LIVATAR-1: REAL-TIME TALKING HEADS GENERATION WITH TAILORED FLOW MATCHING

Haiyang Liu* **Xiaolin Hong*** **Xuancheng Yang*** **Yudi Ruan***
Xiang Lian **Michael Lingelbach** **Hongwei Yi** **Wei Li†**
Hedra Inc.

ABSTRACT

We present **Livatar**, a real-time audio-driven talking heads videos generation framework. Existing baselines suffer from limited lip-sync accuracy and long-term pose drift. We address these limitations with a flow matching based framework. Coupled with system optimizations, **Livatar** achieves competitive lip-sync quality with a 8.50 LipSync Confidence on the HDTF dataset, and reaches a throughput of 141 FPS with an end-to-end latency of 0.17s on a single A10 GPU. This makes high-fidelity avatars accessible to broader applications. Our project is available at https://www.hedra.com/ with with examples at https://h-liu1997.github.io/Livatar-1/.

1 INTRODUCTION

Recent breakthroughs in Large Language Models (LLMs) (Touvron et al., 2023; OpenAI, 2023; Anil et al., 2023) and real-time Text-to-Speech (TTS) (Kim et al., 2021; Ren et al., 2020; Shen et al., 2018) systems have paved the way for highly interactive, streaming AI agents. To truly unlock their potential, these AI agents require visual embodiments, enabling new applications in education, sales, and virtual companionship.

A typical scenario involves LLMs and TTS systems generating a streaming audio response based on a user's input. The remaining key problem to realizing these visualized agents is a real-time and streaming model that can generate talking-head videos from a single image and the streaming audio.

Current approaches have two key problems: limited lip-sync accuracy and long-term pose drift, where cumulative errors cause the head's pose and shape to deviate over time. In this work, we present **Livatar**[1], a system designed to address these challenges and achieve production-ready quality and performance.

2 EXPERIMENTS

We focus on automated, no-reference metrics for evaluation. Following the evaluation protocol of recent video generation benchmarks (Huang et al., 2024), we assess our method across four key dimensions: lip-sync quality (Chung & Zisserman, 2016), content similarity (Radford et al., 2021), image quality (Huang et al., 2024), and motion dynamics (Huang et al., 2024).

We compare **Livatar** with several leading talking-head generation models: SadTalker (Zhang et al., 2023), Real3DPortrait (Ye et al., 2024), Hallo3 (Cui et al., 2024), Sonic (Ji et al., 2025), and our reproduced INFP (Zhu et al., 2025). For evaluation, we construct a unified test set by randomly sampling 100 clips each from the HDTF and our internal datasets, these clips were filtered out from training. All input faces are cropped and resized to 512x512.

As shown in Table 1, our method shows best lip-sync performance over all baselines. More video results are available on our project page.

*Equal Contribution
†Project Lead
[1]This technical report is a shorten version only summarizing the performance of the Livatar due to the intellectual property policy, the complete version was finished earlier.

Table 1: **Comparison with existing methods.** We compare our method with state-of-the-art video diffusion models (*offline*) and other methods (*realtime*) on the HDTF-100 (*left*) and our Internal-100 (*right*) test sets. $*$ denotes our reproduced version. CS and WR are Content Similarity and user study Win Rate (others vs. ours), respectively.

	Cost	Sync-C↑	CS↑	Quality↑	Dynamic↑	Sync-C↑	CS↑	Quality↑	Dynamic↑	WR%
GroundTruth		7.614	0.928	0.642	0.660	6.995	0.885	0.656	0.782	-
Hallo3 (Cui et al., 2024)	*offline*	6.814	0.915	0.638	**0.870**	6.093	0.895	0.633	0.792	26.8
Sonic (Ji et al., 2025)	*offline*	8.495	0.935	0.626	0.600	7.998	0.916	0.616	**0.832**	42.4
SadTalker (Zhang et al., 2023)	*realtime*	6.704	**0.965**	**0.697**	0.080	6.547	**0.961**	**0.687**	0.020	4.5
Real3DPortrait (Ye et al., 2024)	*realtime*	6.811	0.943	0.637	0.030	6.529	0.934	0.602	0.000	6.8
INFP* (Zhu et al., 2025)	*realtime*	7.357	0.928	0.633	0.780	6.635	0.907	0.612	0.690	28.9
Livatar (Ours)	*realtime*	**8.501**	0.944	0.645	0.800	**8.014**	0.940	0.636	0.772	**50.0**

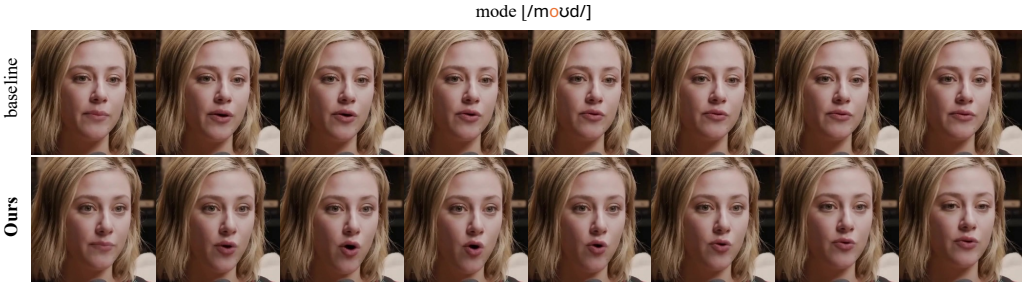

Figure 1: **Lip Synchronization Comparison.** Compare with baseline method, Livatar better handles mouth movements for sounds with strong lip closures, like plosives.

Figure 2: **Long Video Generation Comparison.** Compare with baseline method, Livatar generates long videos with improved appearance consistency.

We implement several system-level optimizations to achieve real-time performance. These optimizations cumulatively reduce the inference latency for a single chunk (generating 24 new frames) from a baseline of 1.1s to 0.17s on an A10 GPU. This achieves a final throughput of 141 FPS. Compared with offline methods, which take around 20 seconds of end-to-end latency and achieve a 0.1 FPS throughput on an H100 (Kong et al., 2024), our method is more efficient for real-time interaction.

3 Conclusion

We presented **Livatar**, a system that generates real-time, streaming talking-head videos from a single image and an audio signal. Our system addresses two limitations of existing models, *i.e.,* limited lip-sync accuracy and long-term pose drift. Our extensive experiments show that **Livatar** achieves competitive performance in both lip-sync quality and inference speed on consumer-grade hardware, demonstrating its suitability for practical, real-world applications.

REFERENCES

Karan Anil, Hyung Won Chung, Zoubin Ghahramani, Slav Petrov, Jing Yu Koh, Tian Lan, Aditya Siddhant, Jonathan Geisler, et al. Palm 2 technical report. arXiv preprint arXiv:2305.10403, 2023.

J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. arXiv preprint arXiv:2412.00733, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21807–21818, 2024.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 193–203, 2025.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the 38th International Conference on Machine Learning (ICML), 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603, 2024.

OpenAI. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pp. 8748–8763. PMLR, 2021.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In Proceedings of Interspeech, 2020.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4779–4783, 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503, 2024.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Wenping Wang, and Qifeng Chen. 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 (CVPR), 2023.

Yucheng Zhu, Anpei Chen, Lingjie Liu, Zhixin Piao, Duygu Ceylan, Nicolas Chappuis, Tuanfeng Wang, and Christian Theobalt. Infp: Identity-neutral facial performance. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.