Appendix: ARIG: Autoregressive Interactive Head Generation for Real-time Conversations

Ying Guo*, Xi Liu*, Cheng Zhen†, Pengfei Yan, Xiaoming Wei
Vision AI Department, Meituan
{guoying16, liuqian70, zhencheng02, yanpengfei03, weixiaoming}@meituan.com

A. Implementation Details

A.1. Network Details

Due to the page limitation, we show network details of modules in this Appendix.

Bidirectional-learning In the bidirectional learning block, we treat the audio-visual behavior of the agent and the user as two independent individuals and use independent model parameters to first understand their own behavior and exchange information through shared attention. This independent understanding is conditioned on the updated audio, which is injected by the modulation block. We embed the audio in 512-dim and, as adaLN in DiT[4], we compute the scale, shift, and gate parameters in the modulation. The depth of the bidirectional learning block is 2. The detailed block structure is shown in Fig. 1.

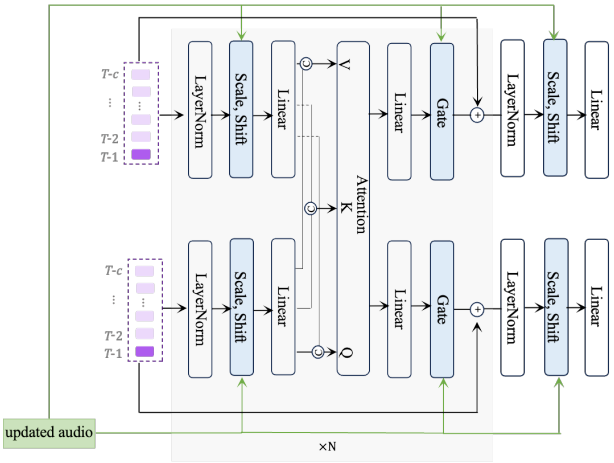


Figure 1. The structure of the Bidirectional-learning.

Integrated-learning In the integrated learning stage, we concatenate the outputs of both parties in the Bidirectional-learning to perform unified learning. We also inject the update audio and perform parallel learning of attention and MLP to improve efficiency. Finally, we extract the information of the agent part as the interaction summary. The detailed structure is shown in Fig. 2.

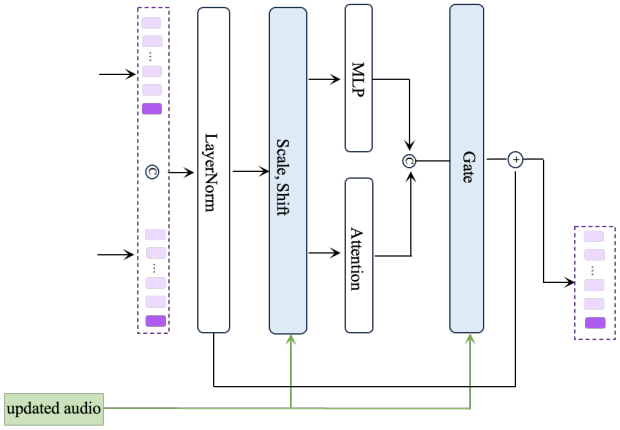


Figure 2. The structure of the Integrated-learning.

Progressive Motion Prediction In this PMP module, we first generate an outline feature for the current motion based on audio information. We set the latest audios from three frames as conditions to provide more complete word pronunciation, and use the cross-attention to predict the coarse outline feature. Then, we utilize the contextual interaction summary (cis-token) and the state feature to perform fine-grained prediction via the condition block which includes cross-attention and feedforward function. We also combine A_T^a to enhance the audio part. Then we refer to the 5-frame motions at the temporal layer to further enforce inter-frame continuity. We then use the output of the temporal layer as a condition for denoising in the DiffusionMLP to sample motions. The DiffusionMLP is a lightweight network composed of 3 MLP blocks, with conditional injection also performed via adaLN[4]. The specific structure of DiffusionMLP is shown in Fig. 3.

*Equal Contribution
†Corresponding Author

Methods	RPCC ↓	CSIM ↑	SyncScore ↑	PSNR ↑	SSIM ↑	FID ↓	SID↑	Var↑
w/o continuous AR modeling	0.129	0.887	7.036	27.62	0.813	23.78	2.261	2.154
w/o bidirectional-integrated learning	0.173	0.841	6.813	24.17	0.749	25.36	2.138	2.016
Ours	**0.125**	**0.901**	**7.218**	**29.67**	**0.827**	**21.64**	**2.428**	**2.397**

Table 1. Ablation study for continuous AR modeling as well as the bidirectional-integrated learning. **Bold** represents the best.

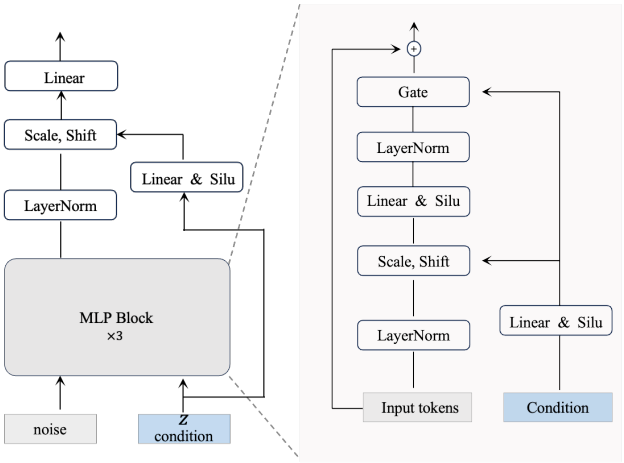


Figure 3. The structure of the DiffusionMLP.

Methods	Overall Naturalness	User-agent Coordination	Motion Diversity	Lip Synchronization
DIM [5]	2.48	2.04	2.12	2.57
Ours	**4.43**	**4.18**	**4.52**	**4.36**

Table 2. User study. The best results are highlighted in bold.

A.2. Inference Details

The dimensions of the input audio and motion are 768 and 262. We first encode them into 512 dimensions and put them into the historical input. In the initial stage, we repeat the motion vector of the agent's reference image and the audio corresponding to the first frame to initialize each cache. The embedding dimension in Bidirectional-learning, Integrated-learning, Contextual-understanding and State-prediction is 512. The dim of feedforward function is 2048. The condition dim in DiffusionMLP is projected into 262.

B. More Experiments

B.1. More Ablation

Continuous AR Modeling To validate the effectiveness of continuous autoregressive (AR) modeling, we utilize discrete indices in an N-sized codebook to represent the motions, and employ the Softmax function as the sampler in PMP. The ablation results are shown in Table 1, and the visual ablation is shown in Figure 4 (a). As shown in Table 1, metrics related to video realism (*e.g.* FID, PSNR and SSIM) and motion diversity (*e.g.* SID and Var) are greatly worsened when using discrete AR modeling. In Figure 4 (a), we can also see that compared to continuous AR modeling, the discrete AR modeling may reduce facial expressiveness significantly, which demonstrates that the discrete modeling is not sufficient for the representation of rich facial expressions, and thus fails to predict realistic and diverse motions.

Bidirectional-integrated Learning We also ablate the network structure of our proposed IBU module and replace the bidirectional and integrated learning module with simple linear layers. As shown in Table 1, all metrics have deteriorated significantly without bidirectional-integrated learning, which validates that our designed bidirectional and integrated learning module is effective in interactive behavior understanding and modeling. In Figure 4 (b), we can see that our method can significantly improve the accuracy of facial motion prediction.

B.2. User Study

We asked 25 people to rate 20 different videos (on a scale of 1-5, the higher the better) generated by DIM[5] and our method across four dimensions: overall naturalness, user-agent coordination, motion diversity, and lip synchronization. To be specific, the 20 test videos are randomly chosen from RealTalk[2]. As shown in Table 2, our method outperforms DIM[5] in all aspects.

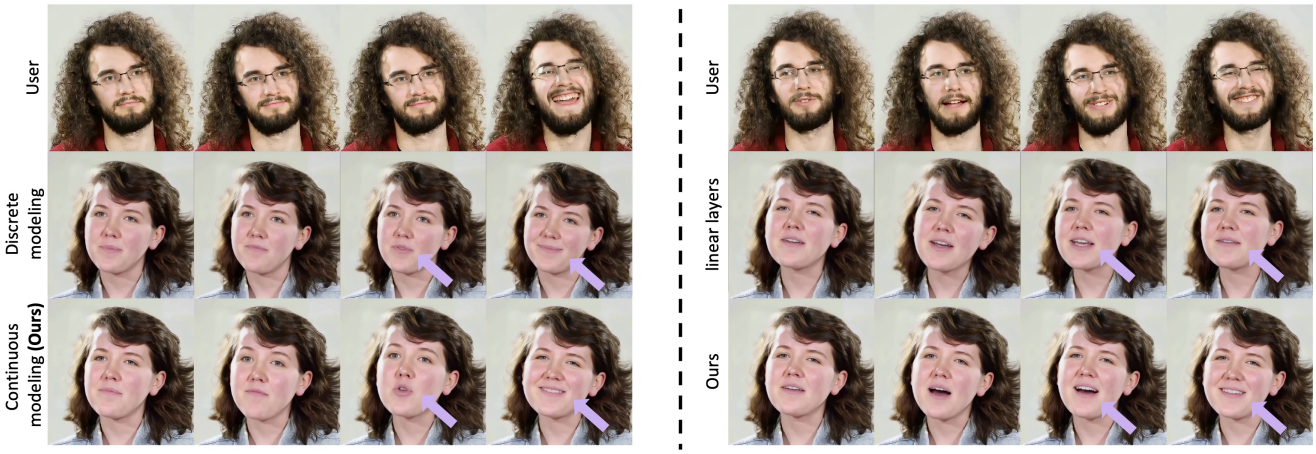
C. Supplementary Visual Results

C.1. Interactive Head Generation

We present visual comparisons of our method with DIM[5] on RealTalk[2] dataset in Figure 5. It can be seen that the agent videos generated by our method are closer to GT and have a significant improvement compared to DIM[5].

C.2. Talking Head Generation

In Figure 6, we compare our results with SadTalker[7], Hallo[6] and EchoMimic[1] on HDTF[8] based on the same reference image and audio. It can be seen that compared to other methods, our method exhibits the closest lip movements and head motions to the ground truth, indicating that the motions produced by our method are the most natural and photorealistic.



(a) Ablation of continuous AR modeling

(b) Ablation of IBU structure

Figure 4. Ablation Study.

Figure 5. Qualitative comparisons with DIM[5] on RealTalk[2] dataset.

C.3. Listening Head Generation

We compare our method with the SOTA listening head generation methods (e.g., RLHG[9], L2L[3], DIM[5] and INFP[10]) on ViCo[9] based on the same speaker and reference image. As shown in Figure 7, the results generated by our method are the most similar to the ground-truth videos, demonstrating great superiority in the realism of facial motions.

D. Limitations and Social Impact

Although our method can achieve real-time interactive motion generation, its application scope is limited to the head and cannot cover the body generation, which deserves further research in the future. In practical applications, our generation method has many positive effects, such as the character animation in movies, creating virtual hosts for advertising, and developing interactive teaching tools to provide immersive experience. However, it may be abused in some scenarios (*e.g.* creating false content for bullying others, creating deceptive videos for spreading misinformation.). In order to prevent the technology from

Figure 6. Qualitative comparisons with state-of-the-art talking head generation methods on HDTF[8] dataset.

being abused, we can ensure that the technology serves a positive purpose by forcibly adding watermarks to the generated content and managing the way the code is obtained.

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. 2

[2] Scott Geng, Revant Teotia, Purva Tendulkar, Sachit Menon, and Carl Vondrick. Affective faces for goal-driven dyadic communication. *arXiv preprint arXiv:2301.10939*, 2023. 2, 3

[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 (CVPR)*, pages 20395–20405, 2022. 3

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 1

[5] Minh Tran, Di Chang, Maksim Siniukov, and Mohammad Soleymani. Dim: Dyadic interaction modeling for social behavior generation. In *European Conference on Computer Vision*, pages 484–503. Springer, 2024. 2, 3

[6] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 2

[7] 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. 2

[8] 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*

Figure 7. Qualitative comparisons with state-of-the-art listenning head generation methods on ViCo[9] dataset.

Recognition, pages 3661–3670, 2021. 2, 4

[9] 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. 3, 5

[10] Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. *arXiv preprint arXiv:2412.04037*, 2024. 3