

# Supplementary of

## *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

### 1. Overview

This document provides additional supporting materials for the main paper. We start with a quantitative comparison of our method against baselines in Sec. 2. Implementation details for both the text-to-garment diffusion model and Gaussian avatar optimization are provided in Sec. 3. Subsequently, we discuss our layer-wise training strategy that encourages body part disentanglement in Sec. 4. Ablation studies and further qualitative results are presented in Sec. 5 and Sec. 6, respectively. *We strongly encourage readers to view the accompanying video for animated avatars.*

### 2. Quantitative Evaluations

**VQAScore** We quantitatively compare our method against baseline methods using the VQAScore [3]. 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 [3] 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 1, 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 1. 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 [3]. In Fig. 2, 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 $\uparrow$ | 0.45  | 0.38       | 0.44    | 0.53  | <b>0.75</b> |
| CLIP $\uparrow$      | 33.50 | 37.44      | 30.74   | 30.81 | 33.39       |

Table 1. **Quantitative Evaluations.** We quantitatively compare the VQAScore [3] and CLIP-score with TADA [2], Fantasia3D [1], GAvatar [5] and HumanGaussians (HG) [4].

### 3. 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 Section 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  $2.6 \times 10^5$  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.

### 4. 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. 5. 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.

## 5. Ablation Studies

In this section, we evaluate the effectiveness of different modules and present qualitative results in Fig. 1.

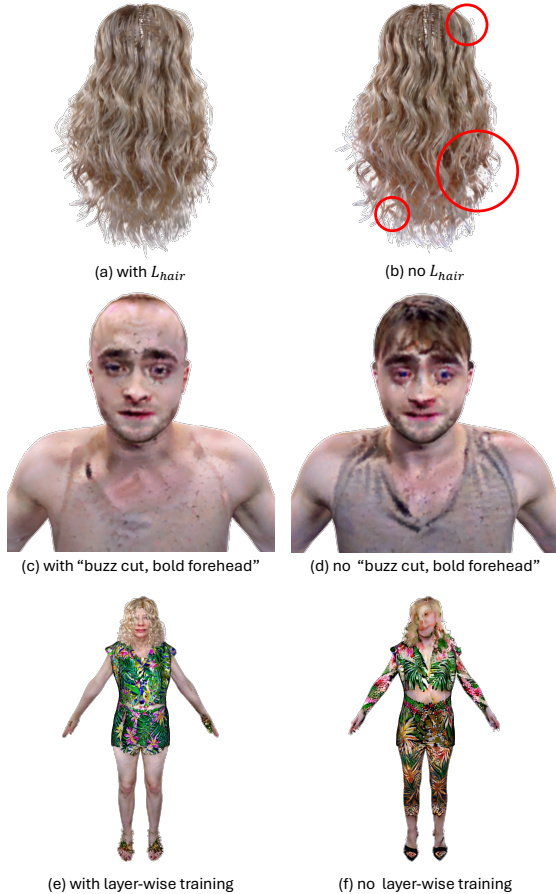


Figure 1. **Ablation Studies.** See Sec. 5 for details.

**Hair constraint.** To address the issue of broken hairs during Gaussian avatar optimization, we introduced a hair constraint, as described in Section 3.3 of the main paper. As illustrated in Fig. 1(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. 1(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.

**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. 1(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).

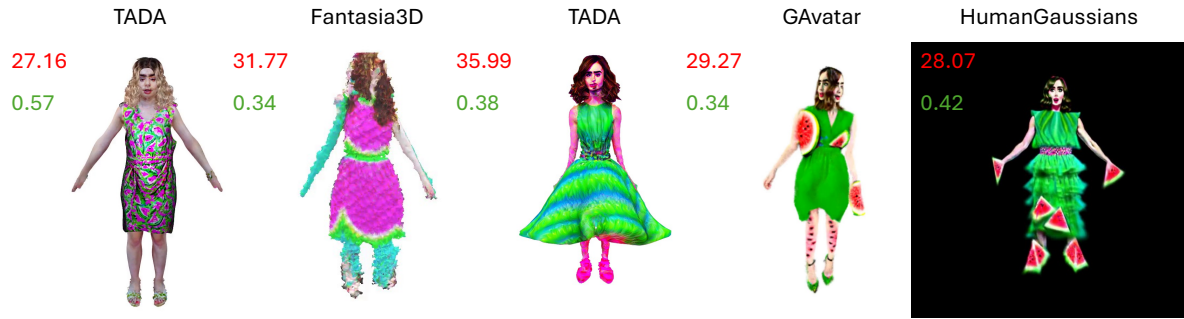
## 6. More Qualitative Results

We present comparisons with baselines in Fig. 2 and Fig. 3. For a more comprehensive evaluation of motion, we strongly encourage readers to refer to the accompanying video. Additionally, in Fig. 4 and Fig 5, 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.

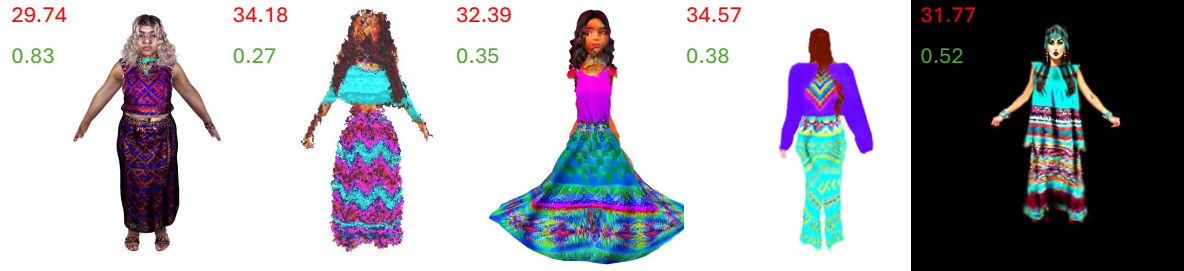
## References

- [1] 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. 1
- [2] 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. 1
- [3] 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. 1
- [4] 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. 1

- [5] 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.



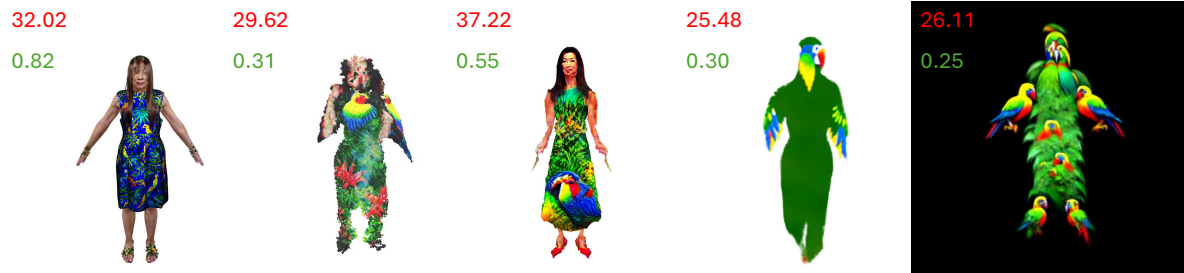
Lily Collins with long layered waves hairstyle wearing a sleeveless dress with cinched waist, knee-length, the garment has a playful watermelon pattern, with vibrant pink watermelon slices and green rinds on a light blue fabric.



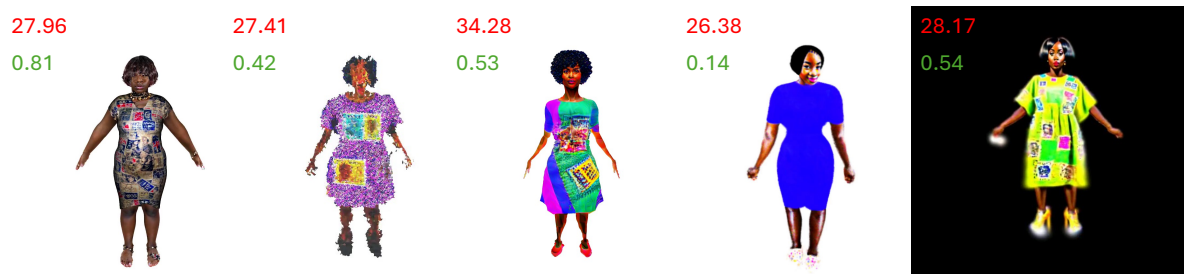
Young Latino female Teenager with long layered waves hairstyle wearing a sleeveless top with wide neckline, flared waist and maxi skirt, high-waisted, the garment has an intricate Aztec-inspired pattern, featuring geometric shapes in rich tones of burgundy and turquoise.



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.



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.



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 2. **More qualitative results.** Red and green numbers indicate CLIP and VQA score, respectively.





Figure 3. 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 hip length, wide garment, round neckline and a knee-length skirt, the garments have a vintage-inspired polka-dot pattern.

Figure 4. 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 5. Layer-wise visualization of SimAvatar.