Supplementary Document for Fresco: Frequency–Spatial Consistent Optimization for Fine-Grained Head Avatar Modeling

Shikun Zhang[1] Yong Li[2*] Yiqun Wang[2*] Qiuhong Ke[1] Cunjian Chen[1]
[1]Department of Data Science and AI, Monash University, Australia
[2]College of Computer Science, Chongqing University, China

In addition to this PDF, we include separate video files, which cover novel-view synthesis, self-reenactment, and cross-identity reenactment. In this supplementary material, we first present additional training details of our method (**A**), followed by qualitative comparisons on cross-identity reenactment (**B**). We then provide ablation studies examining the effects of our progressive frequency curriculum and the stage-wise scheduling of UV-space consistency (**C**), and further include the training configurations of all baseline models for completeness (**D**). To further validate the generality and stability of our framework, we additionally present cross-representation validation results (**E**) and experiments beyond the NeRSemble dataset (**F**). We also analyze early training dynamics to illustrate how delaying high-frequency optimization suppresses pseudo high-frequency artifacts during the initial training stage (**G**). Additional per-subject quantitative results are reported in Tab. 4 and Tab. 5, and the comparison of average inference time on a single NVIDIA RTX A6000 (48GB) GPU is summarized in Tab. 3.

A. Implementation details

All baseline methods and our model are trained on the same hardware environment, using a single RTX A6000 GPU. Our method is trained for 110 epochs with a batch size of 16, using the Adam optimizer with an initial learning rate of 0.001. The three-stage schedule described in Sec. 3.4 is implemented at the epoch level: the first 10 epochs activate only the low-frequency branch; UV-consistency is enabled in the second stage (epochs 10–80) and remains active in the final stage (epochs 80–110), where the high-frequency branch is additionally introduced. The base weights of different loss terms are set to $\lambda_{\text{low}} = 0.25$, $\lambda_{\text{high}} = 0.005$, $\lambda_{\text{UV}} = 0.03$, $\lambda_{\text{pair}} = 0.02$, and $\lambda_{\text{tv}} = 0.004$. After the high-frequency branch is activated, the UV-consistency weight is slightly reduced to $\lambda_{\text{UV}} = 0.02$ to balance detail enhancement and global coherence. α and β respectively denote the weights for the $\lambda_{\text{low/high}}$ and the λ_{UV}. Cosine-gated activation (Eq. (9)) of the high-frequency branch with $T_{\text{high}} : 80$. Frequency filters: Gaussian ($k : 5$, $\sigma : 0.8$) and DoG ($k : 5$, $\sigma_1 : 0.8$, $\sigma_2 : 1.2$). All parameters are fixed across subject IDs, with no per-ID tuning, and consistent improvements over the baseline across all IDs.

B. Cross-identity Reenactment

We also provide comparative results on cross-identity reenactment. As shown in the Fig. 1, our method accurately reproduces mouth movements and better preserves fine-grained wrinkle details.

C. Ablation on Progressive and Stage-wise Scheduling

To evaluate the effectiveness of our staged training strategy, we ablate two key design choices: (1) enforcing both low- and high-frequency constraints throughout the entire training process (Static All-Fre), and (2) enabling UV-space supervision from the beginning rather than activating it in the intermediate stage (Full-time UV). As summarized in Tab. 1, both variants consistently underperform our full model on novel-view synthesis and self-reenactment: PSNR and SSIM decrease, while LPIPS increases. These results indicate that removing either the progressive frequency curriculum or the stage-wise UV scheduling leads to less stable optimization and degraded reconstruction quality.

Fig. 2 further provides a visual comparison of these variants. In the Static All-Fre setting, applying high-frequency before the geometry is stabilized forces the model to fit local details prematurely, leading to noise and over-sharpened artifacts in sensitive regions such as the ear and neck boundaries. In the Full-time UV setting, UV-space consistency is enforced when cross-view correspondences are still unreliable, causing the boundary between the face and hair to become noticeably blurred. In contrast, our staged design first stabilizes low-frequency geometry and only then intro-

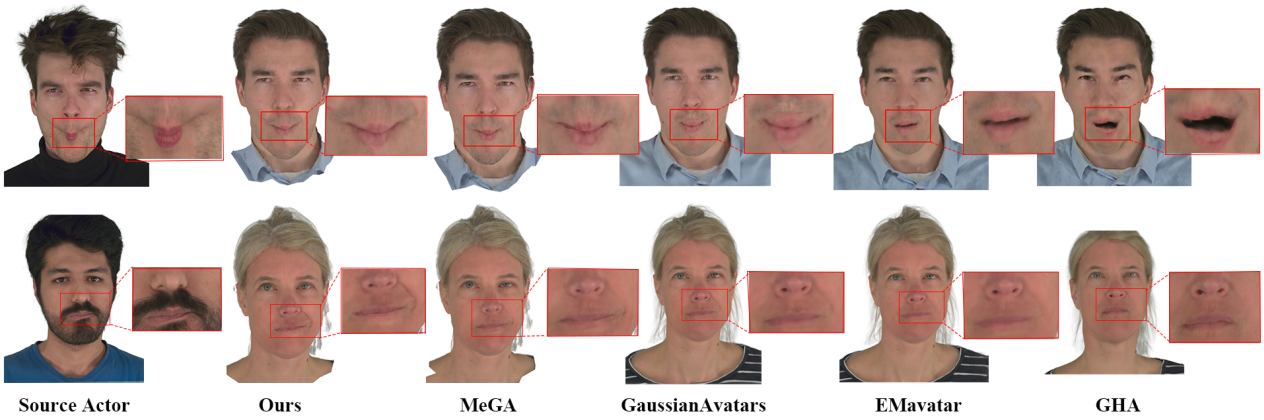

Figure 1. Qualitative comparison on Cross-identity Reenactment of head avatars. Compared with the baselines, which exhibit visible artifacts and struggle on novel expressions, our method produces cleaner renderings and more natural expression transfer.

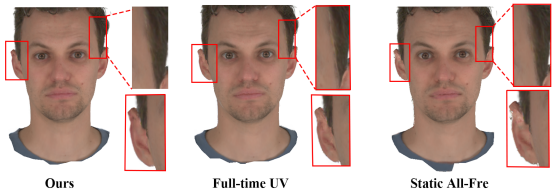

Figure 2. Visual comparison of different training schedules. Our staged strategy produces cleaner structures, sharper boundaries, and more coherent textures, while the Static All-Fre and Full-time UV variants introduce noise, blurring, or boundary artifacts.

Table 1. Additional ablation studies on subject #306: ablation of progressive frequency scheduling and stage-wise UV activation, showing the performance degradation when either design is removed.

Method	Novel-View			Self-Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours	**33.71**	**0.967**	**0.034**	**32.15**	**0.958**	**0.041**
Full-time UV	33.18	0.963	0.036	32.02	0.956	0.043
Static All-Fre	33.06	0.963	0.037	31.90	0.955	0.043

duces high-frequency and UV constraints, yielding cleaner structures, sharper and more coherent boundaries, and an overall more stable appearance.

D. Baseline Configurations

All baseline models are trained from scratch and adopt the identical train/test partition as GaussianAvatars [2]. As detailed in Sec. 4, quantitative evaluation metrics (PSNR, SSIM, and LPIPS) are computed under the same masking scheme for all methods.

GaussianAvatars (GA). For GaussianAvatars (GA), we adopt the publicly released codebase and replicate the original configuration to obtain comparable results. Each subject model is trained for 600,000 iterations to achieve stable convergence.

MeGA. For MeGA [3], we reproduce the method using the official public codebase and strictly adhere to the hyperparameter settings provided by the authors. We train the model for 110 epochs until full convergence is reached. In addition, we evaluate the publicly released checkpoints and observe results closely matching our reproduced model (e.g., 306-subject PSNR: 33.10 vs. 33.13), which confirms the correctness and reliability of our reproduction.

Gaussian Head Avatar (GHA). For GHA, we follow the official implementation and reconstruct the evaluation on the NeRSemble dataset [1]. Specifically, we download the raw NeRSemble videos and generate the training and evaluation data using the same expression subsets as in our setting (EMO-1–4, EXP-2–5, and EXP-8–9), following the preprocessing pipeline released by the authors. Each subject model is trained for 200 epochs until convergence. In the original GHA paper [4], the training data almost fully covers the available expression sequences and the model is mainly evaluated on the free sequence. In contrast, our protocol uses a more restricted set of expression sequences for training and performs self-reenactment on disjoint expression sequences. This configuration places a stronger emphasis on generalization to unseen expressions and may make GHA appear less robust under our more challenging setting.

EMavatar. For EMavatar [5], we adopt the same NeRSemble-based data protocol as described for GHA above. In particular, we process the raw NeRSemble videos using the official preprocessing pipeline and use the same expression subsets as in our experiments (EMO-1–4, EXP-

2–5, and EXP-8–9). Each subject model is trained for 200 epochs until convergence, so that EMAvatar is evaluated under conditions consistent with the other baselines.

E. Cross-representation validation

Fresco consists of two optimization components. The frequency curriculum is representation-agnostic and can be applied without mesh support, whereas the UV-based constraint requires a mesh. Accordingly, we transfer only the frequency component to GaussianAvatars (GA), a pure Gaussian representation. Improved results are reported in Tab. 2 and Fig. 3.

Table 2. Frequency strategies on GA for Self-reenactment.

ID	Method	PSNR↑	SSIM↑	LPIPS↓	Time (h)	Mem (MB)
104	GaussianAvatars	25.48	0.903	0.101	∼**7.29**	∼**2307**
	+ Ours_Fre	**25.76**	**0.903**	**0.099**	∼7.48	∼2313
460	GaussianAvatars	31.85	0.952	0.042	∼**6.46**	∼**2183**
	+ Ours_Fre	**32.11**	**0.954**	**0.040**	∼6.77	∼2215

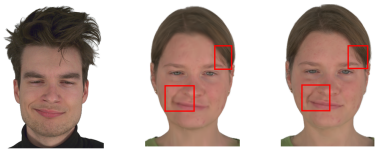

Figure 3. Frequency transfer to GA for Cross-id reenactment.

F. Validation beyond NeRSemble

We further evaluate our method on the INSTA and IMavatar datasets using subjects bala and yufeng. Fig. 4 shows qualitative comparisons with GaussianAvatars (GA) on the self-reenactment task, with PSNR↑ and LPIPS↓ values reported for each result. As shown in the figure, our method reconstructs more accurate expressions with finer details around the mouth and eyes.

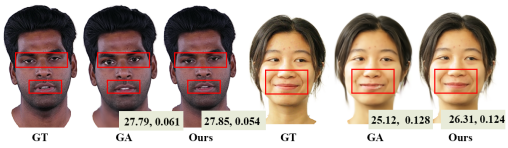

Figure 4. Results on the bala and yufeng subjects.

G. The results of early training

Fig. 5 shows early-stage results (30 epochs). In the MeGA, early high-frequency optimization amplifies small local intensity variations into spurious details before structural stabilization. Our method suppresses this effect during early training by delaying high-frequency optimization.

Table 3. Comparisons on the averaged inference time.

Methods	Ours	MeGA	GaussianAvatars	EMavatar	Gaussian Head Avatar
Inference Time	55ms	55ms	**17ms**	80ms	79ms

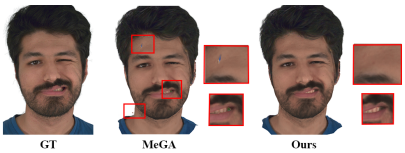

Figure 5. The optimization results of early training (30 epochs).

References

[1] 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. 2

[2] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2

[3] Cong Wang, Di Kang, Heyi Sun, Shenhan Qian, Zixuan Wang, Linchao Bao, and Song-Hai Zhang. Mega: Hybrid mesh-gaussian head avatar for high-fidelity rendering and head editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26274–26284, 2025. 2

[4] Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic gaussians. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2024. 2

[5] Shikun Zhang, Cunjian Chen, Yiqun Wang, Qiuhong Ke, and Yong Li. Eavatar: Expression-aware head avatar reconstruction with generative geometry priors. *arXiv preprint arXiv:2508.13537*, 2025. 2

Table 4. Comparisons with State-of-the-Art Methods on novel view synthesis. We use bold for the best results and underline for the second best.

Subject	**Fresco (Ours)**			MeGA			GaussianAvatars			EMavatar			Gaussian Head Avatar		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
074	**30.36**	0.916	0.107	29.90	0.911	0.106	29.89	**0.921**	**0.105**	28.32	0.905	0.121	27.81	0.904	0.124
104	**28.92**	**0.923**	**0.091**	28.64	0.918	0.097	26.90	0.913	0.103	27.73	0.885	0.094	27.52	0.883	0.094
253	**34.02**	**0.957**	0.036	31.00	0.946	0.056	33.69	0.957	**0.035**	31.01	0.911	0.063	30.82	0.908	0.065
264	**30.98**	**0.954**	**0.060**	30.36	0.952	0.066	30.31	0.933	0.068	30.85	0.903	0.086	30.64	0.900	0.086
302	**33.64**	**0.949**	**0.041**	32.76	0.942	0.046	31.98	0.939	0.052	31.71	0.914	0.070	31.42	0.913	0.071
304	27.40	0.871	0.076	26.14	0.860	0.093	**28.81**	**0.884**	**0.069**	29.69	0.824	0.097	29.30	0.822	0.099
306	33.71	0.967	0.034	33.10	0.963	0.037	**34.16**	**0.969**	**0.030**	30.07	0.910	0.078	29.69	0.906	0.080
460	35.35	0.967	0.025	34.47	0.962	0.028	**36.24**	**0.971**	**0.019**	31.92	0.931	0.051	31.59	0.927	0.052
avg.	**31.80**	**0.938**	**0.058**	30.79	0.932	0.066	31.50	0.935	0.060	30.16	0.898	0.082	29.84	0.895	0.083

Table 5. Comparisons with State-of-the-Art Methods on Self-Reenactment. We use bold for the best results and underline for the second best.

Subject	**Fresco (Ours)**			MeGA			GaussianAvatars			EMavatar			Gaussian Head Avatar		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
074	27.32	0.897	0.112	27.09	0.896	0.117	**27.89**	**0.907**	**0.098**	21.30	0.824	0.158	21.24	0.826	0.158
104	**27.29**	**0.911**	**0.097**	27.19	0.910	0.099	25.48	0.903	0.101	20.42	0.817	0.165	20.31	0.816	0.167
253	**32.46**	**0.949**	**0.043**	29.87	0.940	0.061	30.20	0.946	0.047	25.55	0.873	0.123	25.31	0.871	0.124
264	30.04	**0.950**	0.060	29.60	0.948	0.067	**31.51**	0.943	**0.058**	21.47	0.883	0.121	21.34	0.880	0.123
302	**32.38**	**0.937**	**0.046**	31.53	0.931	0.053	31.59	0.934	0.054	22.83	0.865	0.132	22.64	0.861	0.134
304	27.84	0.873	0.084	27.20	0.870	0.092	**28.96**	**0.904**	**0.079**	21.20	0.798	0.166	20.91	0.797	0.166
306	**32.15**	**0.958**	**0.041**	31.93	0.956	0.043	30.50	0.951	0.049	21.31	0.859	0.177	21.20	0.859	0.180
460	**32.97**	**0.960**	**0.036**	32.55	0.958	0.038	31.85	0.952	0.042	26.45	0.902	0.114	25.79	0.899	0.114
avg.	**30.30**	0.929	**0.064**	29.62	0.926	0.071	29.75	**0.93**	0.066	22.56	0.852	0.145	22.34	0.851	0.145