2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.00041

GaussianIP: Identity-Preserving Realistic 3D Human Generation via Human-Centric Diffusion Prior

Zichen Tang[1] Yuan Yao Miaomiao Cui Liefeng Bo Hongyu Yang[1,2*]

[1]School of Artificial Intelligence, Beihang University, Beijing, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China

{zctang, hongyuyang}@buaa.edu.cn

Abstract

Text-guided 3D human generation has advanced with the development of efficient 3D representations and 2D-lifting methods like Score Distillation Sampling (SDS). However, current methods suffer from prolonged training times and often produce results that lack fine facial and garment details. In this paper, we propose GaussianIP, an effective two-stage framework for generating identity-preserving realistic 3D humans from text and image prompts. Our core insight is to leverage human-centric knowledge to facilitate the generation process. In stage 1, we propose a novel Adaptive Human Distillation Sampling (AHDS) method to rapidly generate a 3D human that maintains high identity consistency with the image prompt and achieves a realistic appearance. Compared to traditional SDS methods, AHDS better aligns with the human-centric generation process, enhancing visual quality with notably fewer training steps. To further improve the visual quality of the face and clothes regions, we design a View-Consistent Refinement (VCR) strategy in stage 2. Specifically, it produces detail-enhanced results of the multi-view images from stage 1 iteratively, ensuring the 3D texture consistency across views via mutual attention and distance-guided attention fusion. Then a polished version of the 3D human can be achieved by directly perform reconstruction with the refined images. Extensive experiments demonstrate that GaussianIP outperforms existing methods in both visual quality and training efficiency, particularly in generating identity-preserving results. Our code is available at: https://github.com/silence-tang/GaussianIP.

1. Introduction

Creating high-quality 3D human avatars based on user inputs is essential for a variety of applications, including Virtual Try-On [10, 11, 26, 71] and immersive telepresence [27, 28, 50]. It also plays an pivotal role in emerging technologies like AR and VR. Text-guided 3D human generation is a task which involves synthesizing a character's geometry and appearance from text prompts. Score Distillation Sampling (SDS) proposed in DreamFusion [46] paves the way for generating 3D objects by distilling from 2D diffusion priors [13, 49], simplifying the process of creating 3D objects. It also inspires researchers to design various pipelines tailored for creating 3D humans.

A common paradigm is to learn a SDS-guided conditional Neural Radiance Field (NeRF) [43] by integrating parametric human body models like SMPL [36] or imGHUM [1] into the framework to model body-related geometry and appearance [2, 17, 24, 69]. However, NeRF-based methods suffer from slow rendering speed and fall short in delivering high-resolution results.

Recently, as 3D Gaussian Splatting (3DGS) [22] unlocks new possibilities for high-fidelity 3D reconstruction with real-time rendering, a few works [34, 68] introduce 3DGS to achieve efficient creation of 3D avatars. Despite achieving promising results, all the NeRF-based or 3DGS-based methods rely solely on text prompts, limiting their diversity in generation and real-world applicability, such as creating identity-preserving 3D avatars from user portraits.

In contrast to these 3D-based methods, recent 2D generative diffusion models [13, 49] have exhibited pronounced superiority in terms of visual fidelity. Built upon these powerful generative base models, several works concentrate on human-centric generation tasks, e.g. Virtual Try-On (VTON) [23, 55, 73] and identity-driven photo personalization [29, 60, 65]. Although [3, 58, 62] attempt to edit 3D human bodies with the aid of 2D human-centric diffusion models, the potential of leveraging these models to generate detailed identity-preserving 3D humans from multi-modal user inputs, remains largely untapped.

In this work, we propose a novel two-stage framework GaussianIP, which is capable of generating realistic

*Corresponding author.

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.00041

identity-preserving 3D human avatars from both text and image prompts. Our key intuition is twofold: (a) We can develop an effective method to distill human-centric diffusion priors, enabling the generation of 3D avatars with high facial identity consistency to the input image and rich clothing details; (b) The powerful generative capability of diffusion models can be leveraged to further refine the results of the distillation process. In stage 1, we train the 3D human Gaussians with our proposed **Adaptive Human Distillation Sampling (AHDS)** guidance, consisting of Human Distillation Sampling (HDS) and an adaptive timestep scheduling strategy. The HDS is designed by decomposing the original score difference and incorporating identity conditions to ensure identity-preserving distillation. In timestep scheduling, we treat the entire HDS process as three consecutive phases, from coarse geometry to fine facial details, and assign different diffusion timestep sampling strategies to each phase to accelerate the generation process. Since the distilled results may exhibit subtle texture smoothing, we design an elaborative refinement mechanism in stage 2 to further improve the intricate details of the clothed human body. Instead of diffusing and denoising the multi-view images directly, which may jeopardizing 3D consistency, we propose a **View-Consistent Refinement (VCR)** mechanism. Specifically, four main views are denoised first and their intermediate attention features are stored to guide the denoising process of the k key views through mutual attention. For other ordinary views, we apply relative distance guided attention fusion to ensure texture consistency with their neighboring key views. Afterwards, the refined images can be utilized to optimize the 3D human Gaussians directly in a reconstruction way, which is highly efficient. Our key contributions include:

- We propose a novel identity-preserving 3D human generation task, where the human avatar should align with the given text and image prompts while maintaining highly-realistic face and clothes appearance. We design GaussianIP, a two-stage framework based on 3D Gaussian Splatting and 2D human-centric prior to achieve this goal.
- We propose an Adaptive Human Distillation Sampling (AHDS) strategy to efficiently generate high-quality 3D human models in stage 1. Compared to the traditional SDS strategy, our AHDS better supports the human-specific generation process, yielding results with enhanced visual quality while reducing training steps by approximately 30%.
- We introduce a View-Consistent Refinement (VCR) mechanism to enhance the visual details of multi-view images from stage 1. Leveraging the refined images, the 3D human model can be further improved in a short time.
- Extensive experiments show that GaussianIP outperforms existing methods in both visual quality and training efficiency, particularly excelling at generating identity-preserving facial details and rich garment textures.

2. Related Work

2D Human-centric Diffusion Models. 2D generative diffusion models [13, 38, 49] have demonstrated remarkable performance in the realm of conditional image generation [7, 12, 54, 70], though they are typically designed for general object generation. Recently, researchers fine-tune these models on human-related datasets [63, 72], creating models that perform exceptionally well in generating content such as clothed human body. A prominent line of work is identity-preserving image customization [29, 30, 60, 65]. These methods achieve single-ID personalization by leveraging global [47] or local [6] facial features to condition the generation process on the identity of a single individual. Another noteworthy direction explores Virtual Try-On (VTON) [55, 73], aiming to generate an outfitted human wearing the given garment. These methods typically involve the integration of garment warping modules to generate deformed garments [9] or designing specialized UNet or injection blocks [23] to align the garment features with the human body. However, lifting these 2D methods to 3D scenes, thereby expanding their applications in AR/VR, remains largely unexplored.

Text-guided 3D General Object Generation. Traditional methods [18, 44, 51] rely on CLIP [47] to optimize the 3D representations while decent results cannot be achieved due to the low expressivity of CLIP. Score Distillation Sampling (SDS) proposed in DreamFusion [46] opens the door for high-quality 3D object generation by distilling from text-to-image diffusion models and motivates various concurrent works [4, 33, 48, 57]. To improve the original SDS loss, several novel distillation techniques [5, 16, 21, 37, 67] have been introduced, such as Variational Score Distillation (VSD) [61], Internal Score Matching (ISM) [31], Asynchronous Score Distillation (ASD) [40], etc., by reexamining the SDS process or analyzing the behavior of 2D diffusion models. Considering the time-consuming NeRF training process, several works [41, 56, 66] adopt 3D Gaussian Splatting (3DGS) [22] as their 3D representation to achieve efficient training and rendering.

Text-guided 3D Human Generation. AvatarCLIP [14] first realizes zero-shot text-driven 3D avatar generation by leveraging CLIP and NeuS [59]. Inspired by SDS, DreamWaltz [17], AvatarVerse [69] and other concurrent works [2, 24] utilize the parametric SMPL model [36] and auxiliary networks like densepose ControlNet [70] to guide the learning process, improving the visual fidelity and 3D consistency of the generated 3D avatars. Apart from these NeRF-based methods, some works integrate other 3D representations into their frameworks to attain more fine-grained control of geometric shapes or textures. AvatarCraft [19], TADA [32] and X-Oscar [39] employ optimized SMPL-

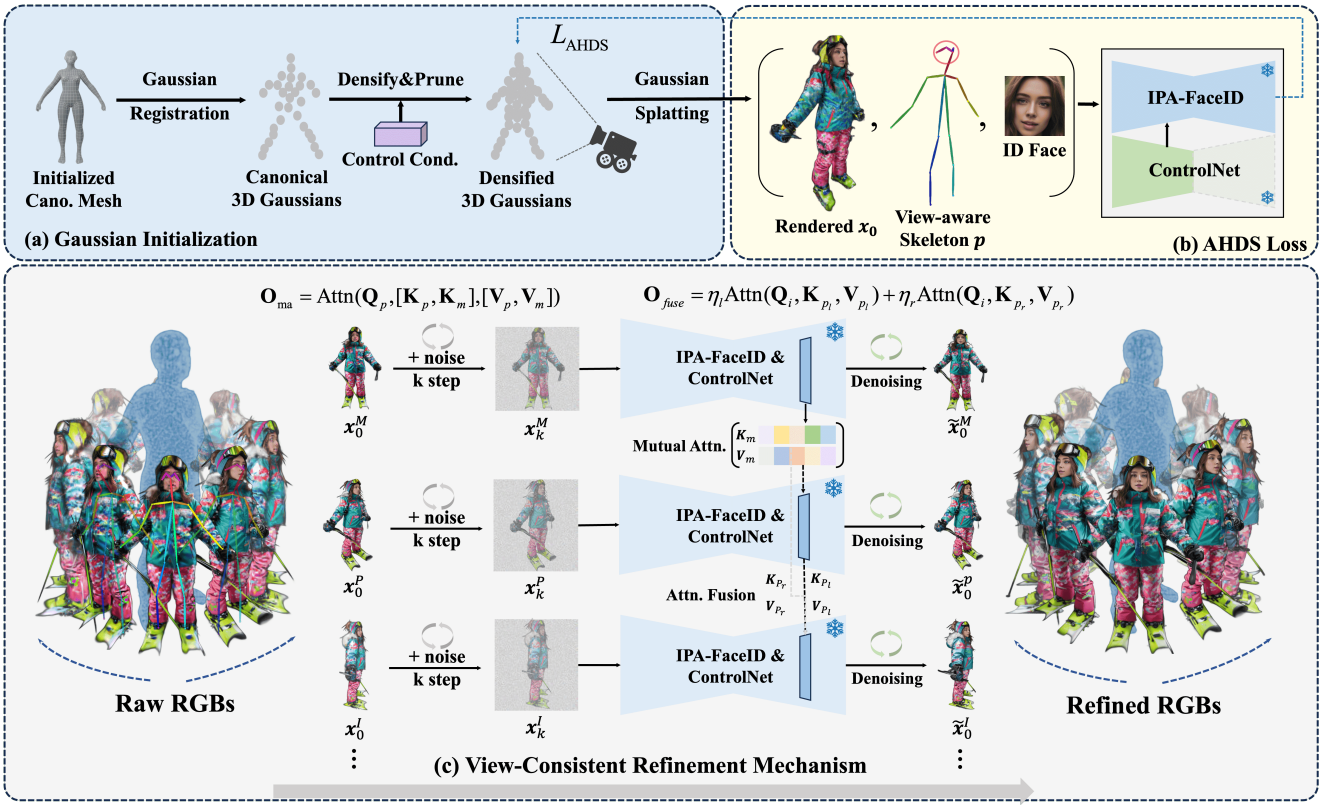


Figure 1. Overview of the GaussianIP framework. We combine 3D Gaussian Splatting (3DGS) with a human-centric diffusion prior to realize high-fidelity 3D human avatar generation. (a) We initialize 3D human Gaussians by densely sample from a SMPL-X mesh. Afterward, (b) a human-centric diffusion model is combined with a pose-guide ControlNet to produce AHDS guidance. The AHDS guidance consists of an HDS guidance, which is proposed to achieve better identity-preserving generation, and an Adaptive Human-specific Timestep Scheduling strategy, which accelerates the HDS training. Furthermore, we propose (c) a View-Consistent Refinement Mechanism to further enhance the delicate texture of faces and garments. We guide the denoising of key views $\boldsymbol{x}_0^P$ with attention features from main views $\boldsymbol{x}_0^M$ through Mutual Attention. Next, we align the denoising of an intermediate view $\boldsymbol{x}_0^I$ with that of its neighbor key views via distance-guided attention fusion. Finally, the refined multi-view images are leveraged to optimize the current 3DGS.

X mesh to represent 3D human bodies. AvatarStudio [42] and HumanNorm [15] achieve enhanced geometric details by utilizing DMTet [53] for 3D representation, combined with multi-stage or decoupled optimization strategies. More recently, HumanGaussian [34], GAvatar [68] and other works [8, 35] introduce 3DGS to create high-resolution realistic 3D avatars and real-time rendering. However, these methods fail to handle image prompts such as user portraits, which limits their real-world applicability.

3. Method

Problem statement. Given a text prompt describing the clothing and an user portrait providing facial identity, the model should swiftly generate a identity-preserving 3D human avatar with refined facial features and intricate clothing details. In the following sections, preliminaries will be present first. Next, we will introduce our Adaptive Human Distillation Sampling (AHDS) process. Following that, a simple yet effective View-Consistent Refinement mechanism (VCR) will be detailed.

3.1. Preliminaries

3D Gaussian Splatting [22] is an effective point-based representation consisting of a set of anisotropic Gaussians. Each 3D Gaussian is parameterized by its center position $\boldsymbol{\mu} \in \mathbb{R}^3$, covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^7$, opacity $\alpha \in \mathbb{R}$ and color $\boldsymbol{c} \in \mathbb{R}^3$. By splatting 3D Gaussians onto 2D image planes, we can perform point-based rendering:

$$
\begin{aligned}
G(\boldsymbol{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) &= \exp(-\frac{1}{2}(\boldsymbol{p}-\boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{p}-\boldsymbol{\mu}_i)), \\
\boldsymbol{c}(\boldsymbol{p}) &= \sum_{i \in \mathcal{N}} \boldsymbol{c}_i \alpha_i' \prod_{j=1}^{i-1}(1-\alpha_j'), \alpha_i' = \alpha_i G(\boldsymbol{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i).
\end{aligned} \tag{1}
$$

Here, $\boldsymbol{p}$ is the coordinate of the queried point. $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{c}_i$, α_i, and α_i' denote the center, covariance, color, opacity, and

density of the i-th Gaussian, respectively. $G(\boldsymbol{p}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ represents the value of the i-th Gaussian at position $\boldsymbol{p}$. $\mathcal{N}$ is a sorted list of Gaussians in this tile.

Score Distillation Sampling [46] leverages a powerful 2D text-to-image diffusion model, ϵ_ϕ, to guide the training of 3D scenes. By distilling gradients from the 2D diffusion model, SDS ensures the 3D scene's rendered images align with the input text prompt. Given a 3D representation parameterized by θ, an image $\boldsymbol{x} = g(\theta)$ can be rendered with a differentiable renderer g. SDS calculates the gradients with respect to the 3D representation θ by:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, \boldsymbol{x}) = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon}\right)\frac{\partial \boldsymbol{x}}{\partial \theta}\right], \quad (2)$$

where y is the text prompt, t is the sampling timestep in a 2D diffusion model, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is the sampled noise and $\boldsymbol{x}_t$ is the noised image. $w(t)$ is a weighting function depending on the timestep t.

3.2. Adaptive Human Distillation Sampling

3D human generation task has distinct characteristics compared to 3D object generation tasks, which typically involve initializing 3D Gaussians [22] with Shape-e [20] and optimizing the 3D object representation through distillation from 2D general-purpose diffusion models [13, 49]. For instance, the human parametric model [45] provides a reliable prior for human body geometry, which can facilitate the training process by starting with smaller diffusion timestep. By leveraging 2D human-centric diffusion models [60, 64, 65], we can enhance the visual quality of generated 3D avatars and ensure identity-preserving generation through effective distillation techniques. In addition, inspired by the coarse to fine nature of 2D human generation process, we can design a more effective timestep sampling strategy tailored for human-specific generation tasks.

Gaussian Initialization. A proper initialization method which can provide a rough geometry is crucial for the success of training. Traditional methods [22, 66] rely on tools like Structure-from-Motion (SfM) [52] and Shape-e [20], which may result in overly sparse points or inconsistent body structures. [34] first proposes to initialize 3D Gaussians on a neutral SMPL-X [45] mesh. We adopt a similar way to prepare for the subsequent training stages.

Distillation Sampling with Human-centric Prior. Since we focus on 3D human generation, distilling from a general diffusion prior like [34, 35, 68] may not be the most optimal approach. Instead, we combine a face-focused version of [65] D_{ip} with a pose-conditioned ControlNet [70] C_{pose} to form a human-specific diffusion prior D^p_{ip}. For each training step, we apply a view-dependent pose skeleton trimming strategy to ensure a pose condition map $\boldsymbol{p}$ which accurately represents the visibility of facial features (eyes, ears, etc.) from different views. For example, when the azimuth exceeds $60°$, the left ear keypoint will be masked. Then the left and right eye will be sequentially masked as the azimuth increases. For the back view, only the ears remain visible. The trimmed $\boldsymbol{p}$ contains pose clue across views, which is essential for mitigating the "Janus" problem.

To fully leverage the given text and image condition when training with SDS loss, we propose a novel human distillation sampling (HDS) guidance by decomposing and redesigning the original score difference. Inspired by [21], we can rewrite the common score difference [46] as:

$$\epsilon^s_\phi(\boldsymbol{x}_t; y, t) - \boldsymbol{\epsilon} = \delta_{\text{rect}} + \delta_{\text{noise}} + \gamma\delta_{\text{cond}} - \boldsymbol{\epsilon}, \quad (3)$$

where $\delta_{\text{rect}} + \delta_{\text{noise}} = \epsilon_\phi(\boldsymbol{x}_t; y_\phi, t)$ and $\delta_{\text{cond}} = \epsilon_\phi(\boldsymbol{x}_t; y, t) - \epsilon_\phi(\boldsymbol{x}_t; y_\phi, t)$ denote the unconditional and conditional terms, respectively, with γ as the classifier-free guidance (CFG) [12] coefficient and y_ϕ the null text. Here, δ_{rect} acts as a rectifier, guiding the rendered image $\boldsymbol{x} = g(\theta)$ towards a denser region within the manifold of real images, while δ_{noise} serves as a denoiser, pushing $\boldsymbol{x}$ towards a cleaner image. However, [61] indicates that the residual $\delta_{\text{noise}} - \boldsymbol{\epsilon}$ is noisy and may cause blurry textures, so we define the identity-preserving score difference $\delta_{ip} = \epsilon^s_{ip}(\boldsymbol{x}_t; y, \boldsymbol{I}_{ip}, t) - \boldsymbol{\epsilon}$ in our HDS as:

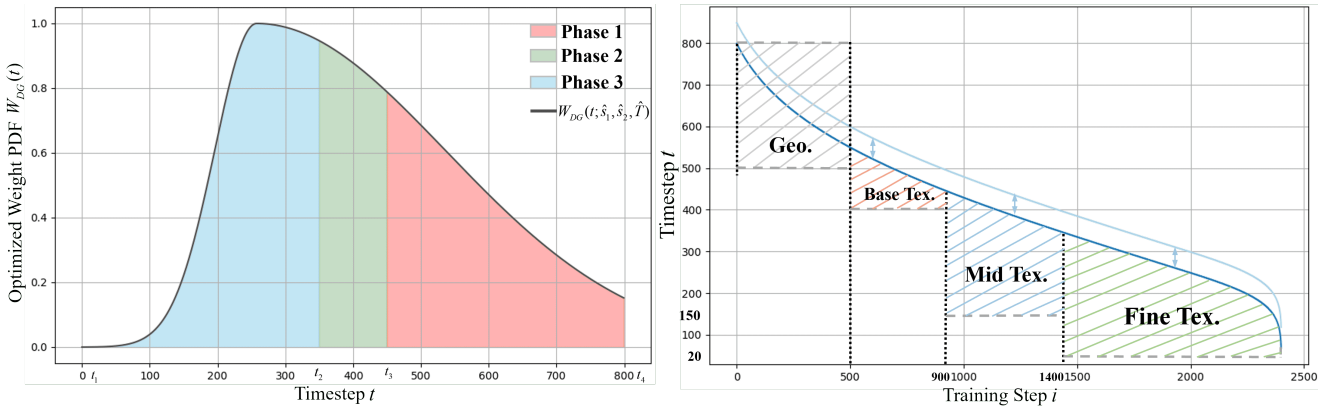
$$\begin{aligned}\delta_{ip} &= \delta^{ip}_{\text{rect}} + \gamma\delta^{ip}_{\text{cond}} \\ &= \delta^{ip}_{\text{rect}} + \gamma(\epsilon_{ip}(\boldsymbol{x}_t; y, \boldsymbol{I}_{ip}, t) - \epsilon_{ip}(\boldsymbol{x}_t; y_\phi, \boldsymbol{I}_\mu, t)),\end{aligned} \quad (4)$$

where ϵ_{ip} is the $\boldsymbol{\epsilon}$-prediction UNet used in the diffusion prior D^{pose}_{ip} and $\boldsymbol{I}_\mu$ can be a face irrelevant to $\boldsymbol{I}_{ip}$. As regards the rectifying direction δ_{rect}, it can be estimated as $\epsilon_\phi(\boldsymbol{x}_t; y_\phi, t)$ when the timestep t is rather small, given the low noise level at these timesteps making the denoising direction negligible. We then use a repelling score to model δ_{rect} at other diffusion timesteps, so we get:

$$\delta^{ip}_{\text{rect}} = \begin{cases} \epsilon_{ip}(\boldsymbol{x}_t; y_\phi, \boldsymbol{I}_\mu, t) & t < \tau \\ \epsilon_{ip}(\boldsymbol{x}_t; y_\phi, \boldsymbol{I}_\mu, t) - \epsilon_{ip}(\boldsymbol{x}_t; y_-, \boldsymbol{I}_\phi, t) & t \geq \tau, \end{cases} \quad (5)$$

where y_- represents the negative prompt, $\boldsymbol{I}_\phi$ denotes an all-zero image, and τ is the timestep threshold. With our redesigned score difference, the generated avatar is more realistic without the issue of over-saturation and exhibits a strong alignment with the given identity image prompts.

Adaptive Human-specific Timestep Scheduling. Despite an identity-preserved 3D avatar with decent appearance can be generated with HDS, it still requires a long training time. Aims to achieve faster HDS optimization within fewer training steps, we propose an adaptive diffusion timestep scheduling strategy motivated by [16, 25], which results in a non-increasing timestep against training step $(t - i)$ curve tailored for 3D human generation tasks. Inspired by the denoising process in 2D human generation, we may naturally divide the entire 3D HDS process into three synergistic phases: geometry and base textures (phase 1), middle-level textures (phase 2) and fine facial and garment details



(a) Optimized Weight PDF for Sampling Timesteps.

(b) Timestep vs. Training Step (t-i) Curve

Figure 2. Illustration of the optimized weight PDF for sampling timesteps and the corresponding timestep vs. training step (t-i) curve. a) Phase 1, 3 occupy the majority of the training steps, while Phase 2 occupies only a small portion, allowing a quick transition to the detailed texture learning in Phase 3. b) We sample the final timestep between the lower bound and t_{DG} for each phase. Note that for the geometry phase ($i < 500$), we sample between 500 and the maximum timestep to ensure a smooth start.

(phase 3). Each phase is assigned with a timestep range, $\text{range}_i = [t_i, t_{i+1}], i = 1, 2, 3$. Intuitively, more training steps should be assigned to phase 1, 3 to model human geometry and intricate details. Phase 2 can be seen as a transitional stage, requiring slightly fewer steps. To derive a $t - i$ curve that accurately exhibits these trends, we start by optimizing a weighting PDF function of timesteps t then map it to the final $t-i$ curve. Empirically, a dual-piecewise Gaussian function $W_{DG}(t; s_1, s_2, T)$ is employed to represent the probability distribution:

$$W_{\text{DG}}(t) = \begin{cases} \frac{1}{\sqrt{2\pi s_1^2}} \exp\left(-\frac{(t-T)^2}{2s_1^2}\right) & t \leq T \\ \frac{1}{\sqrt{2\pi s_2^2}} \exp\left(-\frac{(t-T)^2}{2s_2^2}\right) & t > T. \end{cases} \tag{6}$$

The long-tail property may cause a swift timestep decrease at phase 1 to avoid over-large t (SMPL-X provides a rough geometry so the training can start with smaller t) and suppress over-small t at phase 3, which introduces high gradient variance. Given the training steps assigned to the three phases $n_i, i = 1, 2, 3$ and the total step N, We can estimate the parameters of W_{DG} by solving an optimization problem:

$$\hat{s_1}, \hat{s_2}, \hat{T} = \underset{s_1, s_2, T}{\text{argmin}} \sum_{k=1}^{3} \left(\sum_{t \in \text{range}_k} W_{\text{DG}}(t) - \frac{n_k}{N} \right)^2, \tag{7}$$

with the results presented in Fig. 2 (a). This objective ensures that the accumulative probability of each range aligns with the desired training step proportion of each phase, facilitating the mapping between W_{DG} and the $t - i$ curve. Finally, the corresponding timestep for each training step i can be solved by:

$$t_{\text{DG}}(i) = \underset{\tau}{\text{argmin}} \left| \sum_{t=\tau}^{T} W_{DG}(t) - \frac{i}{N} \right| + \epsilon, \tag{8}$$

where ϵ is a small tuning factor. To prevent texture oversaturation and ensure smooth transitions between stages, we set a lower bound for each phase and sample the final timestep within the range between this lower bound and t_{DG} (shown in Fig. 2 (b)). This adaptive timestep scheduling function can be seamlessly integrated into our HDS process to form the AHDS guidance, ensuring a faster and smoother training process that produces detail-enhanced results.

3.3. View-consistent Refinement Mechanism

With the assistance of the AHDS guidance in stage 1, 3D avatars with rich garment details and better identity consistency can be generated efficiently. However, the distillation process itself may inevitably cause slight texture smoothing. To further enhance detail articulation based on the AHDS training results, a straightforward approach is to add noise to the rendered multi-view images, denoise them, and then optimize the 3DGS using the denoised images. The independent denoising process, however, may lead to texture inconsistencies across multi-view images, as there is no exchange of view information during processing. To address this issue, we design an elegant refinement strategy, which is shown in Fig. 1 (c). The entire refinement process consists of two sub-process, key view refinement and intermediate features propagation. We first guide the denoising of k key views with the attention features of main views (front, back, left, right), ensuring a roughly consistent appearance among key and main views, then we propagate the refinement effects to other views via weighted attention fusion.

Key Views Refinement. For a specific key view P, we inject the attention keys $\boldsymbol{K}_m$ and values $\boldsymbol{V}_m$ of its nearest main view M into the denoising process. Directly applying this injection may cause texture drift, as some unwanted features (from invisible areas) may be queried. To mitigate this, we inflate the self-attention keys and values for the main view, so the keys and values from the two views act as a mutual reference. Denote the attention operation as $\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}$, where d is the model dimension, we can denote the main view guided mutual attention as:

$$\boldsymbol{O}_{ma} = \text{Attn}(\boldsymbol{Q}_p, [\boldsymbol{K}_p, \boldsymbol{K}_m], [\boldsymbol{V}_p, \boldsymbol{V}_m]), \tag{9}$$

where $\boldsymbol{K}_p$, $\boldsymbol{V}_p$ denote the keys and values of view P, respectively, and $[\cdot, \cdot]$ is the concatenate operation. This mutual reference attention ensures that all the key views share an aligned appearance with the main views, laying a solid foundation for the following process.

Intermediate Features Propagation. Finally, to obtain a smooth transition of refinement effects between two adjacent key views (we may regard main views as key views here), we "propagate" the attention features of two key adjacent views to their intermediate views. Given such a view I, we first query its nearest neighbor key views P_l and P_r and calculate the relative distance based on their azimuths, that is: $\text{Dist}(I, P_l) = \frac{|\phi_I - \phi_{P_l}|}{|\phi_{P_r} - \phi_{P_l}|}$, where ϕ_V is the azimuth of view V. We can naturally regard $\eta_l = 1 - Dist(I, P_l)$ as the influence of P_l on I. To properly fuse the attention features $\boldsymbol{K}_{p_l}, \boldsymbol{V}_{p_l}$ of P_l and $\boldsymbol{K}_{p_r}, \boldsymbol{V}_{p_r}$ of P_r into the denoising process of I, we introduce a novel weighted attention fusion strategy guided by relative distances. Specifically, the attention output is combined with two parts, pure self-attention $\boldsymbol{O}_{\text{sa}} = \text{Attn}(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i)$ and fused attention, which is obtained by $\boldsymbol{O}_{\text{fa}} = \eta_l \text{Attn}(\boldsymbol{Q}_i, \boldsymbol{K}_{P_l}, \boldsymbol{V}_{P_l}) + \eta_r \text{Attn}(\boldsymbol{Q}_i, \boldsymbol{K}_{P_r}, \boldsymbol{V}_{P_r})$. The output of the complete attention operation can then be derived as:

$$\boldsymbol{O}_{\text{final}} = \lambda_{\text{self}} \boldsymbol{O}_{\text{sa}} + (1 - \lambda_{\text{self}}) \boldsymbol{O}_{\text{fa}}, \tag{10}$$

where $\lambda_{\text{self}} \in [0, 1]$ is a weight factor. Through this attention fusion process, the refined image of I will have high texture consistency with the results of its neighboring views.

3D Human Gaussian Optimization. With the refined multi-view images which are aligned with each other in textures and semantics, we can optimize the 3D Human Gaussians from stage 1 by directly applying reconstruction losses. At each training step, we randomly select b views from those used in the refinement stage and compute the reconstruction loss between the rendered images $\hat{\boldsymbol{X}}$ and the ground truth $\boldsymbol{X}_{\text{gt}}$:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{L1}} L_1(\hat{\boldsymbol{X}}, \boldsymbol{X}_{\text{gt}}) + \lambda_{\text{lpips}} L_{\text{lpips}}(\hat{\boldsymbol{X}}, \boldsymbol{X}_{\text{gt}}), \tag{11}$$

where λ_{L1} and λ_{lpips} are weight factors. To improve efficiency, we crop each rendered image to the rectangular region of the human body, downsample it by half, and then calculate the reconstruction loss against similarly processed ground truth images.

4. Experiments

4.1. Implementation Details

AHDS Guidance and Stage 1 Training Setups. We adopt IP-Adapter-FaceID-PlusV2 [65] as our human-centric diffusion prior. When solving the time scheduling $t - i$ curve, we set $t_1 = 20$, $t_2 = 350$, $t_3 = 450$ and $t_4 = 800$, respectively. The training steps assigned to the three phases are $n_1 = 900$, $n_2 = 500$ and $n_3 = 1000$, respectively, and the total step of phase 1 training is 2400. During the training process, we set the classifier-free guidance coefficient to $\gamma = 7.5$ and the time threshold in Eq. 6 to $\tau = 170$. The lower bounds for the three stages are 400, 150 and 20, respectively. The densification & pruning operation is conducted from 200 to 1700 steps at an interval of 800 steps while the prune-only operation is applied once at step 1800. Stage 1 is trained with Adam optimizer, with the learning rates of each Gaussian attribute is scheduled following [34].

Refinement and Stage 2 Training Setups. The total denoising step is set to 8 during the refinement process and the number of key views is 4. When conducting intermediate view refinement, the weighted attention fusion factor λ_{self} is set to 0.55. We use the refined images to optimize the 3D human Gaussians for 800 steps with a batch size $b = 8$. The weight factors λ_{L1} and λ_{lpips} are set to 10 and 15, respectively. All experiments are conducted on a single NVIDIA V100 GPU.

Baselines. For fair comparison, we compare with recent SOTA text-to-3D human works instead of general text-to-3D methods, which include DreamWaltz [17], TADA [32], AvatarVerse [69], X-Oscar [39] and HumanGaussian [34].

Criteria. We conduct a user study to evaluate the 3D human humans generated by different methods. We randomly selected 20 prompts to generate 3D human models using each of the comparison methods. 24 participants were asked to evaluate the generated models based on 4 criteria: Facial Detail, Clothing Texture Richness, Overall Visual Quality, and Text Prompt Alignment. We also have ChatGPT-4o rate the images comprehensively on these aspects. Each criterion was rated on a scale from 1 to 5 (higher is better). Additionally, we compare the training time of each method to assess training efficiency.

4.2. Text-guided 3D Human Generation

Qualitative Results. Our method demonstrates clear advantages over [2, 24, 32, 34] in two key aspects. As shown in Fig. 3, it excels in facial detail, capturing sharper and

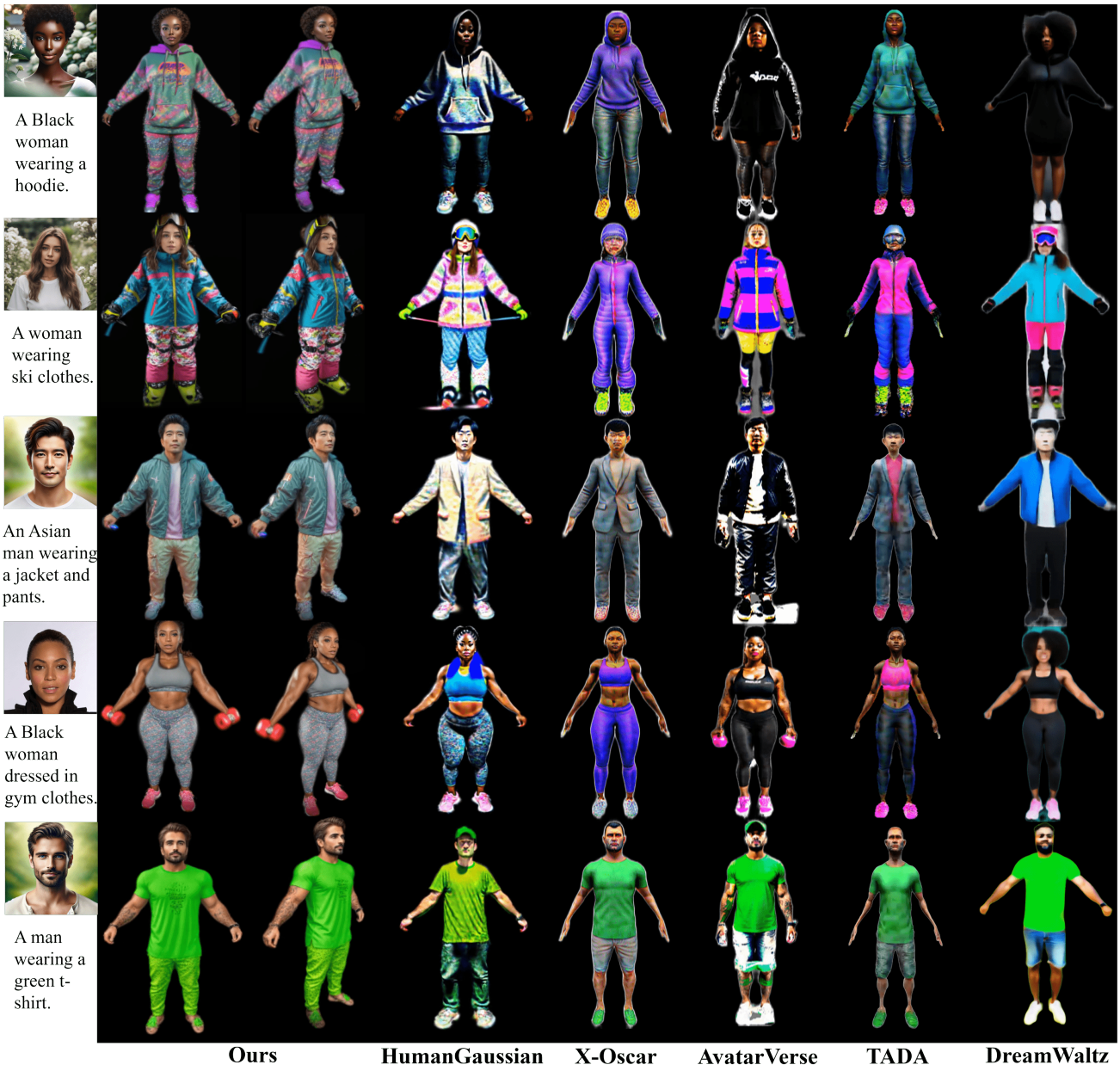


Figure 3. Qualitative comparison results with SOTA text-guided 3D human generation models. Please zoom in for better observation. Note that the baselines cannot handle image prompts, so we compare their text-to-3D results instead. Due to space limitations, please refer to the supplementary materials for the video comparison results.

more realistic facial features. For clothing texture richness, our method preserves intricate patterns and fabric details that others often miss. Additionally, our method generates results with high identity consistency to the given user portraits, which broadens its potential application scenarios.

Quantitative Results. As shown in Table 1, our method achieves consistently higher scores across all five criteria, demonstrating superior fidelity and alignment with prompts. Additionally, We utilize an off-the-shelf commercial face verification algorithm (Face++) to evaluate the performance of our method with regard to identity preservation. All faces of the generated humans successfully match the corresponding user portraits with an average confidence level exceeding 83%. Regarding the training time, all baseline methods necessitate over one hour to complete their training tasks, while our method manages to accomplish the same tasks in approximately 40 minutes, thereby demonstrating its superior training efficiency.

Methods	Fac. Det.	Cloth. Tex.	Vis. Qual.	Text Align.	GPT Score	Training Time	ID Pres.
DreamWaltz [17]	1.33	1.46	1.38	1.58	1.82	1.3h	✗
TADA [32]	2.21	2.46	2.58	3.13	3.24	2h	✗
AvatarVerse [69]	3.67	3.29	2.63	2.96	3.05	1h	✗
X-Oscar [39]	3.29	3.83	3.58	3.63	3.64	3h	✗
HumanGaussian [34]	4.29	4.17	3.96	4.42	4.08	1.2h	✗
Ours	**4.71**	**4.50**	**4.17**	**4.62**	**4.52**	**40min**	✓

Table 1. **Quantitative comparison results.** We conduct evaluations on generation quality from five aspects: (1) Facial Detail; (2) Clothing Texture Richness; (3) Overall Visual Quality; (4) Text Prompt Alignment; (5) ChatGPT Score. We also compare the training time and identity preserving ability here.

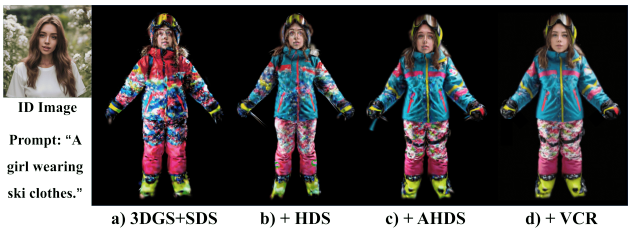


Figure 4. Ablation studies on various module designs. We present the generation results of the human frontal view under four ablation settings: (a) baseline; (b) + HDS; (c) + AHDS; (d) + View-consistent Refinement Mechanism. Detailed ablation settings and result analysis are depicted in Sec. 4.3.

Figure 5. Ablation study on our VCR module. When the multi-view images are denoised independently, the results will loss cross-view texture consistency. In contrast, images refined with our VCR module maintain high 3D texture consistency.

4.3. Ablation Studies

Ablations for AHDS and Refinement Mechanism. We evaluate the visual results of **a)** our baseline approach (3DGS+SDS) and progressively add **b)** HDS, **c)** AHDS (HDS with adaptive timestep scheduling), and **d)** the View-Consistent Refinement (VCR) mechanism. The results in Fig. 4 show that the avatar generated with the baseline forms basic shapes but tends to produce an over-saturated appearance and lacks fine details. Incorporating HDS leads to improvements in identity preservation and clothing details. With the addition of timestep scheduling in AHDS, the training time required for HDS is significantly reduced (with training steps reduced from 3600 to 2400) while achieving enhanced generative quality. The generated results display more intricate textures and detailed facial features. Finally, the refinement mechanism further improves multi-view consistency, achieving high-fidelity, coherent textures across views.

Ablation for View-consistent Refinement. To assess the importance of multi-view consistency, we perform an ablation on the refinement mechanism. As illustrated in Fig. 5, when the image of each view undergoes independent denoising process, the result lacks some texture consistency across various views, showing misaligned and blurred details when we change the camera views. In contrast, applying our view-consistent refinement ensures texture alignment across all views, maintaining consistent and realistic details, especially in critical facial and garment features.

5. Conclusion

In this paper, we present GaussianIP, a novel two-stage framework for generating realistic 3D human avatars from text and image prompts while preserving critical facial identity features. We propose AHDS guidance to facilitate the learning of identity attributes and accelerate the generation process. Additionally, we introduce a VCR mechanism to enhance the texture fineness of 3D human avatars without sacrificing 3D multi-view consistency. Extensive experiments validate the superiority of GaussianIP, highlighting its advantages in both training efficiency and visual fidelity.

Limitations and future work. Our method faces challenges in rendering highly complex poses or extreme garment textures. Further research is required to generalize the framework to broader applications, such as human-object and human-human interactions. We aim to investigate these aspects in future work, enabling more immersive and richer interactive experiences in VR/AR environments.

Acknowledgment

This work is partly supported by the National Key R&D Program of China (2022ZD0161902), and the National Natural Science Foundation of China (No. 62202031). We also give specical thanks to Alibaba Group for their contribution to this paper.

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. 1

[2] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 1, 2, 6

[3] 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. 1

[4] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 2

[5] Zixuan Chen, Ruijie Su, Jiahao Zhu, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Vividdreamer: Towards high-fidelity and efficient text-to-3d generation. *arXiv preprint arXiv:2406.14964*, 2024. 2

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[8] Jia Gong, Shenyu Ji, Lin Geng Foo, Kang Chen, Hossein Rahmani, and Jun Liu. Laga: Layered 3d avatar generation and customization via gaussian splatting. *arXiv preprint arXiv:2405.12663*, 2024. 3

[9] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. Taming the power of diffusion models for high-quality virtual try-on with appearance flow. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7599–7607, 2023. 2

[10] 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. 1

[11] 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*, pages 3470–3479, 2022. 1

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 4

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2, 4

[14] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2

[15] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. 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*, pages 4568–4577, 2024. 3

[16] Yukun Huang, Jianan Wang, Yukai Shi, Boshi Tang, Xianbiao Qi, and Lei Zhang. Dreamtime: An improved optimization strategy for diffusion-guided 3d generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 4

[17] 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. 1, 2, 6, 8

[18] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876, 2022. 2

[19] Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Avatarcraft: Transforming text into neural human avatars with parameterized shape and pose control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14371–14382, 2023. 2

[20] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 4

[21] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 2, 4

[22] 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. 1, 2, 3, 4

[23] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8176–8185, 2024. 1, 2

[24] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 6

[25] Kyungmin Lee, Kihyuk Sohn, and Jinwoo Shin. Dreamflow: High-quality text-to-3d generation by approximating probability flow. *arXiv preprint arXiv:2403.14966*, 2024. 4

[26] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pages 204–219. Springer, 2022. 1

[27] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, New York, NY, USA, 2020. Association for Computing Machinery. 1

[28] Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. Monocular real-time volumetric performance capture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 49–67. Springer, 2020. 1

[29] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024. 1, 2

[30] Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, and Yi Yang. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6400–6409, 2024. 2

[31] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6517–6526, 2024. 2

[32] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. In *2024 International Conference on 3D Vision (3DV)*, pages 1508–1519. IEEE, 2024. 2, 6, 8

[33] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 2

[34] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6646–6657, 2024. 1, 3, 4, 6, 8

[35] Yufei Liu, Junshu Tang, Chu Zheng, Shijie Zhang, Jinkun Hao, Junwei Zhu, and Dongjin Huang. Clothedreamer: Text-guided garment generation with 3d gaussians. *arXiv preprint arXiv:2406.16815*, 2024. 3, 4

[36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 1, 2

[37] Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Campagnolo Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin Solomon. Score distillation via reparametrized ddim. *arXiv preprint arXiv:2405.15891*, 2024. 2

[38] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[39] Yiwei Ma, Zhekai Lin, Jiayi Ji, Yijun Fan, Xiaoshuai Sun, and Rongrong Ji. X-oscar: A progressive framework for high-quality text-guided 3d animatable avatar generation. *arXiv preprint arXiv:2405.00954*, 2024. 2, 6, 8

[40] Zhiyuan Ma, Yuxiang Wei, Yabin Zhang, Xiangyu Zhu, Zhen Lei, and Lei Zhang. Scaledreamer: Scalable text-to-3d synthesis with asynchronous score distillation. In *European Conference on Computer Vision*, pages 1–19. Springer, 2025. 2

[41] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *arXiv preprint arXiv:2402.08682*, 2024. 2

[42] Mohit Mendiratta, Xingang Pan, Mohamed Elgharib, Kartik Teotia, Ayush Tewari, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Avatarstudio: Text-driven editing of 3d dynamic human head avatars. *ACM Transactions on Graphics (ToG)*, 42(6):1–18, 2023. 3

[43] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[44] Nasir Mohammad Khalid, Tianhao Xie, Eugene Belilovsky, and Tiberiu Popa. Clip-mesh: Generating textured meshes from text using pretrained image-text models. In *SIGGRAPH Asia 2022 conference papers*, pages 1–8, 2022. 2

[45] 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. 4

[46] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1, 2, 4

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[48] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, et al. Dreambooth3d: Subject-driven text-to-3d generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2349–2359, 2023. 2

[49] 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, 2, 4

[50] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 84–93, 2020. 1

[51] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 2

[52] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 4

[53] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 3

[54] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8543–8552, 2024. 2

[55] Ke Sun, Jian Cao, Qi Wang, Linrui Tian, Xindi Zhang, Lian Zhuo, Bang Zhang, Liefeng Bo, Wenbo Zhou, Weiming Zhang, et al. Outfitanyone: Ultra-high quality virtual try-on for any clothing and any person. *arXiv preprint arXiv:2407.16224*, 2024. 1, 2

[56] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 2

[57] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[58] 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. 1

[59] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[60] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2, 4

[61] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[62] 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. 1

[63] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 2

[64] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 4

[65] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 2, 4, 6

[66] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6796–6807, 2024. 2, 4

[67] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023. 2

[68] Ye Yuan, Xueting Li, Yangyi Huang, Shalini De Mello, Koki Nagano, Jan Kautz, and Umar Iqbal. Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 896–905, 2024. 1, 3, 4

[69] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Daniel Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7124–7132, 2024. 1, 2, 6, 8

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4

[71] 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. 1

[72] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18697–18709, 2022. 2

[73] 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. 1, 2