, it can restrict the ability of image prompts to effectively steer geometry generation. To enforce geometric constraints on the generated 3D model, we apply a pseudo silhouette loss to shape the contours of the 3D human. Specifically, as shown in Fig. 3, we use two condition maps to guide ControlNet: a partial segmentation map of $\mathcal{H}_I$, excluding the regions to be transferred, and a pose map of $\mathcal{H}_I$, extracted with OpenPose (Cao et al. 2017). This process generates a human image $\mathcal{H}_{I'}$ that combines the body from $\mathcal{H}_I$ with the garment shapes from $\mathcal{H}_g$. We then use SAM to generate the corresponding mask, $\mathcal{H}_{I'}^{\mathcal{M}}$. The pseudo silhouette loss can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{PSL}} = \|\mathcal{H}_{I'}^{\mathcal{M}} - I^{\mathcal{M}})\|_2^2 + \\ \sum_{k \in \text{Edge}(I^{\mathcal{M}})} \min_{\hat{k} \in \text{Edge}(\mathcal{H}_{I'}^{\mathcal{M}})} \|k - \hat{k}\|_1.\end{aligned} \tag{6}$$

It ensures that both the edges and the silhouette mask of $I^{\mathcal{M}}$ align with those of $\mathcal{H}_{I'}^{\mathcal{M}}$. Moreover, we can estimate the normal maps of $\mathcal{H}_I$ and $\mathcal{H}_{I'}$ using ICON (Xiu et al. 2022). By combining partial normal maps from $\mathcal{H}_I$ and $\mathcal{H}_{I'}$ based on segmentation maps, we can also obtain a pseudo normal map ground truth $\mathcal{H}_I^{n'}$ to further constrain the geometry:

$$\mathcal{L}_{norm} = \|\mathcal{H}_I^{n'} - I^n\|_2^2. \tag{7}$$

Note that the camera views used to render $I^{\mathcal{M}}$ and I^n in Eq. 6 and Eq. 7 are specifically the front and back views. The overall geometry loss functions are calculated as follows:

$$\begin{aligned}\mathcal{L}_{geo} = \lambda_{PSL}\mathcal{L}_{PSL} + \lambda_{norm}\mathcal{L}_{norm} \\ + \lambda_{SDS}^{norm}\mathcal{L}_{SDS}^{norm} + \lambda_{lap}\mathcal{L}_{lap},\end{aligned} \tag{8}$$

where $\lambda_{\{PSL,norm,SDS(norm),lap\}}$ denotes the weights used to balance the geometry losses, and $\mathcal{L}_{lap}$ represents the Laplacian smoothing term (Ando and Zhang 2006), applied for surface regularization.

Texture Optimization Stage

Despite the guidance provided by text prompts, accurately capturing the target garment's texture remains challenging, as text descriptions often fail to convey its brightness and saturation. We extract the image prompt embedding of the target garment as $y_r = \mathcal{E}_{ip}(I^g)$. The texture SDS loss $\mathcal{L}_{SDS}^{tex}$ is obtained as follows:

$$\begin{aligned}\nabla_{\phi_c}\mathcal{L}_{SDS}^{tex}(\phi', z^r) = \\ \mathbb{E}_{t,\epsilon}\left[w(t)\left(\hat{\epsilon}_{ip}\left(z_t^r; y, y_r, t\right) - \epsilon\right)\frac{\partial z^r}{\partial \phi_c}\right],\end{aligned} \tag{9}$$

where z^r represents the latent codes encoded from I^r.

Similar to the geometry optimization, we apply mask m to the image prompt features y_r to concentrate the garment texture on the target region. Given the query features $\hat{\mathbf{Z}}$, the output of cross-attention for y_r is denoted as $\mathbf{Z''}$:

$$\mathbf{Z''} = m\,\text{Softmax}\left(\frac{\mathbf{Q'}\mathbf{K}_{ip'}^{\top}}{\sqrt{d}}\right)\mathbf{V}_{ip'}, \tag{10}$$

where $\mathbf{Q'} = \hat{\mathbf{Z}}\mathbf{W}_q, \mathbf{K}_{ip'} = y_r\mathbf{W'}_k$ and $\mathbf{V}_{ip'} = y_r\mathbf{W'}_v$ represent the query, key, and value matrices of the cross-attention module for image prompt features y_r.

To retain the appearance of the source human image in regions unaffected by the garment transfer, we employ $\hat{m}$ to constrain the local texture as follows:

$$\mathcal{L}_{recon} = \|\hat{m}(\mathcal{H}_I - I^r)\|_2^2, \tag{11}$$

where $\hat{m}$ represents the mask extracted from $\mathcal{H}_I$ that excludes the regions to be transferred. The overall texture loss functions are calculated as follows:

$$\mathcal{L}_{tex} = \lambda_{SDS}^{tex}\mathcal{L}_{SDS}^{tex} + \lambda_{recon}\mathcal{L}_{recon}, \tag{12}$$

where $\lambda_{\{recon,SDS(tex)\}}$ denotes the weights used to balance the texture losses.

Experiment

Implementation Details. We train both Ω_g and Ω_c for 100 iterations with one GeForce RTX 2080 Ti GPU. During the geometry optimization stage, we set λ_{PSL}, λ_{norm}, and λ_{lap} to 10,000, and set λ_{SDS}^{norm} to 1. During the texture optimization stage, we set λ_{recon} and λ_{SDS}^{tex} to 10,000 and 1, respectively.

Datasets. Our method does not require paired human and clothing images for training. We select 12 full-body, front-facing human images of different individuals from the DeepFashion dataset (Shen et al. 2021). For each human image, we choose two garment templates from the VITON-HD dataset (Choi et al. 2021), covering various types such as tank tops, short sleeves, and long sleeves. Descriptive text prompts for both the human and garment images are generated using ChatGPT-4o. More details are provided in the supplementary material.

Baselines. To demonstrate the effectiveness of our proposed IPVTON, we conduct a comparative analysis with the following baseline methods. 1) *TEXTure* (Richardson et al. 2023) generates and edits the texture of 3D objects based on text prompts. 2) *TeCH* (Huang et al. 2024c) leverages SDS loss with fine-tuned DreamBooth (Ruiz et al. 2023) for text-guided 3D human reconstruction. 3) *IPDreamer* (Zeng et al. 2023) utilizes IP-Adapter to control both the geometry and appearance of 3D objects. Since TEXTure can generate textures but not geometry, we downsample the 3D human model generated by TeCH for faster UV unwrapping with an atlas, using it as the target for texture generation. To ensure a fair comparison, we apply the mask $\hat{m}$ to the reconstruction loss used in TeCH, so that only the texture of the try-on regions is affected.

Figure 4: **Qualitative comparisons.** Our IPVTON is able to generate realistic 3D try-on results with high-quality textures, viewable from multiple angles.

Qualitative Comparison. As shown in Fig. 4, TEXTure struggles to generate accurate garment textures and correctly position them on the body according to the target prompt. While TeCH can produce try-on results, it faces challenges in reshaping the human body to match the desired garment shapes. IPDreamer, by leveraging IP-Adapter, captures garment features effectively, accurately reflecting the garment's color, length, and style. However, this method, designed for general 3D objects, results in a coarse human appearance and fails to distinguish between the front and back of the person. In contrast, our IPVTON generates realistic 3D try-on results that capture the desired garment shapes and textures while preserving the source identity in non-try-on areas.

Quantitative Comparison. We employ CLIP (Radford et al. 2021) metrics to quantitatively evaluate the faithfulness of the generated 3D try-on results to the target text prompts. We select 8 sets of human images with different identities, each paired with a garment template distinct from the clothing worn by the individuals. For each generated 3D model, we render images from six uniformly sampled angles. CLIP scores are computed by comparing the CLIP embeddings of these images with the target text prompt embeddings. To evaluate geometry faithfulness, we remove texture-related words from the target text prompt and prepend '$the\ normal\ map\ of$'. As shown in Tab. 1, our method achieves the best scores for both geometry and texture faithfulness. We also conduct a user study to further evaluate our method. Using 8 sets of 3D try-on results generated by four methods, we invite 15 volunteers to rank each set according to their preferences for geometry and texture. For each set, users are presented with the source human image, the garment image, and the target text prompt. Participants rank the results separately for geometry and texture

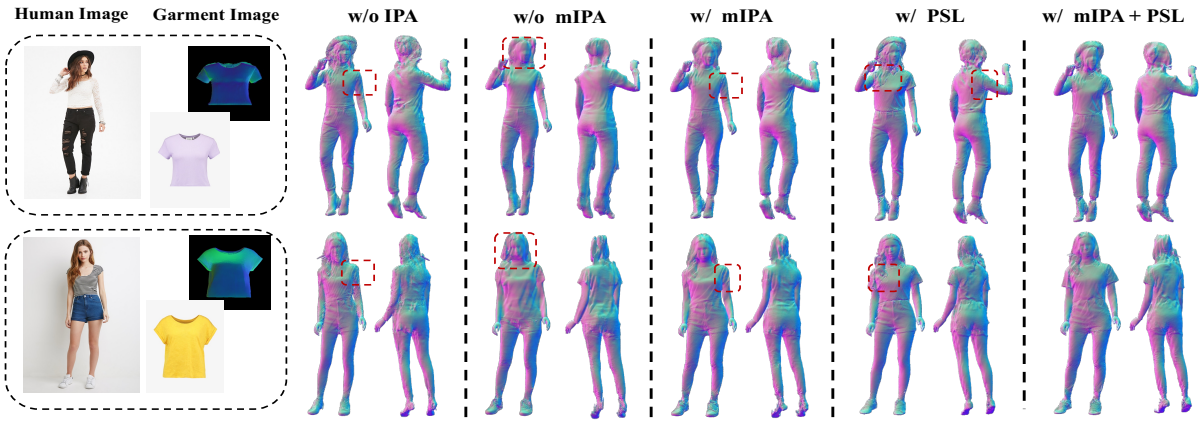

Figure 5: **Ablation study for geometry optimization.** 'mIPA' denotes mask-guided image prompt embeddings.

Methods	CLIP ↑		User ↑	
	Geo-Faith	Tex-Faith	Geometry	Texture
TEXTure	30.85	28.31	2.34	1.53
TeCH	31.41	32.34	2.55	2.6
IPDreamer	28.53	30.49	2.28	1.86
IPVTON	**31.77**	**33.60**	**2.63**	**3.52**

Table 1: Quantitative evaluation of the results obtained from different methods. 'Geo-Faith' and 'Tex-Faith' respectively denote the geometry and texture faithfulness. The best scores are highlighted in **bold**.

on a scale from 4 (highest) to 1 (lowest), without repeating scores. The final report presents the average scores across all sets. As shown in Table 1, our method achieves the highest human preference in both geometry and texture.

Ablation Study. As shown in Fig. 5, the geometry of the human model generated without using the image prompt adapter retains the geometry of the source human image but fails to reflect the desired garment shapes. In the second column, using global image prompt embeddings allows the body shape to adopt the garment's contours, but this also affects other parts, leading to a blurred face. In comparison, mask-guided image prompt embeddings preserve the source body shapes with sharp details, even though the garment shapes are not fully realized. Note that the results in the third column differ from those in the first because the third column includes garment features like the collar. Relying solely on PSL can cause noisy seams and inaccurate shapes, as seen in the back views of the fourth column, due to potential inaccuracies in the generated pseudo silhouette. Combining mask-guided image prompt embeddings with PSL supervision, IPVTON accurately generates the desired garment contours while maintaining well-defined human body shapes. As shown in Fig. 6, when texture is generated solely from text prompts, the resulting texture corresponds to the text prompt but deviates from the garment image. For instance, in the first row of the first column, the lavender crop top generated without using the image prompt adapter is slightly

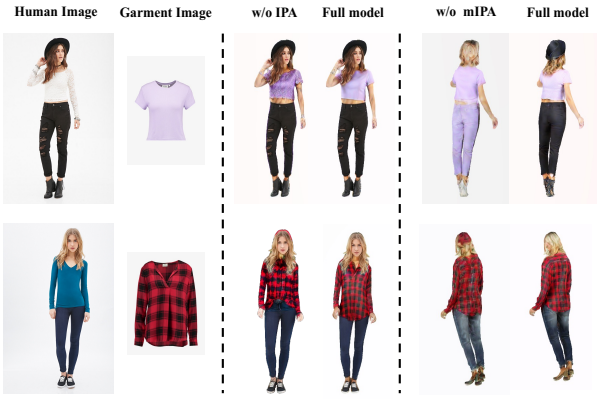

Figure 6: **Ablation study for texture optimization.** 'mIPA' denotes mask-guided image prompt embeddings.

darker than the garment image. In the second column, when only the top is meant to be changed, the pants and hair are also affected if image prompt embeddings are used without mask guidance.

Limitation

Reconstructing an extremely loose target garment may fail, likely due to inherent limitations of the SMPL-X initialization. Additionally, since we use a pre-trained image prompt adapter designed for high-level semantic features, the resulting features may not accurately capture the garment's complex patterns or logos. More details are provided in the supplementary material.

Conclusion

We propose an image-based 3D virtual try-on framework that optimizes 3D models by integrating garment features via a customized diffusion model with an image prompt adapter. Mask-guided prompt embeddings focus on try-on regions, minimizing interference. A pseudo silhouette loss constrains the 3D geometry, shaping the human form with the desired garment and source identity.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) 62272172, Guangdong Basic and Applied Basic Research Foundation 2023A1515012920. Zhuhai Science and Technology Plan Project(2320004002758). This research is partly supported by the MoE AcRF Tier 2 grant (MOE-T2EP20223-0001).

References

Ando, R.; and Zhang, T. 2006. Learning on graph with Laplacian regularization. *Advances in neural information processing systems*, 19.

Bhatnagar, B. L.; Tiwari, G.; Theobalt, C.; and Pons-Moll, G. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5420–5430.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chen, R.; Chen, Y.; Jiao, N.; and Jia, K. 2023. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22246–22256.

Choi, S.; Park, S.; Lee, M.; and Choo, J. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14131–14140.

Choi, Y.; Kwak, S.; Lee, K.; Choi, H.; and Shin, J. 2024. Improving diffusion models for virtual try-on. *arXiv preprint arXiv:2403.05139*.

Doi, A.; and Koide, A. 1991. An efficient method of triangulating equi-valued surfaces by using tetrahedral cells. *IEICE TRANSACTIONS on Information and Systems*, 74(1): 214–224.

Feng, Y.; Choutas, V.; Bolkart, T.; Tzionas, D.; and Black, M. J. 2021. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, 792–804. IEEE.

Ge, Y.; Song, Y.; Zhang, R.; Ge, C.; Liu, W.; and Luo, P. 2021. Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8485–8493.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hong, F.; Zhang, M.; Pan, L.; Cai, Z.; Yang, L.; and Liu, Z. 2022. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, X.; Shao, R.; Zhang, Q.; Zhang, H.; Feng, Y.; Liu, Y.; and Wang, Q. 2024a. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4568–4577.

Huang, Y.; Wang, J.; Zeng, A.; Cao, H.; Qi, X.; Shi, Y.; Zha, Z.-J.; and Zhang, L. 2024b. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36.

Huang, Y.; Yi, H.; Xiu, Y.; Liao, T.; Tang, J.; Cai, D.; and Thies, J. 2024c. Tech: Text-guided reconstruction of lifelike clothed humans. In *2024 International Conference on 3D Vision (3DV)*, 1531–1542. IEEE.

Kim, J.; Gu, G.; Park, M.; Park, S.; and Choo, J. 2024. 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*, 8176–8185.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Laine, S.; Hellsten, J.; Karras, T.; Seol, Y.; Lehtinen, J.; and Aila, T. 2020. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (ToG)*, 39(6): 1–14.

Li, Y.; Chen, H.-y.; Larionov, E.; Sarafianos, N.; Matusik, W.; and Stuyck, T. 2024. Diffavatar: Simulation-ready garment optimization with differentiable simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4368–4378.

Liao, T.; Yi, H.; Xiu, Y.; Tang, J.; Huang, Y.; Thies, J.; and Black, M. J. 2024. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, 1508–1519. IEEE.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Mir, A.; Alldieck, T.; and Pons-Moll, G. 2020. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7023–7034.

Pang, H. E.; Cai, Z.; Yang, L.; Tao, Q.; Wu, Z.; Zhang, T.; and Liu, Z. 2024. Towards robust and expressive whole-body human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36.

Patel, C.; Liao, Z.; and Pons-Moll, G. 2020. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7365–7375.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. 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*, 10975–10985.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Qian, G.; Mai, J.; Hamdi, A.; Ren, J.; Siarohin, A.; Li, B.; Lee, H.-Y.; Skorokhodov, I.; Wonka, P.; Tulyakov, S.; et al. 2023. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ran, L.; Cun, X.; Liu, J.-W.; Zhao, R.; Zijie, S.; Wang, X.; Keppo, J.; and Shou, M. Z. 2024. X-adapter: Adding universal compatibility of plugins for upgraded diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8775–8784.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179–12188.

Richardson, E.; Metzer, G.; Alaluf, Y.; Giryes, R.; and Cohen-Or, D. 2023. Texture: Text-guided texturing of 3d shapes. In *ACM SIGGRAPH 2023 conference proceedings*, 1–11.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22500–22510.

Shen, T.; Gao, J.; Yin, K.; Liu, M.-Y.; and Fidler, S. 2021. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34: 6087–6101.

Shi, X.; Wu, Z.; Lin, G.; Cai, J.; and Joty, S. 2021. Remember what you have drawn: Semantic image manipulation with memory. *arXiv preprint arXiv:2107.12579*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36.

Wu, Z.; Lin, G.; Tao, Q.; and Cai, J. 2019. M2e-try on net: Fashion from model to everyone. In *Proceedings of the 27th ACM international conference on multimedia*, 293–301.

Wu, Z.; Tao, Q.; Lin, G.; and Cai, J. 2020. Exploring bottom-up and top-down cues with attentive learning for webly supervised object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12936–12945.

Xie, Z.; Dong, H.; Gao, Y.; Ma, Z.; and Liang, X. 2024. DreamVTON: Customizing 3D Virtual Try-on with Personalized Diffusion Models. *arXiv preprint arXiv:2407.16511*.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. Icon: Implicit clothed humans obtained from normals. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13286–13296. IEEE.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zeng, B.; Li, S.; Feng, Y.; Li, H.; Gao, S.; Liu, J.; Li, H.; Tang, X.; Liu, J.; and Zhang, B. 2023. Ipdreamer: Appearance-controllable 3d object generation with image prompts. *arXiv preprint arXiv:2310.05375*.

Zhang, H.; Chen, B.; Yang, H.; Qu, L.; Wang, X.; Chen, L.; Long, C.; Zhu, F.; Du, D.; and Zheng, M. 2024. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7124–7132.

Zhao, F.; Xie, Z.; Kampffmeyer, M.; Dong, H.; Han, S.; Zheng, T.; Zhang, T.; and Liang, X. 2021. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13239–13249.

Zhong, X.; Huang, X.; Wu, Z.; Lin, G.; and Wu, Q. 2023a. Sara: Controllable makeup transfer with spatial alignment and region-adaptive normalization. *arXiv preprint arXiv:2311.16828*.

Zhong, X.; Huang, X.; Yang, X.; Lin, G.; and Wu, Q. 2025. Deco: Decoupled human-centered diffusion video editing with motion consistency. In *European Conference on Computer Vision*, 352–370. Springer.

Zhong, X.; Su, Y.; Wu, Z.; Lin, G.; and Wu, Q. 2023b. DI-Net: Decomposed Implicit Garment Transfer Network for Digital Clothed 3D Human. *arXiv preprint arXiv:2311.16818*.

Zhong, X.; Wu, Z.; Tan, T.; Lin, G.; and Wu, Q. 2021. Mv-ton: Memory-based video virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, 908–916.

Zhu, L.; Yang, D.; Zhu, T.; Reda, F.; Chan, W.; Saharia, C.; Norouzi, M.; and Kemelmacher-Shlizerman, I. 2023. Tryondiffusion: A tale of two unets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4606–4615.