

INFP: Audio-Driven Interactive Head Generation in Dyadic Conversations

Supplementary Material

1. Detail for Motion-Based Head Imitation

Loss Function for training stage 1 is as follows:

$$L_{s1} = w_{per}L_{per} + w_{adv}L_{adv} + w_{cyc}L_{cyc}, \quad (1)$$

where L_{per} is perceptual loss, L_{adv} is GAN loss, and L_{cyc} is cycle consistency loss proposed in VASA-1 [5] for disentangling between motion and 3D appearance features.

Model details. The model architectures of face encoder, face decoder, motion encoder and motion flow estimator are shown in Fig. 1.

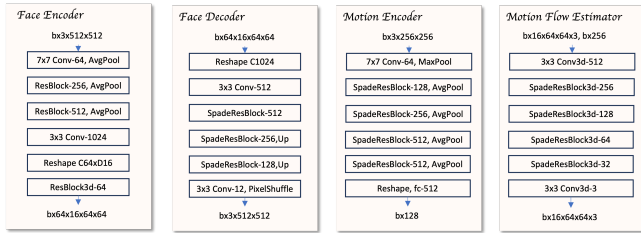


Figure 1. Model details for Section 3.1.

2. Ablation Study

Motion Memory Bank. In stage 2, we remove the verbal/non-verbal motion memory banks from the interactive motion guider, and directly use the dyadic audio feature after several MLP layers as the condition for the motion-attention layer. This leads to noticeable lip motion inconsistencies with ground-truth (Fig. 2). The experiment result of different memory bank size is shown in Tab. 1, it shows that as the memory bank size increases, the generation quality does not increase significantly.

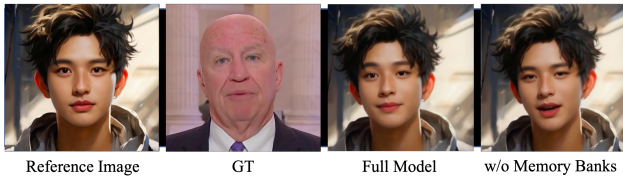


Figure 2. Ablation study on motion memory banks.

Dataset Size. Experiments on varying-scale datasets (Tab. 1) show large-scale data better captures subtle facial expressions and head movements, though weakly correlated with speech.

Hybrid Facial Representation. We change the input to E_m in stage 1 from our carefully-designed hybrid facial

Methods	SSIM↑	PSNR↑	FID↓	SyncScore↑	LPIPS↓	CSIM↑	SID↑	Var↑
INFP	0.834	31.562	15.727	7.188	0.257	0.904	2.613	2.386
10h data	0.811	27.310	19.871	7.092	0.288	0.852	1.861	1.776
100h data	0.830	30.639	15.802	7.163	0.270	0.894	2.437	2.119
d=128	0.828	30.488	16.091	6.782	0.269	0.812	2.270	1.851
d=1024	0.830	32.011	15.722	7.143	0.260	0.909	2.609	2.384
GT	1.000	N/A	0.000	7.261	0.000	0.967	2.891	2.435

Table 1. Ablation study on dataset size and memory bank size.

representation to the original intact image or 2D landmarks map. Results are shown in Fig. 3. It can be seen that there is degradation in generation quality and a leakage of appearance information using original image or landmarks map as input.

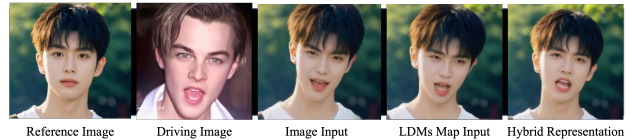


Figure 3. Ablation study on the input to the motion encoder.

3. Inference Speed and Latency

Our method is extremely fast with float-16 precision (for a 3s input audio clip, the entire inference process takes only 0.5s on an L20 GPU). Therefore, we can adopt a sliding-window strategy and reduce the length of each audio segment to control the latency within an acceptable range.

4. Additional Qualitative Results

Since our method can naturally generalize to the task of listening head and talking head generation, we directly use our model without any modification to conduct additional experiments.

We first compare our framework with SOTA listening head generation methods, including L2L [3], RLHG [8] and DIM [4]. ViCo [8] is selected as the benchmark. Results are shown in Fig. 4 (a), which demonstrate that INFP achieves more expressive and diverse motions.

For talking head generation, we select SadTalker[6], AniTalker[2], and EchoMimic[1] as comparing methods. HDTF [7] is selected as the benchmark. Results are shown in Fig. 4 (b), which reveal that INFP can generate more accurate lip movement.

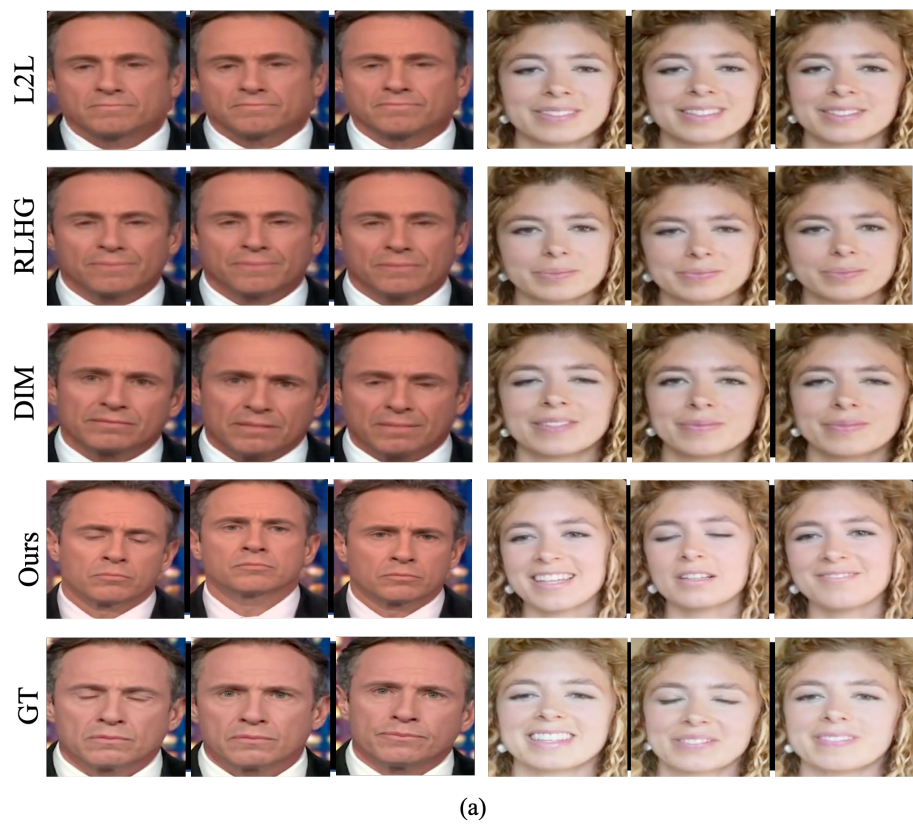


Figure 4. Qualitative comparisons with Listening Generation methods (a) and Talking-Head Generation methods (b).

References

- [1] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. [1](#)
- [2] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding, 2024. [1](#)
- [3] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20395–20405, 2022. [1](#)
- [4] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dyadic interaction modeling for social behavior generation. *arXiv preprint arXiv:2403.09069*, 2024. [1](#)
- [5] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Bain-ing Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024. [1](#)
- [6] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023. [1](#)
- [7] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. [1](#)
- [8] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. [1](#)