SimAvatar: Simulation-Ready Avatars with Layered Hair and Clothing

Xueting Li, Ye Yuan, Shalini De Mello, Gilles Daviet, Jonathan Leaf, Miles Macklin,
Jan Kautz, Umar Iqbal

NVIDIA

https://nvlabs.github.io/SimAvatar

Adele with long layered waved hairstyle wearing a sleeveless dress of tea-length, the garment has a delicate butterfly print, with small, colorful butterflies.

(a) simulation-ready avatar with layered hair, garment and body

(b) simulated avatar in different poses

(c) example simulation-ready avatars.

Figure 1. (a) SimAvatar synthesizes simulation-ready 3D avatars with layered hair, body and clothing. (b) By leveraging physics-based simulators, SimAvatar produces realistic pose-dependent motion effects such as flowing hair strands and garment wrinkles. (c) SimAvatar produces simulation-ready 3D avatars with diverse identities, garments, and hairstyles.

Abstract

We introduce SimAvatar, a framework designed to generate simulation-ready clothed 3D human avatars from a text prompt. Current text-driven human avatar generation methods either model hair, clothing, and the human body using a unified geometry or produce hair and garments that are not easily adaptable for simulation within existing simulation pipelines. The primary challenge lies in representing the hair and garment geometry in a way that allows leveraging established prior knowledge from foundational image diffusion models (e.g., Stable Diffusion) while being simulation-ready using either physics or neural simulators.

To address this task, we propose a two-stage framework that combines the flexibility of 3D Gaussians with simulation-ready hair strands and garment meshes. Specifically, we first employ three text-conditioned 3D generative models to generate garment mesh, body shape and hair strands from the given text prompt. To leverage prior knowledge from foundational diffusion models, we attach 3D Gaussians to the body mesh, garment mesh, as well as hair strands and learn the avatar appearance through optimization. To drive the avatar given a pose sequence, we first apply physics simulators onto the garment meshes and hair strands. We then transfer the motion onto 3D Gaussians through carefully designed mechanisms for each body part. As a result, our

synthesized avatars have vivid texture and realistic dynamic motion. To the best of our knowledge, our method is the first to produce highly realistic, fully simulation-ready 3D avatars, surpassing the capabilities of current approaches.

1. Introduction

The creation of photo-realistic and dynamic human avatars has a wide range of applications, including virtual try-on, film and gaming production, virtual assistants, AR/VR, and telepresence. Traditionally, this process has required specialized training, making it inaccessible to general users. Recently, advancements in foundational diffusion models have accelerated research efforts aimed at democratizing 3D human avatar creation, enabling easy user control through text [16, 46, 51, 88] or images [39].

Earlier approaches for 3D human avatar creation represent the hair, body, and clothing as a single layer, making it difficult to independently simulate or edit each region due to their entangled geometry. To address this limitation, recent works have used layered structures to separately represent the body, garments, or hair [27, 36, 82, 96]. However, many of these approaches rely on implicit representations, such as NeRF [58], to define garment or hair geometry. Although implicit representations facilitate leveraging prior knowledge from foundational diffusion models, they are challenging to animate within existing simulators, limiting the realistic motion of hair and garments due to body movements. As a result, these methods struggle to produce avatars that look realistic when animated.

Thus, a natural question arises: can we design a 3D avatar generation pipeline that can leverage rich prior knowledge from image diffusion models, while being readily compatible with existing simulation pipelines? The key challenge in addressing this question lies in connecting the different representations used in current simulators and text-driven avatar generation pipelines. The former typically requires a smooth and clean non-watertight mesh or specifically designed hair strands, whose topologies are inflexible for optimization and hard to constrain. The latter often adopt implicit representations (such as NeRF [58] or SDF [83]) which, while optimizable by noisy supervision signals from diffusion models, cannot be easily converted to an open mesh or hair strands suitable for simulation.

To address these issues, we present a novel framework, SimAvatar, which generates a 3D human avatar from a text prompt that can be readily animated by existing hair and garment simulators. The key idea is to adopt suitable representations for different human parts (e.g., hair, body, and garments) and leverage the prior knowledge of both image diffusion models and simulators. To this end, we propose representing the geometry of human hair, body, and garment using hair strands, parametric body model SMPL [55],

and meshes, respectively. While these representations are simulation-ready, they pose challenges to learning realistic texture details. To resolve this issue, we further attach 3D Gaussians to these coarse geometric representations. 3D Gaussians are not only flexible and can be easily driven by the mesh or hair strands, but they also show impressive capacity to model realistic appearance when combined with image diffusion models [54, 88].

Given a text prompt, we first generate hair strands, body mesh, and garment mesh using three separate text-conditioned 3D generative models trained specifically for each body region. We then attach and learn 3D Gaussians through Score Distillation Sampling (SDS) based optimization using a powerful diffusion model trained on human images [1]. To drive the avatar using a given pose sequence, we animate the body using linear blend skinning while the hair strands and garment mesh are simulated using their respective simulators [23, 30]. Next, we transfer the motion from the strands or meshes onto 3D Gaussians through a carefully designed mechanism to ensure they conform to the mesh or strand motion. During inference, the avatar can be easily animated by novel pose sequences. Our novel design yields avatars with both realistic appearance and dynamic motion with pose-dependent deformations of the hairs and garment as shown in Fig. 1. Our contributions are summarized as follows:

- To the best of our knowledge, this is the first work that generates fully simulation-ready 3D avatars with separate layers for the body, garment, and hair.
- We propose a text-conditioned diffusion model to generate plausible garment geometry given text inputs and show how existing text-conditioned hair strands and body shape generators can be efficiently leveraged.
- We use 3D Gaussians on top of the body mesh, hair strands, and garment mesh to synthesize high-fidelity appearance details. We optimize the properties of Gaussians using carefully designed SDS-based optimization [87] and ensure plausible texture disentanglement using additional regularization and prompt engineering.
- Our approach produces avatars with detailed geometry, realistic texture, as well as dynamic garment and hair motion, resulting in significantly better quality than existing methods.

2. Related Work
2.1. Text to 3D Avatar Generation

The domain of text-to-3D generation has witnessed significant advancements [18, 52, 63, 66, 78, 85] driven by the advent of large text-to-image models [21, 67, 68, 97]. Following the success in generating static 3D objects, various methods have also been developed to create animatable human avatars [16, 37–40, 42, 46, 51, 95, 99]. ClipMatrix [40] was among the first to create animatable avatars

based on textual descriptions, using a CLIP-embedding loss to optimize a mesh-based representation. AvatarClip [34] followed a similar approach but with a NeRF-based representation [83]. DreamAvatar [16] and AvatarCraft [42] used SDS loss instead of CLIP and learned the NeRF representation in canonical space by integrating human body priors from SMPL [55]. DreamHumans [46] introduced a pose-conditioned NeRF representation to also model pose-dependent deformations. DreamWaltz [38], AvatarVerse [95] and HumanNorm [97] leverage ControlNets [97] and demonstrate improved avatar quality with conditional SDS. AvatarBooth [92] fine-tunes region-specific diffusion models, highlighting that using dedicated models for distinct body regions improves avatar quality. Other methods explore different representations to model the avatars. TADA [51] demonstrates that a mesh-based approach with adaptive mesh subdivision can create high-quality avatars. More recently, GAvatar [88] adopted a primitive-based 3D Gaussian representation, allowing faster rendering in higher resolution and showing higher quality. All of the aforementioned methods, however, represent the human body, hairs, and garment geometry using a single layer and animate the avatar using linear blend skinning. Hence, they cannot model physically plausible and pose-dependent geometric deformations of the avatar's clothing and hair resulting in avatars that do not appear realistic when animated.

There are only a few works that generate avatars with separate body and garment or hair layers. SO-SMPL [82] uses SMPL+D [62] to separately represent the cloth layers, but cannot capture loose clothing due to the inherent limitations of SMPL+D representation. HumanLiff [36], HumanCoser [84], and TELA [27] use NeRF-based representations and progressively generate minimal-clothed human body and layer-wise clothes. GALA [45] takes an existing single-layer avatar and disentangles it into separate layers using DMTet representation [29] and SDS [63] optimization. TECA [96] aims to model hairs in a separate layer but adopts a NeRF-based representation. However, the meshes extracted from NeRF or DMTet tend to be very noisy and cannot be simulated easily using cloth or hair simulators. Our approach, on the other hand, generates avatars with simulation-ready garments and hair. It can model loose clothing, realistic hair with varied lengths and hairstyles, realistic texture, and dynamic motion.

2.2. Garment Modeling and Simulation

One of the most common methods for modeling garments is SMPL+D [3–5, 14, 15, 57, 62, 94], which adds vertex displacements to the SMPL body mesh to capture the geometry of clothed human bodies. However, this approach struggles to represent loose clothing and dresses due to its fixed topology. To address this limitation, many methods propose modeling clothing separately. BCNet [41] directly predicts

garment type and shape parameters using a learned regressor, while DeepCloth [77] represents garments through a canonical UV-representation, encoding them in a shared latent space. More recent methods utilize implicit unsigned distance fields for their flexibility in modeling arbitrary topologies and handling open surfaces, which are crucial for cloth simulators [17, 20, 25, 32, 65]. However, these methods typically focus solely on modeling clothing geometry, neglecting the underlying human body, hair, and texture details. In this work, we build on these methods and propose a novel text-conditioned diffusion model to generate garment meshes using text prompts, while also generating avatars with complete texture and appearance details for garment, hair, and body layers.

For body motion-driven garment simulation, traditional methods use physics simulators to drape garments on the human body while modeling their material properties [13, 19, 80, 81], and avoiding collisions with the underlying body [7, 79]. Differentiable physics simulators have also been proposed to integrate physics into learning and optimization-based frameworks [49, 50]. While these methods produce highly realistic garment animations, physics simulation is generally computationally expensive and can become unstable in cases of severe interpenetration between the garment and body layers. To overcome these limitations, more recent works propose neural simulators, which are neural networks trained to generate garment deformations based on body pose and shape information [30, 31, 57, 60, 69, 71, 72]. Earlier methods were trained in a garment-specific manner and could not generalize across different garment types [59, 70]. However, more recent methods can handle various garment topologies, including loose clothing and dresses [10–12, 30, 72, 73, 90]. While our method can utilize both physics-based and neural simulators, here we use HOOD [30] as the neural simulator.

2.3. Hair Modeling and Simulation

For hair modeling, strand-based representations are widely favored due to their compatibility with physics simulators and ease of geometric manipulation [23, 28, 35, 61, 74, 89]. The main challenge in image- or text-guided generation of hair strands lies in capturing the hair's complex appearance, especially the inner strands that are often occluded. For geometry, recent approaches use data-driven priors based on diffusion models to estimate plausible inner strands despite their occlusion in observed data [75, 76]. For appearance, existing methods employ a strand-aligned 3D Gaussian representation to effectively model intricate hair textures and varied strand thickness [56, 91]. In this work, we build upon these methods for full-body avatar generation. Specifically, we use a diffusion-based text-to-hair generation model [76] to create hair strands from text prompts. We then attach 3D Gaussians [91] to the generated strands to model hair

appearance and refine hairstyle according to the given text prompt. Finally, we incorporate a hair simulator [23] during avatar creation to simulate hairs for varied poses which allows for optimizing the appearance of the inner strands and ensures proper disentanglement of face and hair textures.

3. Method

Given a text prompt $\mathcal{T}$, we introduce a novel framework, SimAvatar, for synthesizing simulation-ready 3D avatars with layered representations of the body, hair, and garment. Unlike traditional text-to-avatar methods that use a single geometry for the entire avatar [18, 51, 88], our approach decomposes the avatar geometry into distinct layers, each leverages a representation tailored to its simulation and geometry modeling. This layered structure not only enhances realism but also enables robust simulation of each component, dramatically improving flexibility in animation and customization. Specifically, our avatar geometry is represented by three distinct layers: (1) **Body mesh**, defined by the SMPL parametric model [55], providing a robust foundation for realistic body shapes and movements; (2) **Garment mesh**, generated using a text-conditioned diffusion model, which produces smooth clothing meshes that are ready for simulation (see Sec. 3.2); (3) **Hair strands**, modeled using a text-conditioned diffusion model, allowing for the generation of complex hairstyles with strand-level detail (see Sec. 3.1). This approach allows our method to readily handle dynamic scenarios. Given a target body pose sequence $\mathcal{P} = \{p_1, ..., p_n\}$, the body mesh is deformed through linear blend skinning [55] while both the garment mesh and hair strands are simulated using advanced physics-based simulators (Sec. 3.1), ensuring realistic motion under various conditions.

To further enable our layered representation to model detailed avatar appearance, including body, hair, and garment, we propose a 3D Gaussian Splatting (3DGS)-based appearance model that builds on top of the layered representation. In particular, we attach 3D Gaussians to the layered geometry of body, garment, and hair, while customizing the Gaussians for each body part to capture the unique structural and dynamic properties of the part. These 3D Gaussians are optimized using Score Distillation Sampling (SDS) [63, 88], allowing us to refine appearance details in a way that leverages powerful diffusion priors (Sec. 3.3). As a result, the generated avatars not only look more realistic but also maintain high fidelity during animation. An overview of our framework is shown in Fig. 2.

3.1. Preliminaries

Garment simulation. In this work, we utilize a neural simulator (i.e., HOOD [30]) to produce a simulated garment mesh sequence. Given a garment mesh g_0 and the target body pose sequence $\mathcal{P}$, HOOD first computes the

SMPL body mesh corresponding to the SMPL parameters. By treating the body meshes as obstacles, HOOD utilizes a GNN to predict the physical status (e.g., position, velocity, etc.) of each vertex on the garment mesh and produce a simulated garment sequence denoted as $\mathcal{G} = \{g_1, ..., g_n\}$. More details are discussed in [30].

Strand-based hair representation. In this work, we choose strand-based hair representation to capture the thin and layer-wise structure of hairs. We represent hair as a point cloud $h_0 \in \mathbb{R}^{N_s \times N_l \times 3}$ that is composed of N_s hair strands, with each strand includes N_l line segments.

Hair simulation. Given hair strands h_0, a target body mesh sequence $\mathcal{P}$ and the simulated garment sequence $\mathcal{G}$, we treat both the body and garment in each time step as obstacles and leverage a hair simulator [8, 9, 22, 24] to produce the animated hair strand sequence denoted as $\mathcal{H} = \{h_1, ..., h_n\}$. This sequence provides a strong prior that drives the 3D Gaussians attached to the strands to produce high-fidelity hair motion, as will be discussed in Sec. 3.3.

3.2. Text-based garment diffusion model

To ensure compatibility with existing simulators, in this work, we choose to use meshes as the coarse garment geometry representation. A straightforward approach to modify the garment mesh to match the given text prompt is through SDS-based optimization [51, 82]. However, optimizing the non-watertight garment mesh in a zero-shot manner is a non-trivial task. On one hand, the topology of garment mesh is hard to change with noisy supervision from foundational diffusion models. On the other hand, the garment simulators all require a clean, compact, and smooth garment mesh as input. To resolve this issue, inspired by recent successes in 3D generative models [86, 93, 98], we utilize large-scale garment datasets [10, 47] and learn a text-based diffusion model for garment mesh generation. Specifically, we first train a Variational Auto-encoder (VAE) to learn a compact latent space capturing garment geometry distribution. We then learn a conditional latent diffusion model within the latent space to generate garment meshes from text prompts.

VAE for garment mesh reconstruction. We adopt the vectorized latent space representation [86, 93, 98] for its simplicity and efficiency. As shown in Fig. 3 (a), given an input garment mesh, we first uniformly sample a set of points (denoted as $X \in \mathbb{R}^{10000 \times 3}$). Next, we encode the sampled points to a set of vectors (denoted as $Z \in \mathbb{R}^{512 \times 16}$) by applying cross-attention layers to X and its downsampled version. To effectively reconstruct garment meshes from latent vectors, we adopt the Unsigned Distance Field
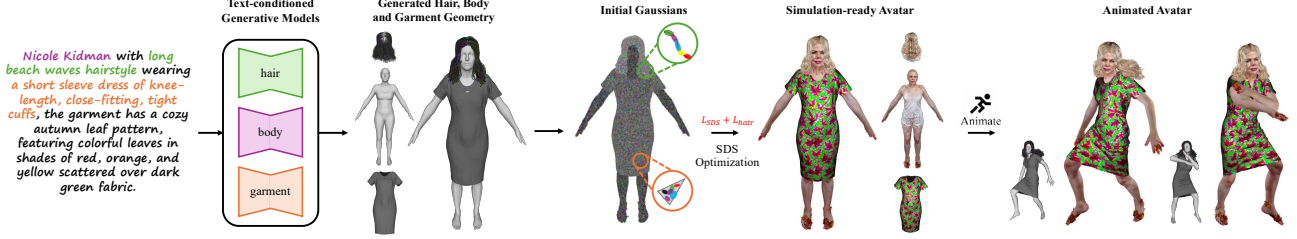
Figure 2. **Overview.** Given a text prompt, SimAvatar first generates hair strands, body mesh, and garment using their respective text-conditioned generative models. We then bring them together using physics simulation, and attach 3D Gaussians to learn their appearances and to adapt them according to the text prompts. We assign one 3D Gaussian to each line segment in the hair strands with their length significantly larger than their diameter (green circle), whereas the Gaussians for meshes are defined within a local coordinate system of each face (orange circle). We optimize the properties of all 3D Gaussians through the Score Distillation Sampling (SDS) loss using image-based diffusion models, and a novel regularization L_{hair} for hairs to ensure plausible hair structure. The generated 3D avatar can be simulated by any pose sequence showing realistic and dynamic motion effects such as flowing hairs and garment wrinkles.
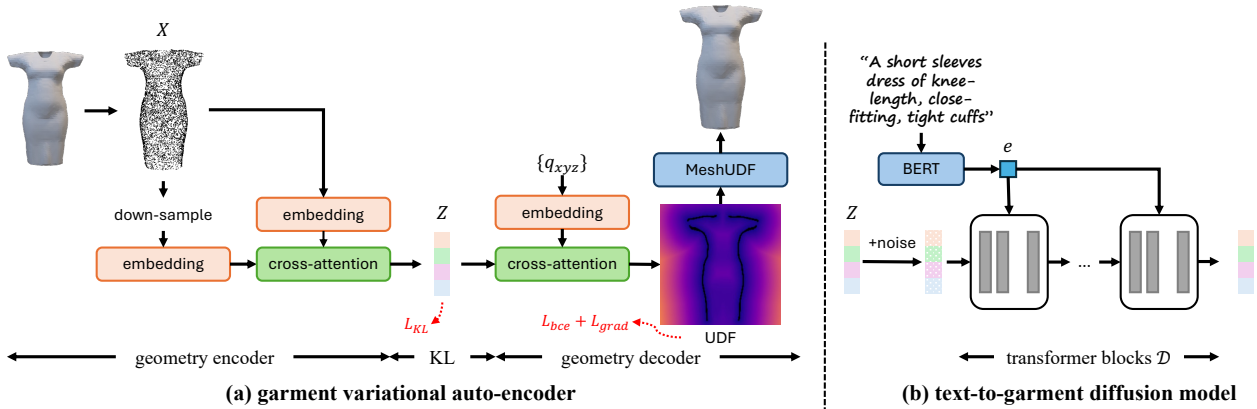

Figure 3. **Text-based garment diffusion model.** See Sec. 3.2 for details.

(UDF) representation [32]. The UDF is designed for non-watertight meshes and uses the gradient at each point to determine the presence of mesh surface. Specifically, we feed query points ($\{q_{xyz}\}$) uniformly sampled from a 3D grid and the latent code Z to few cross-attention layers to predict the Unsigned Distance for all $\{q_{xyz}\}$, which forms the UDF for the garment. The garment mesh can then be extracted from the predicted UDF using MeshUDF [32]. To train the VAE model, we follow prior works [17, 25, 32] and apply: a) a binary cross-entropy loss (L_{bce}) to the predicted Unsigned Distance, b) the L2 loss (L_{grad}) to the gradient at each query point and c) the KL divergence loss (L_{KL}) to the latent code Z. The full loss for the VAE training is $L = L_{bce} + \lambda_{grad}L_{grad} + \lambda_{KL}L_{KL}$, where $\lambda_{grad}, \lambda_{KL}$ are empirically set to $0.0001, 0.1$ in our experiments.

Text-based garment diffusion model. Given the learned vectorized latent space, we train a text-conditioned latent diffusion model for garment mesh generation. As shown in Fig. 3(b), for a latent vector Z, we add random noise ϵ sampled from a Normal Distribution to Z and adopt a

transformer-based network (denoted as $\mathcal{D}$) to recover the added noise. To inject text prompts as conditions, we use the BERT model [26] to extract a text embedding $e \in \mathrm{R}^{768}$ and fuse it with latent features through cross-attention layers. The objective to train the diffusion model follows the EDM [43, 93] and can be formulated as:

$$L(\theta_{\mathcal{D}}) = \mathbb{E}_{\epsilon \in \mathcal{N}(0, \sigma^2 \mathbf{I})} \| \mathcal{D}(Z + \epsilon, \sigma, e)) - Z \|_2^2 \quad (1)$$

where $\theta_{\mathcal{D}}$ denotes the trainable parameters of the denoiser network $\mathcal{D}$ and σ is a randomly chosen noise level. During inference, a randomly sampled noise vector is gradually denoised to a latent code Z matching the given text prompt, which is then decoded to a garment mesh using the decoder in the VAE model introduced earlier.

3.3. Gaussian Avatar Optimization

To further obtain reasonable geometry of hair strands and body shape, we utilize HAAR [76] which is a text-conditioned diffusion model, and the BodyShapeGPT [6] which is a GPT based LLM model to predict hair strands and SMPL shape parameters from the given text. Given the garment mesh, hair strands, and SMPL body mesh, we

attach 3D Gaussians to each component and optimize the attributes of these Gaussians using foundational diffusion models. We choose 3D Gaussians for texture modeling for two reasons. First, compared to implicit representations such as SDF, 3D Gaussians are more flexible for animation and can be easily driven by meshes [64, 88]. Second, 3D Gaussians have shown impressive capacity to capture high-fidelity appearance, especially when combined with diffusion models priors [54, 88]. Next, we discuss how to customize 3D Gaussians for each body part to ensure they maximally conform to the structure and motion of that part.

3D Gaussians on body and garment meshes. We use a similar strategy to combine 3D Gaussians with the body and garment part since their coarse geometry is both represented as 3D meshes. For the body mesh, we utilize a SMPL [55] mesh $\Omega = \text{LBS}(\theta, \beta)$, where θ and β are the SMPL pose and shape parameters, and LBS is the linear blend skinning function. The garment mesh is generated and simulated as discussed in Sec. 3.2 and Sec. 3.1. Below we discuss how to attach and deform a 3D Gaussian to a mesh $\mathcal{M}$, which could be a body or garment mesh. Specifically, we associate each Gaussian $G_i = \{\mu_i, r_i, s_i, f_i, o_i\}$ with a face of the mesh $\mathcal{M}$ and define its position $\mu_i \in \mathbb{R}^3$, rotation $r_i \in \mathbb{R}^3$, and scaling $s_i \in \mathbb{R}^3$ in the face's local coordinate, as well as its color features $f_i \in \mathbb{R}^{d_c}$ and opacity o_i, where d_c is the dimension of the Spherical Harmonic coefficients. The face's coordinate $\{P(\theta), R(\theta), k\}$ is defined similarly to [64], where the origin $P(\theta) \in \mathbb{R}^3$ is computed as the mean position of the face vertices, and the rotation matrix $R(\theta) \in \mathbb{R}^{3 \times 3}$ is formed by concatenating one edge vector of the face, the normal vector, and their cross product. We also compute a scalar k by the mean length of the edges. The global Gaussian position, rotation, and scale $\{\hat{\mu}_i, \hat{r}_i, \hat{s}_i\}$ can be computed by applying the local-to-global transform:

$$
\begin{aligned}
\hat{\mu}_i(\theta) &= kR(\theta) \cdot p_i + P(\theta) \\
\hat{r}_i(\theta) &= R(\theta) \cdot r_i \\
\hat{s}_i(\theta) &= ks_i
\end{aligned} \tag{2}
$$

We initialize the 3D Gaussians by uniformly sampling points on the mesh surface, the face correspondences are maintained throughout the Gaussian densification process [44]. As observed in [88], directly optimizing Gaussian attributes with high-variance diffusion prior-based objective leads to noisy avatars. Thus, we adopt an implicit field $\mathcal{F}_\phi$ with parameters ϕ to model the Gaussian attributes. Specifically, we query the color features f_i, opacity o_i of each Gaussian using its global position $\hat{p}_i(\tilde{\theta})$ under a canonical pose $\tilde{\theta}$ by $(f_i, o_i) = \mathcal{F}_\phi(\hat{\mu}_i(\tilde{\theta}))$. Note that we learn two separate implicit fields for the body $\mathcal{F}_\phi^b$ and garment $\mathcal{F}_\phi^g$ to prevent texture entanglement. The canonical garment mesh is a garment draped on the SMPL body in T-pose. As a re-

sult, the 3D Gaussians attached to the body or garment mesh can be smoothly driven by the avatar's motion by Eq. 2.

3D Gaussians on hair strands. Different from the body and garment meshes, hair strands have a more delicate and complex structure. As a result, the Gaussians associated with the strands are expected to have a long and thin shape. Inspired by existing works that adopt 3D Gaussians for hair reconstruction [56, 91], we assign one 3D Gaussian to each line segment in each hair strand. As shown in the green circle in Fig. 2, for a line segment defined by two endpoints l_i and l_{i+1}, we assign a Gaussian denoted as $G_i = \{\mu_i, r_i, s_i, f_i, o_i\}$ to model the segment. To enforce the Gaussians to be consistent with the hair strands, we explicitly compute the position, rotation, and scale of G_i by:

$$
\begin{aligned}
\mu_i &= (l_i + l_{i+1})/2 \\
s_i &= [\|l_{i+1} - l_i\|/2, \gamma, \gamma] \\
r_i &= [1 + \mu_i \cdot d_i, u \times d_i],
\end{aligned} \tag{3}
$$

where d_i is the direction of a hair strand, and u are unit vectors along the x-axis. We set $\gamma = 0.001$ to keep the hair Gaussians thin. To learn opacity $\{o_i\}$ and color $\{f_i\}$ for the hair Gaussians, we similarly train an implicit field $\mathcal{F}_\phi^h$ as discussed in the garment and body part above. During inference, the hair Gaussians can be driven by the hair strands smoothly and consistently using Eq. 3.

Shading Model. Lighting plays an important role in modeling appearance details in motion such as garment wrinkles. To encourage the 3D Gaussians to capture pose-independent albedo without baked-in shading, we incorporate a Phong shading model [82] into our pipeline. Since the normal for each Gaussian is noisy, we instead use the normal of its corresponding face (denoted as n_p) in the lighting model. To mimic random lighting, we sample the point light position $l_p \in \mathbb{R}^3$, color $l_c \in \mathbb{R}^3$, as well as an ambient light color $l_a \in \mathbb{R}^3$. The shaded color of each 3D Gaussian can then be computed by $f_i' = f_i \cdot (\max(0, n_p \cdot (l_p - \hat{\mu}_i)/\|l_p - \hat{\mu}_i\|) \cdot l_c + l_a)$, where $\hat{\mu}_i$ is the global coordinate of the Gaussian computed by Eq. 2.

Gaussian avatar optimization. To learn the implicit fields for the hair, body, and garment parts, we utilize the Score Distillation Sampling (SDS) [63] objective with a pre-trained text-to-image diffusion model [1] to supervise the Gaussian splatting-based rendering $I(\eta)$ of our model, where η denotes all learnable parameters *i.e.*, $\eta = \{\mathcal{F}_\phi^b, \mathcal{F}_\phi^g, \mathcal{F}_\phi^h\}$. Given a text prompt $\mathcal{T}$ and the noise prediction $\hat{\epsilon}(I_t; \mathcal{T}, t)$ of the diffusion model, SDS optimizes parameters η by minimizing the difference between the noise ϵ added to the rendered image $I(\eta)$ and the predicted noise

Black Teenager girl with afro hairstyle short hair wearing a short sleeve shirt of cropped length, buttonup front, relaxed fit, and midi length skirt with high-waisted, the garment has a vintage floral wallpaper pattern, featuring roses and vines in soft, muted tones.

David Beckham with Pompadour hairstyle wearing a long sleeve sweater, round neckline and Shorts, midthigh length, the sweater features a classic holiday pattern with snowflakes, reindeer, and evergreen trees in traditional colors like red, green, and white.

Figure 4. **Qualitative comparison with the state-of-the-art methods.** See the supplementary video for more comparisons.

$\hat{\epsilon}$ by the diffusion model:

$$\nabla_\eta \mathcal{L}_{\text{SDS}} = E_{t,\epsilon} \left[w(t)(\hat{\epsilon}(I_t; \mathcal{T}, t) - \epsilon) \frac{\partial I}{\partial \eta} \right], \qquad (4)$$

where t is the noise level, I_t is the noised image, and $w(t)$ is a weighting function. Additionally, to prevent broken hairs caused by transparent Gaussians in the middle of a hair strand, we propose a regularization on the opacity of hair Gaussians. Our key intuition is that along a hair strand, the Gaussians closer to the scalp (e.g., hair roots) should have larger opacity values than the Gaussians further away from the scalp (e.g., hair ends). This way the optimization process is able to trim the extra hair to match the prompt while avoiding the broken hair artifacts. Specifically, given a hair point cloud $h_0 \in \mathrm{R}^{N_s \times N_l \times 3}$ as discussed in Sec. 3.1, we sample the opacity ($o \in \mathrm{R}^{N_s \times N_l}$) of its associated Gaussians from the learned implicit fields (i.e., $\mathcal{F}_\phi^h$ discussed in Sec. 3.3) and add the following regularization:

$$L_{hair} = \frac{1}{N_s N_l} \sum_{i=1}^{N_s} \sum_{j=2}^{N_l} (o_{i,j-1} - o_{i,j}) \qquad (5)$$

Our final objective is $L = L_{SDS} + \lambda_{hair} L_{hair}$, where λ_{hair} is empirically set to 1.0 in our experiments.

4. Experiments

4.1. Datasets.

To learn the text-based garment diffusion model discussed in Sec. 3.2, we use the Garment Pattern Generator (GPG) dataset [48] and the CLOTH3D [10] dataset processed by [25]. For text prompts, we leverage the prompt annotations for the GPG dataset from [33]. Since the CLOTH3D dataset does not include paired prompt annotation, we render each garment on top of the SMPL body mesh and query the GPT4v [2] with predefined questions to generate a prompt for each garment mesh, describing its type, shape, length, and width. Together we build a dataset containing around 20000 meshes with paired prompt annotations. The dataset covers common garment types such as t-shirts, tank tops, jackets, shorts, pants, skirts, dresses, etc.

4.2. Qualitative Evaluations.

We compare our synthesized avatars with start-of-the-art text-driven 3D avatar generation methods, including Fantasia3D [18], TADA [51] and GAvatar [88]. First, we present qualitative comparisons of synthesized avatars in a canonical pose in Fig. 4. Our method not only captures diverse garment geometry (e.g., skirts, tops, and bottoms) but also produces superior texture compared to baseline models. Next, we demonstrate animated avatars in Fig. 5. Note that since

Ours **TADA** **GAvtar**

Emma Watson with shoulder-length wavy hair wearing a sleeveless dress of cocktail-length, the garment has a bold leopard print, with classic black and brown spots on a beige background.

Sandra Bullock with long shag hairstyle wearing a sleeveless tunic dress of normal length, the garment has a painterly floral watercolor pattern, featuring soft washes of pink, yellow, and lavender flowers.
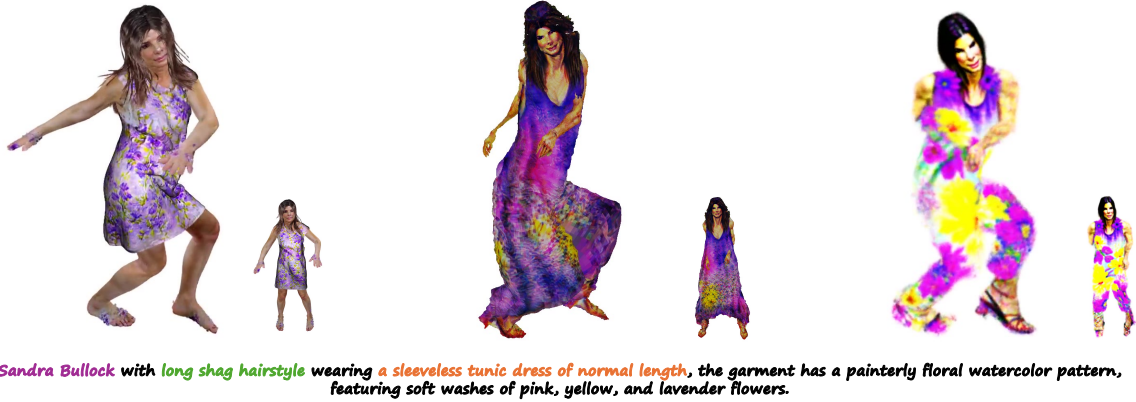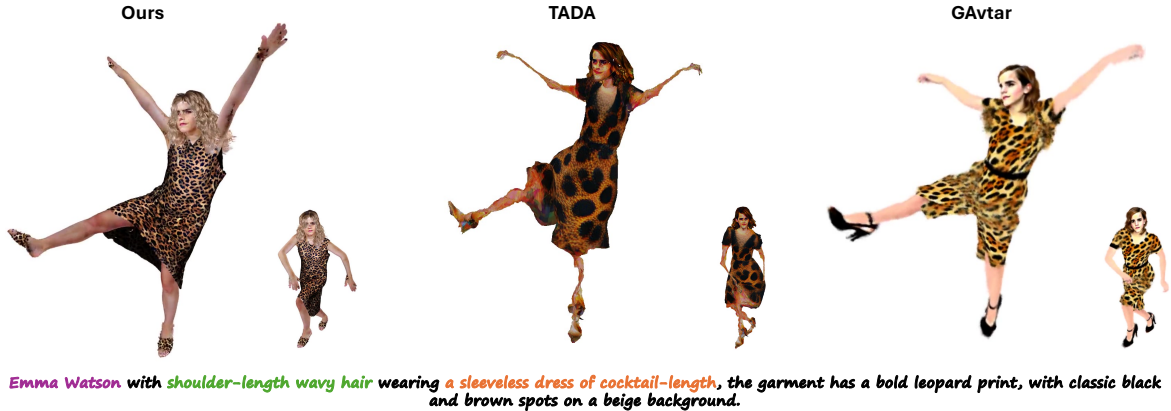
Figure 5. **Qualitative comparison of animated avatars.** See the supplementary video for more animation comparisons.

Fantasia3D utilizes implicit representations, it cannot be readily animated. GAvatar and TADA model the human body and garment by a single geometry representation. For animation, they attach garments to the human body and use linear blending skinning for deformation. As a result, they fail to produce realistic motion for loose garments such as dresses. As shown by the first row of Fig. 5, when the avatar raises her legs, only our model produces physically plausible motion for the dresses, whereas both baselines undesirably separate the dresses into two pieces. We provide abundant animation comparisons in the supplementary video to showcase the difference.

4.3. Quantitative Evaluations.

Metrics	TADA	Fantasia3D	GAvatar
Appearance preference	89.55	100	87.03
Motion preference	91.87	100	94.47

Table 1. **User study.** We provide the preference percentage of our method against state-of-the-art methods. The higher score indicates that more users favor our method. SimAvatar significantly outperforms other methods in terms of user preference. We encourage looking at the supplementary video.

User study. We follow existing works [46, 51, 88] and quantitatively evaluate SimAvatar through an extensive A/B user study. Specifically, we run SimAvatar and baseline methods on 22 prompts. For comparison, we generate a static 360 video of the avatar in A-pose and a video showing the avatar driven by a motion sequence. At each round, we present a pair of results by our method and one of the baselines to a user. For the static 360 video pairs, we ask the user to choose the method with a better appearance. For the animated video pairs, we ask the user to pay more attention to the motion. We collected 540 votes from 18 users and present the results in Table. 1. For both appearance and motion, SimAvatar is consistently favored by users, demonstrating its superior performance compared to baselines.

5. Conclusions and Future Work

In this work, we proposed a novel framework for generating simulation-ready avatars with layered hair and clothing from a text prompt. Unlike existing methods for 3D human avatar generation, our approach represents the human body, hair, and garments in a layered and simulation-ready structure. To this end, we first generated garment mesh, hair strands, and body shape parameters from text-conditioned generative models. We then optimized 3D Gaussians attached to each region to learn realistic appearance. The generated avatars can be seamlessly animated using physics simulators. Both qualitative and quantitative results demonstrate the effectiveness of our method in generating realistic,

animatable clothed avatars, allowing cloth, and hair simulation with dynamic effects such as wrinkles and flowing hair, which surpasses state-of-the-art techniques.

While SimAvatar provides avatars with significantly enhanced realism, there are still areas for further improvements. Currently, our hair and garment generation models are trained on specific datasets, which may bottleneck the diversity due to the used training data. In future work, we aim to explore methods for generating simulation-ready avatars that extend beyond the existing dataset diversity. Additionally, SimAvatar simulates garments and hair sequentially, which can fail in certain cases, such as avatars wearing hoods. Implementing joint simulation for hair and garments would allow us to accommodate more complex garments. Finally, accessories and footwear remain entangled with the body or garment layers; in the future, we plan to focus on generating fully disentangled avatars.

References

[1] Realistic vision. https://huggingface.co/stablediffusionapi/realistic-vision-51, 2023. 2, 6

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 3

[4] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, 2019.

[5] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 3

[6] Baldomero R Árbol and Dan Casas. Bodyshapegpt: Smpl body shape manipulation with llms. In *ECCVW*, 2024. 5

[7] David Baraff and Andrew Witkin. Large steps in cloth simulation. In *Conference on Computer Graphics and Interactive Techniques*, 1998. 3

[8] Miklós Bergou, Max Wardetzky, Stephen Robinson, Basile Audoly, and Eitan Grinspun. Discrete elastic rods. In *SIGGRAPH*, 2008. 4

[9] Miklós Bergou, Basile Audoly, Etienne Vouga, Max Wardetzky, and Eitan Grinspun. Discrete viscous threads. *ACM Transactions on graphics (TOG)*, 2010. 4

[10] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Cloth3d: clothed 3d humans. In *ECCV*, 2020. 3, 4, 7

[11] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. PBNS: Physically based neural simulator for unsupervised garment pose space deformation. *ACM Transaction on Graphics (TOG)*, 2020.

[12] Hugo Bertiche, Meysam Madadi, Emilio Tylson, and Sergio Escalera. Deepsd: Automatic deep skinning and pose space deformation for 3d garment animation. In *ICCV*, 2021. 3

[13] Kiran S Bhat, Christopher D Twigg, Jessica K Hodgins, Pradeep Khosla, Zoran Popovic, and Steven M Seitz. Estimating cloth simulation parameters from video. *SIGGRAPH*, 2003. 3

[14] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *ICCV*, 2019. 3

[15] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *NeurIPS*, 2020. 3

[16] 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 *CVPR*, 2024. 2, 3

[17] Honghu Chen, Yuxin Yao, and Juyong Zhang. Neural-ABC: Neural parametric models for articulated body with clothes. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3, 5

[18] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation. In *ICCV*, 2023. 2, 4, 7, 13

[19] Kwang-Jin Choi and Hyeong-Seok Ko. Stable but responsive cloth. In *ACM SIGGRAPH Courses*. 2005. 3

[20] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 3

[21] Dalle2. https://openai.com/dall-e-2, 2022. 2

[22] Gilles Daviet. Simple and scalable frictional contacts for thin nodal objects. *ACM Transactions on Graphics (TOG)*, 2020. 4

[23] Gilles Daviet. Interactive hair simulation on the gpu using admm. In *SIGGRAPH*, 2023. 2, 3, 4

[24] Gilles Daviet. Interactive hair simulation on the gpu using admm. In *SIGGRAPH*, 2023. 4

[25] Luca De Luigi, Ren Li, Benoit Guillard, Mathieu Salzmann, and Pascal Fua. DrapeNet: Garment Generation and Self-Supervised Draping. In *CVPR*, 2023. 3, 5, 7

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[27] Junting Dong, Qi Fang, Zehuan Huang, Xudong Xu, Jingbo Wang, Sida Peng, and Bo Dai1. TELA: Text to layer-wise 3d clothed human generation. In *ECCV*, 2024. 2, 3

[28] Yun (Raymond) Fei, Henrique Teles Maia, Christopher Batty, Changxi Zheng, and Eitan Grinspun. A multi-scale model for simulating liquid-hair interactions. *ACM Transaction on Graphics (TOG)*, 2017. 3

[29] Jun Gao, Wenzheng Chen, Tommy Xiang, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Learning deformable tetrahedral meshes for 3d reconstruction. In *NeurIPS*, 2020. 3

[30] Artur Grigorev, Bernhard Thomaszewski, Michael J Black, and Otmar Hilliges. HOOD: Hierarchical graphs for generalized modelling of clothing dynamics. In *CVPR*, 2023. 2, 3, 4

[31] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *SIGGRAPH*, 2012. 3

[32] Benoit Guillard, Federico Stella, and Pascal Fua. Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *ECCV*, 2022. 3, 5

[33] Kai He, Kaixin Yao, Qixuan Zhang, Jingyi Yu, Lingjie Liu, and Lan Xu. Dresscode: Autoregressively sewing and generating garments from text guidance. *ACM Transactions on Graphics (TOG)*, 2024. 7

[34] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. AvatarCLIP: Zero-Shot Text-Driven Generation and Animation of 3D Avatars. *SIGGRAPH*, 2022. 3

[35] Jerry Hsu, Tongtong Wang, Zherong Pan, Xifeng Gao, Cem Yuksel, and Kui Wu. Sag-free initialization for strand-based hybrid hair simulation. *ACM Transaction of Graphics*, 2023. 3

[36] Shoukang Hu, Fangzhou Hong, Tao Hu, Liang Pan, Haiyi Mei, Weiye Xiao, Lei Yang, and Ziwei Liu. Humanliff: Layer-wise 3d human generation with diffusion model. *arXiv preprint*, 2023. 2, 3

[37] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, and Ying Feng. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *CVPR*, 2024. 2

[38] 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. In *NeurIPS*, 2023. 3

[39] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *3DV*, 2024. 2

[40] Nikolay Jetchev. ClipMatrix: Text-controlled creation of 3d textured meshes. *arXiv preprint arXiv:2307.05663*, 2023. 2

[41] Boyi Jiang, Juyong Zhang, Yang Hong, Jinhao Luo, Ligang Liu, and Hujun Bao. Bcnet: Learning body and cloth shape from a single image. In *ECCV*, 2020. 3

[42] 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 *ICCV*, 2023. 2, 3

[43] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 5

[44] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *SIGGRAPH*, 2023. 6

[45] Taeksoo Kim, Byungjun Kim, Shunsuke Saito, and Hanbyul Joo. GALA: Generating animatable layered assets from a single scan. In *CVPR*, 2024. 3

[46] Nikos Kolotouros, Thiemo Alldieck, Andrei Zanfir, Eduard Gabriel Bazavan, Mihai Fieraru, and Cristian Sminchisescu. Dreamhuman: Animatable 3d avatars from text. In *NeurIPS*, 2023. 2, 3, 8

[47] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3d garments with sewing patterns. In *NeurIPS*, 2021. 4

[48] Maria Korosteleva and Sung-Hee Lee. Generating datasets of 3d garments with sewing patterns. In *NeurIPS*, 2021. 7

[49] Yifei Li, Tao Du, Kui Wu, Jie Xu, and Wojciech Matusik. Diffcloth: Differentiable cloth simulation with dry frictional contact. *SIGGRAPH*, 2022. 3

[50] Junbang Liang, Ming C. Lin, and Vladlen Koltun. Differentiable cloth simulation for inverse problems. In *NeurIPS*, 2019. 3

[51] Tingting Liao, Hongwei Yi, Yuliang Xiu, Jiaxiang Tang, Yangyi Huang, Justus Thies, and Michael J Black. TADA! text to animatable digital avatars. In *3DV*, 2024. 2, 3, 4, 7, 8, 13

[52] 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 *CVPR*, 2023. 2

[53] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. *arXiv preprint arXiv:2404.01291*, 2024. 13

[54] 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 *CVPR*, 2024. 2, 6, 13

[55] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ToG*, 34(6):248:1–248:16, 2015. 2, 3, 4, 6

[56] Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483*, 2024. 3, 6

[57] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. Learning to dress 3d people in generative clothing. In *CVPR*, 2020. 3

[58] 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. In *ECCV*, 2020. 2

[59] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *SIGGRAPH*, 2022. 3

[60] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *CVPR*, 2020. 3

[61] Emmanuel Piuze, Paul G. Kry, and Kaleem Siddiqi. Generalized helicoids for modeling hair geometry. *Computer Graphics Forum*, 2011. 3

[62] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. ClothCap: Seamless 4d clothing capture and retargeting. *SIGGRAPH*, 2017. 3

[63] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2, 3, 4, 6

[64] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *CVPR*, 2024. 6

[65] Li Ren, Benoit Guillard, Edoardo Remelli, and Pascal Fua. DIG: Draping Implicit Garment over the Human Body. In *ACCV*, 2022. 3

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

[67] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[68] 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. In *NeurIPS*, 2022. 2

[69] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. *Computer Graphics Forum*, 38(2):355–366, 2019. 3

[70] Igor Santesteban, Nils Thuerey, Miguel A. Otaduy, and Dan Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *CVPR*, 2021. 3

[71] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. Self-Supervised Collision Handling via Generative 3D Garment Models for Virtual Try-On. In *CVPR*, 2021. 3

[72] Igor Santesteban, Miguel A Otaduy, and Dan Casas. SNUG: Self-Supervised Neural Dynamic Garments. In *CVPR*, 2022. 3

[73] Igor Santesteban, Miguel A. Otaduy, Nils Thuerey, and Dan Casas. ULNeF: Untangled layered neural fields for mix-and-match virtual try-on. In *NeurIPS*, 2022. 3

[74] Yuefan Shen, Shunsuke Saito, Ziyan Wang, Olivier Maury, Chenglei Wu, Jessica Hodgins, Youyi Zheng, and Giljoo Nam. CT2Hair: High-fidelity 3d hair modeling using computed tomography. *Transactions on Graphics*, 2023. 3

[75] Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, and Egor Zakharov. Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction. In *ICCV*, 2023. 3

[76] Vanessa Sklyarova, Egor Zakharov, Otmar Hilliges, Michael J. Black, and Justus Thies. Text-conditioned generative model of 3d strand-based human hairstyles. In *CVPR*, 2024. 3, 5

[77] Zhaoqi Su, Tao Yu, Yangang Wang, and Yebin Liu. Deepcloth: Neural garment representation for shape and style editing. *TPAMI*, 2023. 3

[78] 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

[79] B. Thomaszewski, Simon Pabst, and Wolfgang Straßer. Asynchronous cloth simulation. 2008. 3

[80] Pascal Volino, Nadia Magnenat-Thalmann, and Francois Faure. A simple approach to nonlinear tensile stiffness for accurate cloth simulation. *Transaction of Graphics (TOG)*, 2009. 3

[81] Huamin Wang, James F. O'Brien, and Ravi Ramamoorthi. Data-driven elastic models for cloth: Modeling and measurement. In *SIGGRAPH*, 2011. 3

[82] Jionghao Wang, Yuan Liu, Zhiyang Dou, Zhengming Yu, Yongqing Liang, Xin Li, Wenping Wang, Rong Xie, and Li Song. Disentangled clothed avatar generation from text descriptions. *arXiv preprint:2312.05295*, 2023. 2, 3, 4, 6

[83] 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. *NeurIPS*, 34:27171–27183, 2021. 2, 3

[84] Yi Wang, Jian Ma, Ruizhi Shao, Qiao Feng, Yu-Kun Lai, Yebin Liu, and Kun Li. Humancoser: Layered 3d human generation via semantic-aware diffusion model. *arXiv preprint: 2312.05804*, 2023. 3

[85] 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. In *NeurIPS*, 2023. 2

[86] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. *arXiv preprint arXiv:2405.14832*, 2024. 4

[87] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[88] 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 *CVPR*, 2024. 2, 3, 4, 6, 7, 8, 13

[89] Cem Yuksel, Scott Schaefer, and John Keyser. Hair meshes. *ACM Trans. Graph.*, 2009. 3

[90] Ilya Zakharkin, Kirill Mazur, Artur Grigorev, and Victor Lempitsky. Point-based modeling of human clothing. In *CVPR*, 2021. 3

[91] Egor Zakharov, Vanessa Sklyarova, Michael J Black, Giljoo Nam, Justus Thies, and Otmar Hilliges. Human hair reconstruction with strand-aligned 3d gaussians. In *ECCV*, 2024. 3, 6

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

[93] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions on Graphics (TOG)*, 2023. 4, 5

[94] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *CVPR*, 2017. 3

[95] 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. In *AAAI*, 2024. 2, 3

[96] Hao Zhang, Yao Feng, Peter Kulits, Yandong Wen, Justus Thies, and Michael J. Black. TECA: Text-guided generation and editing of compositional 3d avatars. In *3DV*, 2024. 2, 3

[97] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2, 3

[98] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 2024. 4

[99] Xuanmeng Zhang, Jianfeng Zhang, Chacko Rohan, Hongyi Xu, Guoxian Song, Yi Yang, and Jiashi Feng. Getavatar: Generative textured meshes for animatable human avatars. In *ICCV*, 2023. 2

A. Quantitative Evaluations

VQAScore We quantitatively compare our method against baseline methods using the VQAScore [53]. Instead of treating the text prompt as a set of unordered words, the VQAScore assesses alignment between prompts and generated assets by posing targeted questions to foundational models. This enables evaluation of more complex prompts involving multiple entities and relationships, providing results that align more closely with human perceptual evaluation [53] compared to the CLIP score. The VQAScore is particularly well-suited to our task, which involves generating avatars with compositional text prompts specifying different parts (e.g., hair, identity, and garments). As shown in Table 2, our model achieves a significantly higher VQAScore than baseline methods, demonstrating its superior ability to align with input prompts.

CLIP-Score Following prior work, we report the CLIP score of our method and baselines in Table 2. However, it is important to note that the CLIP score has been widely observed to be unreliable for accurately assessing visual quality and alignment with text prompts [53]. In Fig. 7, we additionally present both the VQA-Score (green) and the CLIP score (red) for each avatar. Overall, the VQA-Score demonstrates a closer alignment with human perception. That said, since visual quality is inherently subjective, we encourage readers to examine the figures and accompanying videos for a more holistic evaluation.

Metrics	TADA	Fantasia3D	GAvatar	HG	Ours
VQA score ↑	0.45	0.38	0.44	0.53	**0.75**
CLIP ↑	33.50	37.44	30.74	30.81	33.39

Table 2. **Quantitative Evaluations.** We quantitatively compare the VQAScore [53] and CLIP-score with TADA [51], Fantasia3D [18], GAvatar [88] and HumanGaussians (HG) [54].

B. Implementation Details

Text-to-garment diffusion model training. To generate garment meshes from text prompts, we train the VAE and diffusion model sequentially, as described in Sec 3.2 of the main paper. The VAE is trained with a batch size of 14, while the diffusion model uses a batch size of 24. The entire training process takes approximately one week on four V100 GPUs.

Gaussian avatar learning. We uniformly sample 2.6×10^5 and 10^6 Gaussians on the body and garment mesh surface as the initial Gaussian positions. During optimization, we combine the rendered avatar image with a solid background of random color. We optimize the implicit fields (i.e., $\{\mathcal{F}_\phi^b, \mathcal{F}_\phi^g, \mathcal{F}_\phi^h\}$) with a learning rate of 0.001. The optimization takes around six hours per avatar on a single V100 GPU. To facilitate stable training, we first optimize

the Gaussians attached to inner layer (i.e., hair and body) for 4000 iterations. We then include the garment layer and optimize the avatar for another 6000 iterations.

C. Layer-wise Training Strategy

To facilitate disentanglement of the hair, body and garment, we separately render each layer and pair them with different prompts. For example, given the prompt: "Adele with long layered waves hairstyle wearing a sleeveless dress of tea-length, gathered waist, the garment has a delicate butterfly print, with small, colorful butterflies scattered across the neckline and sleeves.", we render the hair, body, and garment layers individually. The corresponding prompts used are: "long layered waves hairstyle" for the hair, "Adele in a tank top and shorts" for the body, and "a sleeveless tea-length dress with a gathered waist, featuring a delicate butterfly print with small, colorful butterflies scattered across the neckline and sleeves" for the garment. To further disentangle face and hair (e.g., preventing hair textures from appearing in the face region), we render the avatar's head in a zoomed-in view and pair it with the prompt: "Adele with short buzz hair and a bold forehead." This approach significantly reduces the likelihood of hair textures being learned by the body region, as will be demonstrated in Sec. D. In addition to rendering views of the face, body, hair, and garment, we also zoom in on specific parts such as the hands, feet, lower body, and upper body to enhance the quality of the avatar's details in these regions.

D. Ablation Studies

In this section, we evaluate the effectiveness of different modules and present qualitative results in Fig. 6.

Hair constraint. To address the issue of broken hairs during Gaussian avatar optimization, we introduced a hair constraint, as described in Sec. 3.3 of the main paper. As illustrated in Fig. 6(a)(b), avatars trained without the hair constraint exhibit broken hairs and floating Gaussians around the head. These results highlight the effectiveness of the hair constraint in producing cohesive and realistic hair representations.

Prompt engineering. To achieve complete disentanglement of hair, body, and garment, we propose optimizing each layer separately while pairing the optimization with distinct prompts. Fig. 6(c)(d) compares the results of using the prompt "buzz cut, bold forehead" for face-view optimization versus not using it. Without this prompt, the model generates hair textures on the body, hindering the full disentanglement of the different parts.
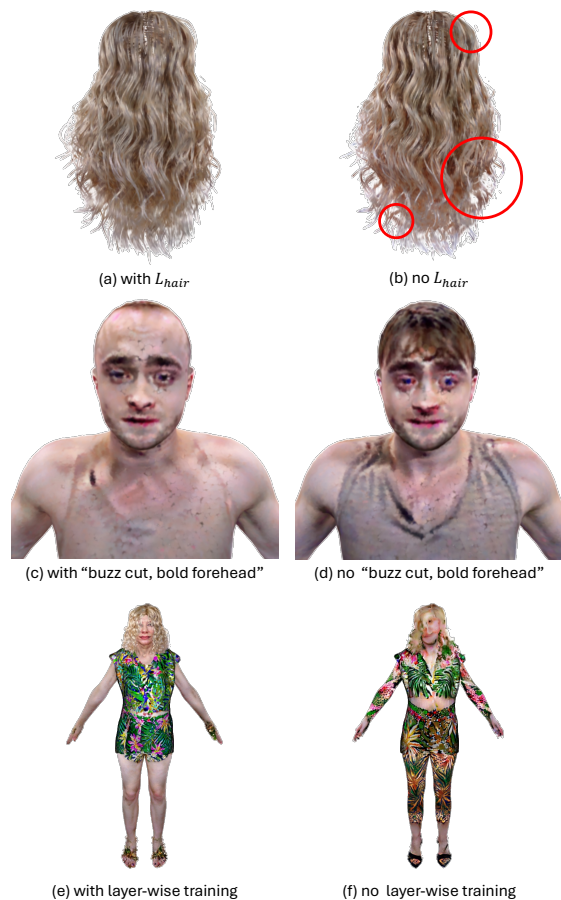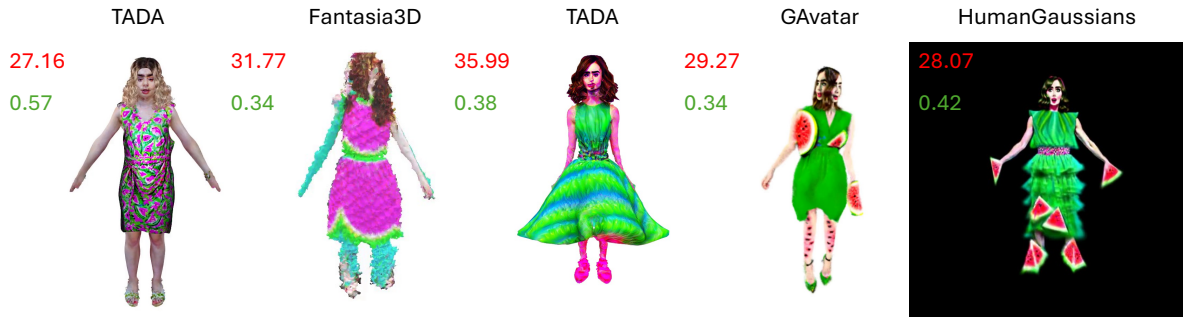
(a) with L_{hair} (b) no L_{hair}

(c) with "buzz cut, bold forehead"   (d) no "buzz cut, bold forehead"

(e) with layer-wise training   (f) no layer-wise training

Figure 6. **Ablation Studies.** See Sec. D for details.

Disentangled training strategy. As outlined above, during training, we optimize each layer (i.e., hair, body, and garment) separately using distinct prompts. To validate the effectiveness of this training strategy, we conducted an experiment where only the full avatar was rendered and optimized by the SDS loss without layer separation. As shown in Fig. 6(e)(f), this approach fails to produce a coherent avatar and results in entanglement between body and garments (e.g., garment textures appearing on the body).

E. More Qualitative Results

We present comparisons with baselines in Fig. 7 and Fig. 8. For a more comprehensive evaluation of motion, we strongly encourage readers to refer to the accompanying video. Additionally, in Fig. 9 and Fig 10, we provide further results of our method, illustrating the geometry and texture of each individual layer, including hair, face, garment, body, and the full avatar.

	TADA	Fantasia3D	TADA	GAvatar	HumanGaussians

27.16   0.57   31.77   0.34   35.99   0.38   29.27   0.34   28.07   0.42

Lily Collins with long layered waves hairstyle wearing a sleeveless dress with cinched waist, knee-length, the garment has a playful watermelon pattern, with vibrant pink watermelon slices and green rinds on a light blue fabric.

29.74   0.83   34.18   0.27   32.39   0.35   34.57   0.38   31.77   0.52
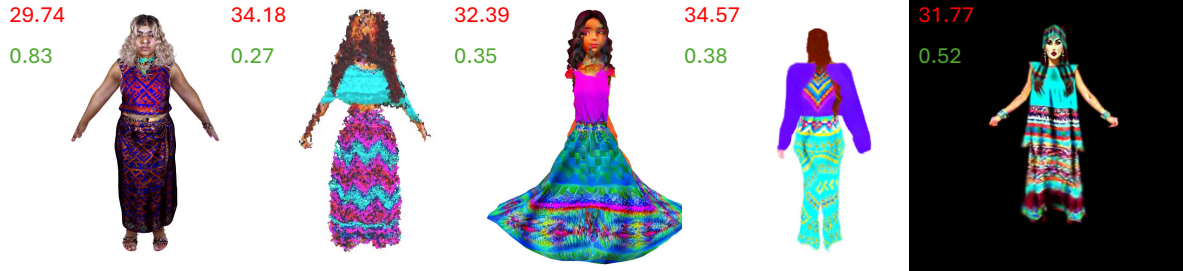

Young Latino female Teenager with long layered waves hairstyle wearing a sleeveless top with wide neckline, flared waist and maxi skirt, high-waisted, the garment has an intricate Aztec-inspired pattern, featuring geometric shapes in rich tones of burgundy and turquoise.

32.77   0.86   32.93   0.23   40.89   0.43   37.86   0.37   39.39   0.64

Middle-Aged Asian Woman with straight lob hairstyle wearing a short sleeve tunic of long length and midi skirt of mid-calf length, A-line silhouette, the garment has a botanical fern pattern, featuring lush green ferns cascading down the fabric.

32.02   0.82   29.62   0.31   37.22   0.55   25.48   0.30   26.11   0.25

Michelle Yeoh with long straight hair wearing a sleeveless dress of midi length, fitted waist, the garment has a bold tropical bird pattern, featuring parrots and toucans perched among vibrant jungle leaves.

27.96   0.81   27.41   0.42   34.28   0.53   26.38   0.14   28.17   0.54
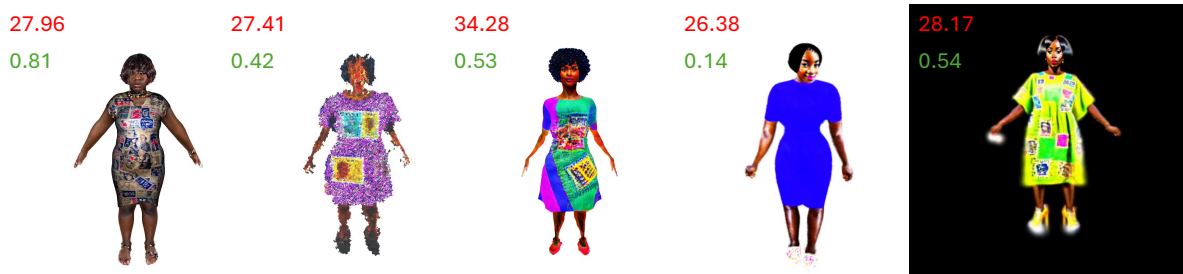
Young Black Woman with bob hairstyle wearing a short sleeve dress of knee length, the garment has a vintage postage stamp print, featuring colorful stamps from around the world.

Figure 7. **More qualitative results.** Red and green numbers indicate CLIP and VQA score, respectively.
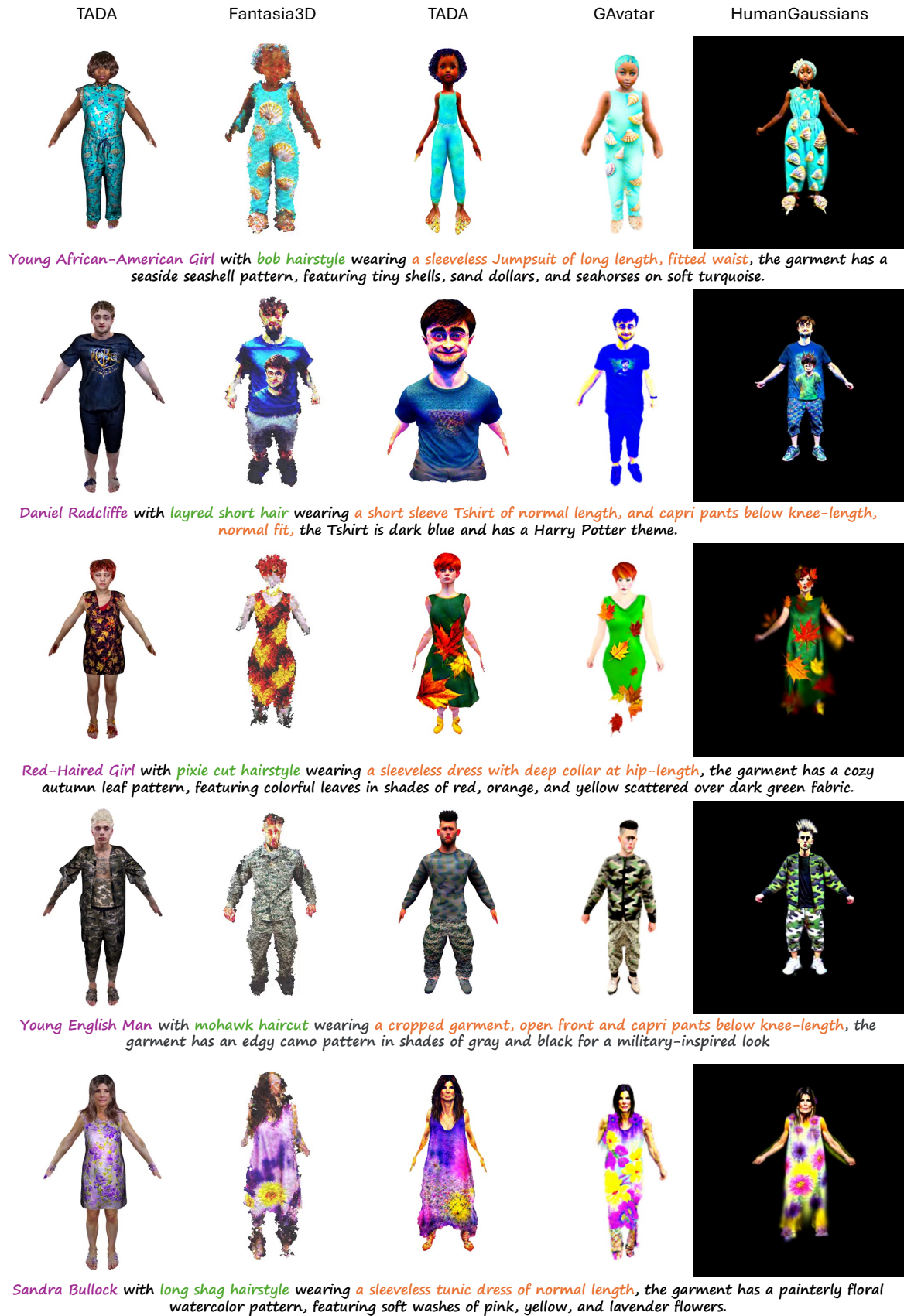
TADA Fantasia3D TADA GAvatar HumanGaussians

Young African-American Girl with bob hairstyle wearing a sleeveless Jumpsuit of long length, fitted waist, the garment has a seaside seashell pattern, featuring tiny shells, sand dollars, and seahorses on soft turquoise.

Daniel Radcliffe with layred short hair wearing a short sleeve Tshirt of normal length, and capri pants below knee-length, normal fit, the Tshirt is dark blue and has a Harry Potter theme.

Red-Haired Girl with pixie cut hairstyle wearing a sleeveless dress with deep collar at hip-length, the garment has a cozy autumn leaf pattern, featuring colorful leaves in shades of red, orange, and yellow scattered over dark green fabric.

Young English Man with mohawk haircut wearing a cropped garment, open front and capri pants below knee-length, the garment has an edgy camo pattern in shades of gray and black for a military-inspired look

Sandra Bullock with long shag hairstyle wearing a sleeveless tunic dress of normal length, the garment has a painterly floral watercolor pattern, featuring soft washes of pink, yellow, and lavender flowers.

Figure 8. **More qualitative results.**

Sandra Bullock with long shag hairstyle wearing a sleeveless tunic dress of normal length, the garment has a painterly floral watercolor pattern, featuring soft washes of pink, yellow, and lavender flowers.

Nicole Kidman with long Beach Waves hairstyle wearing a short sleeve dress of knee-length, close-fitting, tight cuffs, the garment has a cozy autumn leaf pattern, featuring colorful leaves in shades of red, orange, and yellow scattered over dark green fabric.

Lily Collins with long layered waves hairstyle wearing a sleeveless dress with cinched waist, knee-length, the garment has a playful watermelon pattern, with vibrant pink watermelon slices and green rinds on a light blue fabric.

Young African-American Girl with bob hairstyle wearing a sleeveless Jumpsuit of long length, fitted waist, the garment has a seaside seashell pattern, featuring tiny shells, sand dollars, and seahorses on soft turquoise.

Young Latino female Teenager with long layered waves hairstyle wearing a sleeveless top with wide neckline, maxi skirt, high-waisted, the garment has an intricate Aztec-inspired pattern, featuring geometric shapes in rich tones of burgundy and turquoise.

Young white girl with Blunt haircut wearing a long blouse of hiplength, wide garment, round neckline and a knee-length skirt, the garments have a vintage-inspired polka-dot pattern.

Figure 9. **Layer-wise visualization of SimAvatar.**

African-American with voluminous curls hairstyle wearing a lilac cardigan and taupe shorts.

Anne Hathaway with Curtain Bangs hairstyle wearing sleeveless jumpsuit of full-length, the garment has a subtle leaf pattern, where delicate vines run vertically from top to bottom

Black Teenager girl with afro hairstyle short hair wearing a short sleeve shirt of cropped length, buttonup front, and midi length skirt with high-waisted, the garment has a vintage floral wallpaper pattern, featuring roses and vines in soft, muted tones.

Cate Blanchett with curly short hairstyle wearing a sleeveless dress of cropped length, the garment has an exotic jungle print, showcasing palm leaves, birds, and tropical flowers in vibrant colors.

Daniel Radcliffe with layered short hair wearing a short sleeve Tshirt of normal length, and capri pants below knee-length, normal fit, the Tshirt is dark blue and has a Harry Potter theme.

Emma Watson with shoulder-length wavy hair wearing a short sleeve dress of cocktail-length, the garment has a bold leopard print, with classic black and brown spots on a beige background.

Figure 10. **Layer-wise visualization of SimAvatar.**