

# Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation

## –Supplementary Material–

Anonymous CVPR submission

Paper ID 7282

### 1. Detailed Experimental Setup

**Datasets.** For training, we used the AV Speech [1] dataset (filtered to 662 hours) and the VFHQ [7] dataset (filtered to 2 hours) for pretraining, along with additional talking-head video data collected from the internet (32 hours) for supervised fine-tuning (SFT). For validation, we used the HDTF [8] dataset (filtered to 0.83 hours), RAVDESS [3] dataset (filtered to 0.55 hours, public high-definition indoor talking scene dataset with rich emotions) and supplementary internet-sourced data (0.49 hours). *Note that the supplementary internet-sourced data is only used for qualitative comparison.* To ensure data quality for both training and validation, we first applied the Mediapipe [4] face detection tool to detect face regions, filtering out instances where facial movement exceeded 50%. We further refined the data using Sync-C and Sync-D to exclude samples with low lip-sync scores.

**Metrics.** The evaluation metrics for the portrait image animation approach include Fréchet Inception Distance [2] (FID), Fréchet Video Distance [6] (FVD), Synchronization-C [5] (Sync-C), and Synchronization-D [5] (Sync-D). FID and FVD assess the similarity between generated images/videos and real data, with lower values indicating better, more realistic outputs. Sync-C and Sync-D measure lip synchronization in terms of content and dynamics, with higher Sync-C and lower Sync-D scores indicating better audio alignment.

### 2. Detailed Human Evaluation

To assess the quality of the generated talking head animations, a human evaluation was conducted, focusing on participants' subjective assessments of lip synchronization, body movement realism, and temporal coherence.

**Participants.** The study included 30 participants, with 66.7% aged 24-30 and 33.3% aged 30-40. The gender distribution was 30% male and 70% female, and 83.3% had prior experience with AIGC models.

**Task and Measurement.** Participants rated each animation on a 5-point Likert scale, assessing its coherence with the input and the quality of reasoning in the animation. A total of 100 videos were presented in random order to ensure unbiased evaluation, providing insights into subjective perceptions of animation quality and alignment with natural expressions.

**Results.** As shown in Fig. 1, participants consistently rated Teller as superior in lip synchronization, body movement realism, and temporal coherence compared to the four benchmark methods. Teller also showed low variance in scores for lip synchronization and body movement realism, demonstrating its robustness and consistency in producing realistic movements.

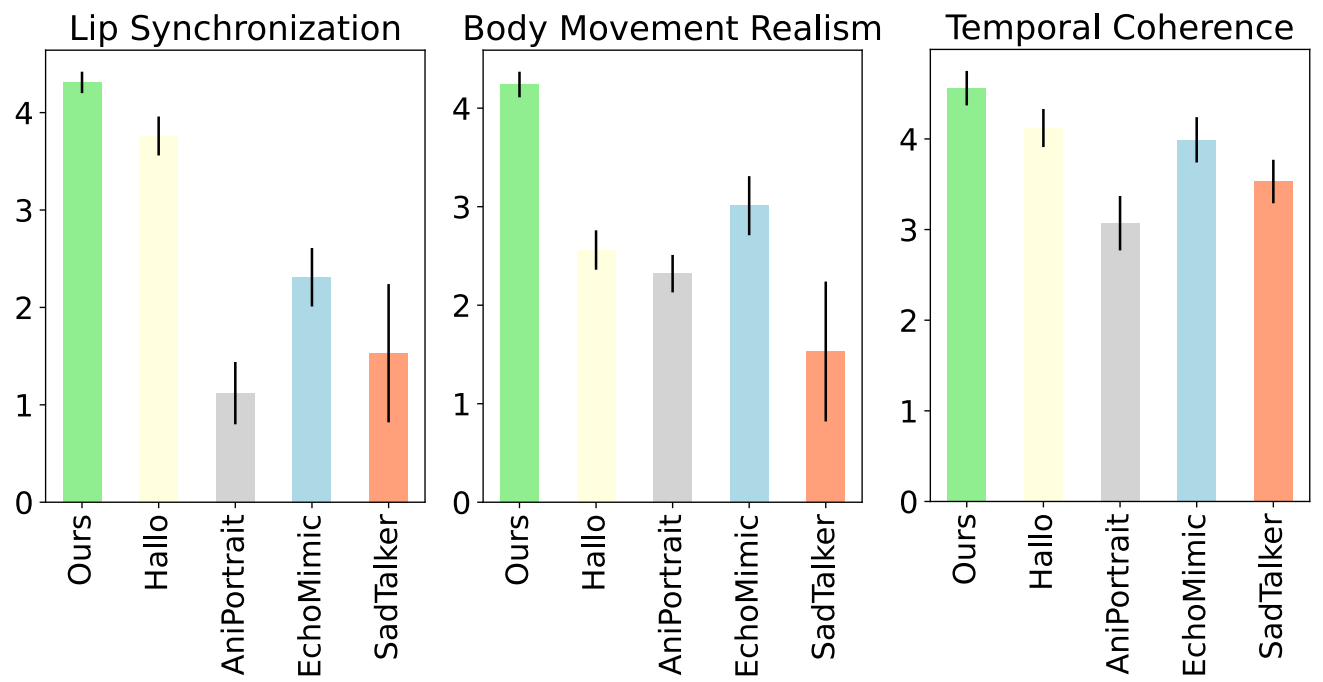


Figure 1. Human evaluation results among our proposed Teller and other SoTA methods.

**References**

- [1] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):1–11, 2018. 1
- [2] 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. *Advances in neural information processing systems*, 30, 2017. 1
- [3] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 1
- [4] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1
- [5] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 1
- [6] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1
- [7] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 1
- [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 Recognition*, pages 3661–3670, 2021. 1