HumanGaussian: Text-Driven 3D Human Generation with Gaussian Splatting

Xian Liu[1], Xiaohang Zhan[2], Jiaxiang Tang[3], Ying Shan[2], Gang Zeng[3], Dahua Lin[1], Xihui Liu[4], Ziwei Liu[5]

[1]CUHK   [2]Tencent AI Lab   [3]PKU   [4]HKU   [5]NTU

Project Page: https://alvinliu0.github.io/projects/HumanGaussian

Figure 1. We propose **HumanGaussian**, an efficient yet effective framework that generates high-quality 3D humans with fine-grained geometry and realistic appearance. Our method adapts 3D Gaussian Splatting into text-driven 3D human generation with novel designs.

Abstract

*Realistic 3D human generation from text prompts is a desirable yet challenging task. Existing methods optimize 3D representations like mesh or neural fields via score distillation sampling (SDS), which suffers from inadequate fine details or excessive training time. In this paper, we propose an efficient yet effective framework, **HumanGaussian**, that generates high-quality 3D humans with fine-grained geometry and realistic appearance. Our key insight is that 3D Gaussian Splatting is an efficient renderer with periodic Gaussian shrinkage or growing, where such adaptive density control can be naturally guided by intrinsic human structures. Specifically, 1) we first propose a Structure-Aware SDS that simultaneously optimizes human appearance and geometry. The multi-modal score function from both RGB and depth space is leveraged to distill the Gaussian densification and pruning process. 2) Moreover, we devise an Annealed Negative Prompt Guidance by decomposing SDS into a noisier generative score and a cleaner classifier score, which well addresses the over-saturation issue. The floating artifacts are further eliminated based on Gaussian size in a prune-only phase to enhance generation smoothness. Extensive experiments demonstrate the superior efficiency and competitive quality of our framework, rendering vivid 3D humans under diverse scenarios.*

1. Introduction

Creating high-quality 3D humans from user condition is of great importance to a wide variety of applications, ranging from virtual try-on [27, 66, 67, 75] to immersive telepresence [37, 64]. To this end, many researchers explore the task of text-driven 3D human generation, which synthesizes the character's appearance and geometry based on text prompts. Traditional methods resort to a hand-crafted pipeline, where 3D models are first regressed from multi-view human captures, and then undergo a series of manual processes like rigging, skinning, texturing, and animating [3, 26, 33, 36]. To ease human labor for 3D asset creation of diverse layouts, the exemplar work DreamFusion [56] proposes score distillation sampling (SDS) to harness rich 2D text-to-image prior (*e.g.*, Stable Diffusion [62], Imagen [63]) by optimizing 3D scenes to render samples that reside on the manifold of higher likelihood. Though accomplishing reasonable results on single objects [6, 47, 61, 80], it is hard for them to model detailed human bodies with complex articulations.

To incorporate structural guidance, recent text-driven 3D human studies combine SDS with body shape models such as SMPL [43] and imGHUM [1]. In particular, a common paradigm is to integrate human priors into representations like mesh and neural radiance field (NeRF), either by taking the body shape as mesh/density initialization [20, 34, 90],

or by learning a deformation field based on linear blend skinning (LBS) [4, 82, 89]. However, they mostly compromise to trade-off between efficiency and quality: the mesh-based methods [25, 39, 87] struggle to model fine topologies like accessories and wrinkles; while the NeRF-based methods [22, 23, 91] are time/memory-consuming to render high-resolution results. How to achieve fine-grained generation efficiently remains an unsolved problem.

Recently, the explicit neural representation of 3D Gaussian Splatting (3DGS) [31] provides a new perspective for real-time scene reconstruction. It enables multi-scale modeling across multiple granularities, which is suitable for 3D human generation. Nevertheless, it is non-trivial to exploit such representation in this task with two challenges: **1)** 3DGS characterizes a tile-based rasterization by sorting and α-blending anisotropic splats within each view frustum, which only back-propagates a small set of high-confidence Gaussians. However, as verified in the 3D surface-/volume-rendering studies [17, 29, 49, 73], sparse gradient could hinder network optimization of geometry and appearance. Therefore, structural guidance is required in 3DGS, especially for the human domain that demands hierarchical structure modeling and generation controllability. **2)** The naive SDS necessitates a large classifier-free guidance (CFG) [18] scale for image-text alignment (*e.g.*, 100 as used in [56]). But it sacrifices visual quality with over-saturated patterns, making realistic human generation difficult. Besides, due to the stochasticity of SDS loss, the original gradient-based density control in 3DGS is unstable, which incurs blurry results with floating artifacts.

In this paper, we propose an efficient yet effective framework, **HumanGaussian**, that generates high-quality 3D humans with fine-grained geometry and realistic appearance. Our intuition lies in that 3D Gaussian Splatting is an efficient renderer with periodic Gaussian shrinkage or growing, where *such adaptive density control can be naturally guided by intrinsic human structures*. The key is to incorporate explicit structural guidance and gradient regularization to facilitate Gaussian optimization. Specifically, we first propose a *Structure-Aware SDS* that jointly learns human appearance and geometry. Unlike previous studies [7, 31, 86] that adopt the generic priors like Structure-from-Motion (SfM) points or Point-E [51], we instead anchor the Gaussian initial positions on the SMPL-X mesh. The subsequent densification and pruning processes thus focus on regions around the body surface, effectively capturing geometric deformations like accessories and wrinkles. Additionally, we extend the pre-trained Stable Diffusion [62] to simultaneously denoise the image RGB and depth as our SDS source model. Such dual-branch design distills the joint distribution of two spatially-aligned targets (*i.e.*, RGB and depth), which boosts the Gaussian convergence with both structural guidance and textural realism. To further improve

renderings with natural appearance, we devise an *Annealed Negative Prompt Guidance*. In particular, we decompose SDS into a noisier generative score and a cleaner classifier score, where equipping the latter term with a decreasing negative prompt guidance enables realistic generation under nominal CFG scales (*e.g.*, 7.5), as also proven in concurrent text-to-3D studies [30, 88]. In this way, we manage to avoid over-saturated patterns with appropriate CFG scales that well balance sample quality and diversity. Moreover, due to the high variance of SDS loss, directly relying on gradient information to control densities results in blurry geometry [31]. In contrast, we propose to eliminate floating artifacts based on Gaussian size in a prune-only phase.

To summarize, our main contributions are three-fold: **1)** We propose an efficient yet effective framework **HumanGaussian** for high-quality 3D human generation with fine-grained geometry and realistic appearance. As one of the earliest attempts in taming Gaussian Splatting for text-driven 3D human domain, we hope to pave the way for future research. **2)** We propose the *Structure-Aware SDS* to jointly learn human appearance and geometry with explicit structural guidance. **3)** We devise the *Annealed Negative Prompt Guidance* to guarantee realistic results and eliminate floating artifacts. Extensive experiments demonstrate the superior efficiency and competitive quality of our framework, rendering vivid 3D humans under diverse scenarios.

2. Related Work

3D Neural Representations. Diverse 3D scene representations are proposed for spatial geometry and texture modeling, such as voxel, point cloud, mesh, and neural field. With the trade-off among training time, memory efficiency, rendering capability, and network compatibility, different representations are chosen based on problem setting: **1)** Voxel, a Euclidean representation that stores scene information in a grid manner [9, 46, 84], can be easily adapted for CNNs, but is limited in render resolution due to the cubic computational cost. **2)** Point cloud, a discrete point set sampled from 3D surface, is efficient to render [38, 57, 58]. However, it fails to capture the fine-grained details due to its discontinuous nature. **3)** Mesh, a compact representation expressing the connectivity among vertices, edges, and faces, inherits time efficiency from the well-rounded graphic pipelines [16, 78, 81], but struggles to create accurate topology. **4)** Neural field, an implicit function of each 3D position's attributes, is capable of modeling complex structures in arbitrary resolution [41, 49, 53, 83, 85], yet the optimization and inference are slow. Recently, 3D Gaussian Splatting (3DGS) [31, 45] has shown impressive results in the 3D reconstruction, surpassing previous representations with better quality and faster convergence. In this work, we try to unlock the potential of 3D Gaussian splatting on the challenging task of text-driven 3D human generation.

Text-to-3D Generation. Recent diffusion-based text-to-3D works can be grouped into two types: **1)** 3D native pipelines, which directly capture the distribution of 3D data [28, 44, 51] or the reconstructed intermediate features [5, 8, 15, 52] on specific domains. Although some recent works [2, 21] extend the model's capacity by training on large-scale 3D datasets like Objaverse [10], they are still confined to single objects. **2)** Optimization-based 2D lifting pipelines, which optimize 3D scene representations in a per-prompt manner by distilling from rich prior learned in the 2D domain. For example, some early attempts use CLIP guidance [59] to boost multi-view image-text alignment [24, 48, 50, 76], while recent methods resort to score distillation sampling (SDS) to inherit unprecedented rendering quality from exemplar text-to-image models [6, 47, 56, 77, 80]. Notably, the heavy computation burden of NeRF incurs long training time, which motivates concurrent works to adapt the representation of Gaussian splatting for text-to-3D generation [7, 74, 86]. In this work, we choose 3D Gaussian due to its efficiency and efficacy, but focus on the text-driven 3D human domain that demands both fine detail capturing and realistic texture generation.

Text-Driven 3D Human Generation. By incorporating human prior like SMPL [43] and imGHUM [1] models, the 3D text-to-human literature typically adapts general text-to-3D techniques with domain-specific designs [4, 12, 19, 25, 40, 90, 92, 94]. For example, AvatarCLIP [20] integrates NeuS [79] with SMPL prior and utilizes CLIP for guidance. DreamHuman [34] utilizes a pose-conditioned NeRF based on imGHUM to learn albedo and density fields with SDS. AvatarVerse [91] fine-tunes the ControlNet [93] branch with DensePose [13] as SDS source for view-consistent generation. TADA [39] deforms the SMPL-X [55] shape with displacement and optimizes texture UV-map by hierarchical rendering. Concurrent to our work, HumanNorm [22] also adds explicit structural constraint with the fine-tuned text-to-depth/normal models. However, they characterize a two-stage pipeline with DMTet [69] representation, which fails to capture the fine-grained human details efficiently.

3. Our Approach

We present **HumanGaussian** that generates high-quality 3D humans with fine-grained geometry and realistic appearance. The overall framework is illustrated in Fig. 2. To make the content self-contained and narration clearer, we first introduce some pre-requisites in Sec. 3.1. Then, we present the *Structure-Aware SDS* which jointly learns human appearance and geometry with explicit structural guidance. The multi-modal score function from both RGB and depth space is leveraged to distill the Gaussian densification and pruning process (Sec. 3.2). Finally, we elaborate the *Annealed Negative Prompt Guidance* to guarantee realistic results and eliminate floating artifacts in Sec. 3.3.
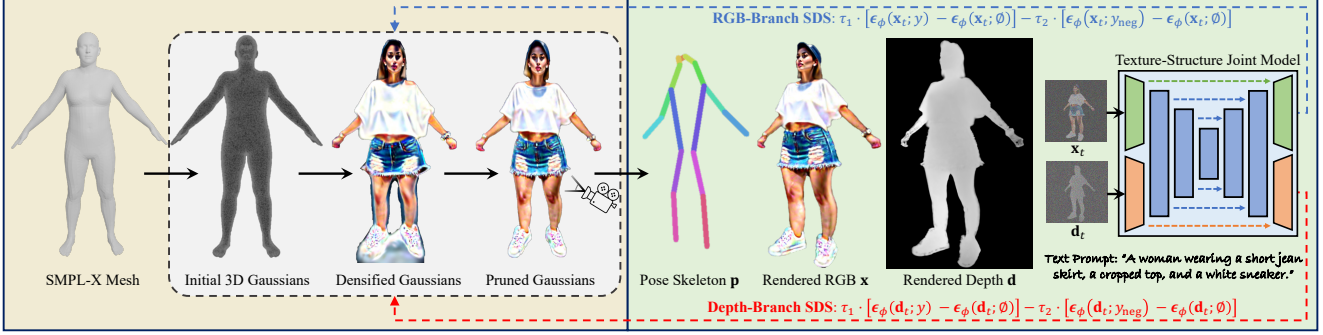
Figure 2. **Overview of the proposed HumanGaussian Framework.** We generate high-quality 3D humans from text prompts with the neural representation of 3D Gaussian Splatting (3DGS). In *Structure-Aware SDS*, we start from the SMPL-X prior to densely sample Gaussians on the human mesh surface as initial center positions. Then, a Texture-Structure Joint Model is trained to simultaneously denoise the image **x** and depth **d** conditioned on pose skeleton **p**. Based on this, we design a dual-branch SDS to jointly optimize human appearance and geometry, where the 3DGS density is adaptively controlled by distilling from both the RGB and depth space. In *Annealed Negative Prompt Guidance*, we use the cleaner classifier score with an annealed negative score to regularize the stochastic SDS gradient of high variance. The floating artifacts are further eliminated based on Gaussian size in a prune-only phase to enhance generation smoothness.

3.1. Preliminaries

SMPL-X [55] is a 3D parametric human model that defines the shape topology of body, hands, and face. It contains $10,475$ vertices and 54 keypoints. By assembling pose parameters θ (consists of body pose θ_b, jaw pose θ_f, and finger pose θ_h), shape parameters β, and expression parameters ψ, we can represent 3D SMPL-X human model $M(\beta, \theta, \psi)$ as:

$$T(\beta, \theta, \psi) = \bar{T} + B_s(\beta) + B_p(\theta) + B_e(\psi),$$
$$M(\beta, \theta, \psi) = \text{LBS}(T(\beta, \theta, \psi), J(\beta), \theta, \mathcal{W}), \quad (1)$$

where $\bar{T}$ is the mean template shape; B_s, B_p, and B_e are the blend shape functions for shape, pose, and expression, respectively; $T(\beta, \theta, \psi)$ is the non-rigid deformation from $\bar{T}$; $\text{LBS}(\cdot)$ is the linear blend skinning function [35] that transforms $T(\beta, \theta, \psi)$ into target pose θ, with the skeleton joints $J(\beta)$ and blend weights $\mathcal{W}$ defined on each vertex.

Score Distillation Sampling (SDS) is proposed in Dream-Fusion [56] to distill the 2D pre-trained diffusion prior for optimizing 3D representations. Specifically, we represent a 3D scene parameterized by θ and use a differentiable rendering function $g(\cdot)$ to obtain an image $\mathbf{x} = g(\theta)$. By pushing samples towards denser regions of the real-data distribution across all noise levels, we make renderings from each camera view resemble the plausible samples derived from the guidance diffusion model ϕ. In practice, DreamFusion uses Imagen [63] as the score estimation function $\boldsymbol{\epsilon}_\phi(\mathbf{x}_t; y)$, which predicts the sampled noise $\boldsymbol{\epsilon}_\phi$ given the noisy image $\mathbf{x}_t$, text embedding y, and timestep t. SDS optimizes 3D scenes using gradient descent with respect to θ:

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

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a Gaussian noise; $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$ is the noised image; α_t, σ_t, and w_t are the terms that control the noise schedule decided by the diffusion sampler.

3D Gaussian Splatting [31] features an effective representation for novel-view synthesis and 3D reconstruction. Different from those implicit counterparts like NeRF [49], 3D Gaussian Splatting represents the underlying scene through a set of anisotropic Gaussians parameterized by their center position $\mu \in \mathbb{R}^3$, covariance $\Sigma \in \mathbb{R}^7$, color $c \in \mathbb{R}^3$, and opacity $\alpha \in \mathbb{R}$. By projecting 3D Gaussians onto the camera's imaging plane, the 2D Gaussians are assigned to the corresponding tiles for point-based rendering [95]:

$$G(p, \mu_i, \Sigma_i) = \exp(-\frac{1}{2}(p - \mu_i)^{\mathsf{T}} \Sigma_i^{-1}(p - \mu_i)),$$
$$\mathbf{c}(p) = \sum_{i \in \mathcal{N}} c_i \sigma_i \prod_{j=1}^{i-1} (1 - \sigma_j), \sigma_i = \alpha_i G(p, \mu_i, \Sigma_i), \quad (3)$$

where p is the location of queried point; μ_i, Σ_i, c_i, α_i, and σ_i are the center position, covariance, color, opacity, and density of the i-th Gaussian, respectively; $G(p, \mu_i, \Sigma_i)$ is the value of the i-th Gaussian at point p; $\mathcal{N}$ denotes the set of 3D Gaussians in this tile. Besides, 3D Gaussian Splatting improves a GPU-friendly rasterization process with better quality, faster rendering speed, and less memory usage.

3.2. Structure-Aware SDS

To adapt Gaussian Splatting for text-driven 3D human generation, the simplest way is by substituting scene representations of mesh or implicit neural fields with explicit 3DGS [31]. However, three problems remain: **1)** The original 3DGS training process heavily relies on center position initialization, even for the comparatively easier 3D reconstruction with dense multi-view image supervision. How

to initialize Gaussians in the text-to-3D generation setting, especially for the highly-detailed human domain. **2)** To ease 3DGS learning with both appearance and geometry guidance, we have to supplement structural knowledge like text-to-depth/normal to text-to-image. While previous studies mostly capture a singular modality, how to simultaneously learn the joint distribution of structure and texture remains challenging. **3)** The commonly-used SDS [56] distills solely from RGB space, yet it is hard to optimize the point-based 3DGS renderings with sparse gradients. How to add explicit structural supervision for better convergence.

Gaussian Initialization with SMPL-X Prior. Our solution to the first problem is to initialize 3D Gaussian center positions based on the SMPL-X mesh shape. We choose it as human domain-specific structural prior due to two reasons: **1)** Previous studies [7, 31, 86] use either Structure-from-Motion (SfM) points [71] or generic text-to-point-cloud priors of Shap-E [28] and Point-E [51]. However, such methods typically fall short in the human category, resulting in over-sparse points or incoherent body structures. **2)** As an extension to SMPL [43], SMPL-X complements the shape topology of face and hands, which are beneficial to intricate human modeling with fine-grained details. Based on such observations, we propose to uniformly sample points over SMPL-X mesh surface as 3DGS initialization. Specifically, $100k$ 3D Gaussians are instantiated with unit scaling, mean color, and no rotation, which are significantly denser than SMPL-X-defined vertices to model local details. We scale and transform the 3DGS to make it reasonable human size and located in the center of 3D space. To further extract 2D skeleton as structural condition, we map SMPL-X joints to the COCO-style human keypoints. When drawing keypoints on canvas as body skeleton maps, we carefully cull the occluded joints for more accurate visualization, such as removing the left/right eye and ear keypoints from the right-/left-side view, and hiding face keypoints from back view.

Learn Texture-Structure Joint Distribution. Since the SMPL-X prior merely serves as an initialization, more comprehensive guidance is needed to facilitate 3DGS training. Instead of distilling the 3D scene from a single-modality diffusion model that solely learns the appearance [63] or geometry [15], we propose to harvest an SDS source model capturing the joint distribution of both texture and structure. In particular, we take inspiration from a recent work [42] to extend the pre-trained Stable Diffusion [62] with structural expert branches to denoise the image RGB and depth simultaneously. By trading off between the spatial alignment and accurate distribution learning, we replicate the diffusion UNet backbone's `conv_in`, the first `DownBlock`, the last `UpBlock`, and `conv_out` layers to deal with the denoising of each target. To further enable flexible skeleton control, we additionally take pose map $\mathbf{p}$ as an input condition via channel-wise concatenation. The overall network is trained

with $\mathbf{v}$-prediction [65] to minimize the simplified objective:

$$\mathbb{E}_{\mathbf{v},t}\left[||\mathbf{v}_\phi(\mathbf{x}_t;\mathbf{p},y) - \mathbf{v}_t^{\mathbf{x}}||_2^2 + ||\mathbf{v}_\phi(\mathbf{d}_t;\mathbf{p},y) - \mathbf{v}_t^{\mathbf{d}}||_2^2\right], \quad (4)$$

where $\mathbf{x}$ and $\mathbf{d}$ are the image-depth pairs annotated from large dataset; $\mathbf{x}_t = \alpha_t\mathbf{x} + \sigma_t\epsilon_{\mathbf{x}}$ and $\mathbf{d}_t = \alpha_t\mathbf{d} + \sigma_t\epsilon_{\mathbf{d}}$ are the noised features of RGB and depth; $\epsilon_{\mathbf{x}}, \epsilon_{\mathbf{d}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are independently sampled noise; $\mathbf{v}_t^{\mathbf{x}} = \alpha_t\epsilon_{\mathbf{x}} - \sigma_t\mathbf{x}$ and $\mathbf{v}_t^{\mathbf{d}} = \alpha_t\epsilon_{\mathbf{d}} - \sigma_t\mathbf{d}$ are the $\mathbf{v}$-prediction learning targets at time-step t for the RGB and depth, respectively. In this way, we obtain a unified model that captures both the image texture of appearance and structure of fore-/back-ground relationship, which can be used in SDS to boost 3DGS learning.

Dual-Branch SDS as Optimization Guidance. With the extended diffusion model that produces the spatially aligned image RGB and depth, we can guide the 3DGS optimization process from both structural and textural aspects. Specifically, the per-view depth map at each pixel can be computed by accumulating depth value overlapping the pixel of $\mathcal{N}$ ordered Gaussian instances via point-based α-blending:

$$\mathbf{d}(p) = \sum_{i\in\mathcal{N}} d_i\sigma_i \prod_{j=1}^{i-1}(1 - \sigma_j), \sigma_i = \alpha_i G(p, \mu_i, \Sigma_i), \quad (5)$$

where d_i is the projected depth of the i-th Gaussian center μ_i in the current camera view; $G(p, \mu_i, \Sigma_i)$ is the value of the i-th Gaussian at the queried point p as defined in Eq. 3. Afterward, all the rendered depth maps $\mathbf{d}$ are normalized to the data range of $[0, 1]$. Combined with image renderings $\mathbf{x}$, we can optimize 3DGS with an improved dual-branch SDS:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \lambda_1 \cdot \mathbb{E}_{\epsilon_{\mathbf{x}},t}\left[w_t(\epsilon_\phi(\mathbf{x}_t;\mathbf{p},y) - \epsilon_{\mathbf{x}})\frac{\partial\mathbf{x}}{\partial\theta}\right]$$
$$+ \lambda_2 \cdot \mathbb{E}_{\epsilon_{\mathbf{d}},t}\left[w_t(\epsilon_\phi(\mathbf{d}_t;\mathbf{p},y) - \epsilon_{\mathbf{d}})\frac{\partial\mathbf{d}}{\partial\theta}\right], \quad (6)$$

where λ_1 and λ_2 are coefficients balancing the effects from the structural and textural branches; $\epsilon_\phi(\cdot)$ are the ϵ-predictions derived from joint texture-structure model in Eq. 4. Such structural regularization helps reduce geometry distortions, thus benefiting the optimization of 3DGS with sparse gradient information. Note that although a concurrent work [22] also adds structural constraints by finetuning two *individual* text-to-normal and text-to-depth models for SDS, they suffer from misaligned depth and normal predictions, which could potentially mislead the network training.

3.3. Annealed Negative Prompt Guidance

Decompose SDS with Classifier-Free Guidance. To enforce text-3D alignment of high quality, DreamFusion [56] uses a large classifier-free guidance scale to update the score matching difference term δ for 3D scene optimization:

$$\delta = \underbrace{[\epsilon_\phi(\mathbf{x}_t;y) - \epsilon]}_{\text{generative score } \delta_g} + \tau \cdot \underbrace{[\epsilon_\phi(\mathbf{x}_t;y) - \epsilon_\phi(\mathbf{x}_t;\varnothing)]}_{\text{classifier score } \delta_c}, \quad (7)$$

where τ is the classifier-free guidance scale; $\varnothing$ denotes the null condition of an empty prompt. In this formulation, we can naturally decompose score matching difference into two parts $\delta = \delta_g + \tau \cdot \delta_c$, where the former term is a generative score that pushes images towards more realistic manifold; and the latter term is a classifier score that aligns samples with an implicit classifier [18]. However, as the generative score contains Gaussian noise ϵ of high variance, it provides stochastic gradient information that harms the training stability. To deal with this, Poole *et al.* [56] deliberately intensify CFG scale τ to make classifier score dominate the optimization, leading to over-saturated patterns. Instead, we only leverage the cleaner classifier score δ_c as SDS loss.

Regularize Gradient with Negative Prompts. The negative prompts are widely used in text-to-image to bypass undesired properties. In view of this, we propose to augment classifier score for better 3DGS learning. Specifically, we substitute the null condition $\varnothing$ with negative prompts y_{neg} to derive the improved negative classifier score δ_{nc}:

$$
\begin{aligned}
\delta_{nc} &= \epsilon_\phi\left(\mathbf{x}_t; y\right) - \epsilon_\phi\left(\mathbf{x}_t; y_{\text{neg}}\right), \\
&= \underbrace{\left[\epsilon_\phi\left(\mathbf{x}_t; y\right) - \epsilon_\varnothing\right]}_{\text{classifier score } \delta_c} - \underbrace{\left[\epsilon_\phi\left(\mathbf{x}_t; y_{\text{neg}}\right) - \epsilon_\varnothing\right]}_{\text{negative score } \delta_n}, \quad (8)
\end{aligned}
$$

where $\epsilon_\varnothing = \epsilon_\phi\left(\mathbf{x}_t; \varnothing\right)$ is the abbreviation for formula conciseness. In this way, we regularize SDS gradients from two aspects: the rendered image is not only guided to align with the implicit classifier, but also forced to repel from the negative mode. Empirically, we find the negative score harm quality at small timesteps. So we use an annealed negative prompt guidance to combine both scores for supervision:

$$
\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{\epsilon, t}\left[w_t\left(\tau_1 \cdot \delta_c - \tau_2 \cdot \delta_n\right)\frac{\partial \mathbf{x}}{\partial \theta}\right], \quad (9)
$$

where τ_1 and τ_2 balance the effects from classifier and negative scores, with τ_2 gradually drops at smaller timesteps. Note that such guidance can also be applied to depth. In practice, we use Eq. 9 for both image and depth branches.

Size-Conditioned Gaussian Pruning. Although the proposed annealed negative prompt guidance provides cleaner gradients for 3DGS training, the SDS-based supervision is still much more stochastic than reconstruction-based error, leading to vaporous blurriness near the human surface. One observation is that such floating artifacts emerge as tiny-size Gaussians during the 3DGS densification process of cloning and splitting. To this end, we propose to eliminate them based on the scaling factor in a prune-only phase (Gaussian covariance $\Sigma \in \mathbb{R}^7$ can be decoupled as a scaling factor $s \in \mathbb{R}^3$ and a rotation quaternion $q \in \mathbb{R}^4$). Particularly, after the adaptive density control that heavily relies on high-variance SDS, we extend a prune-only phase to remove Gaussian instances whose scaling factor is below a certain threshold. Note that a potential problem is that some

useful Gaussians could be mistakenly eliminated. Despite this, in practice we find such a mechanism robust to maintain fine-grained geometry for two reasons: **1)** Those tiny-scale Gaussians contribute little to point-based α-blending, which are negligible in the rendered results. **2)** Thanks to the dense initialization based on SMPL-X prior, the Gaussians are redundant near the human body surface, which could compensate for the detail loss during optimization.

4. Experiments

4.1. Implementation Details

SDS Guidance Model Setups. As elaborated in Sec. 3.2, we extend the pretrained SD to capture the joint distribution of texture and structure by simultaneously denoising RGB and depth. The depth maps are labeled by MiDaS [60] on LAION [68]. The model is finetuned from *SD 2.0* with **v**-prediction [65] in 512 resolution. The DDIM scheduler [72] is used with classifier-score weight τ_1 as 7.5. We gradually drop the negative-score weight τ_2 from 1.0 to 0.0 at timestep 200 for annealed negative prompt SDS guidance.

3D Gaussian Splatting Setups. The 3D Gaussians are initialized with $100k$ instances evenly sampled on SMPL-X mesh surface with opacity of 0.1. The color is represented by Spherical Harmonics (SH) coefficients [11] of degree 0 following [31]. The whole 3DGS training takes 3600 iterations, with the densification & pruning from 300 to 2100 iterations at an interval of 300 steps. The prune-only phase is conducted at a scaling factor threshold of 0.008 from 2400 to 3300 every 300 steps. The overall framework is trained using Adam optimizer [32], with the betas of $[0.9, 0.99]$ and the learning rates of $5e-5$, $1e-3$, $1e-2$, $1.25e-2$, and $1e-2$ for the center position μ, scaling factor s, rotation quaternion q, color c, and opacity α, respectively.

Training and Implementation Setups. The framework is implemented in PyTorch [54] based on ThreeStudio [14]. We use the camera distance range of $[1.5, 2.0]$, fovy range of $[40°, 70°]$, elevation range of $[-30°, 30°]$, and azimuth range of $[-180°, 180°]$. During the 1200 to 3600 iterations, we zoom into the head region with camera distance range of $[0.4, 0.6]$ at 25% probability to enhance facial quality. The dual-branch SDS loss weights for RGB and depth λ_1, λ_2 are both set as 0.5. We use the training resolution of 1024 with a batch size of 8 and the whole optimization process takes one hour on a single NVIDIA A100 (40GB) GPU.

4.2. Text-Driven 3D Human Generation

Comparison Methods. We compare with two categories of recent SOTA works: **1)** General text-to-3D methods, including DreamGaussian [74] and GaussianDreamer [86], which are two concurrent models that also use 3DGS as neural representation. **2)** Specialized text-to-3D human models, where we show visual comparisons with TADA [39] and

Figure 3. **Visual Comparisons with Text-to-3D and 3D Human Models.** We compare with recent state-of-the-art baselines on five different prompts, each showing two camera views. Note that the textural unrealism and blurriness are highlighted with **yellow** arrows; the geometric artifacts are highlighted with **green** rectangles. Please kindly **zoom in for best view** and refer to demo video for more results.

DreamHuman [34]. Note that since the pre-trained model or training code of DreamHuman is not officially released, we directly use the results from their paper and project page. **Qualitative Analysis.** The visualized generation results are shown in Fig. 1, where the proposed framework can generate high-quality 3D humans with realistic appearance and fine-grained geometry (*e.g.*, the cloth wrinkles and accessories of ear-ring in the 4-th row). To further validate the

effectiveness of our method, we show visual comparisons with recent works [34, 39, 74, 86] in Fig. 3. Note that we augment two 3DGS-based baselines with multi-view [70] or Shap-E [28] prior to make their results stronger, following the authors' instructions. It can be seen that **HumanGaussian** achieves superior performance, rendering more realistic human appearance, more coherent body structure, better view consistency, and more fine-grained detail capturing.
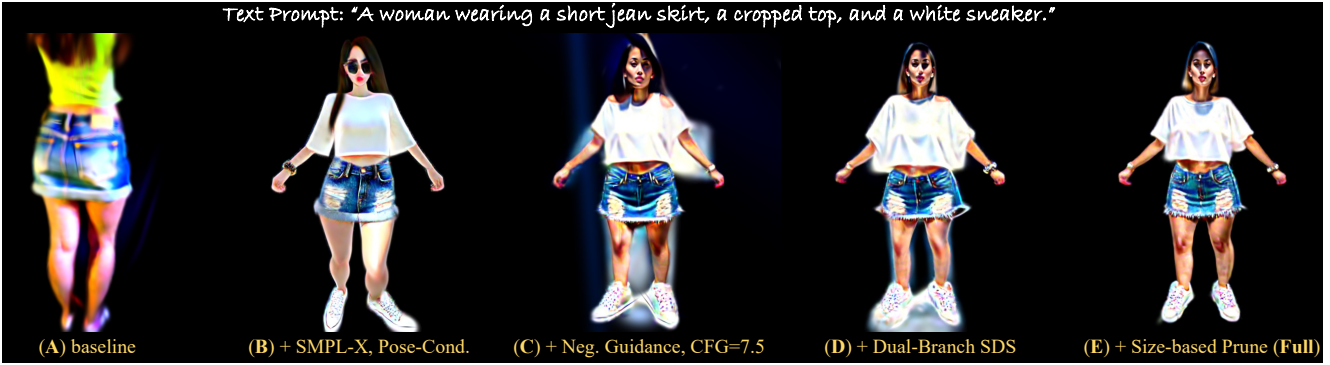
Figure 4. **Ablation Studies on HumanGaussian Module Design.** We present generation results of the human frontal view under five ablation settings for better visualization and comparisons: **(A)** *baseline*; **(B)** *+SMPL-X, Pose-Cond.*; **(C)** *+Neg. Guidance, CFG*=7.5; **(D)** *+Dual-Branch SDS*; **(E)** *+Size-based Prune*. The detailed ablation setting designs and result analysis are elaborated in Sec. 4.3.

User Study. We conduct a user study to better reflect 3D human quality. For fair comparisons, we randomly sample 30 prompts from DreamHuman [34] and involve 17 participants for subjective evaluation. The users are asked to rate three aspects: (1) *Texture Quality*; (2) *Geometry Quality*; (3) *Text Alignment*. The rating scale is 1 to 5, with the higher the better. The results are reported in Table 1, where our method performs the best on all three aspects.

4.3. Ablation Study

We present the ablation studies on two key modules of our framework in Figure 4. In particular, we explore the following aspects: **1)** structural prior from SMPL-X and pose condition; **2)** annealed negative prompt guidance; **3)** dual-branch SDS; and **4)** size-conditioned Gaussian pruning, by *gradually adding each component* as our ablation settings: **(A)** *baseline* that naively generates 3D human with 3DGS; **(B)** *+SMPL-X, Pose-Cond.*, which uses SMPL-X for initialization and skeleton for SDS guidance model conditioning; **(C)** *+Neg. Guidance, CFG*=7.5 that incorporates annealed negative prompt guidance with nominal CFG scale of 7.5; **(D)** *+Dual-Branch SDS* that extends the depth-branch SDS instead of solely relying on supervisions from pixel-space; **(E)** *+Size-based Prune* that removes tiny Gaussian artifacts.

Ablation Analysis. By comparing the visual quality among different settings, we can clearly see the effectiveness of each design: **1)** Different from **Config A** that suffers from incorrect articulations with multi-face Janus problem, explicit body prior and pose guidance in **Config B** guarantee coherent body structures. **2)** With the cleaner negative classifier score in **Config C**, we manage to bypass cartoonish styles or over-saturated patterns caused by large CFG scales and achieve realistic appearance. **3)** The dual-branch SDS (**Config D**) further provides joint texture-structure guidance to regularize geometric errors near limb and hair. **4)** Thanks to the size-conditioned Gaussian removal in a prune-only phase, in the full model (**Config E**) we eliminate floating artifacts near the human surface. Please kindly refer to our demo video for clearer ablation result comparisons.

Methods	Text. Qual.	Geo. Qual.	Text Align.
TADA [39]	3.76	3.53	4.35
DreamHuman [34]	3.41	3.65	4.24
DreamGaussian [74]	2.18	2.88	2.94
GaussianDreamer [86]	2.94	3.00	3.06
HumanGaussian (Ours)	**4.24**	**3.88**	**4.71**

Table 1. **User Study Results.** We conduct subjective evaluations on 3D human generation quality from three aspects: (1) *Texture Quality*; (2) *Geometry Quality*; (3) *Text Alignment*. The Mean Opinion Scores (MOS) protocol is adopted to rate based on the score scale of 1 to 5, with 1 being the poorest and 5 being the best.

5. Discussion

Conclusion. In this paper, we propose an efficient yet effective framework **HumanGaussian** for high-quality 3D human generation with fine-grained geometry and realistic appearance. We first propose *Structure-Aware SDS* that simultaneously optimizes human appearance and geometry. Then we devise an *Annealed Negative Prompt Guidance* to guarantee realistic results without over-saturation and eliminate floating artifacts. Extensive experiments demonstrate the superior efficiency and competitive quality of our framework, rendering vivid 3D humans under diverse scenarios.

Limitations and Future Work. As an early attempt in taming 3DGS for text-driven 3D human, our method creates realistic results. However, due to the limited performance of existing T2I models for hand and foot generation, we find it sometimes fails to render these parts faithfully. Besides, the back-view texture may look blurry, which arises from the fact that 2D pose-conditioned models are mostly trained on human frontal views, while the prior on human back-view is scarce. We will explore these problems in future work.

Ethical Considerations. Although our method can generate realistic 3D humans, it could be misused for malicious purposes. As part of our responsibility, we strongly advocate the proper use of this technique, which we believe can enhance research community and industrial applications.

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. 2, 3

[2] Anonymous. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. In *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review. 3

[3] Bernd Bickel, Mario Botsch, Roland Angst, Wojciech Matusik, Miguel Otaduy, Hanspeter Pfister, and Markus Gross. Multi-scale capture of facial geometry and motion. *ACM transactions on graphics (TOG)*, 26(3):33–es, 2007. 2

[4] 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. *arXiv preprint arXiv:2304.00916*, 2023. 2, 3

[5] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. *arXiv preprint arXiv:2304.06714*, 2023. 3

[6] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. *arXiv preprint arXiv:2303.13873*, 2023. 2, 3

[7] Zilong Chen, Feng Wang, and Huaping Liu. Text-to-3d using gaussian splatting. *arXiv preprint arXiv:2309.16585*, 2023. 2, 3, 5

[8] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023. 3

[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 628–644. Springer, 2016. 3

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 6

[12] William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. Textdeformer: Geometry manipulation using text guidance. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 3

[13] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[14] Yuan-Chen Guo, Ying-Tian Liu, Chen Wang, Zi-Xin Zou, Guan Luo, Chia-Hao Chen, Yan-Pei Cao, and Song-Hai Zhang. threestudio: A unified framework for 3d content generation, 2023. 6

[15] Anchit Gupta, Wenhan Xiong, Yixin Nie, Ian Jones, and Barlas Oğuz. 3dgen: Triplane latent diffusion for textured mesh generation. *arXiv preprint arXiv:2303.05371*, 2023. 3, 5

[16] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (ToG)*, 38(4):1–12, 2019. 3

[17] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996. 2

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 6

[19] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 3

[20] 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, 3

[21] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[22] 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. *arXiv preprint arXiv:2310.01406*, 2023. 2, 3, 5

[23] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *arXiv preprint arXiv:2305.12529*, 2023. 2

[24] 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. 3

[25] 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. *arXiv preprint arXiv:2303.17606*, 2023. 2, 3

[26] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 2

[27] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided

diffusion model for human image generation. *arXiv preprint arXiv:2304.04269*, 2023. 2

[28] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 3, 5, 7

[29] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 2

[30] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 3

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2, 3, 4, 5, 6

[32] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[33] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 2

[34] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. *arXiv preprint arXiv:2306.09329*, 2023. 2, 3, 7, 8

[35] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 811–818. 2023. 4

[36] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 2

[37] Ruilong Li, Kyle Olszewski, Yuliang Xiu, Shunsuke Saito, Zeng Huang, and Hao Li. Volumetric human teleportation. In *ACM SIGGRAPH 2020 Real-Time Live!*, pages 1–1. 2020. 2

[38] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 3

[39] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. Tada! text to animatable digital avatars. *arXiv preprint arXiv:2308.10899*, 2023. 2, 3, 6, 7, 8

[40] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022. 3

[41] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. In *European Conference on Computer Vision*, pages 106–125. Springer, 2022. 3

[42] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 5

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 2, 3, 5

[44] Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nicholas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, and James Lucas. Att3d: Amortized text-to-3d object synthesis. *arXiv preprint arXiv:2306.07349*, 2023. 3

[45] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 3

[46] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 922–928. IEEE, 2015. 3

[47] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2, 3

[48] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 3

[49] 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. 2, 3, 4

[50] 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. 3

[51] Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. Point-e: A system for generating 3d point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 2, 3, 5

[52] Evangelos Ntavelis, Aliaksandr Siarohin, Kyle Olszewski, Chaoyang Wang, Luc Van Gool, and Sergey Tulyakov. Autodecoding latent 3d diffusion models. *arXiv preprint arXiv:2307.05445*, 2023. 3

[53] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 3

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming

Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 6

[55] 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. 3, 4

[56] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3, 4, 5, 6

[57] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3

[58] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 3

[59] 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. 3

[60] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 6

[61] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 2

[62] 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. 2, 5

[63] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2, 4, 5

[64] 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. 2

[65] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 5, 6

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

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

[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 6

[69] 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

[70] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 7

[71] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 5

[72] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[73] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 2

[74] 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. 3, 6, 7, 8

[75] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 589–604, 2018. 2

[76] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3

[77] 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. 3

[78] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 3

[79] 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. 3

[80] 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. *arXiv preprint arXiv:2305.16213*, 2023. 2, 3

[81] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1042–1051, 2019. 3

[82] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022. 2

[83] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022. 3

[84] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 3

[85] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 3

[86] Taoran Yi, Jiemin Fang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussian splatting with point cloud priors. *arXiv preprint arXiv:2310.08529*, 2023. 2, 3, 5, 6, 7, 8

[87] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision*, pages 173–191. Springer, 2022. 2

[88] 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. 3

[89] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 2

[90] Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864*, 2023. 2, 3

[91] Huichao Zhang, Bowen Chen, Hao Yang, Liao Qu, Xu Wang, Li Chen, Chao Long, Feida Zhu, Kang Du, and Min Zheng. Avatarverse: High-quality & stable 3d avatar creation from text and pose. *arXiv preprint arXiv:2308.03610*, 2023. 2, 3

[92] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J Black. Text-guided generation and editing of compositional 3d avatars. *arXiv preprint arXiv:2309.07125*, 2023. 3

[93] 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. 3

[94] Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, and Changjie Fan. Zero-shot text-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21023, 2023. 3

[95] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa volume splatting. In *Proceedings Visualization, 2001. VIS'01.*, pages 29–538. IEEE, 2001. 4