AURA: AUdio-dRiven streaming Avatar

Nathaniel Cohen[1,2], Nicolas Dufour[1], Amélie Royer[1], Alasdair Newson[2], and Patrick Perez[1]

[1] Kyutai, France {nathaniel.cohen, nicolas.dufour, amelie, patrick}@kyutai.org
[2] ISIR, Sorbonne Université, France anewson@isir.upmc.fr

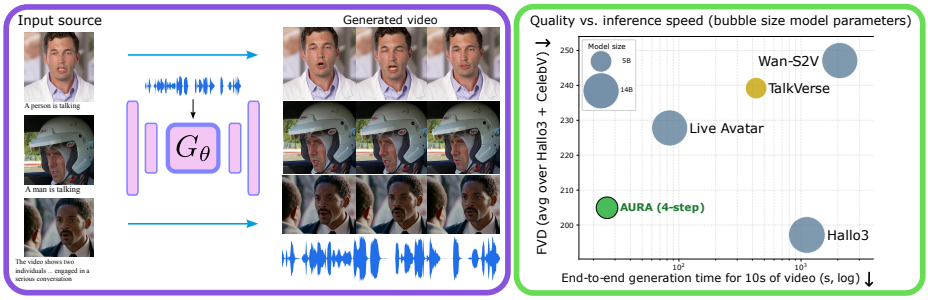

Fig. 1: AURA generates 720p talking-head videos at 9.7 fps on a single GPU in a causal streaming manner (0.64 s chunks). *Left:* three samples from our speech-to-video model G_θ for the same audio but different reference identities and text prompts, temporally aligned across rows. *Right:* quality–speed trade-off on CelebV-HQ and Hallo3, reporting FVD vs. end-to-end time for 10 s of video (log scale, lower is better on both axes); circle area is proportional to parameter count. AURA and Live Avatar are streaming; TalkVerse, Hallo3, and Wan-S2V are non-causal.

Abstract. Current audio-driven video diffusion models synthesize realistic talking-head videos but require dozens of denoising steps over multi-second chunks, making them unsuitable for streaming applications such as voice assistants and virtual avatars, where low latency and online causal generation are essential. AURA generates 720p video chunk-by-chunk at 9.7 fps with a rolling KV cache, requiring only the current chunk's audio and thus introducing no delay between audio capture and generation, making it well-suited to interactive avatar applications. Our two-stage distillation pipeline first trains the causal student to match the teacher's trajectories under a chunk-wise causal attention mask, then applies Distribution Matching Distillation (DMD) with backward simulation to close the few-step quality gap. On standard talking-head benchmarks, AURA matches the teacher's quality while running 17× faster.

1 Introduction

Recent speech-to-speech models [6,16] enable real-time full-duplex voice interaction, but generate audio only. Since human communication also relies on visual

cues, lip motion shapes perceived phonemes [14], and an embodied face aids comprehension and enhances social presence [2, 18], audio-driven talking-head models synthesize video from a reference image and speech, enabling virtual avatars, synthetic presenters, and game characters. All such applications share one constraint: the avatar must be generated on-the-fly, from an unbounded audio stream.

State-of-the-art talking-head models build on large pretrained video diffusion transformers: Hallo3 [4] adapts CogVideoX [23], and TalkVerse [21], our teacher, adapts Wan2.2 to produce 720p video. All rely on bidirectional attention and dozens of denoising steps over multi-second chunks, taking minutes per chunk on a single GPU, precluding streaming. Distillation compresses such teachers into causal few-step students [10, 26], but existing work targets text- or image-to-video, or, for audio-driven generation, relies on 14–18B models needing multiple GPUs for real-time 480p [11, 12, 17]. The concurrent work Avatar Forcing [5] further requires a future-audio look-ahead, incurring an incompressible delay. A further obstacle is data scarcity: public clips are too short for distillation from teachers like TalkVerse.

We introduce AURA, a 5B-parameter causal few-step speech-to-video model that streams 720p talking-head video at 9.7 fps on a single GPU, a 17× speedup over its teacher. AURA is trained via a two-stage distillation pipeline: Stage 1 aligns the causal student with the teacher's ODE trajectories; Stage 2 closes the few-step gap with Distribution Matching Distillation (DMD) and backward simulation. Inference relies on a rolling KV cache with bounded memory. On standard benchmarks, AURA matches its teacher on visual quality (FVD, JEDi) and audio-visual alignment while achieving the highest throughput among single-GPU audio-driven baselines. Figure 1 illustrates qualitative results and the speed–quality trade-off against prior work.

2 Method

AURA distills the non-causal audio-driven teacher TalkVerse into a 4-step causal student that streams 720p avatars in constant memory, via the two-stage pipeline of Fig. 2.

AURA Architecture. TalkVerse [21] is a 5B DiT built on Wan2.2-5B, conditioned on a reference image, Wav2Vec2 speech features [1] injected via cross-attention, and a T5-XXL prompt [15]; its VAE applies $(t, h, w)=(4, 16, 16)$ compression for efficient 720p generation. We introduce two modifications turning this 50-step bidirectional teacher into a streaming 4-step causal student. *(i) Chunked causal attention.* The student generates $n_c=4$ latent frames ($\sim$0.64 s) per chunk; a block-wise causal mask blocks attention to future chunks while keeping it bidirectional within a chunk. *(ii) Per-chunk mixed timesteps.* A naive KV cache stores one entry per denoising step, scaling memory linearly with step count. We instead train the student to handle chunks at heterogeneous noise levels, each with its own timestep t_i, so that inference caches each past chunk at a single low-noise level, one shared cache across steps, a 4× memory reduction.

Stage 1 — ODE Teacher Forcing. Applying DMD directly to a causal student is unstable when its architecture differs from the teacher's [26]. We therefore initialize the student by regressing it onto teacher ODE trajectories. The dataset stores latents at five timesteps $\{0, 0.25, 0.5, 0.75, 1\}$ along the teacher's 50-step ODE; the student is queried at per-chunk mixed timesteps sampled from $\{0.25, 0.5, 0.75, 1\}$ and trained to predict the clean latent x_0 via MSE:

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}_{t,z_t,c}\big[\|\hat{x}_0(z_t, t, c) - x_0\|_2^2\big]. \tag{1}$$

Stage 2 — Backward-Simulation DMD. Stage 2 closes the few-step gap with DMD [25], whose gradient is a score gap evaluated on student samples. We keep three networks: the frozen teacher f_{teacher}, the student G_θ (from Stage 1), and a critic f_ϕ initialized from the teacher and updated $5\times$ per student step. Following DMD2's [24] *backward simulation*, each iteration (i) unrolls G_θ without gradients into intermediate states along its own trajectory, (ii) re-feeds them through G_θ with gradients under mixed timesteps, and (iii) evaluates the DMD gradient at the resulting $\hat{x}_0$:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{t,\,\hat{x}_0}\left[\big(f_\phi(\hat{x}_t, t) - f_{\text{teacher}}(\hat{x}_t, t)\big)\frac{\partial \hat{x}_0}{\partial \theta}\right], \tag{2}$$

where $\hat{x}_0 = G_\theta(z, c)$, z is the input noise, c is the conditioning (audio, prompt, reference image), and $\hat{x}_t = \Psi(\hat{x}_0, t)$ forward-diffuses $\hat{x}_0$ to noise level t. This trains the student on the states it meets at inference.

Inference — Rolling KV Cache. The student generates one chunk at a time in 4 steps with a fixed-size *rolling* KV cache retaining two entries: the **first chunk** as a fixed anchor (akin to an attention sink [22]) for long-term identity consistency, and the **last chunk** for immediate temporal context. Discarding all others gives constant memory regardless of duration; mixed-timestep training lets cached entries sit at a single noise level.

Training Dataset. As public clips are too short for the teacher's $\sim$5 s window, we build a 30K-clip synthetic set by running TalkVerse on Hallo3 reference images and prompts [4] with TTS speech [27]. **No real video is used at any stage.**

3 Experiments

3.1 Experimental Setup

Architecture and training. The student is fine-tuned from a frozen TalkVerse (5B) with LoRA adapters, generating 720p video in 4-frame chunks at 4 steps per chunk, via ODE teacher forcing (Stage 1) then backward-simulation DMD (Stage 2). We use 64 H100 GPUs. **Baselines.** We compare against our teacher TalkVerse [21] (5B), the non-causal Wan-S2V [8] and Hallo3 [4] ($\sim$14B), and the streaming few-step Live Avatar [11] (14B), our closest competitor. All use the authors' released checkpoints under recommended settings. **Benchmarks and metrics.** We evaluate on HALLO3 TEST (100 clips, 10 s, real English speech)

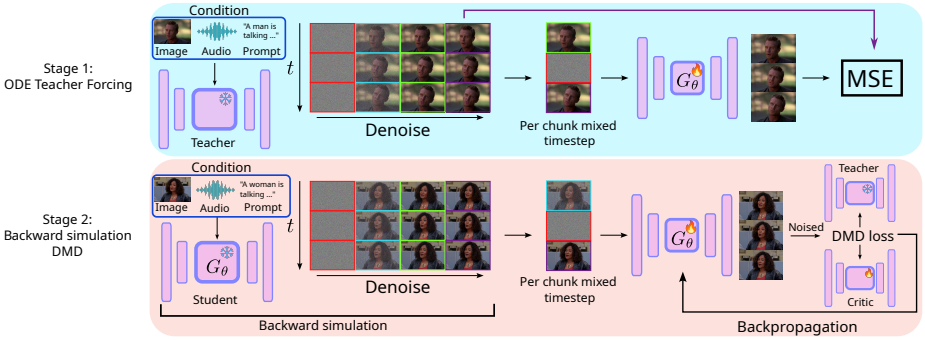

Fig. 2: Two-stage distillation pipeline. *Stage 1 (top)*: the frozen teacher produces an ODE trajectory; a per-chunk mixed-timestep state is fed to the causal student G_θ, trained under MSE. *Stage 2 (bottom)*: G_θ generates a video without gradients (backward simulation); the DMD loss compares teacher and critic scores on states from the student's own trajectory.

Table 1: Main results on HALLO3 TEST (H) and CELEBV-HQ TEST (C). Steps: 50 (TalkVerse, Hallo3), 40 (Wan-S2V), 4 (Live Avatar, AURA); TalkVerse and AURA are 5B, others ∼14B. **Bold**: best, underlined: second.

Method	Test	fps↑	FID↓	FVD↓	JEDi↓	CSIM↑	Sync-C↑	Sync-D↓	WE↓
Hallo3	H	0.22	**43.1**	186.3	**0.16**	0.74	5.06	9.9	0.0036
Wan-S2V	H	0.08	55.2	200.5	0.27	0.71	**5.40**	**9.0**	0.0032
TalkVerse	H	0.58	55.7	215.5	0.42	0.74	3.82	10.5	**0.0019**
Live Avatar	H	1.9	50.8	**182.7**	0.47	0.77	4.90	9.3	0.0039
AURA	H	**9.7**	57.2	191.0	0.37	**0.80**	4.09	10.0	0.0034
Hallo3	C	0.22	**36.0**	**208.0**	**0.257**	0.71	**6.20**	9.4	**0.0038**
Wan-S2V	C	0.08	46.7	293.7	0.640	0.62	**6.20**	**8.9**	0.0061
TalkVerse	C	0.58	56.3	262.9	0.600	0.59	4.80	10.5	0.0059
Live Avatar	C	1.9	43.1	272.9	1.087	0.67	5.90	9.2	0.0057
AURA	C	**9.7**	45.9	218.9	0.477	**0.74**	4.70	10.1	0.0055

and CELEBV-HQ TEST [28] (150 clips, 8–20 s, TTS speech). We report FID [9], FVD [20], and JEDi [13] (MMD on V-JEPA features, more reliable at our sample counts) for visual quality; Sync-C/D [3] for lip-sync; CSIM (ArcFace [7] cosine similarity) for identity; and warping error (RAFT [19]) for temporal consistency. All timings use a single H100.

Results. Table 1 reports our main results. Despite using 4 steps instead of 50, AURA surpasses its TalkVerse teacher on most quality metrics and matches the 14B Wan-S2V on FVD, JEDi, and audio-visual alignment. It leads all baselines on identity preservation (CSIM), which we attribute to the rolling cache anchoring the first generated chunk. Above all, at 9.7 fps end-to-end on one H100, AURA is the fastest by far, a 17× speedup over TalkVerse and 5× over Live Avatar, placing it on the quality–speed Pareto front (Fig. 1, right).

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Cassell, J., Sullivan, J., Prevost, S., Churchill, E.F. (eds.): Embodied Conversational Agents. MIT Press, Cambridge, MA (2000)
3. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Workshop on Multi-view Lip-reading, ACCV (2016)
4. Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., Zhu, S.: Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference, CVPR'25. pp. 21086–21095 (2025)
5. Cui, L., Hu, W., Zhang, W., Yang, Z., Shi, F., Liu, X.: Avatarforcing: One-step streaming talking avatars via local-future sliding-window denoising (2026), https://arxiv.org/abs/2603.14331
6. Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue (2024), https://arxiv.org/abs/2410.00037
7. Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., Zafeiriou, S.: ArcFace: Additive angular margin loss for deep face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 5962–5979 (Oct 2022). https://doi.org/10.1109/tpami.2021.3087709, http://dx.doi.org/10.1109/TPAMI.2021.3087709
8. Gao, X., Hu, L., Hu, S., Huang, M., Ji, C., Meng, D., Qi, J., Qiao, P., Shen, Z., Song, Y., Sun, K., Tian, L., Wang, G., Wang, Q., Wang, Z., Xiao, J., Xu, S., Zhang, B., Zhang, P., Zhang, X., Zhang, Z., Zhou, J., Zhuo, L.: Wan-S2V: Audio-driven cinematic video generation (2025), https://arxiv.org/abs/2508.18621
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
10. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. Advances in Neural Information Processing Systems **38**, 167283–167308 (2026)
11. Huang, Y., Guo, H., Wu, F., Zhang, S., Huang, S., Gan, Q., Liu, L., Zhao, S., Chen, E., Liu, J., Hoi, S.: Live avatar: Streaming real-time audio-driven avatar generation with infinite length (2026), https://arxiv.org/abs/2512.04677
12. Low, C., Wang, W.: Talkingmachines: Real-time audio-driven facetime-style video via autoregressive diffusion models (2025), https://arxiv.org/abs/2506.03099
13. Luo, G.Y., Favero, G.M., Luo, Z.H., Jolicoeur-Martineau, A., Pal, C.: Beyond FVD: An enhanced evaluation metrics for video generation distribution quality. In: International Conference on Learning Representations, ICLR'25. vol. 2025, pp. 66218–66252 (2025)
14. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (Dec 1976). https://doi.org/10.1038/264746a0
15. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)
16. Roy, R., Raiman, J., gil Lee, S., Ene, T.D., Kirby, R., Kim, S., Kim, J., Catanzaro, B.: Personaplex: Voice and role control for full duplex conversational speech models (2026), https://arxiv.org/abs/2602.06053

17. Shen, L., Qiao, Q., Yu, T., Zhou, K., Yu, T., Zhan, Y., Wang, Z., Tao, M., Yin, S., Liu, S.: SoulX-FlashTalk: Real-time infinite streaming of audio-driven avatars via self-correcting bidirectional distillation (2026), https://arxiv.org/abs/2512.23379
18. Sumby, W.H., Pollack, I.: Visual contribution to speech intelligibility in noise. The Journal of the Acoustical Society of America **26**(2), 212–215 (Mar 1954). https://doi.org/10.1121/1.1907309
19. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision, ECCV'20. pp. 402–419. Springer (2020)
20. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges (2019), https://arxiv.org/abs/1812.01717
21. Wang, Z., Wang, J., Ma, K., Lin, D., Zhou, B.: Talkverse: Democratizing minute-long audio-driven video generation (2025), https://arxiv.org/abs/2512.14938
22. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: International Conference on Learning Representations, ICLR'24. vol. 2024, pp. 21875–21895 (2024)
23. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations, ICLR'25. vol. 2025, pp. 83048–83077 (2025)
24. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T.: Improved distribution matching distillation for fast image synthesis. Advances in neural information processing systems **37**, 47455–47487 (2024)
25. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, CVPR'24. pp. 6613–6623 (2024)
26. Yin, T., Zhang, Q., Zhang, R., Freeman, W.T., Durand, F., Shechtman, E., Huang, X.: From slow bidirectional to fast autoregressive video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR'25. pp. 22963–22974 (2025)
27. Zeghidour, N., Kharitonov, E., Orsini, M., Volhejn, V., de Marmiesse, G., Grave, E., Pérez, P., Mazaré, L., Défossez, A.: Streaming sequence-to-sequence learning with delayed streams modeling. Tech. rep. (2025), https://arxiv.org/abs/2509.08753
28. Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., Loy, C.C.: CelebV-HQ: A large-scale video facial attributes dataset. In: European conference on computer vision, ECCV'22. pp. 650–667. Springer (2022)