HairCUP: Hair Compositional Universal Prior for 3D Gaussian Avatars

Supplementary Material

A. Preliminaries: URAvatar [4]

Our method builds on URAvatar [4], a universal 3D avatar model that extends the relightable 3D Gaussian representation of RGCA [9] to multiple subjects. To provide the necessary background, we first describe RGCA [9] for person-specific relightable 3D Gaussian avatars, followed by the multi-subject extension introduced in URAvatar [4].

RGCA [9] proposed a relightable 3D Gaussian head avatar model [9] that learns a latent space of facial expressions using a conditional variational autoencoder (VAE) [3]. The encoder maps an unwrapped UV texture map and tracked mesh vertices to an expression code, which is then used by a set of decoders to generate 3D Gaussian primitives. Given the unwrapped texture map $\mathbf{T}$ and tracked mesh vertices $\mathbf{V}$, the encoder produces the mean μ_e and covariance σ_e of the expression code:

$$\mu_e, \sigma_e = \mathcal{E}(\mathbf{V}, \mathbf{T}; \Theta_e). \tag{1}$$

The decoders reconstruct the tracked mesh vertices $\mathbf{V}$ and generate Gaussian primitives [2], which are splatted [12] to render the avatar. Building on this, URAvatar [4] generalizes the relightable 3D Gaussian avatar to multiple subjects by introducing an identity-conditioned hypernetwork [1]. This hypernetwork, $\mathcal{E}_{\text{id}}$, generates bias maps for avatar decoders and expression-agnostic attributes, given the UV-unwrapped mean albedo and geometry maps of the facial tracked meshes:

$$\Theta^{\text{id}}_{\text{g}}, \Theta^{\text{id}}_{\text{vi}}, \Theta^{\text{id}}_{\text{vd}}, \{o_k, \boldsymbol{\rho}_k\}_{k=1}^{N} = \mathcal{E}_{\text{id}}(\mathbf{T}_{\text{mean}}, \mathbf{G}_{\text{mean}}; \Phi_{\text{id}}), \tag{2}$$

where N is the number of Gaussians, o_k and $\boldsymbol{\rho}_k$ are expression-agnostic opacity and albedo of 3D Gaussians, and $\Theta^{\text{id}}_{\text{g}}, \Theta^{\text{id}}_{\text{vi}}, \Theta^{\text{id}}_{\text{vd}}$ are identity-conditioned bias maps injected into the intermediate feature maps of their respective decoders.

The geometry decoder $\mathcal{D}_{\text{g}}$ predicts tracked mesh vertices:

$$\{\hat{t}_k\}_{k=1}^{N} = \mathcal{D}_{\text{g}}(\boldsymbol{z}, \boldsymbol{e}_{\{\text{l,r}\}}, \boldsymbol{r}_n; \Theta^{\text{id}}_{\text{g}}, \Phi_{\text{g}}), \tag{3}$$

where $\boldsymbol{z}$ is an expression code, $\boldsymbol{e}_{\{\text{l,r}\}}$ are eye gaze direction vectors, and $\boldsymbol{r}_n$ is the axis-angle neck rotation relative to the head. The predicted vertices serve as anchors for Gaussians produced by the appearance decoders. The two Gaussian decoders, $\mathcal{D}_{\text{vi}}$ and $\mathcal{D}_{\text{vd}}$, generate the geometric and appearance attributes needed to evaluate each Gaussian's radiance:

$$\{\delta\boldsymbol{t}_k, \boldsymbol{q}_k, \boldsymbol{s}_k, \boldsymbol{d}_k, \sigma_k\}_{k=1}^{N} = \mathcal{D}_{\text{vi}}(\boldsymbol{z}, \boldsymbol{e}_{\{\text{l,r}\}}, \boldsymbol{r}_n; \Theta^{\text{id}}_{\text{vi}}, \Phi_{\text{vi}}), \tag{4}$$

$$\{\delta\boldsymbol{n}_k, v_k\}_{k=1}^{N} = \mathcal{D}_{\text{vd}}(\boldsymbol{z}, \boldsymbol{e}_{\{\text{l,r}\}}, \boldsymbol{r}_n, \boldsymbol{\omega}_o; \Theta^{\text{id}}_{\text{vd}}, \Phi_{\text{vd}}), \tag{5}$$

where $\delta\boldsymbol{t}_k$ is the position offset, $\boldsymbol{q}_k$ is the orientation, and $\boldsymbol{s}_k$ is the scale of each Gaussian. $\boldsymbol{d}_k$ represents the SH coefficients for color and monochrome components [9], and σ_k is the roughness parameter as defined in Eq. (2) of the main paper. The term $\delta\boldsymbol{n}_k$ denotes the view-dependent delta normal, and v_k represents the visibility term.

To account for eye modeling, URAvatar includes a universal relightable explicit eye model adapted from Saito et al. [9]. The eye hypernetwork $\mathcal{E}_{\text{eye}}$ generates bias maps for the eye Gaussian decoders, ensuring identity-specific adaptation:

$$\Theta^{\text{e}}_{\text{vi}}, \Theta^{\text{e}}_{\text{vd}} = \mathcal{E}_{\text{eye}}(\mathbf{T}_{\text{e}}, \mathbf{G}_{\text{e}}; \Phi^{\text{e}}_{\text{id}}), \tag{6}$$

where $\mathbf{T}_{\text{e}}$ and $\mathbf{G}_{\text{e}}$ correspond to the eye region in the mean texture and geometry maps. The eye Gaussian decoders predict similar attributes as the main avatar decoders, with a unified decoder for the specular visibility map to better preserve eye reflection priors. For further details, we refer readers to the paper [4].

B. Synthetic Bald Image Generation

B.1. Synthetic Bald Image Pairs

To validate the consistency between the original and synthetic bald images used for training, we present example image pairs in Fig. 1. These pairs are constructed using the compositing scheme described in the main paper (Eq. (22)), where the face region is taken from the original image and the occluded scalp region is inpainted with the rendered bald mesh. By carefully processing the hair mask to ensure smooth transitions, our method produces visually coherent bald images across diverse viewpoints and expressions.

B.2. Bald Texture Optimization

Optimization details. We present the details of bald texture optimization (Sec. 3.3). To optimize the bald texture MLP, we use the loss function $\mathcal{L}_{\text{bald}}$ from Eq. (20) in the main paper, running a two-stage optimization over 2500 iterations. For the first 1500 iterations, we apply only the reconstruction loss $\mathcal{L}^{\text{rec}}_{\text{bald}}$ to reconstruct the visible face region. In the next 1000 iterations, we introduce SDS loss [7] to refine the texture in hair-occluded regions while continuing reconstruction loss, using weights $\lambda^{\text{rec}}_{\text{bald}} = 1$ and $\lambda^{\text{sds}}_{\text{bald}} = 10^{-6}$. During the SDS stage, we employ an inpainting image-to-image diffusion model with ControlNet [10], trained on dome-captured human images [5]. For

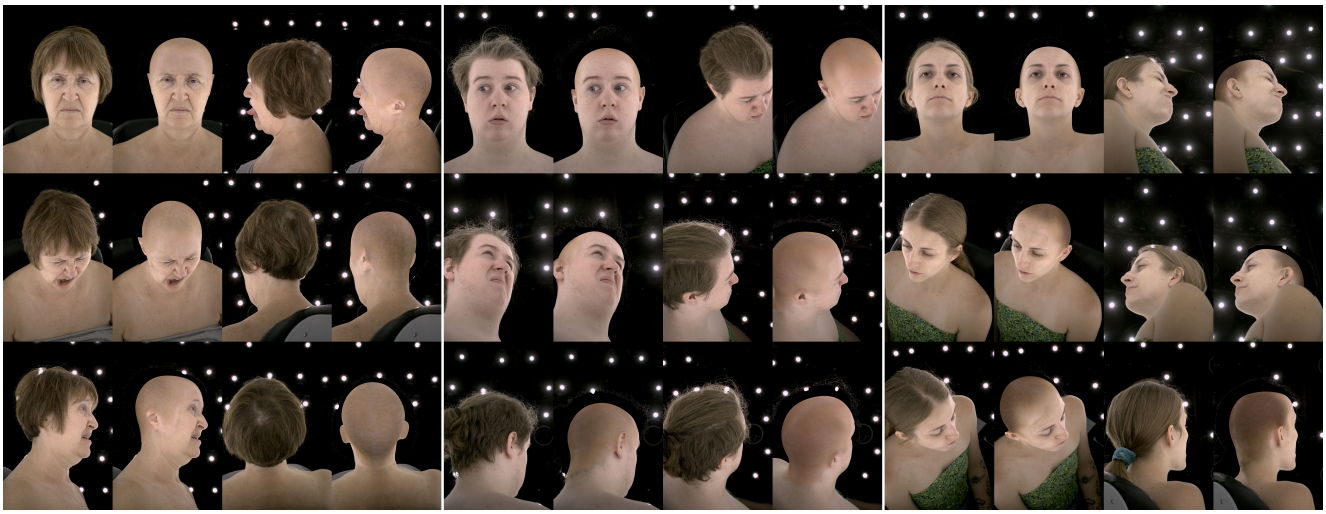

Figure 1. **Synthetic bald image pairs.** Each pair shows (left) the original image and (right) the synthetic bald image generated using our compositing pipeline. The synthetic bald images preserve facial identity while removing hair occlusion, enabling effective supervision for face-hair disentanglement.

(a) Target hairstyle (b) Hair-tied capture (c) Optimized bald mesh

Figure 2. **Auxiliary capture for bald texture optimization.** To minimize occlusion from certain hairstyles, we capture an additional image with the subject's hair tied back (b). This ensures that the optimized bald texture (c) maintains consistent skin color, even when the target hairstyle (a) differs.

the first 500 iterations of SDS loss, we use a bald image prompt generated from a pretrained text-to-image (T2I) inpainting diffusion model [8] as an input image prompt to our diffusion model. In the final 500 iterations, we replace this image prompt with the rendered bald mesh, using its actively optimized texture map for rendering. By this stage, the rendered bald image provides better consistency than the bald image generated from the pretrained T2I model, leading to more coherent texture refinement.

Auxiliary capture for bald texture optimization. Certain hairstyles, such as long hair or bangs, can cause severe occlusions that degrade the quality of the optimized texture map. To mitigate this, we capture subjects with their hair tied back or secured with a thin hairband to minimize occlusion (*e.g.* Fig. 2b). Reducing occlusion maximizes the visible reconstruction region and decreases reliance on the diffusion prior. It is important to note that this capture is used solely for bald texture optimization. For instance, in Fig. 2, although the target hairstyle for training the hair-compositional avatar corresponds to Fig. 2a, we use a separate capture (Fig. 2b) to ensure that the pseudo-bald images maintain consistent skin color beneath the hair.

C. More Qualitative Results

Hairstyle transfer. To demonstrate the robust disentanglement and flexible compositionality of our 3D avatar model, we provide additional qualitative results of hairstyle transfer in this section. As elaborated in the main paper, our framework enables the independent manipulation and transfer of facial and hair components across different identities. This is achieved by defining hair Gaussians relative to a bald mesh anchor, which allows for seamless adaptation to the target subject's head shape without the need for additional scaling or alignment steps. Figure 3 illustrates the capability of our model to transfer various hairstyles onto a single facial identity while preserving the facial characteristics and expression. In this example, a consistent facial identity and expression is combined with different hair attributes sourced from various individuals, showcasing how a single avatar can adopt diverse hairstyles realistically. Conversely, Figure 4 demonstrates the flexibility of our method in transferring a single hairstyle onto multiple distinct facial identities, each maintaining their unique facial features and expressions. This cross-reenactment with hair transfer highlights the generalizability of our hair model, as it adapts a specific hairstyle to different head shapes and facial features, producing visually coherent and high-fidelity results. These examples collectively emphasize the effectiveness of

Figure 3. **Hairstyle Transfer: Single Face, Multiple Hairs.** This figure demonstrates transferring various hairstyles onto a single facial identity. The consistent facial features and expressions highlight the model's ability to seamlessly integrate different hairstyles while preserving identity.

Figure 4. **Hairstyle Transfer: Single Hair, Multiple Faces.** This figure illustrates transferring a consistent hairstyle across multiple distinct facial identities with varying expressions. The results show the adaptability of our hair model to different head shapes, enabling robust cross-reenactment with hair transfer.

our compositional prior model in achieving high-quality, controllable 3D avatar synthesis through disentangled face and hair representations.

Relighting with hairstyle transfer. Our model inherits the relightable 3D Gaussian appearance model from Saito et al. [9] and URAvatar [4], enabling realistic lighting effects on both face and hair. As shown in Fig. 5, our approach is the first to support relightable hairstyle transfer, maintaining consistent illumination across both components. While this aspect builds on existing techniques, it marks a significant step forward by demonstrating relightability in the context of hairstyle transfer, ensuring natural and cohesive lighting under varying conditions.

Compositional 3D avatars. Our approach provides a unified 3D compositional representation of training subjects. Fig. 6 presents the results of our model trained with 64 subjects for compositional rendering, face-only rendering, and hair-only rendering, demonstrating effective separation of face and hair without compromising the quality of the combined 3D avatar. Notably, our model reconstructs a plausible facial appearance even in regions occluded by hair, which is crucial for seamless hairstyle transfer.

Zero-shot and fine-tuned 3D compositional avatars. Our model extends zero-shot inference to a compositional setting, generating 3D avatars with plausible face and hair representations by directly feeding the geometry and albedo maps of a novel identity into the identity-conditioned hypernetworks [1, 4], without requiring fine-tuning. Unlike autodecoder-based models [6], which require latent code inversion to obtain reasonable results for unseen identities [11], our hypernetwork-based design enables zero-shot inference in a simple feed-forward manner. As shown in Fig. 7, our zero-shot compositional avatar successfully reenacts expressions while reconstructing a full 3D appearance, even in regions originally occluded by hair, benefiting from the priors learned during pretraining. However, hair reconstruction in zero-shot results is less detailed compared to the face. This is because hair exists in a significantly higher-dimensional manifold with complex variations in shape and texture, making it more challenging to model. Fine-tuning Sec. 3.5 on a head rotation video with a neutral expression refines both facial and hair details, significantly enhancing visual fidelity.

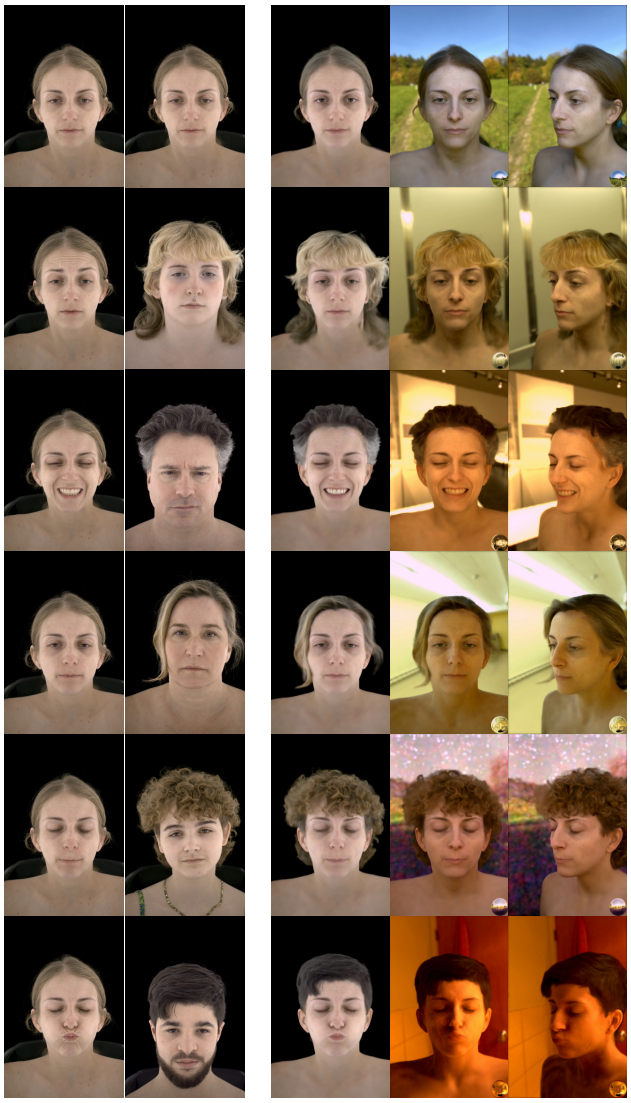

Figure 5. **Relighting with hairstyle transfer.** The leftmost column shows face and expression reference images captured from a real subject (Face/Exp. ID), with expression changing across frames. The second column shows the hair identity image (Hair ID) used for hair transfer. The remaining columns visualize avatar rendering results under different lighting conditions. "Orig. Lighting" corresponds to the original lighting condition under which the subject was captured. "Relighting" corresponds to avatar rendering under novel lighting conditions defined by various environment maps, with each environment map visualized as a reference ball in the bottom-right corner of each image.

References

[1] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 1, 3

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[3] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proc. ICLR*, 2014. 1

[4] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *Proc. ACM SIGGRAPH Asia*, 2024. 1, 3

[5] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 1

[6] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. CVPR*, 2019. 3

[7] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *Proc. ICLR*, 2023. 1

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

[9] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *Proc. CVPR*, 2024. 1, 3

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. ICCV*, 2023. 1

[11] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. *arXiv preprint arXiv:2408.06019*, 2024. 3

[12] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Ewa splatting. *IEEE Trans. Vis. Comput. Graph.*, 8(3):223–238, 2002. 1

Figure 6. **Compositional 3D avatars of the training subjects.**

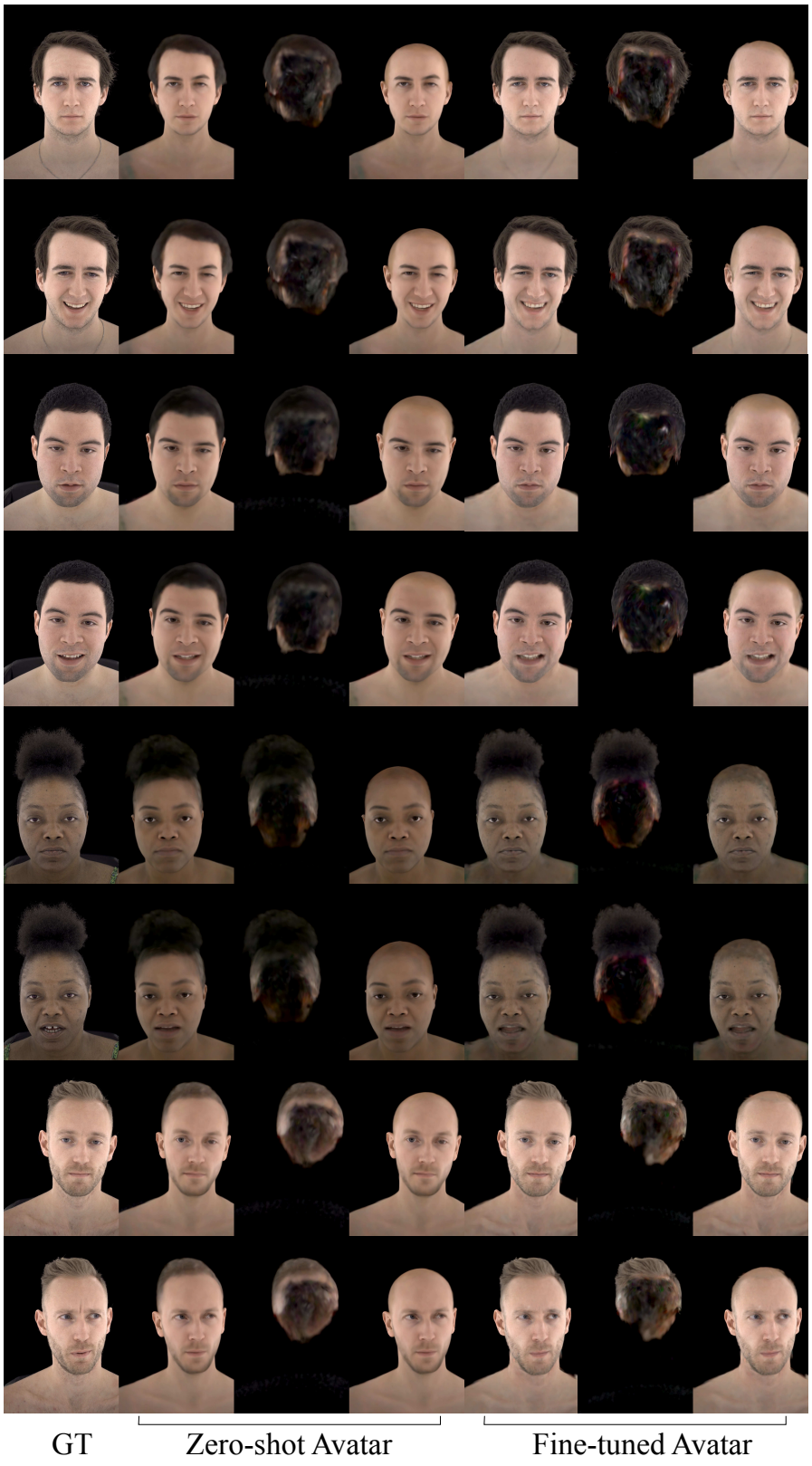

Figure 7. **Zero-shot and Fine-tuned Compositional Avatars.** Our model generates a plausible 3D avatar for a novel identity without fine-tuning (Zero-shot Avatar, middle column), reenacting the facial expression shown in the reference image (GT, left column). We visualize the compositional, hair-only, and face-only renderings for both the zero-shot and fine-tuned avatar results (Fine-tuned Avatar, right column). Fine-tuning improves visual fidelity and consistency while preserving the disentangled structure of face and hair.