FlexAvatar: Learning Complete 3D Head Avatars with Partial Supervision

Supplementary Material

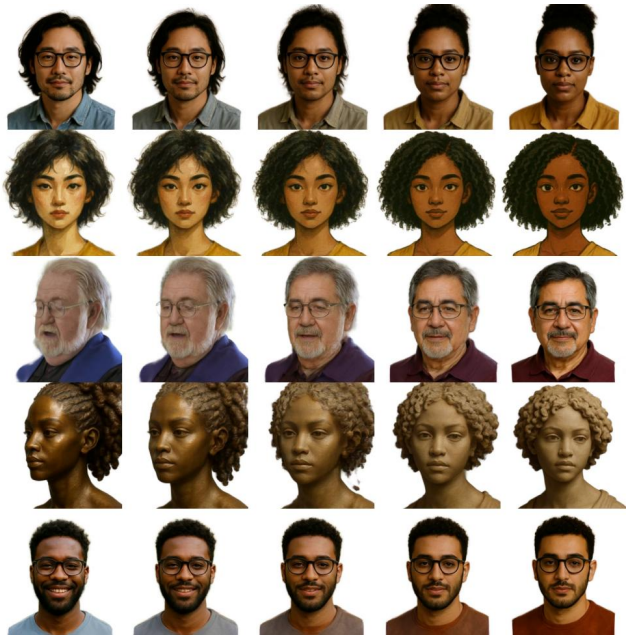

Figure 9. **Interpolation of 3D Head Avatars.** FlexAvatar can produce realistic 3D interpolations between people by interpolating the latent avatar code $\mathcal{A}$, the expression code z_{exp}, and the camera π of two persons.

In this supplementary document, we provide additional comparisons, analysis, and training details. We also highly recommend readers to watch the supplementary video which highlights several aspects of our method, shows plenty of avatars in motion, and features a real-time where a user is walked through the process of creating their own avatar.

A. Additional Comparisons

A.1. Qualitative Comparison on Portrait Animation

Fig. 17 shows qualitative comparisons on the cross-reenectment setting on the VFHQ test split. We compare with the two most recent baselines GAGAvatar [1] and LAM [5]. In both cases, we use the publicly available code to obtain the renderings. Our method produces highly-realistic portrait animations that can capture subtle expressions. Furthermore, our renderings are noticeably sharper than the baselines and contain fewer artifacts, especially under large head rotations of the driver.

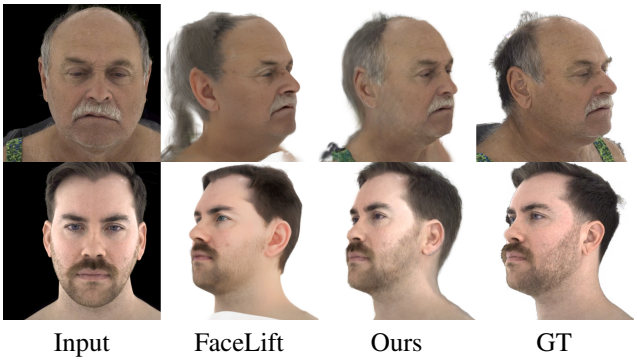

Figure 10. **Visual comparison with FaceLift on the Ava256 dataset.**

	Front					Back		
	PSNR↑	SSIM↑	LPIPS↓	AKD↓	CSIM↑	PSNR↑	SSIM↑	LPIPS↓
FaceLift	12.8	0.715	0.357	6.32	0.658	13.2	0.687	0.411
Ours	**17.2**	**0.786**	**0.265**	**4.72**	**0.771**	**15.2**	**0.709**	**0.408**

Table 6. **Quantitative comparison with FaceLift on Ava256.**

A.2. Qualitative Comparison on Few-shot Setting

Fig. 18 shows qualitative comparisons on the few-shot avatar creation setting following Avat3r [8]. Our method creates artifact-free 3D head avatars that closely resemble the input persons and allow expressive animations.

A.3. Comparison with FaceLift

We compare our method with FaceLift [10] for single-image 3D head reconstruction in two settings:

(i) On Ava256 (Fig. 10 and Tab. 6), we use 4 frontal and 4 back cameras for 5 subjects. Our model slightly outperforms FaceLift quantitatively on back-head renderings and produces noticeably more accurate frontal reconstructions. For fairness, we use a version of our model not trained on Ava256.

(ii) On in-the-wild images (Fig. 11), our method matches FaceLift in completeness while better handling accessories such as caps and glasses.

In contrast to FaceLift, our method also supports head animation, which is part of our core contribution.

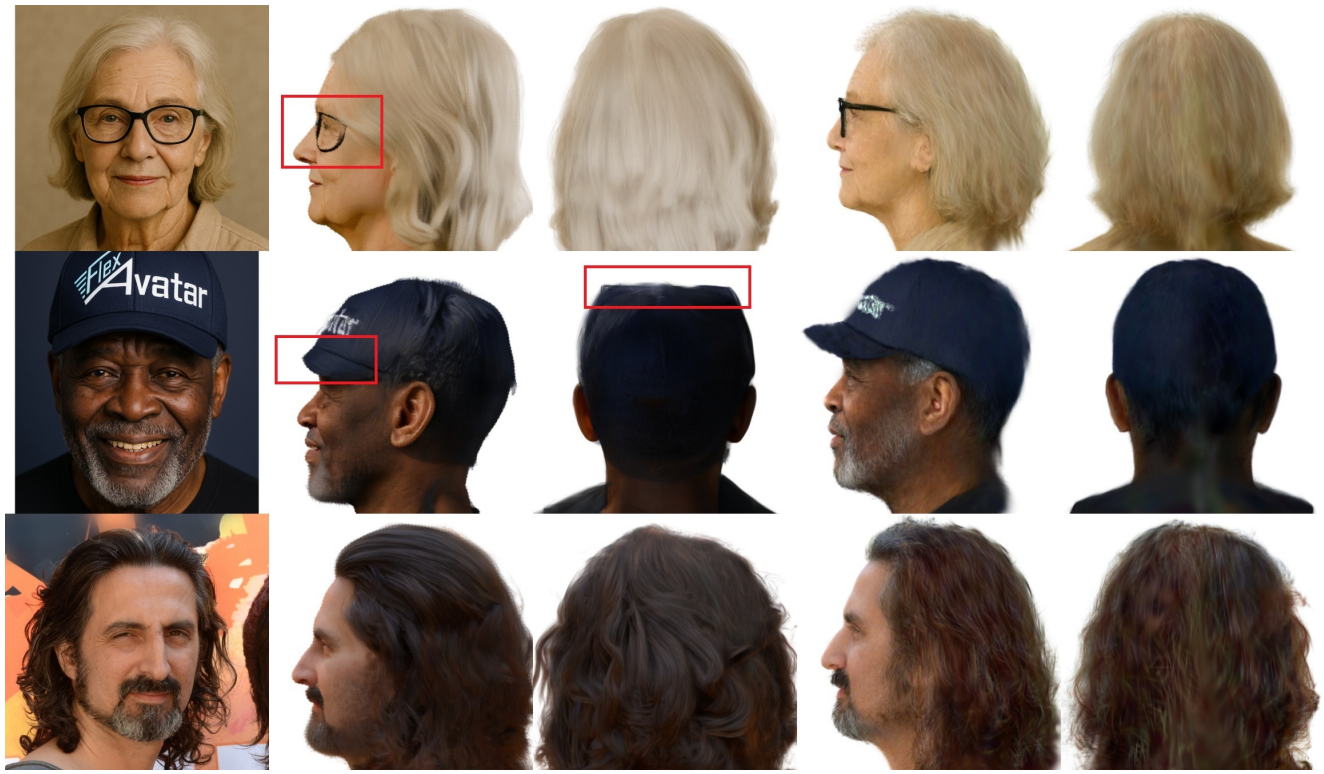

Figure 11. **In-the-wild 360° Comparison with FaceLift.**

B. Additional Analyses

B.1. Interpolation Between Persons

Due to the smooth nature of our avatar latent space, we can produce interpolations between persons. This is done by first obtaining the avatar codes from each portrait and then computing a convex combination between them:

$$\mathcal{A}_1 = E(I_1) \tag{1}$$

$$\mathcal{A}_2 = E(I_2) \tag{2}$$

$$\mathcal{A}_{int} = \alpha\mathcal{A}_1 + (1-\alpha)\mathcal{A}_2 \tag{3}$$

Fig. 9 shows example interpolations.

B.2. Analysis of Bias Sinks

To better understand the effect of bias sinks on the model, we finetune a 2D-only model on the NeRSemble dataset using 1 bias sink per each of the dataset's 16 cameras. As shown in Fig. 12, the model learns that the presence of the "left cam" bias sink correlates with a head that is only complete from the left side. This validates, that the bias sinks are an effective way to make the model mirror the behavior of a specific training data subset during inference without loosing generality.

B.3. Analysis of 3D Data Ratio for Bias Sinks

We analyze how much 3D data is required for the bias sink mechanism to work. To do that, we finetune a 2D-only model with various amounts of multi-view data. Fig. 13 shows that the bias sinks already lead to noticeably more complete heads with only 1% of the 3D training data (=17 different people). Gradually increasing the amount of 3D training data makes the bias sinks more effective with 10% (=186 people) already producing a complete 3D head.

B.4. Analysis of Robustness

In Fig. 14, we show our method on 2 challenging lighting situations. Sole inference with z_{3D} (w/o fitting) attenuates shadows due to the even lighting bias of multi-view data. This is resolved in our full pipeline with fitting. We also refer to our supplementary video that contains 57 avatars from in-the-wild images.

B.5. FPS and VRAM usage

During inference, our model needs 1.7GB of VRAM. Animation and rendering run at 20 fps on an RTX 3090 GPU. Avatar creation, including all processing, takes 2 minutes. For a demonstration, see the live demo in the supplemental video.

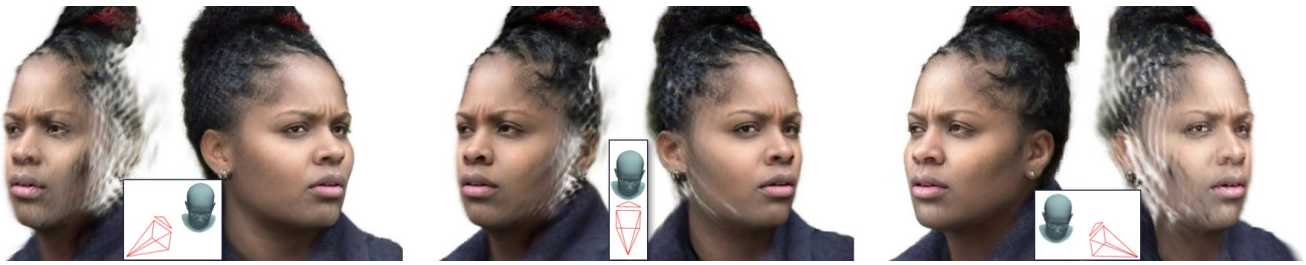

Figure 12. **Analysis of Bias Sinks.** In this ablation experiment, there is one bias sink for each of the 16 cameras of the NeRSemble [7] dataset. Using the bias sink for a left camera during inference results in a head that is only complete when seen from the left side. Analogously, the bias sink for a right camera leads to the opposite effect. As such, each bias sink effectively captures the viewpoint bias of its respective training data subset. In our full method, we exploit this behavior to obtain a bias sink that produces full 360° heads.

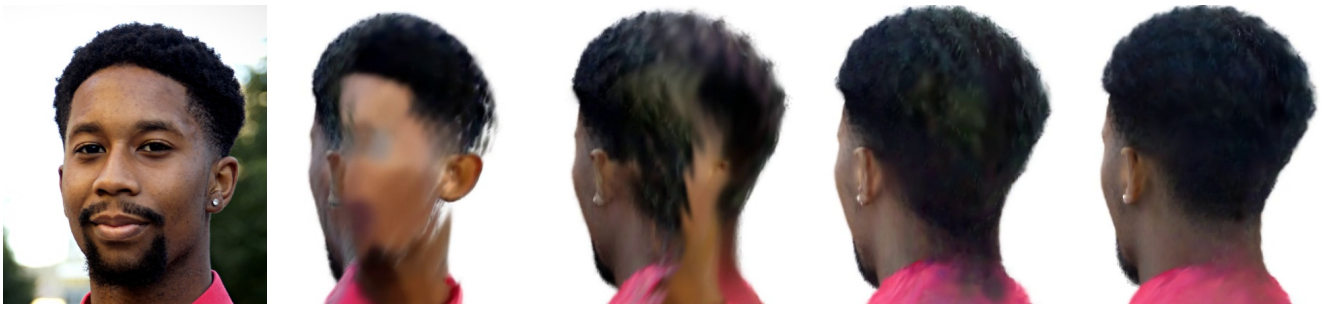

Figure 13. **Amount of 3D data required for bias sinks.** The bias sink mechanism leads to noticeably more complete heads even when only a small proportion of multi-view training data is available.

B.6. Analysis of FLAME dependence

We finetune our model using codes from VOODOOXP's expression encoder [12] instead of FLAME expression codes. Fig. 15 shows that our method is not dependent on FLAME's expression space.

B.7. Analysis of Data Efficiency during Fitting

In Fig. 16, we analyze how the quality of an avatar increases with the number of available input images. To do so, we use the monocular videos from the 5 NeRSemble benchmark [7] persons and apply the fitting procedure as described in the main paper with 2000 optimization steps. It can be seen that both image quality (PSNR) as well as identity preservation (CSIM) greatly increase with the first ∼100 frames and level off after that. We achieve competitive performance on the benchmark with an order of magnitude less input frames required.

C. Training Details

C.1. Data Preparation

To remove the background in the training videos, we use MatAnyone [13]. For single input images during inference, we use MODNet [6]. We also use MODNet to segment out the background in the generations of GAGAvatar [1] in the supplemental video and in Fig. 17. This is because GAGAvatar can only render images with black background due to its use of a screen-space renderer.

Head-centric coordinates. We simplify the models task by always predicting the avatar in FLAME's canonical space, i.e., factoring out the effect of rigid head movement. To do this, the rigid head transformation matrix is instead applied to the cameras. During inference, head movement is then also modelled by factoring the head motion into the rendering viewpoint. As a side effect, it becomes harder for the model to predict the correct torso pose which has to move relative to the canonical head pose.

Expression codes. As it can be seen in Fig. 15, our architecture is agnostic to the specific choice of animation signal. In our experiments, we use FLAME expression codes obtained from Pixel3DMM [4]. However, note that our design allows to train on different animation signals without any change to the architecture itself. Possible animation controls may be expression codes from implicit morphable head models [3] or codes derived from speech.

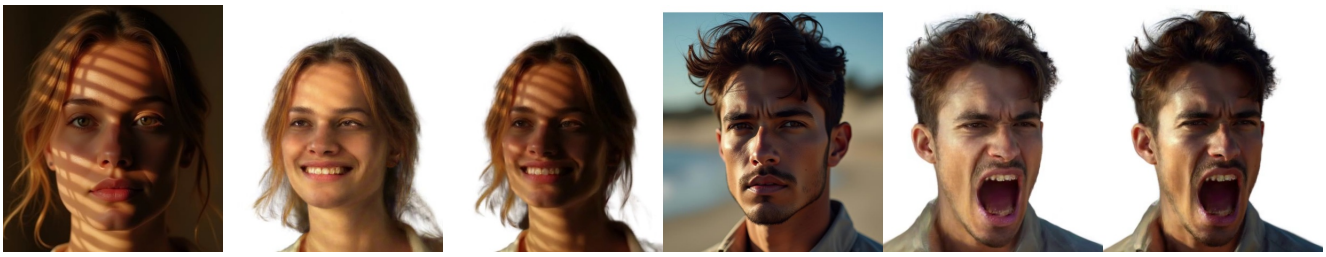

Figure 14. **Performance under challenging lighting.** Example inputs taken from IC-Light [14].

Figure 15. **Analysis of FLAME dependence.** FlexAvatar can be trained with different expression control signals such as expression codes from VOODOOXP [12].

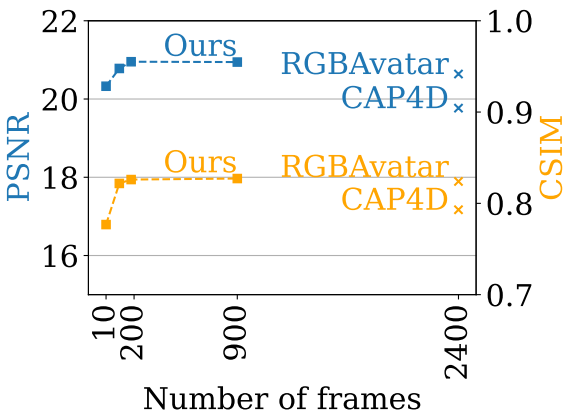

Figure 16. **Analysis of Data Efficiency during fitting.** We plot the performance of our method on the NeRSemble Benchmark [7] in relation to how many frames of a person were used during fitting to create the avatar. Note that the two most competitive baselines on the benchmark, RGBAvatar [9] and CAP4D [11] use all available frames while our method requires only $\sim \frac{1}{10}$ of the frames for a competitive performance. By using $\sim \frac{2}{5}$ of the frames, FlexAvatar outperforms the baselines.

	Hyperparameter	Value
Architecture	ViT patch size	16×16
	hidden dimension D	768
	#cross-attention layers in encoder	8
	#cross-attention layers in decoder	8
	#StyleGAN-PixelShuffle layers	2
	Size of avatar code $\mathcal{A}$	$32 \times 32 \times 768$
In & Out	Input image resolution	512×512
	Train render resolution	512×512
	Gaussian attribute map resolution	256×256
	#3D Gaussians	$\sim$ 58k
Expression MLP	Dimension of expression code	135
	#expression sequence MLP layers	2
	Dimension of expression sequence MLP	256
	Expression sequence MLP activation	ReLU

Table 7. **Hyperparameters.**

References

[1] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024. 1, 3, 5

[2] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 6

[3] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023. 3

[4] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025. 3

[5] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 1, 5

[6] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1140–1147, 2022. 3

[7] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 3, 4

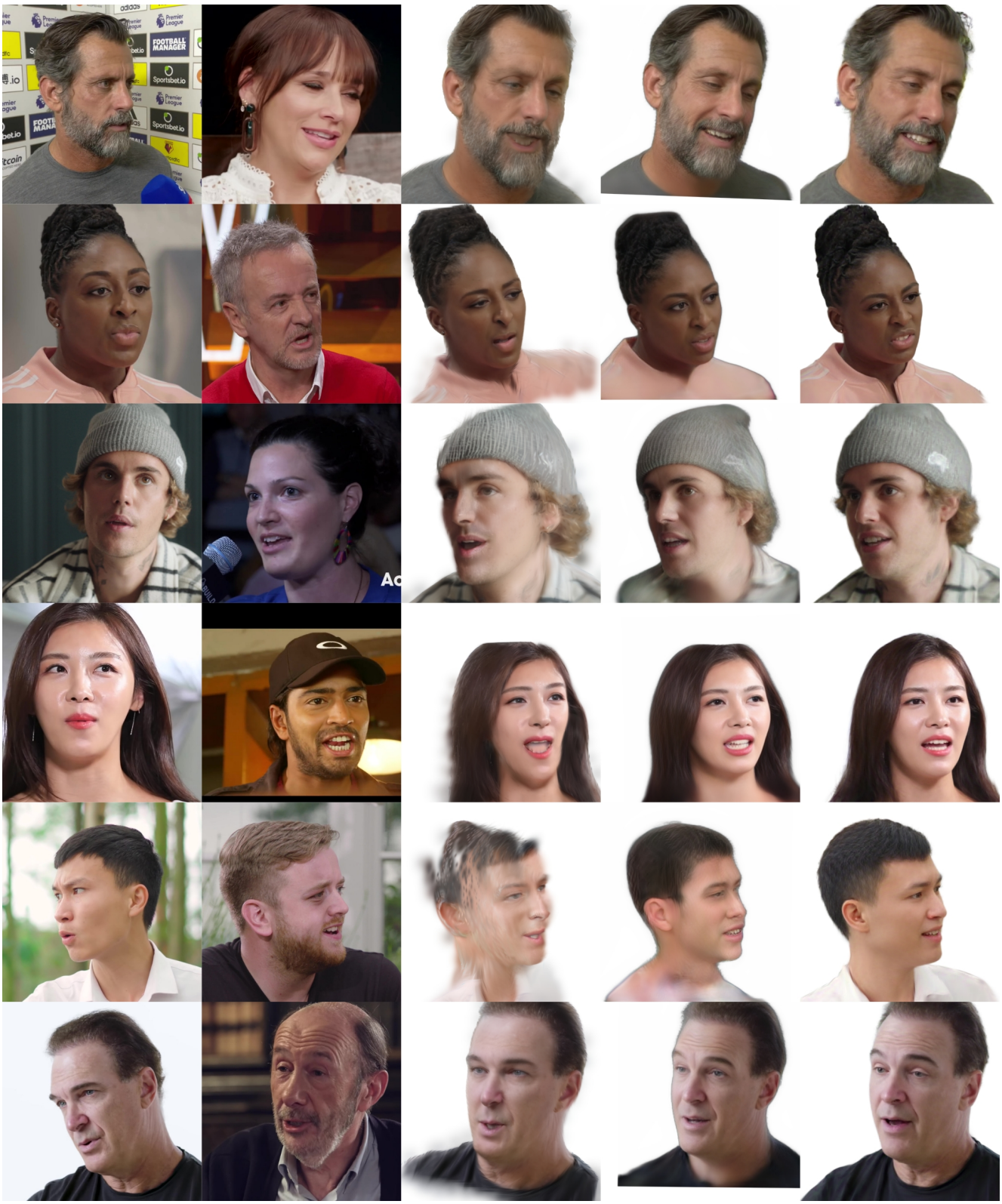

Figure 17. **Qualitative Portrait Animation with cross-reenactment on the VFHQ test split.**

Figure 18. **Qualitative Few-shot Avatar Creation comparison on the Ava256 dataset.**

[8] 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. 1, 6

[9] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 4

[10] Weijie Lyu, Yi Zhou, Ming-Hsuan Yang, and Zhixin Shu. Facelift: Learning generalizable single image 3d face reconstruction from synthetic heads. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12691–12701, 2025. 1

[11] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330. IEEE Computer Society, 2025. 4

[12] Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *arXiv preprint arXiv:2405.16204*, 2024. 3, 4

[13] Peiqing Yang, Shangchen Zhou, Jixin Zhao, Qingyi Tao, and Chen Change Loy. Matanyone: Stable video matting with consistent memory propagation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7299–7308, 2025. 3

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[15] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 6