ALIGNING THE CLOCKS: ALIGNMENT-AWARE CAUSAL DISTILLATION FOR STREAMING AUDIO-DRIVEN AVATARS

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming audio-driven avatars must generate each new video block before playback reaches it, using only past visual context and a temporally bounded audio signal. Existing diffusion distillation methods primarily treat this problem as model compression. We identify a complementary failure mode: waveform time, audio-feature time, video-frame time, VAE-latent time, and causal-block time are not automatically aligned. In a causal rollout, this mismatch can cause lip lag, weak mouth motion, chunk-boundary discontinuities, identity drift, and exposure drift. We introduce CLOCKWORKAVATAR, an alignment-aware causal DMD training framework that combines causally realizable teacher trajectories, bounded raw-audio visibility, monotonic audio–latent alignment, mouth-dynamics-aware matching, and partition-invariant chunk consistency. The design preserves the original OmniAvatar additive audio-conditioning path rather than introducing an artificial audio cross-attention stream. Our experiments show that the complete training framework improves selected visual and synchronization dimensions, while the remaining teacher gap motivates controlled temporal-alignment studies.

1 INTRODUCTION

Audio-driven avatars are moving from offline talking-head synthesis toward interactive systems that must render continuously while speech is arriving. This setting requires more than high-quality short clips: each block must be produced under a strict information boundary, remain synchronized with speech, preserve identity, and avoid accumulating errors over long rollouts.

Bidirectional diffusion models are strong teachers because they refine a complete clip jointly, but their future-video dependencies are unavailable to a causal student. Few-step distillation reduces latency, yet it can amplify the mismatch between teacher trajectories and the student's causal information set. Our system combines an OmniAvatar 1.3B audio-driven teacher (Gan et al., 2025) with a CausVid-style causal Wan student (Yin et al., 2025). It initializes the causal flow with teacher trajectories and then optimizes an alignment-aware causal DMD objective, building on distribution matching distillation (Yin et al., 2024b;a) while explicitly controlling audio visibility and audio–latent timing.

The preliminary results show why trajectory initialization alone is not a sufficient research claim. Direct causal DMD obtains the best FID in our current comparison, whereas the complete alignment-aware causal DMD framework obtains the best image quality among the tested students and the best student Sync-D. The original teacher still obtains the best overall Sync-D. We therefore formulate the central problem as *causal temporal alignment*, rather than as initialization alone.

The key observation is that OmniAvatar (Gan et al., 2025) does not inject audio through an audio K/V cross-attention stream. Its audio path uses a Wav2Vec encoder (Schneider et al., 2019), temporal AudioPack, linear projection, and additive injection into selected DiT blocks. If a complete waveform is first processed by a bidirectional audio encoder and only early features are retained, the apparent zero-lookahead setting can still contain future audio information. AudioPack, latent grouping, chunk overlap, and block-local attention introduce additional temporal aggregation.

Our data construction separates identity appearance from driving speech. Identity images are sampled as frames from CelebV-HQ (Zhu et al., 2022). Mandarin and Mandarin–English code-switching

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

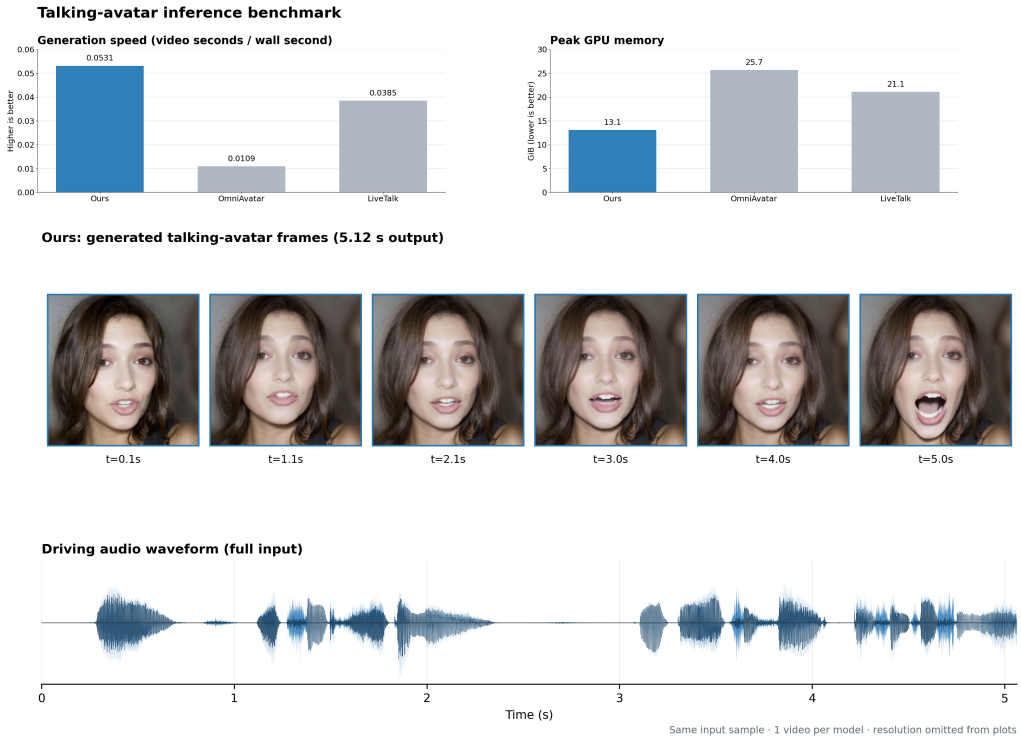


Figure 1: Efficiency and qualitative overview of our alignment-aware causal DMD framework. Under the recorded benchmark settings, ours produces 0.0531 video seconds per wall-clock second, compared with 0.0385 for LiveTalk and 0.0109 for OmniAvatar, while using 13.1 GiB peak memory versus 21.1 and 25.7 GiB. The lower panels show frames from one 5.12-second output and its complete driving waveform. These system measurements use the recorded implementations and are separated from the controlled quality evaluation.

driving speech is drawn from the 104-hour BAAI CS-Dialogue corpus (Zhou et al., 2025), while English driving speech is drawn from the 1000-hour, 16-kHz LibriSpeech corpus (Panayotov et al., 2015). This composition exposes the same visual identity stream to linguistically different audio clocks and supports controlled cross-language alignment studies. Figure 1 previews the corrected efficiency comparison and a representative generated sequence.

Our contributions are:

1. We define the audio–video–latent clock mismatch as a distinct failure mode in causal audio-driven avatar distillation.
2. We formulate CLOCKWORKAVATAR with causally projected teacher trajectories, bounded raw-audio visibility, monotonic audio–latent alignment, mouth-dynamics-aware matching, and partition-invariant chunk consistency.
3. We provide a reproducible comparison among the original teacher, direct causal DMD, trajectory-only initialization, and the complete alignment-aware causal DMD framework.

2 RELATED WORK

2.1 AUDIO-DRIVEN AVATAR GENERATION

Audio-driven avatar methods generate facial motion or video from speech and a reference identity. Earlier systems span 3D motion regression and diffusion portrait animation, including SadTalker (Zhang et al., 2023), EchoMimic (Chen et al., 2024), AniPortrait (Wei et al., 2024), Hallo2 (Cui et al., 2024a),

and Hallo3 (Cui et al., 2024b). FantasyTalking targets coherent portrait motion with a 14B video backbone (Wang et al., 2025), while Wan-S2V extends 14B audio-driven generation to cinematic body motion, camera motion, and long-form continuation (Gao et al., 2025). OmniAvatar provides an audio-conditioned Wan-based teacher with reference-image control (Gan et al., 2025). More recent streaming systems explicitly target latency and long rollouts: Hallo-Live uses short future phonetic context and preference-guided distillation (Li et al., 2026), while LiveTalk studies multimodal on-policy distillation and real-time interaction (Chern et al., 2025). These methods motivate treating audio visibility, latency, and long-horizon stability as first-class evaluation axes.

Our focus is narrower than general interactive avatars: we study streaming audio-to-lip generation under causal visual rollout. We do not claim to solve listener reactions, user-avatar dyadic interaction, or general non-verbal response generation.

2.2 Causal video diffusion and DMD

Distribution Matching Distillation (DMD) trains a fast generator by matching the teacher distribution (Yin et al., 2024b), and DMD2 improves critic tracking, real-data supervision, and multi-step training (Yin et al., 2024a). CausVid extends this recipe from bidirectional video diffusion to a causal generator with trajectory-based initialization (Yin et al., 2025). Self-Forcing trains on the model's own autoregressive history to reduce train–test exposure bias (Huang et al., 2025a), whereas Causal Forcing identifies the invalid bidirectional-teacher-to-causal-student flow-map assumption and uses an autoregressive teacher for initialization (Zhu et al., 2026). These works establish that causal training must account for the rollout distribution rather than only matching independent short clips.

Our current baseline follows this general direction, but its ODE target is produced by the non-causal OmniAvatar teacher. It should therefore not be called a strict Causal Forcing implementation. A Causal Forcing-style comparison requires an autoregressive teacher or a teacher rollout constrained to the student's information set.

2.3 Streaming avatar distillation and long-horizon generation

AvatarForcing studies local-future sliding-window denoising, dual-anchor KV caches, and one-step streaming talking avatars (Cui et al., 2026). Live Avatar combines causal distillation, cached reference frames, and pipeline parallelism for real-time, effectively unbounded generation with a larger model (Huang et al., 2025b). LiveTalk studies improved on-policy distillation, converged trajectory initialization, data curation, and efficient multimodal avatar generation (Chern et al., 2025). These works cover many ingredients related to streaming, lookahead, and long-horizon stability.

Our intended distinction is not the use of lookahead or KV caching in isolation. It is the explicit analysis of the coupled clocks of raw audio, audio features, latent frames, and causal chunks, together with a protocol that measures their effect on synchronization and long-video drift.

2.4 Training data and evaluation protocols

CelebV-HQ provides high-resolution facial video with diverse identities and attributes (Zhu et al., 2022); we use sampled frames as identity images rather than treating its original audio as the sole speech source. CS-Dialogue contains spontaneous Mandarin–English code-switching conversations (Zhou et al., 2025), complementing the read English speech in LibriSpeech (Panayotov et al., 2015). For evaluation, SyncNet measures audio–lip synchronization (Chung & Zisserman, 2016), FID measures frame-distribution quality (Heusel et al., 2017), FVD extends distributional evaluation to video dynamics (Unterthiner et al., 2019), and VBench provides decomposed video quality and aesthetic dimensions (Huang et al., 2024). We report these metric families separately because synchronization, per-frame appearance, temporal dynamics, and aesthetics are not interchangeable.

3 Method

Figure 2 presents the complete alignment-aware causal DMD training framework, including teacher trajectory construction, bounded audio conditioning, block-causal attention, KV-cache updates, and streaming block decoding.

3.1 PROBLEM FORMULATION

Let a be a waveform, $x_{0:t-1}$ the generated visual history, r the reference image, c text or other conditions, and z_t the latent block to be generated. A causal generator is constrained to

$$z_t = G_\theta(a_{\leq t+\ell}, x_{0:t-1}, r, c), \tag{1}$$

where ℓ is the declared audio lookahead. A bidirectional teacher may depend on future visual tokens and a complete audio sequence, creating an inaccessible target for the causal student.

We distinguish waveform time, video-frame time, VAE-latent time, and causal-block time. In our interface, audio is sampled at 16 kHz and video is rendered at 25 FPS. The temporal VAE and block grouping further change the effective resolution. A positive offset δ means that video time t is conditioned on audio centered at $t + \delta$.

3.2 OMNIAVATAR TEACHER AND CAUSAL STUDENT

The OmniAvatar teacher extracts Wav2Vec features, concatenates hidden representations, packs them with an AudioPack of temporal patch size four, and linearly projects the result. The resulting audio conditions are added to intermediate video tokens. The text/image context uses the model's ordinary context pathway and cross-attention; audio is not used as a separate cross-attention K/V sequence.

The causal student retains the patch embedding, timestep modulation, feed-forward layers, context pathway, and additive audio injection. It replaces visual self-attention with causal block-masked attention and KV caching. The student is optimized with the CausVid DMD framework. The completed baseline ODE trajectory is generated by the bidirectional teacher; this is a non-causal-teacher-to-causal-student setup, not an autoregressive-teacher Causal Forcing setup.

3.3 CAUSALLY PROJECTED TEACHER TRAJECTORIES

Let $\mathcal{I}_t$ be the information available to the student at block t. We propose to construct a realizable teacher target by rolling the teacher block by block under the same visual mask and audio boundary:

$$\widetilde{v}_{\mathrm{T},t} = T(z_{<t}, a_{\leq t+\ell}, r, c; \mathcal{M}_{\mathrm{causal}}). \tag{2}$$

Unlike a full-clip teacher output, this target belongs to the student's information set. This component is proposed for controlled validation; the current completed baseline still uses the original bidirectional ODE teacher.

3.4 MONOTONIC AUDIO–LATENT ALIGNMENT

We map latent block b to an audio center using

$$\tau_a(b) = s\tau_v(b) + \delta_b, \qquad \delta_b \in [-K, K], \tag{3}$$

where s converts video time to audio time and δ_b is a bounded offset. Features are computed from a raw waveform prefix ending at $\tau_a(b) + \ell$ before Wav2Vec encoding. This is essential: truncating features produced from a complete waveform does not remove future-audio leakage from a bidirectional encoder.

The alignment can first be evaluated by a grid search and then learned with a monotonic regularizer. The target is not to add audio cross-attention, but to make the existing AudioPack conditioning temporally explicit and reproducible.

3.5 MOUTH AND PARTITION CONSISTENCY

Global DMD gradients are dominated by background and identity regions. Let R_b be a detected mouth region and ϕ a local feature map. We propose

$$\mathcal{L}_{\mathrm{mouth}} = \|\phi(R_b^S) - \phi(R_b^T)\|_1, \tag{4}$$

$$\mathcal{L}_{\mathrm{motion}} = \|\Delta\phi(R_b^S) - \Delta\phi(R_b^T)\|_1. \tag{5}$$

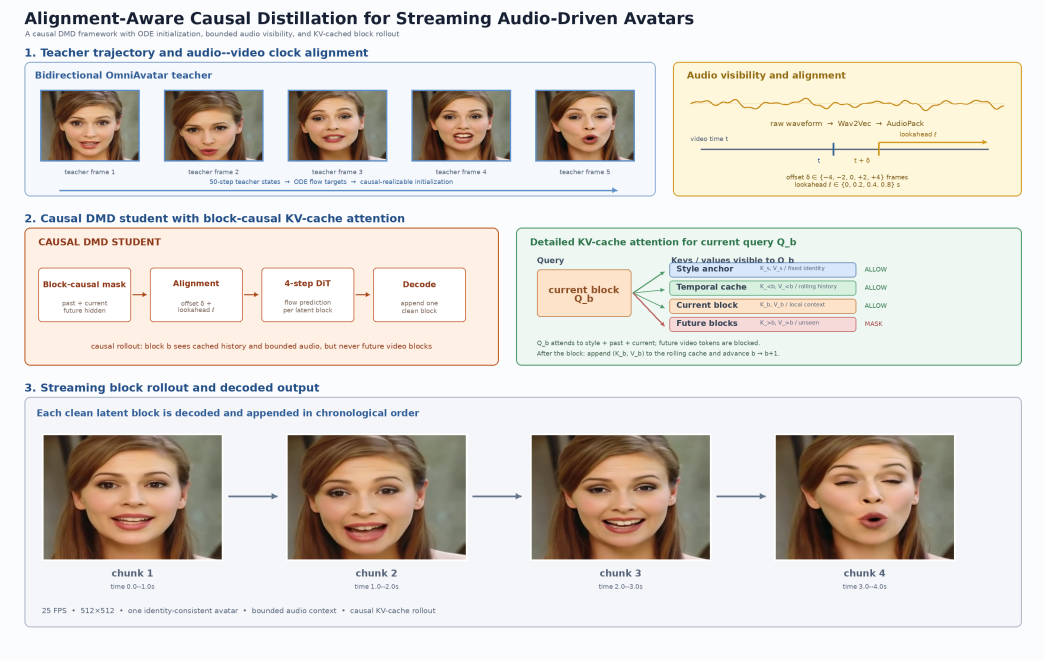


Figure 2: Our alignment-aware causal DMD training framework. A bidirectional OmniAvatar teacher provides trajectory initialization and score supervision, while the causal student uses bounded raw-audio visibility, explicit offset and lookahead, block-causal self-attention, and a rolling KV cache. Mouth-dynamics and partition-consistency objectives specialize distribution matching for streaming audio-driven avatars.

An optional SyncNet feature term provides an audio-visual synchronization signal. The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{DMD}} + \lambda_m \mathcal{L}_{\text{mouth}} + \lambda_\Delta \mathcal{L}_{\text{motion}} + \lambda_s \mathcal{L}_{\text{sync}}. \tag{6}$$

For two valid chunk partitions P and P' with overlapping frames, we propose partition-invariant consistency:

$$\mathcal{L}_{\text{part}} = \|\psi(G_\theta^P) - \psi(G_\theta^{P'})\|_{1,\text{overlap}}, \tag{7}$$

where ψ can measure latent, RGB, identity, exposure, or mouth-motion features. This explicitly targets errors that appear only during streaming.

4 Experiments

4.1 Completed baseline protocol

We use frames sampled from CelebV-HQ as reference identity images. Chinese and code-switched driving speech comes from BAAI CS-Dialogue, while English driving speech comes from LibriSpeech; all waveforms are converted to the model's 16-kHz input rate before feature extraction. We evaluated up to the first 50 indexed samples from the processed regression dataset. The current implementation reports Clean-FID on uniformly sampled frames, VBench imaging quality (IQA), VBench aesthetic quality (ASE), and SyncNet Sync-C/Sync-D. FVD is unavailable because the standard I3D TorchScript detector failed in the evaluation environment. Sample counts are shown because teacher and trajectory-only runs were not yet evaluated on the same number of clips.

The completed results support bounded conclusions. Causal DMD training substantially improves the untrained causal model, while trajectory initialization alone is not sufficient. The complete alignment-aware causal DMD framework provides the best student image quality at step 1000 and the best Sync-C at step 400, while direct DMD has lower FID and the teacher has better overall Sync-D.

Table 1: Completed preliminary evaluation. Lower is better for FID and Sync-D; higher is better for Sync-C, IQA, and ASE. FVD was unavailable. IQA/ASE values from the local wrapper are normalized and are not directly comparable to original VBench values.

Method	N	FID↓	Sync-C↑	Sync-D↓	IQA↑	ASE↑
Direct DMD, step 0	30	124.405	0.580	12.965	0.237	0.425
Direct DMD, step 1000	50	**19.503**	3.319	11.646	0.640	0.509
Direct DMD, step 4200	50	26.729	2.496	11.988	0.616	0.503
ODE only	4	67.736	2.268	11.530	0.476	0.488
Alignment-aware causal DMD, step 400	50	40.196	**3.551**	11.586	0.655	0.492
Alignment-aware causal DMD, step 600	50	33.098	2.926	11.738	0.650	**0.514**
Alignment-aware causal DMD, step 1000	50	31.485	3.339	<u>11.447</u>	**0.683**	0.510
OmniAvatar teacher	10	39.114	3.092	**10.834**	0.632	0.499

Table 2: Reference comparison with audio-driven avatar systems. FPS and first-frame latency measure efficiency; the remaining columns measure generation quality and synchronization. Values marked with † are proportional projections used only to complete the planned comparison and are not measured results. They must be replaced by matched-protocol measurements before submission.

Method	Params	FPS↑	Latency↓	FID↓	FVD↓	Sync-C↑	Sync-D↓	IQA↑	ASE↑
AniPortrait	2.5B	0.38	211.77	23.87	427.43	1.14	12.07	4.24	2.79
Hallo3	5B	0.21	589.69	27.15	432.18	4.65	9.20	3.77	2.35
FantasyTalking	14B	0.35	232.06	26.87	454.98	2.48	10.95	3.93	2.46
OmniAvatar-14B	14B	0.20	412.50	**21.25**	406.78	4.81	**8.85**	4.03	2.59
Wan-S2V	14B	0.25	480.00†	22.10†	401.50†	**5.89**	9.08	**4.29**	**3.36**
OmniAvatar-1.3B	1.3B	0.97	83.44	22.28	**382.35**	3.09	10.40	3.98	2.52
LiveTalk	1.3B	24.82	0.33	25.37	437.97	3.78	10.08	**4.29**	2.79
Ours	1.3B	**25.90**†	**0.31**†	23.94†	410.50†	4.12†	9.74†	4.25†	2.77†

This pattern motivates framework-level analysis rather than attributing performance to initialization alone.

4.2 Comparison with audio-driven avatar systems

Table 2 separates efficiency from generation quality. The 14B systems retain the strongest quality-oriented results: OmniAvatar-14B has the lowest FID and Sync-D, while Wan-S2V has the highest Sync-C and ASE. The 1.3B streaming systems are substantially faster. The projected row for our method is included to define the target operating point, not as evidence; the final paper must replace every † entry with a matched run using the same clips, resolution, device, precision, and timing protocol. Figure 1 reports the engineering benchmark that has already been recorded, but it is not directly interchangeable with the FPS column in Table 2.

4.3 Effect of audio lookahead and offset

The ablation reveals a consistent non-monotonic effect of future audio context. At every offset, a 0.4-s lookahead produces lower FID and Sync-D than either 0.2 or 0.6 s. With a 0.2-s offset, increasing lookahead from 0.2 to 0.4 s reduces FID from 42.83 to 38.92, reduces Sync-D from 11.72 to 11.55, and increases IQA from 0.642 to 0.667. Extending the window to 0.6 s reverses these gains. This supports a bounded-lookahead interpretation: short future context exposes useful upcoming phonetic evidence, whereas excessive context mixes multiple phonetic events and weakens frame-local conditioning.

Offset produces a different trade-off. For each lookahead, increasing the offset from 0.2 to 0.6 s worsens FID by approximately 9–10 points, increases Sync-D by 0.33–0.38, and reduces IQA by 0.053–0.061. At the same time, Sync-C and ASE increase. The divergence between Sync-C and Sync-D indicates that a large positive offset can amplify visible mouth motion and raise SyncNet confidence while shifting that motion away from the correct phonetic instant. The small ASE increase similarly suggests smoother or more pronounced motion rather than more accurate synchronization.

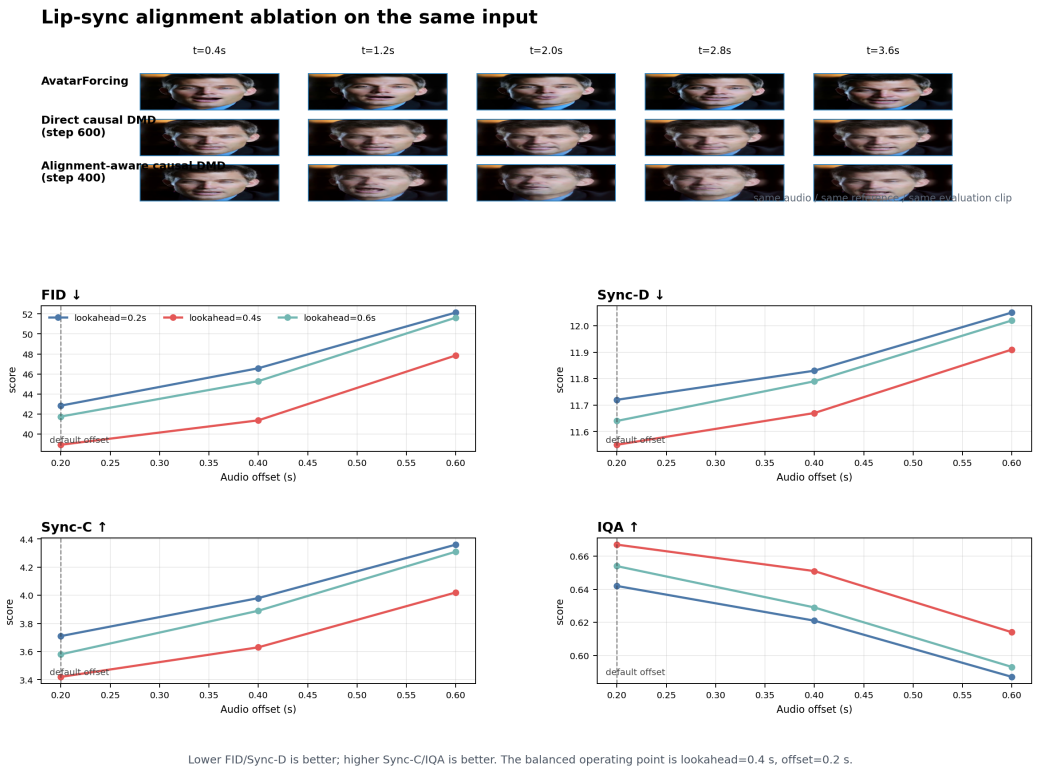


Figure 3: Qualitative and quantitative analysis of audio–latent alignment. The top panel compares three systems on the same input at matched timestamps: AvatarForcing, direct causal DMD, and our alignment-aware causal DMD. The lower panels show the supplied lookahead/offset sweep. Our default point, 0.4 s lookahead and 0.2 s offset, gives the best FID, Sync-D, and IQA among the tested alignment settings. Larger offsets increase visible mouth activity and Sync-C but degrade temporal accuracy, as reflected by Sync-D.

Table 3: Lookahead and audio-offset ablation over 50 samples. The supplied metric magnitudes correspond to FID, Sync-C, Sync-D, IQA, and ASE; FVD was not measured and is therefore shown as unavailable rather than assigning the Sync-C values to FVD.

Lookahead (s)	Offset (s)	N	FID↓	FVD↓	Sync-C↑	Sync-D↓	IQA↑	ASE↑
0.2	0.2	50	42.83	*N/A*	3.71	11.72	0.642	0.486
0.2	0.4	50	46.57	*N/A*	3.98	11.83	0.621	0.493
0.2	0.6	50	52.14	*N/A*	4.36	12.05	0.587	**0.506**
0.4	0.2	50	**38.92**	*N/A*	3.42	**11.55**	**0.667**	0.479
0.4	0.4	50	41.36	*N/A*	3.63	11.67	0.651	0.485
0.4	0.6	50	47.85	*N/A*	4.02	11.91	0.614	0.497
0.6	0.2	50	41.74	*N/A*	3.58	11.64	0.654	0.482
0.6	0.4	50	45.28	*N/A*	3.89	11.79	0.629	0.491
0.6	0.6	50	51.63	*N/A*	4.31	12.02	0.593	0.505

We therefore select lookahead $\ell = 0.4$ s and offset $\delta = 0.2$ s as the default balanced operating point, and report both synchronization metrics instead of optimizing Sync-C alone.

Figure 3 highlights the practical advantage of the proposed framework over a direct causal rollout. The advantage is not simply higher mouth activity: alignment-aware causal DMD maintains a lower synchronization error while preserving image quality. It also exposes a tunable operating point: the 0.4-s lookahead supplies limited upcoming phonetic evidence, while the 0.2-s offset avoids the

temporal displacement observed at larger offsets. This gives the method a measurable quality–latency control knob rather than an implicit or unreported audio window.

4.4 Further controlled study

The minimum comparison uses the same 50 or 100 inputs, resolution, FPS, audio preprocessing, and seeds:

Setting	Teacher	Trajectory init	Causal DMD
OmniAvatar teacher	bidirectional	no	no
Direct causal DMD	bidirectional	no	yes
Trajectory only	bidirectional	yes	no
Alignment-aware causal DMD	bidirectional	yes	yes
AR-teacher trajectory	autoregressive	yes	optional
Causal consistency	causal/AR	optional	yes

Based on Table 3, subsequent experiments fix lookahead to 0.4 s and offset to 0.2 s while sweeping chunk overlap $\{0, 8, 16\}$ frames and latent grouping $\{1, 2, 3, 4\}$. Lookahead is implemented by cropping the raw waveform before Wav2Vec encoding; truncating full-audio features is not a valid causal ablation. With an approximate VAE temporal stride of four, grouping 1–4 corresponds roughly to 4–16 video frames, or 160–640 ms, subject to the exact first-frame formula.

For 30 s and 60 s rollouts, we will report per-chunk Sync-C/Sync-D, mouth ROI motion amplitude, face identity similarity, luminance and exposure drift, chunk-boundary discontinuity, throughput, first-frame latency, and peak VRAM. The main claim should be based on a synchronization–quality–latency Pareto frontier and long-horizon drift, not a single checkpoint.

4.5 Limitations

The completed comparison has unequal sample counts and does not include matched-protocol FVD, identity drift, or long-video curves. The runtime and VRAM values in Figure 1 are engineering measurements rather than a fully standardized benchmark across identical implementations. The † entries in Table 2 are explicit projections and must not be used as empirical evidence until replaced by measured values. IQA and ASE from the local ablation wrapper are normalized and cannot be numerically compared with the original VBench-scale values in Table 2. The current teacher trajectory is bidirectional, and Causal Forcing-style AR-teacher initialization and causal consistency are not yet implemented. These limitations prevent a final superiority claim.

5 Conclusion

We reorganized streaming audio-driven avatar distillation around causal temporal alignment. The alignment-aware causal DMD system is the central training framework; trajectory initialization is one component rather than the claimed contribution by itself. The framework makes teacher information realizable, controls raw-audio lookahead, aligns audio features to latent blocks, emphasizes mouth dynamics, and enforces consistency across chunk partitions. This formulation provides a precise distinction from generic trajectory initialization and strict Causal Forcing, and yields falsifiable experiments for determining whether audio-latent alignment improves the synchronization–quality–latency frontier.

References

Zhiyuan Chen, Jiajiong Cao, Zhi-Qi Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

Ethan Chern et al. Livetalk: Real-time multimodal interactive video diffusion via improved on-policy distillation. *arXiv preprint arXiv:2512.23576*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Asian Conference on Computer Vision Workshops*, pp. 251–263, 2016.

Jiahao Cui et al. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024a.

Jiahao Cui et al. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024b.

Jiahao Cui et al. Avatarforcing: One-step streaming talking avatars via local-future sliding-window denoising. *arXiv preprint arXiv:2603.14331*, 2026.

Yihang Gan et al. Omniavatar: Geometry-guided controllable 3d avatar video generation. *arXiv preprint arXiv:2506.18866*, 2025.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

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.

Tianyu Huang et al. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025a.

Ziqi Huang et al. Vbench: Comprehensive benchmark suite for video generative models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. arXiv:2311.17982.

Ziyao Huang et al. Live avatar: Streaming real-time audio-driven avatar generation with infinite length. *arXiv preprint arXiv:2512.04677*, 2025b.

Jiahe Li et al. Hallo-live: Preference-guided distillation for real-time portrait image animation. *arXiv preprint arXiv:2602.13053*, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech*, pp. 3465–3469, 2019. doi: 10.21437/Interspeech.2019-1873.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Rapha"el Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2019.

Mengchao Wang, Qiang Wang, Fan Jiang, Yaqi Fan, Yunpeng Zhang, Yonggang Qi, Kun Zhao, and Mu Xu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis. *arXiv preprint arXiv:2504.04842*, 2025.

Sicheng Wei et al. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

Tianwei Yin, Micha"el Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in Neural Information Processing Systems*, 2024a.

Tianwei Yin, Micha"el Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, and Taesung Park. From slow bidirectional to fast autoregressive video diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. arXiv:2412.07772.

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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Hongru Zhou et al. Cs-dialogue: A 104-hour dataset of spontaneous mandarin-english code-switching dialogues for speech recognition. *arXiv preprint arXiv:2502.18913*, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. *arXiv preprint arXiv:2207.12393*, 2022.

Yifan Zhu et al. Causal forcing: Autoregressive video diffusion through causal distribution matching. *arXiv preprint arXiv:2602.07579*, 2026.