Synergizing Motion and Appearance: Multi-Scale Compensatory Codebooks for Talking Head Video Generation

Supplementary Material

1. Additional Implementation Details

We perform multi-scale compensation across $N = 4$ scales. We employ the keypoint-based motion flow estimator from FOMM [11]. The multi-scale motion flows are estimated at a size of 64×64. We use convolution layers to encode the motion flows into a latent motion flow space of size $32 \times 32 \times 32$ and set the multi-scale motion codebook size to $K = 1024$ and $d_m = 32$. We also use convolution layers to decode the quantized motion flow features while adopting the motion flow updater in MRFA [12] as our motion flow residual decoder. We employ the image encoder and decoder architecture from VQGAN [2] and further encode the multi-scale appearance features into a size of $32 \times 32 \times 256$. The multi-scale appearance codebook size is set to $T = 1024$ and $d_a = 256$.

We follow the unsupervised training pipeline from FOMM [11], where the source and driving frames are extracted from the same video, and our framework learns to reconstruct the driving frame. For the training objective, we use the perceptual loss from FOMM [11] along with the L1 loss as the image reconstruction loss, and we set the loss weights as $\lambda_{adv} = 0.8$, $\lambda^1 = 0.5$, $\lambda_{recon,m} = 32$ and $\beta = 0.25$. The entire framework is trained end-to-end utilizing the Adam optimizer, with a learning rate set to 8×10^{-5} and a batch size of 16 for 250K iterations on four NVIDIA RTX 3090 GPUs.

2. More Details on Experiments

2.1. Additional Details on the Compared Methods

We evaluate the performance of the compared methods using their released pre-trained models, and we present the training datasets used for each method in Tab. 1. All the GAN-based methods [7, 8, 11, 12, 14] and our method are trained on the VoxCeleb1 [10] training set, while the diffusion-based methods AniPortrait [15] and Follow-Your-Emoji (FYE) [9] are trained on larger-scale datasets, including VFHQ [16], CelebV-HQ [18], HDTF [17], and their own collected dataset [9].

2.2. More Experimental Results

2.2.1. Video Results

We present video results for the ablation study and state-of-the-art comparisons on the project page[1] to demonstrate the effectiveness of our video generation approach.

[1] https://shaelynz.github.io/synergize-motion-appearance/

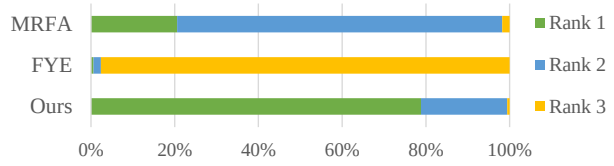


Figure 1. User study results ranking the quality of videos generated by different methods.

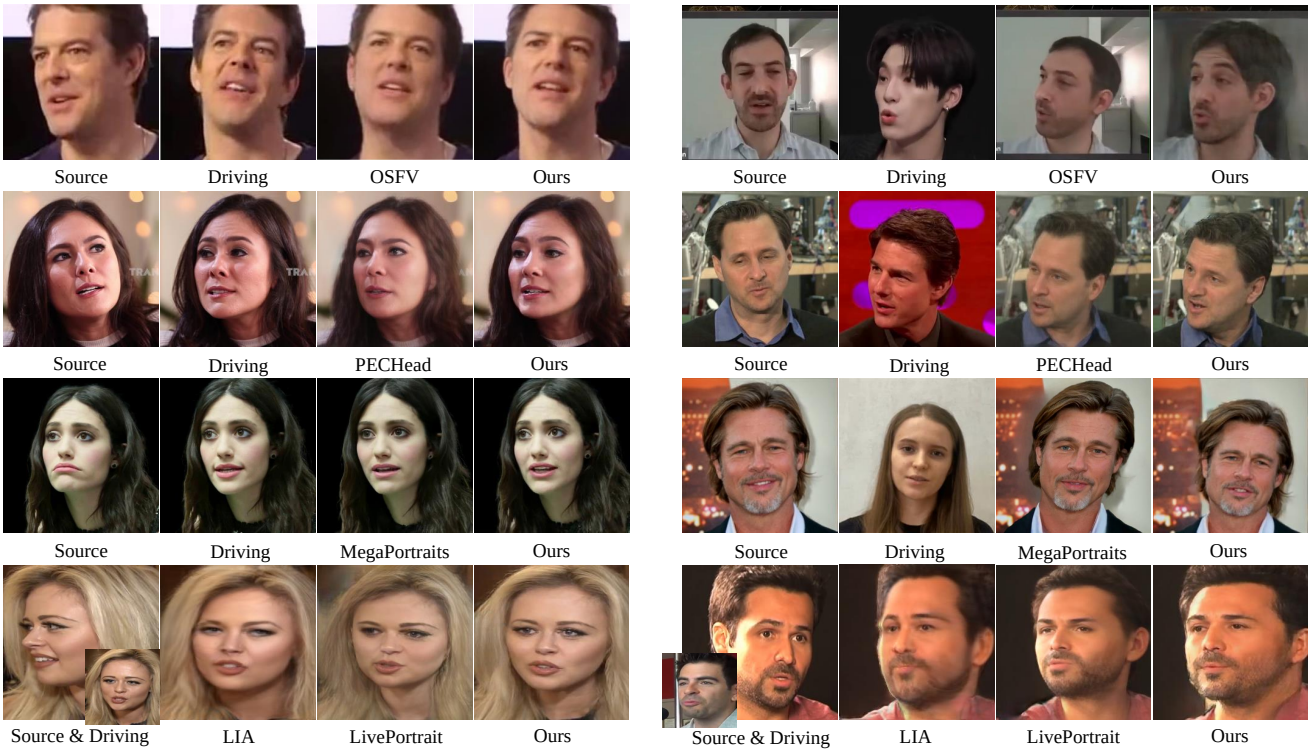
2.2.2. More Comparison Results

Cross-identity Reenactment. In the absence of ground truth for cross-identity reenactment, we conduct a user study comparing our approach to recent state-of-the-art methods, including a GAN-based model (MRFA [12]) and a diffusion-based model (Follow-You-Emoji (FYE) [9]). We randomly selected 10 source-driving pairs and asked 30 participants to evaluate the generated videos based on appearance realism, motion naturalness, and overall quality. The results shown in Fig. 1 indicate that users prefer our method, confirming its superiority.

Method	Training Dataset	FID ↓	CSIM ↑	ARD ↓
AniPortrait [15]	VFHQ [16], CelebV-HQ [18]	*66.61*	*0.7226*	*2.9146*
FYE [9]	HDTF [17], VFHQ [16], their collected dataset [9]	*60.05*	*0.7558*	*3.0822*
FOMM [11]	VoxCeleb1 [10]	80.00	0.6010	1.8331
LIA [14]	VoxCeleb1 [10]	**72.55**	**0.6505**	2.5404
DaGAN [8]	VoxCeleb1 [10]	85.32	0.5743	2.0604
MCNet [7]	VoxCeleb1 [10]	82.72	0.5618	1.6970
MRFA [12]	VoxCeleb1 [10]	77.63	0.5962	**1.5903**
Ours	VoxCeleb1 [10]	<u>76.47</u>	<u>0.6142</u>	<u>1.6234</u>

Table 1. Quantitative comparison for cross-identity reenactment on VoxCeleb1 dataset. AniPortrait [15] and Follow-Your-Emoji (FYE) [9] are trained on much larger-scale datasets and are not suitable for a direct comparison.

We also present quantitative comparison results for cross-identity reenactment in Tab. 1. We use FID [6] for image quality evaluation, Average Rotation Distance (ARD) for motion transfer evaluation following [12], and cosine similarity (CSIM) for identity preservation following [5]. Diffusion-based AniPortrait [15] and Follow-Your-Emoji (FYE) [9] are trained on much larger-scale datasets and are excluded from the comparison. Our method generally demonstrates the highest overall performance, confirming its effectiveness. LIA [14] is slightly better in image quality and identity preservation, as it uses latent codes instead of keypoints as the motion representation, which helps appearance preservation. However, its motion transfer quality is much worse. We also provide a qualitative comparison



(a) Same-identity Reconstruction

(b) Cross-identity Reenactment

Figure 2. Qualitative comparison with more state-of-the-art approaches for (a) same-identity reconstruction and (b) cross-identity reenactment on VoxCeleb1 or examples from the corresponding papers or project pages for closed-source methods (*i.e.*, OSFV [13], PECHead [3], and MegaPortraits [1]). Our method better mimics the driving motion and preserves more facial details.

with LIA in Fig. 2 where our method mimics the driving motion better. Although MRFA [12] can transfer motion well, its output frame quality may be low and it may not preserve source identity effectively. Although diffusion-based methods [9, 15] can achieve even better image quality and identity preservation performance, they struggle to transfer motion faithfully, which leads to undesirable visual quality.

Comparison with more state-of-the-art approaches. We additionally provide comparisons with more state-of-the-art approaches, including an open-source method (*i.e.*, LivePortrait [4]) and several closed-source methods (*i.e.*, OSFV [13], PECHead [3], and MegaPortraits [1]). The qualitative comparison in Fig. 2 highlights the advantages of our method in the preservation of facial details and expression transfer. We also provide a quantitative comparison with the open-source LivePortrait [4] in Tab. 2. Although LivePortrait is trained on significantly larger datasets and is unsuitable for a direct fair comparison, our method can still outperform it on all metrics for same-identity reconstruction.

Inference speed. We evaluate the inference speed using an NVIDIA RTX 3090 and provide the results in Tab. 3. Our approach shows clear advantages upon recent state-of-the-art methods [4, 9, 12, 15], indicating its potential for real-time performance.

Method	# Training Video Frames	Same-identity Reconstruction FID ↓	PSNR ↑	$\mathcal{L}_1$ ↓	LPIPS ↓	AKD ↓	AED ↓	Cross-identity Reenactment FID ↓	CSIM ↑	ARD ↓
LivePortrait [4]	69M	*48.11*	*22.94*	*0.0484*	*0.2213*	*1.5516*	*0.1602*	***75.95***	***0.7260***	***1.3497***
Ours	4.3M	**43.15**	**25.30**	**0.0355**	**0.1846**	**1.2039**	**0.1071**	76.47	0.6142	1.6234

Table 2. Quantitative comparison with LivePortrait [4] on VoxCeleb1. LivePortrait, being trained on *significantly larger* data, is unsuitable for a direct comparison.

	MRFA [12]	AniPortrait [15]	FYE [9]	LivePortrait [4]	Ours
FLOPs ↓	403.05G	9.18T	15.04T	1.31T	**352.91G**
FPS ↑	12.41	0.36	0.39	11.28	**15.13**

Table 3. Inference speed comparison.

2.2.3. Additional Ablation Study

Effect of the code allocation scheme. We propose a novel code allocation scheme for motion and appearance codebooks that assigns different codes to corresponding scales. This allows certain codes to be shared across multiple scales, facilitating the transfer of information between them. To assess the effect of our code allocation scheme, we conduct an ablation study and present results in Tab. 4. We compare with two alternative codebook splitting schemes: sharing all codes across all scales and splitting the codes

Method	FID ↓	PSNR ↑	$\mathcal{L}_1$ ↓	LPIPS ↓	AKD ↓	AED ↓
Sharing all codes	43.23	25.12	0.0359	0.1860	1.2124	**0.1065**
Splitting the codes equally	**42.52**	25.20	0.0358	0.1857	**1.1893**	0.1075
Code Allocation (Ours)	43.15	**25.30**	**0.0355**	**0.1846**	1.2039	0.1071

Table 4. Ablation study on the code allocation scheme.

Method	# Params (M)	FID ↓	PSNR ↑	$\mathcal{L}_1$ ↓	LPIPS ↓	AKD ↓	AED ↓
Baseline*	82.2	48.09	21.64	0.0549	0.2480	2.6798	0.2214
Ours	82.2	**43.15**	**25.30**	**0.0355**	**0.1846**	**1.2039**	**0.1071**

Table 5. Ablation study on the model design.

Number of Codes	FID ↓	PSNR ↑	$\mathcal{L}_1$ ↓	LPIPS ↓	AKD ↓	AED ↓	FLOPs (G) ↓	FPS ↑	Memory (M) ↓
256	47.50	25.18	0.0358	0.1861	**1.1970**	**0.1039**	**351.57**	**15.60**	**6411**
512	46.62	25.11	0.0362	0.1877	1.2190	0.1072	352.01	15.47	**6411**
1024 (Ours)	**43.15**	**25.30**	**0.0355**	**0.1846**	1.2039	0.1071	352.91	15.13	6413

Table 6. Ablation study on the codebook size. We present the results of different code numbers.

equally among the scales. As demonstrated in Tab. 4, our code allocation scheme generally achieves the best overall performance, confirming the superior performance of our code allocation scheme.

Effect of the model design. To verify the source of our performance improvement, we compare with a new "Baseline*", which has parameters comparable to our full model, achieved by increasing the ResBlock channel numbers of our image encoder and decoder. We present the results in Tab. 5. Our method significantly improves upon "Baseline*". The clear performance gap further confirms that the improvement comes from our model design rather than the increased parameters, indicating the effectiveness of our model design.

Codebook size. To assess how the codebook size affects the generation speed and quality, we vary the number of codes in the codebooks to achieve different codebook sizes and present results in Tab. 6. Larger codebooks generally improve generation quality by providing sufficient capacity to learn diverse motion and appearance codes, with only a slight decrease in speed/memory performance. A small codebook of 256 also performs well, likely because codes are retrieved more frequently during training, allowing for better optimization within the same training iterations. However, its image quality remains limited.

3. Limitation

A limitation of our method is the appearance leakage problem in cross-identity reenactment, where the face in the generated video tends to have a shape similar to that of the driving face rather than the source face. This issue arises from the keypoint-based motion flow estimator that we adopt to produce the initial coarse motion flow and the driving keypoints for multi-scale motion codebook compensation. Although this motion flow estimator is robust to non-facial motion, such as hair and neck movement, by learning unsupervised keypoints on talking heads, the keypoints also inherently model facial shapes, which leads to the entanglement of motion and shape. Thus, appearance leakage is a common issue for keypoint-based methods. Our method can effectively alleviate this issue by demonstrating better appearance preservation than other state-of-the-art keypoint-based approaches. As evidenced in Tab. 1, we achieve the highest CSIM score among these approaches, excluding LIA [14], which uses latent codes instead of keypoints for motion representation. This issue can also be mitigated through relative motion transfer [11], which is widely adopted by previous methods [7, 8, 12].

References

[1] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *ACM MM*, 2022. 2

[2] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 1

[3] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *CVPR*, 2023. 2

[4] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2

[5] Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. Marionette: Few-shot face reenactment preserving identity of unseen targets. In *AAAI*, 2020. 1

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 1

[7] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *ICCV*, 2023. 1, 3

[8] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, 2022. 1, 3

[9] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia*, 2024. 1, 2

[10] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 1

[11] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1, 3

[12] Jiale Tao, Shuhang Gu, Wen Li, and Lixin Duan. Learning motion refinement for unsupervised face animation. In *NeurIPS*, 2024. 1, 2, 3

[13] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 2

[14] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *ICLR*, 2022. 1, 3

[15] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 1, 2

[16] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *CVPRW*, 2022. 1

[17] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021. 1

[18] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *ECCV*, 2022. 1