

# FlexAvatar: Flexible Large Reconstruction Model for Animatable Gaussian Head Avatars with Detailed Deformation

## Supplementary Material

### A. Implementation Details

**Training settings.** During training, we randomly configure 1 to 4 input images per batch to enhance model adaptability to varying expression inputs. All input images are sampled from random timesteps and random views, while four different viewpoints from distinct timesteps are randomly selected for supervision.

**Data processing.** For the Nersemble and FaceCap multi-view datasets, we employ the state-of-the-art multi-view FLAME [4] estimation method VHAP [6] for FLAME parameter estimation, use MODNet [2] for mask extraction, and utilize ParsingHuman [5] for human parsing. For wild data, we adopt the advanced monocular FLAME estimation method Pixel3DMM [1] for FLAME tracking.

**Network Architecture.** Tab. 1 presents the architecture of our network and the specific settings of input/output parameters.

	Hyperparameter	Value
Image Encoder	Input image size	$512 \times 512$
	Input image number	1 to 4
	Patch size	$16 \times 16$
	Output token dimension	$1024 \times 512$
Self-Attn Block	Token dimension	512
	Self-attn layers	6
Cross-Attn Block	Head Query Token	$2500 \times 512$
	Cross-attn layers	6
Decoder Block	Upsample ratio	$\times 8$
	Input dimension	$50 \times 50 \times 512$
	ID feature map	$400 \times 400 \times 32$
	Gaussian feature map	$400 \times 400 \times 14$
UNet Block	Position map	$400 \times 400 \times 3$
	Gaussian feature map	$400 \times 400 \times 14$
	Downsample layers	4
	Upsample layers	4

Table 1. Hyperparameters of our network architecture.

**Structure of Head Query Tokens.** Our head query token is a learnable parameter of shape  $50 \times 50 \times 1024$ , initialized with sinusoidal positional encodings for each position. During training, the fixed sampling scheme (predefined by FLAME) from the UV space to the FLAME mesh surface ensures stable convergence to a universal canonical head representation.

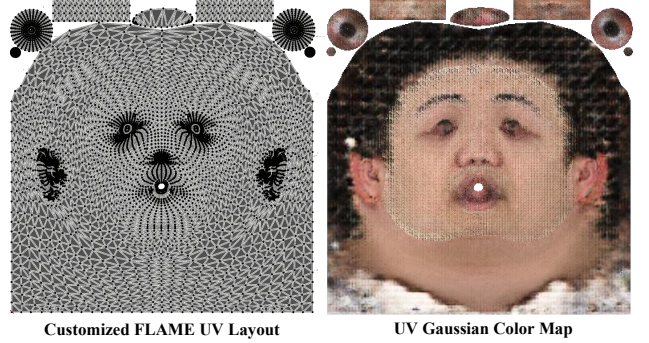


Figure 1. **UV Structure.** Our method use a  $400 \times 400$  UV map structure to establish the mapping relationship between the Gaussian primitives and the FLAME model.

**Structure of UV Map.** Fig. 1 illustrates the structure of the UV map we employed. On the left side of Fig. 1 is the layout of the FLAME vertices and faces in the UV coordinate system. To enhance the stability of the Gaussian points inside the oral cavity, we added a teeth region (the two rectangular areas at the top of the figure) to the original FLAME UV layout. This modification increases the number of vertices from the original 5,023 to 5,143. The right side of Fig. 1 shows a visualization of the UV Gaussian color map. The number of Gaussian primitives is set to  $n = 141,445$ , initialized from the UV map at a resolution of  $400 \times 400$ . Leveraging the regional UV masks provided by FLAME (e.g., for the mouth, face, hair) allows us to selectively fix the Gaussian primitives of certain areas during refinement.

**Anchor Expressions.** To identify the most expressive expressions, we employ a distribution-adjustment scheme that selects 20 anchor expressions for each person from the full dataset. A subset of the carefully curated anchor expressions is presented in Fig. 3. These expressions include pronounced and representative actions (e.g., exaggerated mouth opening, deep furrowing of brows, and lateral mouth stretching). Incorporating these extremes provides the model with clear learning targets for handling challenging expressions, without compromising its performance on more common, subtle ones.

### B. Experiment Results

**Effect of Number of Train Subjects.** Fig. 2 qualitatively demonstrates the improvement of our model with an in-

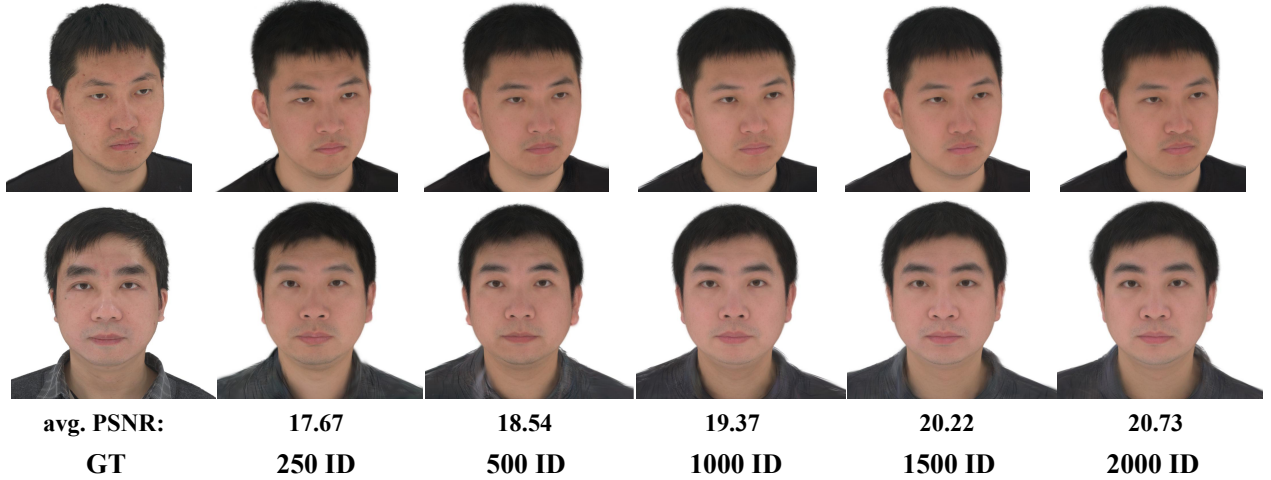


Figure 2. **Qualitative comparison on training id numbers.** The figure illustrates that our identity similarity increases as the number of training identities grows.

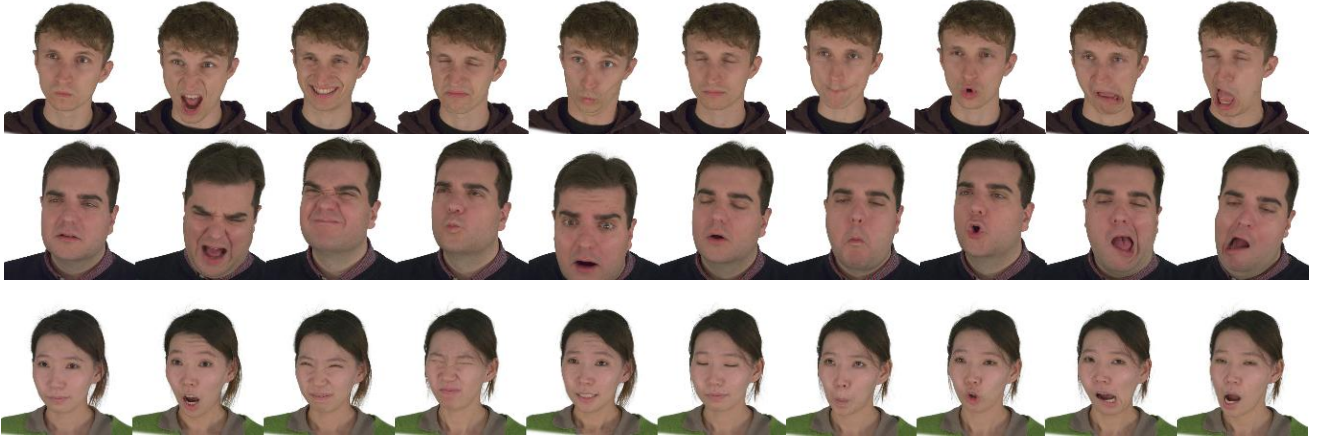


Figure 3. **Anchor Expressions.** Subset of our selected anchor expressions for training.

creasing number of training identities. The results show a clear trend of rising identity consistency with the source image as the number of IDs grows. This trend validates the scalability of our approach. We anticipate that further increasing the scale of training data will continue to enhance model performance, ultimately allowing it to handle various challenging cases through a single feedforward pass.

**More Comparison.** As shown in Fig. 5, we further conducted a qualitative comparison between our method and Avat3r [3], a state-of-the-art 4-view reconstruction method. Since Avat3r is not yet open-sourced, we evaluated our model using the same four input images showcased on their official website. As shown in our results, our method achieves superior performance in terms of identity similarity and finer details like teeth. It is worth noting that while Avat3r employs an expression latent for animation, granting it strong generative capabilities (e.g., for tongue generation), its cross-attention mechanism severely limits in-

ference speed, reportedly reaching only 8 FPS. In contrast, our model, benefiting from a lightweight U-Net architecture, achieves a real-time driving speed of 45 FPS, which is crucial for interactive digital human applications.

Fig. 6 presents the qualitative comparison of self-reenactment results between our method (under single-image input settings) and HeadGAP [7]’s inversion + fine-tuning approach trained on our FaceCap dataset. HeadGAP [7] is a method that performs latent code inversion based on a prior model, followed by input-view refinement. The quantitative metrics for HeadGAP are as follows: PSNR: 20.77, SSIM: 0.8343, LPIPS: 0.2210, CSIM: 0.8489, AKD: 3.6739, AED: 2.1148. These results are comparable to our feedforward output but significantly inferior to our refined results. The comparison demonstrates that, although both methods employ input-view refinement, our approach achieves superior rationality and consistency in the refined outputs compared to HeadGAP.



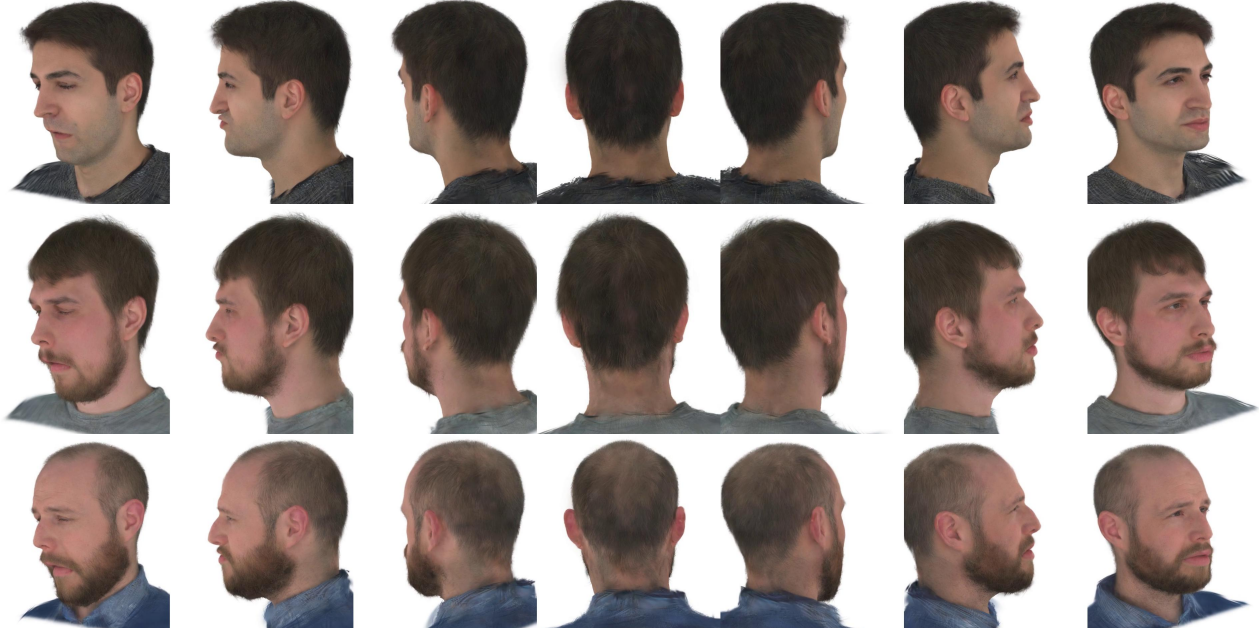


Figure 4. **Back head results.** 360-degree rendering of our head avatars.

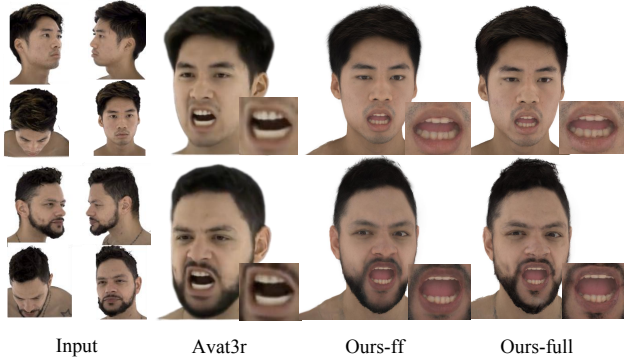


Figure 5. **Qualitative comparisons with Avat3r [3].** Our method outperforms Avat3r in terms of clarity, similarity, and detail authenticity. Please zoom in to see details.

**More Results.** Fig. 7 show the cross-reenactment results driven by IMAvatar dataset[8]. The results demonstrate that the model produces robust animations under in-the-wild conditions. Fig. 4 showcases our full-head rendering results. Benefiting from the full-head data in the FaceCap dataset and our UV-aligned design, our model successfully generates complete head reconstructions. Fig. 10 presents additional self-reenactment results, while Fig. 11 and Fig. 12 display additional cross-reenactment examples. These results collectively demonstrate our model’s strong adaptability to a wide range of expressions. It performs well not only on in-domain datasets like FaceCap and Nersemble but also generalizes effectively to in-the-wild images captured by mobile phones. This provides a practical solution for creating high-quality avatars from a few uncalibrated images.

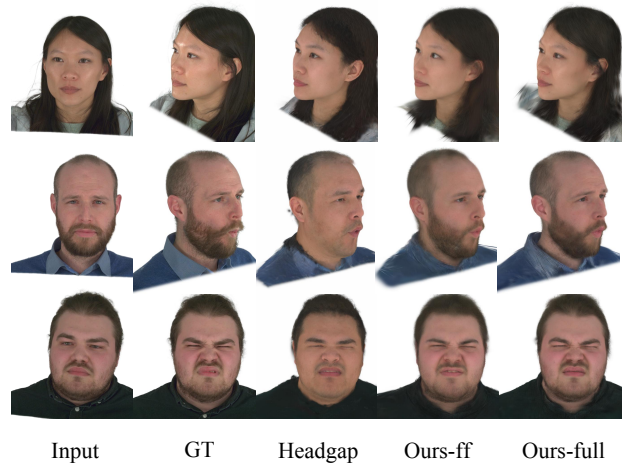


Figure 6. **Qualitative comparisons with Headgap [7].** Our method outperforms Headgap in both feed-forward results and re-fine results. Please zoom in to see details.



Figure 7. **Wild animation results driven by IMAvatar[8].**

**Memory Analysis.** As shown in Fig.8, we report quantitative results of GPU memory and feed-forward PSNR as the number of input views increases: four views achieve a reasonable trade-off between memory cost, efficiency and

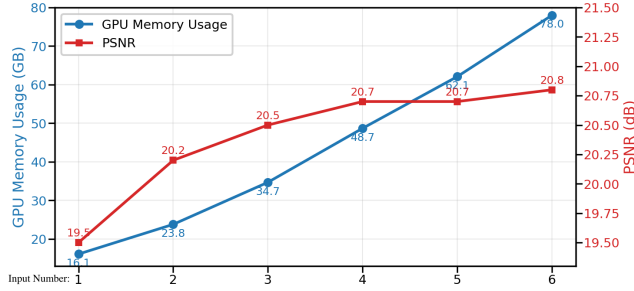


Figure 8. Quantitative results of GPU memory and feed-forward PSNR as the number of input views increases.

the feed-forward reconstruction quality. Memory usage in refinement is mainly governed by the number of supervised images per batch, not the input encoding images. Smaller batch size can be employed to handle arbitrary numbers of images while saving memory. For example, with a supervision batch size of 1 and gradient checkpointing technique, the GPU peak usage is 16.8 GB that is feasible on consumer level GPUs (e.g., 24GB).

**Speed Analysis.** We found that reducing the UV-map resolution (e.g., from 400 to 256) yields a large speedup (from 45 fps to 83 fps) with negligible degradation in animation quality, and thus can be deployed on mobile or VR devices.

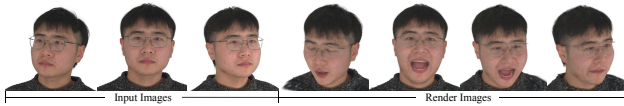


Figure 9. Failure cases of facial accessories (e.g., glasses).

**Failure Case.** As discussed in our limitation section, the current system still exhibits several unresolved challenges. As shown in Fig. 9, due to the lack of relevant training data and the absence of specialized geometric design, our model struggles with examples featuring very fluffy long hair or eyeglasses. We posit that employing additional, dedicated Gaussian modeling for these specific areas could be a viable solution. We leave the investigation of these challenges for future work.

### C. Social impact

Our work presents a paradigm shift for applications reliant on realistic digital humans. By streamlining avatar creation from minimal input, it democratizes high-quality character generation for VR, gaming and telehealth. This efficiency paves the way for scalable, practical, and engaging avatar applications across sectors.

### References

- [1] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction, 2025. 1
- [2] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022. 1
- [3] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12089–12100, 2025. 2, 3
- [4] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1
- [5] Kunliang Liu, Ouk Choi, Jianming Wang, and Wonjun Hwang. Cdgnet: Class distribution guided network for human parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4473–4482, 2022. 1
- [6] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 1
- [7] 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. 2, 3
- [8] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13545–13555, 2022. 3



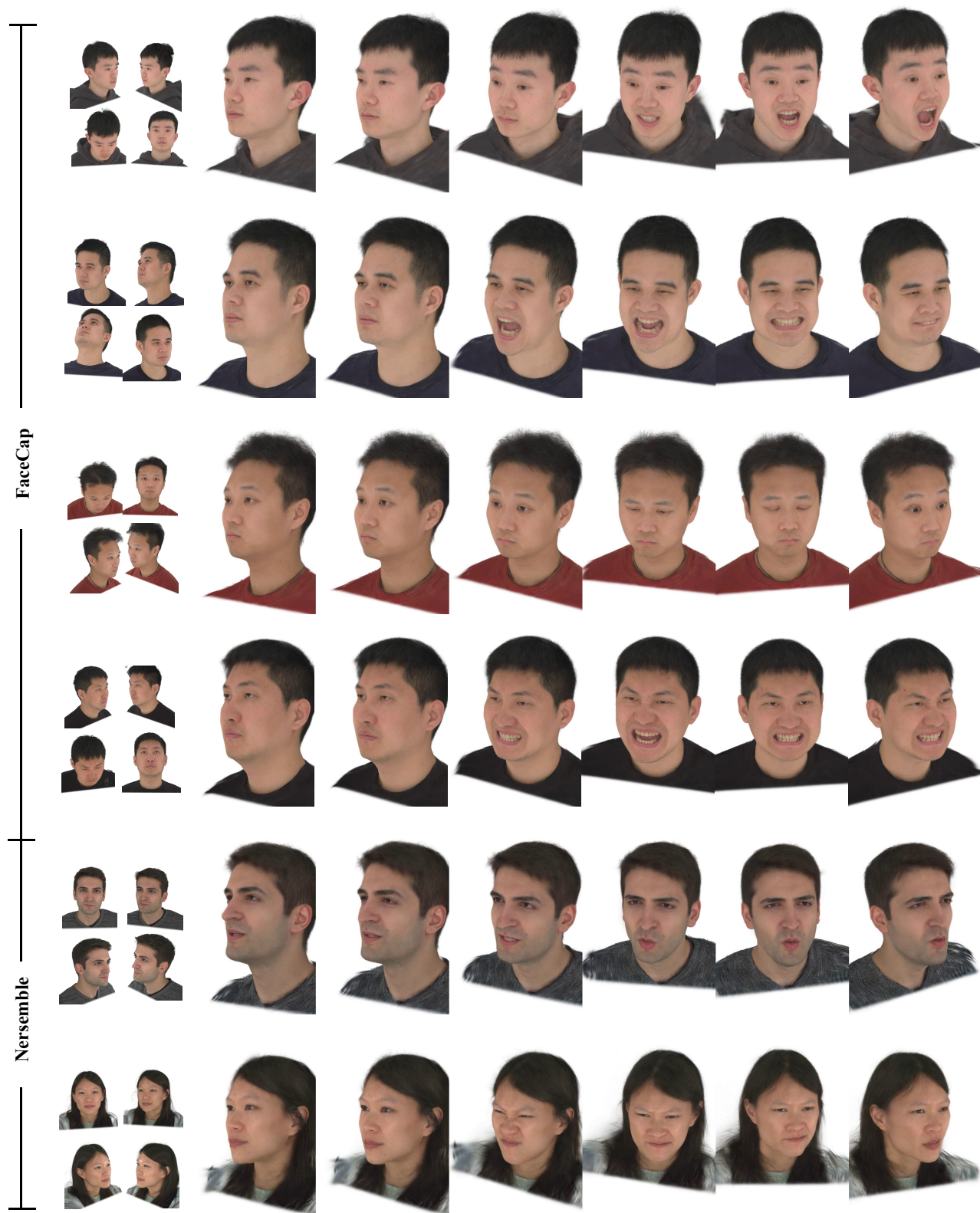


Figure 10. Additional self-reenactment results.



Figure 11. Additional cross-reenactment results.





Figure 12. Additional cross-reenactment results.