INJECTED BUT NOT ACCUMULATED: MEASURING AND STABILIZING SYNCHRONIZATION IN TRAINING-FREE STREAMING AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint audio-video diffusion models typically generate synchronized content within windows of a few seconds. We study whether a training-free, causal rolling-window extension with constant memory introduces cumulative synchronization drift. Our measurement protocol combines a calibrated offset estimator, per-system sensitivity audits, sample-accurate seam diagnostics, content-validity checks, and stratified prompts with a negative control. On a public joint transformer, naive streaming introduces a -120 ms conditioning-anchor offset at 91% of seams. A closed-form analysis predicts seam offsets within 2.5 ms across tested settings. Despite these local errors, measured offsets show no sustained increase in continuations up to five minutes. The results support window-level resynchronization through joint denoising, while identifying abrupt identity changes and boundary artifacts as persistent limitations. We evaluate three training-free stabilizers: modality-aligned conditioning, audio-lead conditioning, and cross-window fixed noise. Their combination reduces the mean absolute offset estimate from 0.285 to 0.253 s on a held-out 200-prompt benchmark, with consistent improvements in the reported quality metrics. Sensitivity audits support the comparison with naive streaming but limit comparisons with an offline consensus baseline. Preliminary responses from a pre-registered human study favor the combined method. On a second backbone, image-conditioned continuation degrades synchronization after the first window, and fixed noise provides a prompt-dependent improvement. We release the protocol, prompt suite, and code.

1 INTRODUCTION

Diffusion transformers jointly generate short video clips and synchronized audio (Ruan et al., 2022; Liu et al., 2025b; Low et al., 2025), but their native windows typically span only a few seconds. We study streaming extension with causal generation, constant memory, and constant cost per generated second. This definition does not imply real-time generation.

Length extension is well studied for video alone (Wang et al., 2023; Qiu et al., 2023; Kim et al., 2024; Yin et al., 2024; Huang et al., 2025). Joint generation also requires temporal alignment between modalities, even when each stream remains locally coherent. In our model, the conditioning rule assigns a non-integer 5.3125 audio steps to each video latent, requiring explicit treatment of the two temporal grids. Concurrent streaming systems address temporal inconsistency through training (Su et al., 2026; Ding et al., 2026). Whether local alignment errors accumulate during training-free extension remains an empirical question (Liu et al., 2026c).

We first characterize synchronization using a gain-audited offset estimator (Iashin et al., 2024), sample-accurate seam cross-correlation, cross-window identity measurements, content-validity checks, and a negative control. Applied to a naive extension of a public 2.1B joint transformer, the protocol distinguishes local seam errors from cumulative drift (Fig. 1). Of 912 seams, 91% have a measured conditioning-anchor offset of -120 ms. Conditioning arithmetic predicts seam lags within 2.5 ms across fifteen tested settings. Over two minutes, the sum of local errors has magnitude

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.

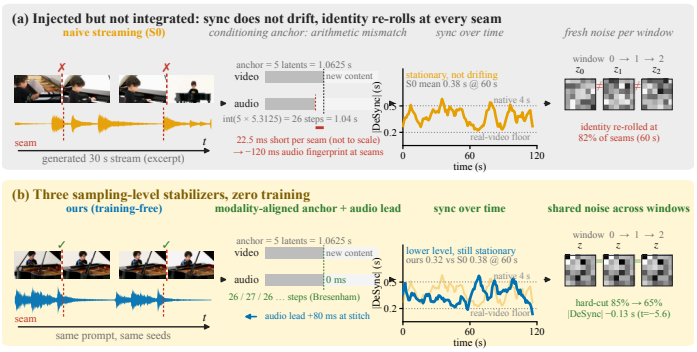


Figure 1: **Injected but not accumulated.** (a) Naive rolling-window streaming of a 4 s joint audio-video model uses each window's tail as the next window's anchor. Rounding leaves a 22.5 ms overlap mismatch, while anchor placement introduces a -120 ms seam offset. The measured trajectory $|\mathrm{DeSync}|(t)$ shows no sustained increase over two minutes, although boundary artifacts and abrupt identity changes remain. The frames illustrate two seams in a 30 s stream. (b) Three sampling-level stabilizers with the backbone frozen: modality-aligned anchor arithmetic (26/27 audio steps by Bresenham rounding), an 80 ms audio lead at stitching, and one shared noise for all windows. With matched prompts and seeds, the combined method reduces boundary artifacts and the offset estimate (0.32 vs. 0.38 s at 60 s), while improving identity continuity.

4.3 s, whereas the median realized end offset is 0.23 s. Continuations up to five minutes show no sustained increase. These observations support window-level re-synchronization, while revealing frequent boundary artifacts and abrupt identity changes.

We then evaluate three sampling interventions. Correcting the anchor arithmetic removes the original seam offset but increases the windowed offset estimate. The baseline's misalignment provides an implicit audio lead, motivating an explicit lead parameter. Reusing initial noise across windows further reduces synchronization variability and identity changes. The combined method improves over naive streaming on a held-out 200-prompt benchmark, with favorable quality metrics and supporting estimator audits. An offline consensus baseline (Park et al., 2026) preserves identity more effectively but has limited estimator sensitivity and requires full-duration denoising before output. Preliminary human responses favor our method. On a second backbone, image-conditioned continuation degrades synchronization after the first window; the benefit of fixed noise varies across prompts.

Our contributions are as follows.

- A measurement protocol for synchronization in long-form audio-video generation: gain-audited offset curves, seam-level, identity and content-validity instruments, and a prompt suite with a negative control (Sec. 3).
- A seam-resolved characterization of streaming synchronization: conditioning arithmetic predicts local errors, while measured offsets show no cumulative increase in the evaluated joint-conditioning setting (Secs. 4, 6.3).
- Three training-free stabilizers that improve synchronization over naive streaming at constant memory, evaluated through paired experiments and a preliminary human study (Secs. 5, 6).

2 Related Work

Joint audio-video generation. Joint audio-video generation commonly uses diffusion transformers that either place video and audio tokens in one sequence under shared self-attention (Ruan et al., 2022; Wang et al., 2024; Liu et al., 2025b; 2026b; Zhang et al., 2025a; Wang et al., 2025) or couple two towers through cross-attention (Low et al., 2025; HaCohen et al., 2026), with production systems at scale (Polyak et al., 2024; Google DeepMind, 2025) and short-window benchmarks (Shimada et al., 2024; Zhou et al., 2026b). These models focus on finite generation windows and use

positional information for alignment (Xu et al., 2025). Our study instead measures synchronization over extended generation time (cf. Polaczek et al., 2026; Gao et al., 2026).

Long-form and streaming video generation. Training-free extension schedules noise or windows at sampling time (Wang et al., 2023; Qiu et al., 2023; Kim et al., 2024; Lu et al., 2024; Chen et al., 2025; Feng et al., 2026; Yang et al., 2026a); trained approaches distill causal students (Henschel et al., 2024; Chen et al., 2024; Yin et al., 2024; Huang et al., 2025; Yang et al., 2025; Liu et al., 2025c), address compounding errors (Ning et al., 2023; Zhang et al., 2025b; Cui et al., 2025; Li et al., 2025; Chen et al., 2026a; Li et al., 2026; Zhang & Lim, 2024), and preserve identity with anchors or memory (Yi et al., 2025; Liu et al., 2026a; Meng et al., 2026). These methods primarily address visual consistency within a single stream.

Long and streaming audio-video generation. Recent work extends joint audio-video generation beyond native windows. Park et al. (2026) harmonize overlapping windows offline, requiring full-duration processing before output; Jung et al. (2026) scale inference within a window; multi-shot and unbounded systems are trained (Chen et al., 2026b; Huang et al., 2026a; Pan et al., 2026). Recent trained streaming systems address general audio-video generation (Su et al., 2026; Ding et al., 2026; Zhou et al., 2026a; Huang et al., 2026b; Bai et al., 2026; Duan et al., 2026; Zhang et al., 2026b;a; Team et al., 2026) or avatars (Tian et al., 2026; Han et al., 2026; Zhu et al., 2026; Zhang et al., 2026c; Xie et al., 2025). Among these, Su et al. (2026) identify temporal asymmetry between modalities and train an audio sink token to address it. Our analysis complements this work by measuring offset trajectories and seam-level errors without additional training. Cascades with a video-to-audio model (Shan et al., 2025) carry no audio state across windows, and Simon et al. (2026) study length generalization for video-to-audio, where the visual stream is given. Our method provides a training-free streaming extension of a joint generator.

Measuring synchronization. Offset estimators trained on real video (Chung & Zisserman, 2017; Iashin et al., 2024) score generated output (Cheng et al., 2024); event-agreement metrics compare onsets with motion peaks (Yariv et al., 2023). A concurrent audit compares their reliability (Sharma & Tapadiya, 2026), while showing why estimator behavior requires scrutiny. Incoherent output can receive favorable quality and synchronization scores (Huang et al., 2026c). Two concurrent benchmarks score trained streaming systems (Liu et al., 2026c) or minute-scale offline generation (Liu et al., 2026d) using interval-averaged offsets. Our protocol additionally combines calibration, per-system gain audits, seam diagnostics, and a negative control. The within-process correspondence drift studied by Hu et al. (2025) differs from the output-level temporal offset measured here.

3 Streaming extension and measurement protocol

3.1 Preliminaries

We study a public 2.1B joint audio-video diffusion transformer (Liu et al., 2026b) that generates a $W = 4.0625$ s window per sampling pass: $T_v = 17$ latent video frames (about 6.6k tokens at 240p) and $T_a = 104$ audio latent steps at 25 per second (about 416 tokens). The tokens share bidirectional self-attention, without an autoregressive state. We therefore extend generation at sampling time. The conditioning rule maps each video latent to 5.3125 audio steps. In the released clamping interface, mask parameters are unused, so clean conditioning latents are processed at the current noise level, degrading generation (cf. Chung et al., 2022; Zhang et al., 2023). We instead condition by replacement (Song et al., 2020; Lugmayr et al., 2022; Zou et al., 2026; Yang et al., 2026b), overwriting the conditioning region at every step with the reference latents noised to the current timestep. This training-free procedure does not apply reconstruction guidance (Ho et al., 2022). All streaming systems in this paper use it (App. C).

3.2 Naive streaming baseline

The naive streaming extension, denoted NS (Fig. 2), rolls the window forward with replacement conditioning: window 0 is generated as released, and window $k > 0$ conditions on the last $c = 5$ latent video frames and the last $c_a = \lfloor 5.3125\, c \rfloor = 26$ audio steps of window $k-1$, initializing the remainder with fresh noise. Each window adds 3.0 s. Those 26 audio steps start 0.12 s later than the

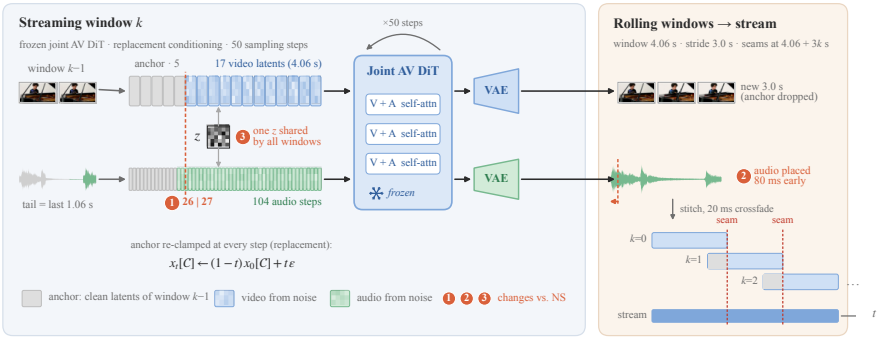


Figure 2: **Training-free streaming with a frozen joint AV DiT.** Left: the previous window's tail is the anchor of the video (5 of 17 latents) and audio (26 or 27 of 104 steps) token strips, re-clamped by replacement at every step; the rest starts from noise z, the frozen DiT denoises both jointly, and the VAEs decode 3.0 s of new content. OURS changes only ① the audio anchor interval (26/27 steps scheduled to match the video overlap on average), ② an 80 ms audio lead at the stitch, ③ one shared z for all windows. Right: 4.06 s windows, stride 3.0 s, seams at $4.06 + 3k$ s.

video span, so the model receives audio context running 120 ms ahead of the video context. We retain this tail-anchor rule as a baseline and analyze it in Sec. 4. Windows are decoded independently, audio is placed at absolute sample positions with a 20 ms crossfade, and a cross-correlation check verifies zero additional stitching lag relative to the reference sample placement.

3.3 MEASUREMENT PROTOCOL

We measure synchronization as a time series and audit estimator sensitivity. A pretrained offset estimator (Iashin et al., 2024) is applied in 4.8 s windows with a 1 s stride. We use the expectation of its offset distribution and track range saturation. Each curve is summarized by its end drift D_{end} and Theil-Sen slope (App. A).

Injecting known offsets into real footage shows a linear but compressed response (slope 0.64, $r = 0.999$), a content-dependent floor of 0.199 ± 0.128 s, and a response that returns toward zero beyond the estimator's range. We report saturation separately and audit injected-offset responses on generated outputs (Sec. 6.2). The event-agreement metric (Yariv et al., 2023) performs at the density-matched random baseline in our calibration and is excluded. All audio measurements use lossless waveform sidecars (App. A).

Complementary measurements capture errors that the offset estimator may miss. Cross-correlation between a window's tail and its regenerated copy measures seam lag at 0.0625 ms resolution; its running sum records local errors. An identity measure reports seams whose CLIP similarity falls more than two standard deviations below the within-window mean. Content-validity checks identify degenerate outputs, including near-silent audio that can produce near-zero offset estimates. Spectral-flux peaks quantify acoustic discontinuities at seams.

The prompt suite has two tiers: SyncProbe-48, stratified over rhythmic performance, vocalization, discrete impacts and ambient textures (the negative control), for diagnosis and selection, and a held-out 200-prompt subset of a public benchmark (Liu et al., 2025b) for the prespecified confirmatory contrast. All arms share seeds, so every comparison is paired.

4 WHY DOES SYNCHRONIZATION NOT DRIFT?

4.1 LOCAL ERRORS WITHOUT CUMULATIVE DRIFT

Naive streaming shows no sustained increase in measured offset (Fig. 3). On SyncProbe-48 at 60 s, $D_{\text{end}} = +0.011 \pm 0.216$ s and the slope is not significantly different from zero. Cross-correlation gives a median end-to-end offset of 0.09 s. For twelve prompts at 120 s, offsets do not increase in the second minute (median end offset 0.23 s). At 300 s, all eight trajectories across four prompts

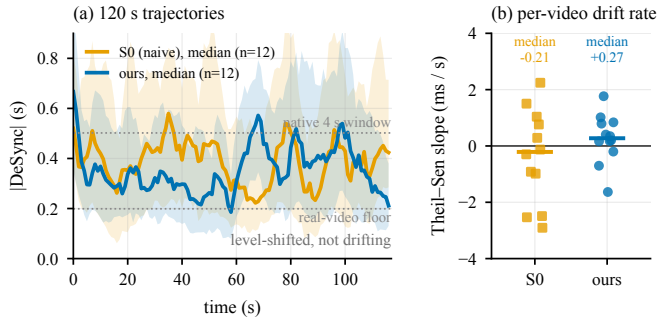


Figure 3: **No sustained increase in measured synchronization offsets** (SyncProbe-48 subset, 120 s, n=12 per arm). (a) Median and inter-quartile band of $|\mathrm{DeSync}|(t)$ for NS and OURS: no upward trend between the real-footage floor and the native single-window level. (b) Per-video Theil-Sen slopes are centred on zero. Signed readings and five-minute continuations are in App. E.

and two methods have slopes within $\pm 10^{-3}$ s per second. These observations provide no evidence of sustained drift over the evaluated durations.

Despite the limited temporal trend, NS has a mean absolute offset estimate of 0.400 ± 0.215 s, above the reference level of 0.199 ± 0.128 s. The ambient control has a lower estimate (0.138 s). The small signed mean (-0.030 s) is consistent with temporal variability rather than a persistent offset in one direction.

Interpretation requires accounting for estimator sensitivity. Injected offsets yield gains of 0.37 to 0.48 on generated outputs, compared with 0.64 on real footage. Beyond 60 s, our assessment also relies on estimator-independent cross-correlation and the cumulative seam-error measure described below.

4.2 PREDICTING SEAM OFFSETS FROM CONDITIONING ARITHMETIC

Under NS, 91% of 912 seams have a measured cross-correlation lag of -120.0 ms between the original and regenerated conditioning spans, consistent with the anchor's delayed start. Low-energy audio can make this measurement unreliable. For the tail-anchor rule, conditioning length c and an explicit audio lead of L latent steps give

$$\mathrm{lag}(c, L) \;=\; -\left[\, \frac{104 - \lfloor 5.3125\, c \rfloor}{25} \;-\; \frac{68 - 4c}{16} \,\right] \;-\; 0.04\, L \quad \text{seconds}, \tag{1}$$

The corresponding interval arithmetic predicts measured lags within 2.5 ms across fifteen settings covering $c \in \{2, 5, 10\}$, three anchor definitions, 33/65/97-frame windows, and leads of both signs (Fig. 6, App. C). This agreement identifies anchor placement as a deterministic source of seam error and motivates the correction in Sec. 5.

4.3 EVIDENCE FOR WINDOW-LEVEL RE-SYNCHRONIZATION

The discrepancy between cumulative seam error and realized offset suggests that local errors are not propagated as cumulative temporal shifts. Three observations support this interpretation. First, B3 generates audio without an audio anchor and has windowed statistics similar to NS, despite irregular seam lags. Joint denoising can therefore maintain within-window alignment even without audio conditioning (cf. Pedersen et al., 2025). Second, conditioned segments have lower offset estimates than unconditioned first-window crops (0.402 vs. 0.515 s), motivating the conditioning interventions in Sec. 5. Third, attention state is not retained across windows. Cross-modal attention is asymmetric: audio queries assign up to 42% of their attention mass to video tokens, whereas video queries assign under 2% to audio. This pattern remains stable across windows. These observations support re-synchronization through joint denoising but do not establish attention as its causal mechanism. We do not intervene on positional re-indexing or attention anchors. A second backbone with image-only continuation exhibits a different offset profile (Sec. 6.3).

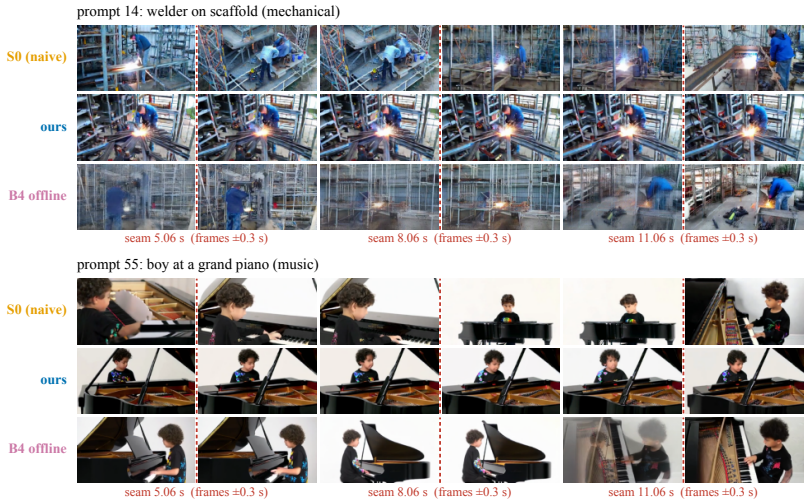


Figure 4: **Abrupt identity changes at window boundaries.** Frames ± 0.3 s around three consecutive seams in two 30 s examples (matched prompts and seeds across methods). NS changes camera, subject or layout at the illustrated seams; B4 preserves identity with visible blurring; OURS better preserves the subject, clothing, and layout while pose changes. Additional examples appear in App. H.

Boundary artifacts and identity discontinuities remain substantial (Fig. 4): seams exhibit excess spectral flux ($z = 5.4$), and 79% meet the identity hard-cut criterion. Similarity to the first window shows little temporal trend. This combination indicates abrupt changes at seams without progressive identity degradation in these measurements (cf. Lu et al., 2026).

5 TRAINING-FREE STABILIZERS

We evaluate three interventions motivated by the diagnosis (Fig. 2). Correcting the conditioning arithmetic removes the systematic anchor mismatch but increases the windowed offset estimate, motivating an explicit audio-lead parameter. Shared initial noise provides a further consistency constraint. All three preserve model weights and constant per-window computational cost.

Notation. Window k holds video latents v_k with $T_v = 17$ frames and audio latents a_k with $T_a = 104$ steps of $\tau_a = 0.04$ s. The 4.0625 s content interval corresponds to 101.6 audio steps; the remainder is padding. The default $c = 5$ video anchor spans 1.0625 s, corresponding to $5.3125\,c$ audio steps. The replacement sampler overwrites the conditioning region at every step t, $x_t[\mathcal{C}] \leftarrow (1-t)\,x_0[\mathcal{C}] + t\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ in the rectified-flow parametrization of the backbone, and restores $x_0[\mathcal{C}]$ after the last step; the seam is the instant at which the anchor ends and new content begins.

① **Modality-aligned conditioning arithmetic.** Equation 1 attributes the systematic seam offset to different audio and video anchor intervals. NS takes the trailing $\lfloor 5.3125\,c \rfloor = 26$ steps of the padded buffer, which start 3 steps (0.12 s) after the video span. The repair anchors both modalities to the same interval in seconds, taking audio steps $[\,s_k,\ s_k + c_a^{(k)})$ from the step s_k at which the video span starts, with a Bresenham schedule (Bresenham, 1965; Demaine et al., 2007),

$$c_a^{(k)} \;=\; \big\lfloor 5.3125\,c\,(k{+}1) \big\rfloor - \big\lfloor 5.3125\,c\,k \big\rfloor \;\in\; \{26, 27\}, \tag{2}$$

which bounds cumulative rounding error by one audio step; floor rounding alone leaves 0.0225 s per window. The measured -120 ms offset drops to zero at 809 of 912 seams and the ledger falls from -4.34 s over two minutes to $+2 \pm 31$ ms. However, the windowed offset estimate increases by 0.076 s relative to NS ($t = +6.52$). Ablations associate this increase with anchor position ($t = +5.80$), rather than the rounding schedule.

Algorithm 1: OURS: training-free streaming with three stabilizers

Input: prompt p; windows K; video anchor $c = 5$; lead $L = 2$; noise z_0
$(v_0, a_0) \leftarrow \text{sample}(p;\ z_0)$;
for $k = 1$ **to** $K-1$ **do**
  $c_a \leftarrow \lfloor 5.3125\, c\,(k{+}1) \rfloor - \lfloor 5.3125\, c\, k \rfloor$; // ① 26/27 alternation
  $v^{\text{ref}} \leftarrow v_{k-1}[-c{:}]$; $a^{\text{ref}} \leftarrow a_{k-1}[\, s_k{-}L : s_k{-}L{+}c_a\,]$; // ② audio lead
  $(v_k, a_k) \leftarrow \text{sample}\big(p \mid \text{replace}(v^{\text{ref}}, a^{\text{ref}});\ z_0\big)$; // ③ shared noise
decode windows; drop anchors; place audio of window k at $3k - L\tau_a$ s with a 20 ms crossfade; assert zero stitching lag;

② Audio-lead conditioning. The misaligned NS anchor supplies audio context 120 ms ahead of the video context, which improves within-window alignment in our measurements. We represent this lead as an explicit parameter L in latent steps on top of the corrected arithmetic: the audio anchor becomes $a_{k-1}[\, s_k - L,\ s_k - L + c_a^{(k)})$, and at stitching the audio of window k is placed $L\,\tau_a$ seconds early so the stitched stream carries no net shift. Audio look-ahead is standard in audio-driven animation (Chen & Liu, 2025; Chu et al., 2026; Seo et al., 2025; Xiao et al., 2025); here it only shifts which audio latents are clamped relative to the video anchor without additional model evaluations. The response is directional (Fig. 6a): negative lead has no statistically significant effect (-80 ms: $t = -1.5$), while positive lead reduces the offset estimate over the tested range ($+80$ ms: -0.050 s, $t = -5.18$), and the NS point $+120$ ms is available under the tail-anchor rule; the corrected anchor permits at most $L = 2$. We use $L = 2$, corresponding to 80 ms. The signed mean remains similar, consistent with reduced temporal variability. This setting introduces a fixed -80 ms seam displacement. Audio-leading offsets have asymmetric perceptual tolerances (International Telecommunication Union, 1998); reported seam artifacts in our preliminary human study are below 2% (Sec. 6.4), without directly measuring detection of the offset.

③ Cross-window fixed noise. Each window starts from the same initial noise, $z_k = z_0$, instead of a fresh draw. Shared noise is a known consistency prior in video-only extension (Ge et al., 2023; Qiu et al., 2023; Tan et al., 2025; Liu et al., 2025d). Here it also reduces audio-video offset estimates (-0.051 s, $t = -2.51$ on NS; -0.128 s on the corrected anchor without lead, $t = -5.62$) and reduces identity hard cuts from 79% to 65% at no conditioning change. A known limitation is texture sticking (Chang et al., 2025). In our short-window setting, degraded outputs receive misleadingly favorable consistency scores (Sec. 6.2). At the native window length, measured image quality remains comparable to the fresh-noise setting; the combined method also improves FVD (577 vs. 634).

Combined system. Algorithm 1 lists the full procedure; other settings match NS. The benefits of lead and shared noise are partially overlapping (Table 3). The combined method has a fixed -80 ms seam offset without observed cumulative drift.

6 EXPERIMENTS

Setup. The primary comparison uses the same public 2.1B backbone, replacement sampler, and per-prompt seeds. SyncProbe-48 at 30 to 300 s is the exploratory tier for diagnosis and selection; a 200-prompt stratified benchmark subset at 30 s is the confirmatory tier with one prespecified contrast, and its first 24 prompts at 60 s a held-out duration tier (App. B). Baselines: B1, the single-window model; B2, direct length extrapolation in one pass; B3, video-conditioned streaming with independent audio generation per segment; B4, offline parallel harmonization of overlapping windows in the style of Park et al. (2026) on the same backbone. Trained streaming systems are outside this controlled comparison because they require different trained models. Primary-backbone comparisons use matched prompts and seeds (tiers in App. D).

6.1 MAIN COMPARISON

Table 1 reports the confirmatory comparison on $n = 200$ paired prompts. The OURS$-$NS difference is -0.032 s ($t = -2.93$), smaller than the exploratory estimate of -0.058 s. Category-level

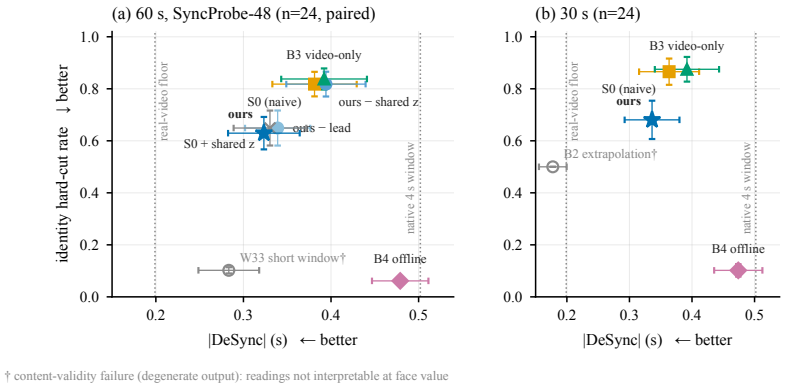


Figure 5: **Sync-identity plane on SyncProbe-48** (mean±SEM, n=24 seed-paired prompts; lower-left is better). OURS improves both measures relative to NS. B4 has fewer identity discontinuities but higher raw offset estimates and requires offline processing. Its low estimator gain limits synchronization comparisons. Grey open markers indicate content-validity failures: B2 produces blurred frames despite a low offset estimate, and W33 receives favorable scores despite poor image quality.

Table 1: Held-out 200-prompt benchmark at 30 s: expectation readout in seconds (real-footage floor 0.199), identity hard-cut rate, and official quality metrics on native-window segments. Memory refers to GPU memory. †: invalidated by the content-validity gate; *: estimator gain collapsed (Sec. 6.2).

Method	Causal	Const. mem.	|DeSync| ↓	Hard-cut ↓	KVD ↓	FAD ↓	MQ ↑	FVD ↓
B2 length extrapolation	✓	–	0.187†	–	1.25	9.89	(1.58)†	1119
B3 video-only streaming	✓	✓	0.268	83%	1.95	9.08	0.268	634
NS naive streaming	✓	✓	0.285	84%	1.96	8.89	0.264	634
B4 offline consensus	–	✓	0.437*	**6%**	1.87	8.18	0.253	629
OURS (ours)	✓	✓	**0.253**	72%	**1.21**	**7.91**	**0.369**	**577**

differences favor OURS for living scenarios (-0.067 s, $t = -3.46$) and musical sounds (-0.051 s, $t = -3.76$). Ambient content remains near the reference level, while industrial scenarios show an increase ($+0.072$ s, $t = +2.05$; App. E). All reported quality metrics improve relative to NS.

The baselines exhibit different limitations (Fig. 5). B2 has severe visual degradation despite an offset estimate below the real-footage reference level, illustrating the need for validity checks. B3 has no detectable windowed synchronization difference from NS (minimum detectable effect, MDE, 0.09 s), but has irregular seam offsets. B4 reduces hard cuts to 6% while increasing the raw offset estimate. Its low estimator gain limits direct synchronization comparisons, and it emits output only after full-duration denoising.

Across evaluated durations (Table 2, Fig. 10), OURS has lower mean offset estimates than NS, including 0.480 vs. 0.553 s at 300 s. On the held-out 60 s tier (Table 5), the difference is not statistically significant (0.237 vs. 0.251 s, MDE 0.07 s). Hard cuts decrease from 92.5% to 80.5% ($t = -2.14$), with favorable quality metrics. Measured peak streaming memory is 19.5 GB across tested durations (Table 10).

6.2 INSTRUMENT AUDIT AND ABLATIONS

Injected-offset audits give gains of 0.417 for OURS and 0.474 for NS, with no statistically significant difference (paired $t = -0.84$). The gain-corrected difference remains favorable (≈ -0.072 s). This supports the comparison without establishing equivalent sensitivity. B4 has a much lower gain of 0.086, restricting interpretation of its raw offset estimates (App. A).

Table 3 isolates lead and shared-noise effects on the corrected anchor. Both reduce offset estimates, with partially overlapping benefits; shared noise accounts for most of the identity improvement.

Table 2: SyncProbe-48 across durations (24 paired prompts; 12 at 120 s; mean±SEM). Paired OURS−NS: -0.027 (t=-0.91), -0.058 (t=-2.39), -0.055 (t=-2.03).

	30 s		60 s		120 s	
Method	|DeSync|	hard-cut	|DeSync|	hard-cut	|DeSync|	hard-cut
NS naive streaming	0.363±.048	87%	0.381±.048	82%	0.502±.039	78%
B4 offline consensus	0.474±.039	**10%**	0.479±.032	**6%**	0.580±.037	**5%**
OURS (ours)	**0.336**±.044	68%	**0.323**±.041	63%	**0.448**±.051	56%

Table 3: Ablation at 60 s on the corrected anchor (①), n=24 paired: 2×2 over lead (②) and shared noise (③), paired effects, and the anchor-length sweep.

Configuration	|DeSync| ↓	hard-cut ↓	Contrast / sweep	Δ|DeSync| (t)	hard-cut
① only	0.467±.045	85%	lead | fresh z	-0.073 (-4.97)	
①+② lead	0.394±.045	82%	shared z | lead 0	-0.128 (-5.62)	-20 pt
①+③ shared z	0.339±.037	65%	shared z on NS	-0.051 (-2.51)	-17 pt
OURS (all three)	**0.323**±.041	**63%**	anchor c=2/5/10	0.458/0.476/0.534	90/81/64%

Longer anchors improve identity at the expense of synchronization. A 33-frame window gives favorable metric values but fails the quality check (CLIP-IQA 0.29; App. E).

6.3 SECOND BACKBONE

We also evaluate Ovi (Low et al., 2025), an 11B twin-tower generator that conditions video continuation on the previous final frame without an audio anchor (App. F, Table 9). Later windows have offset estimates 0.165 s above the first ($t = +2.76$), followed by little further increase. Shared noise reduces the mean stitched estimate from 0.446 to 0.347 s ($t = -2.67$, $n = 16$), primarily on the eight pilot prompts. Hard cuts increase from 23% to 56%. Prompt heterogeneity and unequal estimator gains limit the generality of this result.

6.4 HUMAN EVALUATION

The pre-registered two-alternative forced-choice study (App. I) compares OURS with NS and B4 using 12 s excerpts from 30 held-out prompts, each spanning four streaming seams. Seven of the planned fifteen raters had completed the study at submission; two meet the strict exclusion rule and six meet the sensitivity-analysis rule. Descriptive responses favor OURS on synchronization and preference, with seam artifacts reported in under 2% of its trials (Table 11). These interim results do not constitute the completed pre-registered test.

7 CONCLUSION

Our measurements distinguish predictable seam offsets from cumulative synchronization drift in training-free joint audio-video streaming. Three sampling interventions reduce offset estimates and boundary discontinuities while preserving the evaluated quality metrics. **Limitations:** the seam analysis primarily concerns one backbone at 240p; long-duration evidence is limited to small prompt subsets and complementary seam measurements. Identity continuity remains incomplete, second-backbone effects vary across prompts, and human results are preliminary.

AI USE STATEMENT

Generative AI tools were used to improve the grammar, wording, and readability of the manuscript. The authors reviewed and verified all AI-assisted revisions and take full responsibility for the final content, including the methods, results, and conclusions.

ETHICS STATEMENT

This work develops inference-time procedures for an existing open joint audio-video generator without training a new model. Generation prompts are scene descriptions or entries from a public benchmark; no personal data were used as generation inputs. The perceptual study involved volunteer raters who were informed of the task and instructed to use headphones. Responses are associated with numeric rater identifiers, and no personally identifying information was collected. Real-footage synchronization references use Creative Commons material. Extending generation duration retains the underlying model's risks, including the production of misleading audiovisual content. Use and redistribution remain subject to the applicable model and data licenses.

REPRODUCIBILITY STATEMENT

The streaming protocol, sampling procedures, and implementation settings are described in Secs. 3, 5, and App. C. Prompt selection and seed assignments are specified in App. B; estimator calibration and gain audits are detailed in App. A. Statistical reporting and the pre-registered human-study protocol are provided in Apps. D and I. The supplementary material includes the sampler, evaluation scripts, prompt and seed files, numerical summaries, human-study materials, and demonstration videos.

REFERENCES

Lichen Bai, Tian-Hui Zhang, Shitong Shao, D. Tan, Qi-Yu Zhong, Zheng-Peng Xie, Haopeng Li, Qinghao Huang, D. Shen, Tengjiao Ji, Wei Wang, Peicheng Wu, Yuxuan Zhao, Xiang-Yu Zhu, Welly Luo, Shu-Ruei Yang, and Ze-Ke Xie. Mainecoon: Pursuing a real-time audio-visual social world model. *arXiv preprint arXiv:2606.17800*, 2026. doi: 10.48550/arXiv.2606.17800.

J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1): 25–30, 1965. doi: 10.1147/sj.41.0025.

Pascal Chang, Jing-Wei Tang, Markus Gross, and Vinicius C. Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2504.03072.

Bohong Chen and Hai-Yang Liu. Dystream: Streaming dyadic talking heads generation via flow matching-based autoregressive model. *arXiv preprint arXiv:2512.24408*, 2025. doi: 10.48550/arXiv.2512.24408.

Boyuan Chen, Diego Marti Monso, Yilun Du, M. Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2407.01392.

Jingyuan Chen, Fuchen Long, Jie An, Zhaofan Qiu, Ting Yao, Jie-Bo Luo, and Tao Mei. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. In *AAAI Conference on Artificial Intelligence*, 2025. doi: 10.48550/arXiv.2501.09019.

Taiye Chen, Qi Zhang, and Yi-Sen Wang. Mitigating compounding error via video representation regularization. *arXiv preprint arXiv:2607.27036*, 2026a. doi: 10.48550/arXiv.2607.27036.

Yuheng Chen, Teng Hu, Yu-Ji Wang, Qingdong He, Zhu-Cun Xue, Qianyu Zhou, Jason Li, Lizhuang Ma, Jiangning Zhang, and Da-Cheng Tao. Cinedance: Towards next-generation multi-shot long-form cinematic audio-video generation. *arXiv preprint arXiv:2606.09639*, 2026b. doi: 10.48550/arXiv.2606.09639.

H. Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52734.2025.02691.

Xuan-Geng Chu, Yunkai Han, Wei Mao, and Shih-En Wei. Personalizing causal audio-driven facial motion via dynamic multi-modal retrieval. *arXiv preprint arXiv:2604.23692*, 2026. doi: 10.48550/arXiv.2604.23692.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and J. C. Ye. Improving diffusion models for inverse problems using manifold constraints. In *Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2206.00941.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Lecture Notes in Computer Science*, pp. 251–263, 2017. doi: 10.1007/978-3-319-54427-4_19.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuan-Hao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. doi: 10.48550/arXiv.2510.02283.

Erik D. Demaine, Francisco Gómez-Martín, H. Meijer, D. Rappaport, Perouz Taslakian, G. Toussaint, T. Winograd, and David R. Wood. The distance geometry of music. In *Computational geometry*, 2007. doi: 10.1016/j.comgeo.2008.04.005.

Yan-Bo Ding, Zhi-Zhi Guo, Quan-Yue Song, Yi-Shan He, Zhi-Xiang He, Yong-Xiang Li, and Yali Wang. Ripple: Real-time streaming audio-video generation with cross-modal recurrent memory. *arXiv preprint arXiv:2607.26818*, 2026. doi: 10.48550/arXiv.2607.26818.

Nan Duan, Hao-Yang Huang, Weiyang Jin, Hao-Ran Li, Yaowei Li, Yuming Li, Yijun Liu, Xin Lu, Xiaoxiao Ma, Yanwen Ma, Yaofeng Su, Yilang Sun, Hao-Yu Wang, Ze-Yue Xue, Song-Chun Zhang, and Junhao Zhuang. Long-horizon audio-visual generation for persistent stories and interactive worlds. *arXiv preprint arXiv:2608.23383*, 2026.

Xiao-Kun Feng, Jia-Shu Zhu, Mei-Qi Wu, Chu-Bin Chen, Fang-Yuan Mao, Hai-Yang Guo, Jia-Hong Wu, Xiangxiang Chu, and Kai-Qi Huang. Enhancing train-free infinite-frame generation for consistent long videos. *arXiv preprint arXiv:2605.18233*, 2026. doi: 10.48550/arXiv.2605.18233.

Sheng-Chuan Gao, Teng Hu, Bohao Feng, Luchen Li, Wenqiang Wang, Hong-Qian Deng, and Ran Yi. Efficient audio-visual generation via synchrony-aware cross-modal sparse attention. *arXiv preprint arXiv:2608.15522*, 2026.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Y. Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *IEEE International Conference on Computer Vision*, 2023. doi: 10.1109/ICCV51070.2023.02096.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.00695.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.01457.

Google DeepMind. Veo: A text-to-video generation system (veo 3 tech report). `https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf`, 2025. Technical report.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, A. Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, M. Finkelson, Michael Kupchick, Nir Zabari, N. Guetta, N. Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, V. Kulikov, Yaron Inger, Y. Shiftan, Zeev Melumian, and Zeev Farbman. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. doi: 10.48550/arXiv.2601.03233.

Meng-Lin Han, Yang Ding, Yulei Lu, Hao-Ran Yu, Xin Ma, Junyi Chen, Zhang-Kai Ni, Lin Ma, and Yaohui Wang. Vorch-streamer: Extending human audio-visual generation to real-time long-form streaming. *arXiv preprint arXiv:2608.05663*, 2026.

Roberto Henschel, L. Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52734.2025.00245.

Jonathan Ho, Tim Salimans, A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2204.03458.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zi-Han Su, Zheng-Guang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. *arXiv preprint arXiv:2511.21579*, 2025. doi: 10.48550/arXiv.2511.21579.

Jiehui Huang, Yuechen Zhang, Bin Xia, Jia-Hao Wang, Xu He, Zhen-Chao Tang, Meng Chu, Xin Tao, Peng-Fei Wan, and Jiaya Jia. Unityshots: Memory-driven multi-shot audio-video generation with boundary-aware gating. *arXiv preprint arXiv:2606.21661*, 2026a. doi: 10.48550/arXiv.2606.21661.

Lianghua Huang, Zhi-Fan Wu, Wei Wang, Yupeng Shi, Mengyang Feng, Junjie He, Chenwei Xie, Yu Liu, Jingren Zhou, Ang Wang, Bang Zhang, Baole Ai, Chen Liang, Cheng Yu, Chong-Yang Zhong, Ji Qi, Kai Zhu, Pandeng Li, Peng Zhang, Wenyuan Zhang, Xinhua Cheng, Yi-tong Huang, Yun Zheng, Yu-Zheng Wang, and Zou-Bin Bi. Wan-streamer v0.1: End-to-end real-time interactive foundation models. *arXiv preprint arXiv:2606.25041*, 2026b. doi: 10.48550/arXiv.2606.25041.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2506.08009.

Yin-Ming Huang, Shu-Yuan Tu, Xi Yan, Zi-Han Yang, Jianhua Han, Hang Xu, Kaihang Pan, Yu-Gang Jiang, and Zuxuan Wu. Va-judger: Reward modeling from human preference feedback for joint video-audio generation. *arXiv preprint arXiv:2608.18607*, 2026c.

Yuanhui Huang, Wenzhao Zheng, Yuan Gao, Xin Tao, Peng-Fei Wan, Di Zhang, Jie Zhou, and Jiwen Lu. Owl-1: Omni world model for consistent long video generation. *arXiv preprint arXiv:2412.09600*, 2024. doi: 10.48550/arXiv.2412.09600.

Vladimir E. Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

International Telecommunication Union. Relative timing of sound and vision for broadcasting. Technical Report Rec. BT.1359-1, ITU-R, 1998.

Jaemin Jung, Kyeongha Rho, Inkyu Shin, and Joon Son Chung. Inference-time scaling for joint audio-video generation. In *Trans. Mach. Learn. Res.*, 2026. doi: 10.48550/arXiv.2606.03183.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. In *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2405.11473.

Haodong Li, Shaoteng Liu, Zhe Lin, and M. Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026. doi: 10.48550/arXiv.2602.07775.

Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212*, 2025. doi: 10.48550/arXiv.2510.09212.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Neural Information Processing Systems*, 2025a. doi: 10.48550/arXiv.2501.13918.

Jinzhuo Liu, Jiangning Zhang, Wencan Jiang, Ya-Biao Wang, Dingkang Liang, Zhu-Cun Xue, Ran Yi, and Yong Liu. Advancing narrative long video generation via training-free identity-aware memory. *arXiv preprint arXiv:2605.18733*, 2026a. doi: 10.48550/arXiv.2605.18733.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Rongxin Jiang, Jiebo Luo, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025b. doi: 10.48550/arXiv.2503.23377.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. *arXiv preprint arXiv:2602.19163*, 2026b. doi: 10.48550/arXiv.2602.19163.

Kaiqi Liu, Hao-Xuan Zeng, Jingqi Liu, Jia-Cong Fang, Ziqi Cai, Yun-Yao Mao, He Liu, Yu-Hao Sheng, Shuchen Weng, and Boxin Shi. Streamav-bench: A comprehensive benchmark for streaming audio-video generation. *arXiv preprint arXiv:2608.26336*, 2026c.

Kunhao Liu, Wen-Bo Hu, Jiale Xu, Shan Ying, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025c. doi: 10.48550/arXiv.2509.25161.

Tengfei Liu, Yang Shi, Xuanyu Zhu, Jiafu Tang, Liu Yang, Qixun Wang, Zhuo-Ran Zhang, Yuqi Tang, Fengxiang Wang, Yuhao Dong, Xinlong Chen, Bozhou Li, Bohan Zeng, Yuetong Ding, Xiaohan Zhang, Jia-Lu Chen, Hao-Tian Wang, Yuanxing Zhang, Pengfei Wan, and Leye Wang. Longav-compass: Towards unified evaluation of minute-scale audio-visual generation across t2av, i2av, and v2av. *arXiv preprint arXiv:2605.26244*, 2026d. doi: 10.48550/arXiv.2605.26244.

Zhenjie Liu, Jian Lu, Ren-Jie Lu, Cong Liang, and Shangfei Wang. Consisttalk: Intensity controllable temporally consistent talking head generation with diffusion noise search. In *AAAI Conference on Artificial Intelligence*, 2025d. doi: 10.48550/arXiv.2511.06833.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025. doi: 10.48550/arXiv.2510.01284.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. In *Neural Information Processing Systems*, 2024. doi: 10.48550/arXiv.2407.19918.

Yuxing Lu, Jiayang Sun, Guibo Zhu, and Mingxiang Cao. Asymtalker: Identity-consistent long-term talking head generation via asymmetric distillation. *arXiv preprint arXiv:2605.02948*, 2026. doi: 10.48550/arXiv.2605.02948.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, R. Timofte, and L. Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01117.

Yu Meng, Xiangyang Luo, Le-Tian Li, Wen Jiang, Chen Gao, Xin-Lei Chen, Yong Li, and Xiao-Ping Zhang. Tethercache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026. doi: 10.48550/arXiv.2606.13035.

Mang Ning, Mingxiao Li, Jianlin Su, A. A. Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2308.15321.

Shao-Hua Pan, Jun-Bao Chen, Shengyi He, Jing-Feng Xue, Wen Tao, Hao-Cheng Feng, Si-Ming Fan, Dong-Wei Pan, Yi Yang, Wei He, and Hang Zhou. Encore: Infinite audio-video generation with adaptive signal routing. *arXiv preprint arXiv:2609.04249*, 2026.

Jangho Park, G. Park, Gihyun Kwon, and Jong Chul Ye. Flowlong: Inference-time long video generation via manifold-constrained tweedie matching. *arXiv preprint arXiv:2605.20910*, 2026. doi: 10.48550/arXiv.2605.20910.

Christian Pedersen, L. Zanna, and Joan Bruna. Thermalizer: Stable autoregressive neural emulation of spatiotemporal chaos. In *International Conference on Machine Learning*, 2025. doi: 10.48550/arXiv.2503.18731.

Sagi Polaczek, Noa Kraicer, G. Metzer, Zhuo Ning, Ali Mahdavi-Amiri, Daniel Cohen-Or, and Raja Giryes. The attention triangle in audio-video models. *arXiv preprint arXiv:2609.03586*, 2026.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Is han Misra, Ji Hou, Jialiang Wang, K. Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam S. Tsai, S. Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali K. Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, B. Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, E. Schonfeld, Elliot Blanchard, Felix Juefei-Xu, F. Nord, Jeff Liang, John Hoffman, J. Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Li-Cheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, S. Sampson, Shikai Li, Simone Parmeggiani, S. Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Hao-Nan Qiu, Menghan Xia, Yong Zhang, Yin-Yin He, Xintao Wang, Shan Ying, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2310.15169.

Lu-Dan Ruan, Y. Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and B. Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52729.2023.00985.

Pranab Kumar Sen. Estimates of the regression coefficient based on kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389, 1968. doi: 10.1080/01621459.1968.10480934.

Junyoung Seo, Rodrigo Mira, A. Haliassos, Stella Bounareli, Honglie Chen, L. Tran, Seungryong Kim, Zoe Landgraf, and Jie-Xuan Shen. Lookahead anchoring: Preserving character identity in audio-driven human animation. *arXiv preprint arXiv:2510.23581*, 2025. doi: 10.48550/arXiv.2510.23581.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yue-Hai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025. doi: 10.48550/arXiv.2508.16930.

Jai Kumar Sharma and Peeyush Tapadiya. What do audio-visual synchronization metrics actually measure? *arXiv preprint arXiv:2608.25157*, 2026.

Kazuki Shimada, Christian Simon, Takashi Shibuya, Shusuke Takahashi, and Yuki Mitsufuji. Savgbench: Benchmarking spatially aligned audio-video generation. *arXiv preprint arXiv:2412.13462*, 2024. doi: 10.48550/arXiv.2412.13462.

Christian Simon, Masato Ishii, Wei-Yao Wang, Koichi Saito, Akio Hayakawa, D. Shim, Zhi-Wei Zhong, Shuyang Cui, Shusuke Takahashi, Takashi Shibuya, and Yuki Mitsufuji. Echoes over time: Unlocking length generalization in video-to-audio generation models. *arXiv preprint arXiv:2602.20981*, 2026. doi: 10.48550/arXiv.2602.20981.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, S. Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Yaofeng Su, Yuming Li, Ze-Yue Xue, Jie Huang, Siming Fu, Hao-Ran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 2026. doi: 10.48550/arXiv.2603.11647.

Jiangtong Tan, Hu Yu, Jie Huang, Jie Xiao, and Feng Zhao. Freepca: Integrating consistency information across long-short frames in training-free long video generation via principal component analysis. In *Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.02605.

Vorch Team, Xiaoyu Chen, Yang Ding, Cong Han, Meng-Lin Han, Yuxin Hong, Jiebo Hou, Ze-Qun Jie, Xiang Li, Jing Liu, Qi Liu, Yulei Lu, Si-Yuan Luo, Lin Ma, Xin Ma, Yinlong Qian, Peng Shi, Fang Wan, Siqi Wang, Yao-Hui Wang, Yao-Wei Wang, Yi-Di Wu, Si-Qian Yang, Mingyu Yin, Hao-Ran Yu, Gang Yue, Lisai Zhang, and Yuting Zhang. Vorch-omni: Multi-task orchestration of sight and sound. *arXiv preprint arXiv:2608.05803*, 2026.

Lin-Rui Tian, Qi Wang, and Bang Zhang. Streamchar: Long-horizon streaming character audio-video generation with decoupled orchestration. *arXiv preprint arXiv:2605.25659*, 2026. doi: 10.48550/arXiv.2605.25659.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Duo-Min Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xin-Yao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Da-Xin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025. doi: 10.48550/arXiv.2509.06155.

Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. doi: 10.48550/arXiv.2305.18264.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, 2022. doi: 10.48550/arXiv.2207.12396.

Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation. *arXiv preprint arXiv:2406.07686*, 2024. doi: 10.48550/arXiv.2406.07686.

Yusong Wu, K. Chen, Tian-Yu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and S. Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2022. doi: 10.1109/ICASSP49357.2023.10095969.

Steven Xiao, Xindi Zhang, Dechao Meng, Qi Wang, Peng Zhang, and Bang Zhang. Knot forcing: Taming autoregressive video diffusion models for real-time infinite interactive portrait animation. *arXiv preprint arXiv:2512.21734*, 2025. doi: 10.48550/arXiv.2512.21734.

You Xie, Tianpei Gu, Zenan Li, Chenxu Zhang, Guoxian Song, Xiaochen Zhao, Chao Liang, Jianwen Jiang, Hongyi Xu, and Linjie Luo. X-streamer: Unified human world modeling with audio-visual interaction. *arXiv preprint arXiv:2509.21574*, 2025. doi: 10.48550/arXiv.2509.21574.

Jin Xu, Zhifang Guo, Jinzheng He, Hang-Rui Hu, Ting He, Shuai Bai, Ke qin Chen, Jia-Lin Wang, Yang Fan, K. Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. doi: 10.48550/arXiv.2503.20215.

Haoning Yang, Xinyuan Chen, Yaohui Wang, and Guo Lu. Diff-vf: Training-free high-quality long video generation via diffusion model. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2026a. doi: 10.1145/3838802.

Lingxiao Yang, Liu Liu, Mo-Ran Li, Han Feng, Wenjian Cao, Jiangning Zhang, and Ye Shi. In-context forcing: Uncovering context effects in autoregressive video diffusion. *arXiv preprint arXiv:2608.05237*, 2026b.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, En-Ze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025. doi: 10.48550/arXiv.2509.22622.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI Conference on Artificial Intelligence*, 2023. doi: 10.48550/arXiv.2309.16429.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025. doi: 10.48550/arXiv.2512.05081.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52734.2025.02138.

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, T. Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. In *International Conference on Machine Learning*, 2023. doi: 10.48550/arXiv.2304.03322.

Guozhen Zhang, Zi-Xiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025a. doi: 10.48550/arXiv.2511.03334.

Lisai Zhang, Yi-Di Wu, Qi Liu, Xin Ma, Yang Ding, Gang Yue, Si-Qian Yang, Jingyuan Chen, Lin Ma, and Yao-Hui Wang. Vorch-director: Interactive world story model via noise-aware error rectification. *arXiv preprint arXiv:2608.05776*, 2026a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Neural Information Processing Systems*, 2025b. doi: 10.52202/085713-1024.

Siyang Zhang and Ser-Nam Lim. Towards chunk-wise generation for long videos. *arXiv preprint arXiv:2411.18668*, 2024. doi: 10.48550/arXiv.2411.18668.

Song-Chun Zhang, Yaowei Li, Junhao Zhuang, Weiyang Jin, Hao-Yu Wang, Xin Lu, Yilang Sun, Shiyi Zhang, Hao-Ran Li, Xiaoxiao Ma, Yuming Li, Yijun Liu, Yaofeng Su, Yanwen Ma, Haoyu Wu, Zi-Han Su, Yue Ma, Lvmin Zhang, Hao-Yang Huang, Ze-Yue Xue, A. Rao, and Nan Duan. Echowm: Open and enterable omnimodal world models. *arXiv preprint arXiv:2608.23189*, 2026b.

Yuxuan Zhang, H. Xiong, Jiayi Song, Jinpeng Yu, Yang Shi, Jiaming Liu, Ruihua Huang, and Liwei Wang. Uniswap: Streaming audio-visual identity swapping for talking videos. *arXiv preprint arXiv:2608.11752*, 2026c.

Yu-Peng Zhou, Lianghua Huang, Zhifan Wu, Jia-Bao Wang, Yupeng Shi, Biao Jiang, Daquan Zhou, Yu Liu, Ming-Ming Cheng, and Qibin Hou. Mutual forcing: Dual-mode self-evolution for fast autoregressive audio-video character generation. *arXiv preprint arXiv:2604.25819*, 2026a. doi: 10.48550/arXiv.2604.25819.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yi-Fan Yang, Zhening Xing, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. *arXiv preprint arXiv:2604.08540*, 2026b. doi: 10.48550/arXiv.2604.08540.

Lu Zhu, Xingtong Ge, Fangyu Lin, Yi Zhang, Zhening Liu, Mengfei Li, Yu-Meng Zhang, Guang-Lu Song, Yu Liu, and Jun Zhang. Omni-liveavatar: Minute-level real-time streaming joint audio-video avatar generation. *arXiv preprint arXiv:2608.13602*, 2026.

Kai Zou, Dian Zheng, Hongbo Liu, Tiankai Hang, Bin Liu, and Neng-Hai Yu. Hiar: Efficient autoregressive long video generation via hierarchical denoising. *arXiv preprint arXiv:2603.08703*, 2026. doi: 10.48550/arXiv.2603.08703.

A Calibration and measurement details

Audio path. All audio measurements use the lossless waveform sidecar written by the pipeline. The container's AAC track carries a constant encoder delay of $+64$ ms (1024 samples at 16 kHz), measured with a click train through the encode-decode path. This delay shifts the waveform timing and can bias synchronization and onset measurements.

Readout. Synchformer (Iashin et al., 2024) predicts a class distribution over a ± 2 s grid with 0.2 s steps. We slide its native 4.8 s window along the stream with stride 1 s and read the expectation of the distribution rather than its argmax. This gives a continuous-valued estimate and roughly halves the per-window standard deviation in our calibration. Positions whose reading reaches the range limit are counted as saturated and reported separately, because beyond ± 2 s the estimator folds toward zero. Per-video summaries are the mean absolute reading, the signed mean, the saturation fraction, D_{end} (mean of the last fifth minus the first fifth) and the Theil-Sen slope (Sen, 1968).

Reference calibration. On real, verifiably synchronized footage with injected offsets, the expectation readout responds linearly with slope 0.64 ($r = 0.999$) and a content-dependent floor of 0.199 ± 0.128 s; per-class floors range from 0.14 (ambient) to 0.25 s. The event-agreement metric of Yariv et al. (2023) barely responds to injected offsets up to ± 1 s and, on real footage, reads 0.19 against 0.21 for a density-matched random-peak null, so it is excluded.

Per-arm gain audit. To audit estimator sensitivity, we inject ± 0.2, ± 0.5 and ± 1.0 s into the lossless audio of the generated output and fit the slope of the reported against the injected offset. Gains: Ours 0.417 and NS 0.474 at 30 s (paired $t = -0.84$); 0.368 and 0.475 at 60 s; B4 0.086 [0.043, 0.130] at a floor of 0.433; B2 at 60 s 0.046 [0.041, 0.051]; second backbone (stitched 20 s) 0.59 [0.50, 0.68] for naive continuation and 0.39 [0.19, 0.59] for fixed noise. Silence-conditioned reanalysis excludes 20 ms blocks below -50 dBFS before reading.

Identity. CLIP image embeddings of the frame before and after every seam are compared with the within-window baseline (frames 3 s apart inside one window); a seam is a hard cut when the cross-seam similarity falls more than two standard deviations below the within-window mean. The seam ratio is cross-seam over within-window similarity.

Quality. Following the official JavisBench protocol (Liu et al., 2025b), every long output is cut into consecutive 4.0625 s segments so that the metrics see the same format as native outputs; FVD/KVD (Unterthiner et al., 2018), FAD (Kilgour et al., 2018), video, motion and audio quality (Liu et al., 2025a), ImageBind-AV (Girdhar et al., 2023), CLIP-IQA (Wang et al., 2022) and CLAP (Wu et al., 2022) are averaged over segments. FVD carries an n bias across arms of different length (Ge et al., 2024), so cross-arm comparisons use KVD.

B SyncProbe-48 prompt suite and confirmatory tier

SyncProbe-48 has 48 prompts, 4 sound classes $\times$ 4 sub-types $\times$ 3 prompts: periodic (guitar, drums, violin, claps/steps), vocal (conversation, speech, dog bark, singing), impact (basketball, hammer, chopping, doors) and ambient (rain, waves, wind, fire). Ambient prompts are the negative control: these scenes lack discrete, identifiable events for precise temporal alignment, limiting interpretation of estimated offsets. Every prompt describes a single scene that can continue indefinitely and binds the sound source to a visible motion; the three prompts of a sub-type vary scene, subject and motion amplitude but not the sound source. The 120 s subset is the first prompt of each sub-type of the periodic, vocal and impact classes (12 prompts; ambient prompts, whose synchronization is undefined, are not extended); the 300 s set is the first prompt of each periodic sub-type (guitar, drums, violin, claps/steps). Seeds follow `base_seed = 42 + index` and `seed_k =`

Table 4: SyncProbe-48: the 16 sub-type cells with the first prompt of each (three prompts per cell; full list in the supplementary CSV).

class	sub-type	example prompt
periodic	guitar	a man is playing an acoustic guitar on a busy street corner, strumming steadily as pedestrians walk past
periodic	drums	a drummer is playing a steady rock beat on a drum kit in a rehearsal room, hitting the snare and cymbals
periodic	violin	a violinist is playing a lively tune in a subway station, her bow moving quickly across the strings
periodic	clap_steps	a group of people are clapping their hands in unison at a concert, cheering between claps
vocal	conversation	two friends are chatting animatedly at a sidewalk cafe, gesturing with their hands as they talk
vocal	speech	a man is giving a passionate speech at a podium in front of a small crowd, raising his hand to emphasize points
vocal	dog_bark	a dog is barking loudly at the fence in a backyard, pacing back and forth
vocal	singing	a woman is singing a slow folk song on a quiet stage, holding the microphone with both hands
impact	basketball	a basketball player is dribbling a ball on an outdoor court, the ball bouncing steadily against the concrete
impact	hammer	a carpenter is hammering nails into a wooden plank in his workshop, striking each nail several times
impact	chopping	a person is chopping vegetables on a wooden cutting board in a kitchen, the knife hitting the board with each cut
impact	doors	a woman is opening and closing wooden cabinet doors and drawers while tidying her kitchen
ambient	rain	heavy rain is falling on a quiet street at night, water dripping from the roofs and splashing in puddles
ambient	waves	ocean waves are rolling onto a sandy beach under an overcast sky, foam spreading over the sand
ambient	wind	strong wind is blowing through a golden wheat field, the stalks bending in waves
ambient	fire	a campfire is burning steadily at night, flames crackling and embers rising into the dark

`base_seed`$\cdot 100003 + k$ for window k; the file order is frozen and append-only, and all arms share it. The confirmatory tier is a 200-prompt stratified sample of JavisBench L2 (Liu et al., 2025b), never used during development; the primary contrast, OURS against NS on mean |DeSync| at 30 s, was fixed before the tier was generated. Table 4 lists the cells; the full list is provided in the supplementary material.

C ANCHOR DEFINITIONS, SAMPLER, AND IMPLEMENTATION

Sampler. The released clamping interface passes mask parameters that the model never reads, so clean conditioning latents are treated as fully noised, degrading generation through shared attention. We condition by replacement: at every denoising step the conditioning region is overwritten with the clean latents noised to the current timestep, and the region is restored exactly after the last step. Rectified-flow sampling with 50 steps and classifier-free guidance 5, as released.

Anchors. Video anchor: the last $c = 5$ latent frames (17 frames, 1.0625 s) of the previous window. Audio anchor, three definitions: tail (NS), the last $\lfloor 5.3125c \rfloor = 26$ of the previous window's 104 steps, an interval starting 0.12 s late; upstream arithmetic, the released loop's frame-count rule, 0.48 s off; time-aligned, the steps covering the video span in seconds, 0.0225 s short per window, which the Bresenham schedule (Bresenham, 1965) bounds by scheduling 26 and 27 steps, keeping cumulative rounding error below one audio step. For example, 104 steps cover 4.16 s of padded audio of which 4.0625 s is content; the video span $[3.00, 4.06]$ s corresponds to steps $[75, 101.6)$, so the tail anchor $[78, 104)$ starts 3 steps, 0.12 s, late.

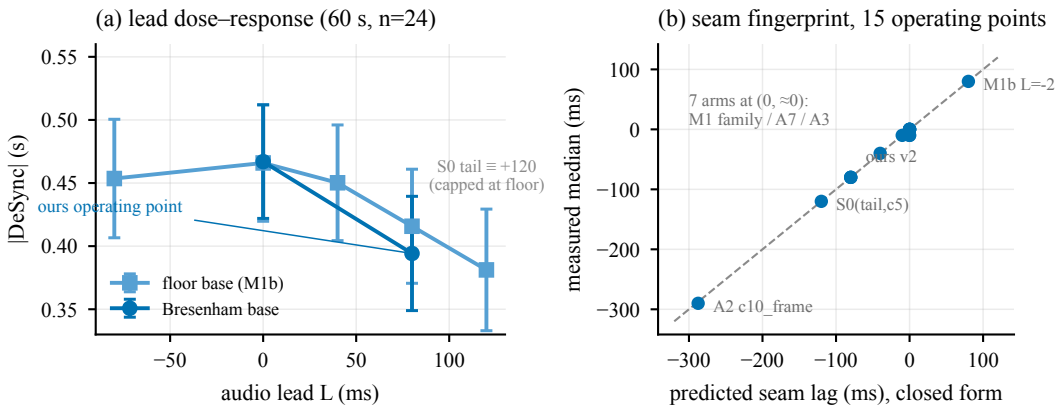


Figure 6: (a) Audio-lead dose-response at 60 s (n=24 paired subsets) on two anchor bases; $+80$ ms is the operating point, and $+120$ ms coincides with the NS tail anchor. (b) Closed-form prediction (Eq. 1) of the per-seam audio lag against the measured cross-correlation median over fifteen operating points; every point lies on $y = x$ within 2.5 ms.

Fingerprint. Eq. 1 gives -0.120 s for the NS tail anchor with c=5 and L=0. After correcting the anchor interval, the remaining lead term is -0.080 s for L=2. Fig. 6 shows all fifteen operating points, including the three anchor definitions, $c \in \{2, 5, 10\}$ and lead doses of both signs, against the measured cross-seam cross-correlation lag (resolution 0.0625 ms).

Baselines. B3 conditions only the video stream by replacement and generates each window's audio from fresh noise without an audio anchor; its audio segments are cropped and cross-faded as for NS. B4 implements offline consensus in the style of Park et al. (2026) on the same backbone. At each rectified-flow step, overlapping regions are reconciled by averaging the windows' Tweedie estimates $\hat{x}_0 = x_t - t\, v_\theta(x_t, t)$ before re-noising. The full set of windows must be retained throughout sampling, and no output is emitted before the final step. B2 generates the full target duration in one pass.

Attention probe. The shared-attention mass of Sec. 4 is read from the released model at four denoising steps per window by summing, per layer and head, the attention weights that audio queries place on video keys and vice versa; the reported numbers are the maximum over layers of the per-layer mean, averaged over windows $k \geq 1$. The probe is observational and introduces no intervention.

Stitching. Windows are decoded independently; video keeps only the latents after the anchor; audio is cropped by seconds, placed at absolute sample positions with a 20 ms equal-power crossfade, and a cross-correlation assertion verifies zero lag against the reference construction on every sample.

D Statistical reporting

We distinguish three reporting tiers. *Confirmatory*: the paired Ours−NS difference in |DeSync| on the held-out 200-prompt tier at 30 s (-0.032 s, $t = -2.93$, $n = 200$), prespecified before evaluation. *Secondary*: effects surviving a Bonferroni correction over roughly 40 development contrasts ($|t| \gtrsim 3.6$): the anchor regression ($t = +6.52$), the lead dose response ($t = -5.18$), the fixed-noise effect without lead ($t = -5.62$), the B4 raw offset contrast ($t = +5.41$), and the second-backbone later-minus-first contrast ($t = +3.78$). *Exploratory*: all duration, class and category contrasts, and the second-backbone noise contrast. Primary-backbone comparisons use paired prompts and seeds; second-backbone comparisons are prompt-paired. Window-level comparisons are identified separately. We report t statistics and confidence intervals where available. Effect estimates are smaller on held-out data: -0.058 s at selection (SyncProbe-48, 60 s), -0.032 s at confirmation, -0.013 s (n.s., MDE 0.07 s) on the 60 s held-out tier. Gain correction gives an estimated 30 s effect of approx-

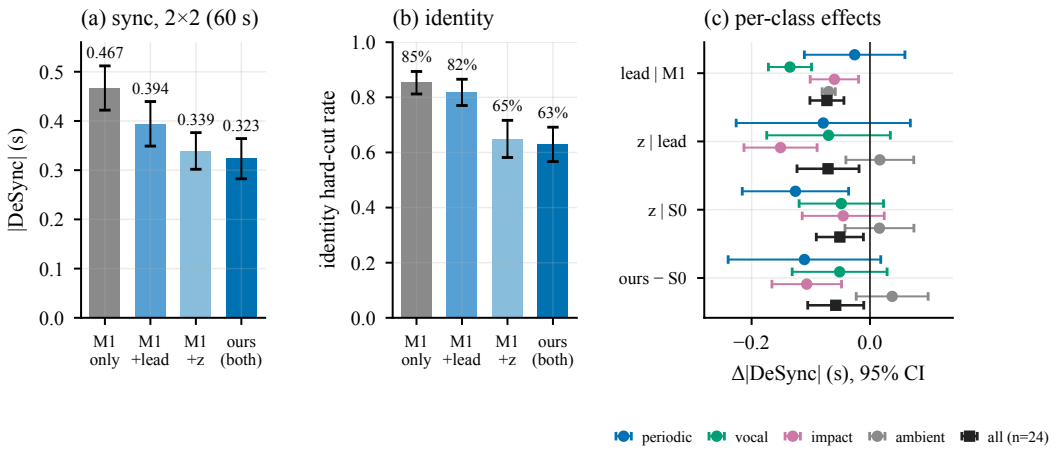


Figure 7: **Component ablation at 60 s** (n=24 seed-paired). (a,b) 2×2 over audio lead and shared noise on top of the modality-aligned anchor; benefits partially overlap, and shared noise accounts for most of the identity improvement. (c) Paired $\Delta|\text{DeSync}|$ with 95% CI per sound class: gains concentrate in periodic and impact sounds. The ambient class has no clearly identifiable synchronization events and serves as a negative control.

Table 5: Held-out 60 s tier, $n = 24$ prompts, seed-paired. Synchronization is reported with its minimum detectable effect; identity and quality follow the 30 s profile.

Method	|DeSync| ↓	Sat. ↓	Hard-cut ↓	Seam ratio ↑	KVD ↓	MQ ↑
NS naive streaming	0.251	8.8%	92.5%	0.837	3.55	0.280
OURS (ours)	0.237	6.9%	**80.5%**	**0.931**	**2.68**	**0.381**
paired t	−0.41	–	−2.14	+6.63	–	–

imately −0.07 s under the linear-response assumption. The results support a modest improvement at 30 s and no observed cumulative drift in the tested continuations. Identity improvements occur at 30, 60 and 120 s and on the 60 s held-out tier, but not on the four 300 s periodic prompts, where both arms average 61% hard cuts (Table 8). The supplementary material provides numerical summaries and 120 s synchronization trajectories.

E ADDITIONAL ABLATIONS AND TABLES

The tested rounding rules (floor, round, Bresenham) produce no statistically significant differences in windowed synchronization. This comparison is distinct from the cumulative rounding error controlled by the Bresenham schedule. The fixed-noise benefit persists across tested CFG scales and step counts. Five-minute continuations are in Table 8: all eight Theil-Sen slopes lie within $\pm 10^{-3}$ s per second and end drifts within ± 0.26 s. Class- and category-resolved paired effects are in Tables 6 and 7. Direct extrapolation at 60 s (n=8) reads 0.213 ± 0.148 with an estimator gain of 0.046, indicating low sensitivity on degraded content.

F SECOND BACKBONE DETAILS

To examine dependence on the backbone and conditioning scheme, we repeat the measurements on Ovi (Low et al., 2025), an 11B twin-tower generator whose continuation conditions each new window on the last frame of the previous one through the image path (Huang et al., 2024) while audio receives no conditioning. Across sixteen prompts and four- or eight-window continuations (Table 9), later windows have offset estimates 0.165 s above the first ($t = +2.76$, $n = 48$; 0.136 s at eight windows, $t = +3.78$), a step at the first seam followed by little further increase. Reusing the noise lowers the stitched offset from 0.446 to 0.347 s at 20 s (paired $t = -2.67$, $n = 16$; $t = -1.73$

Table 6: Paired $\Delta|\text{DeSync}|$ (s) by SyncProbe-48 sound class at 60 s (n=6 per class, mean$\pm$SEM) and overall (n=24, with t).

contrast	periodic	vocal	impact	ambient	all (t)
OURS − NS	-0.111±0.066	-0.051±0.041	-0.107±0.030	+0.038±0.031	-0.058 (-2.39)
shared z on NS	-0.126±0.046	-0.048±0.036	-0.045±0.035	+0.016±0.030	-0.051 (-2.51)
shared z on lead	-0.079±0.075	-0.070±0.053	-0.151±0.032	+0.017±0.029	-0.071 (-2.65)
lead on ①	-0.026±0.043	-0.135±0.019	-0.060±0.021	-0.070±0.006	-0.073 (-4.97)

Table 7: Confirmatory tier (200 prompts, 30 s): paired OURS−NS |DeSync|. The overall row reports the primary comparison; category-level contrasts are exploratory and categories are multi-label.

category	$\Delta\|\text{DeSync}\|$ (s)	t
all (n=200)	-0.032	-2.93
Living scenario	-0.067	-3.46
Musical sounds	-0.051	-3.76
Single event / single subject	≈ -0.05	≈ -2.2
2D animation / virtual (small n)	-0.11 / -0.13	–
Ambient sounds	≈ 0	n.s.
Industrial scenario (n=22)	$+0.072$	$+2.05$

at 40 s), with most of the improvement on the eight pilot prompts. Hard cuts increase from 23% to 56% ($t = +2.83$). The image condition already provides identity information, which may alter the role of shared noise. These results suggest dependence on the backbone and conditioning scheme, but do not isolate their causal contributions.

The sixteen prompts are the eight pilot prompts plus eight SyncProbe-48 prompts, two per class; the first window of the fixed-noise arm is bit-identical to native single-window generation under the same seed. On the first eight prompts the stitched offset falls by 0.186 s ($t = -4.43$) and later windows read 0.222 s better than the first ($t = -3.91$); on the eight added afterwards the stitched contrast is -0.012 s ($t = -0.27$) and later windows read worse at 40 s ($+0.195$, $t = +2.58$). Injected-offset gains: naive 0.59 [0.50, 0.68] and 0.49 [0.35, 0.63]; fixed noise 0.39 [0.19, 0.59] and 0.31 [0.10, 0.52], limiting direct interpretation of the raw differences. Chained last-frame conditioning was reported to be robust for strong silent-video backbones and to fail through bad noise draws rather than slow drift (Zhang & Lim, 2024). Our measurements additionally identify a change in audio-video alignment after the first window.

G EFFICIENCY DETAILS

Our streaming procedure is causal with constant per-window cost; it does not achieve real-time generation. On an RTX 3090, a 4 s window requires approximately one minute (50 steps, two forward passes per step). Peak memory is 19.5 GB across tested durations, and the first window is emitted after one sampling pass. B4 revisits every window at each denoising step. Its measured total runtime is similar, but output is available only after full-duration denoising. Training-free extension addresses a different computational constraint from inference acceleration through distillation.

H CASE STUDIES

Fig. 11 provides additional qualitative examples. NS has a mean spectral-flux excess of $z = 5.4$ at seams. Under OURS, the fixed -80 ms seam displacement is not accompanied by excess spectral flux. B2 at 30 s illustrates metric limitations: largely static, near-silent output has $|\text{DeSync}| = 0.187$ s, FVD 1119, and a favorable VLM quality reward. W33 similarly combines a high consistency score with poor image quality (CLIP-IQA 0.29). Demonstration videos are provided in the supplementary material.

Table 8: Five-minute continuations (the first prompt of each periodic sub-type, same seed trajectory as the 60 s runs): |DeSync| over the whole stream, Theil-Sen slope, and identity hard-cut rate. Identity does not separate the arms on these four prompts.

	|DeSync| (s)		slope (s/s)		hard-cut	
prompt	NS	OURS	NS	OURS	NS	OURS
guitar	0.365	**0.269**	−0.00009	+0.00008	44%	99%
drums	0.835	**0.806**	+0.00065	+0.00013	100%	71%
violin	0.612	**0.570**	−0.00037	−0.00099	61%	60%
claps / steps	0.401	**0.275**	−0.00026	−0.00018	38%	14%
mean	0.553	**0.480**			61%	61%

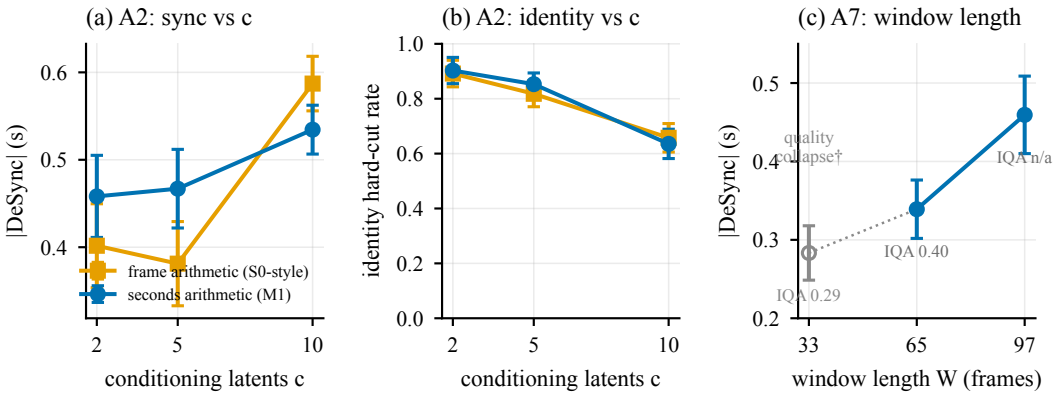


Figure 8: Conditioning-window and window-length sweeps at 60 s (n=24). (a,b) Anchor length c with frame- vs. seconds-arithmetic for the audio anchor. (c) Window length W: W=33 gives favorable scores while image quality collapses (CLIP-IQA 0.29).

I HUMAN STUDY PROTOCOL AND RESULTS

Pre-registration. The protocol was finalized on 2026-09-10 before responses were collected and is included in the supplementary material.

Materials. 30 s outputs of OURS, NS and B4 for held-out benchmark prompts with identical seeds; excerpt seconds 8–20 (12 s, spanning four seams at 10.06/13.06/16.06/19.06 s for the streaming arms), audio remuxed from the lossless sidecar, video re-encoded at native 432×240, 16 fps. Thirty sync prompts whose primary sound class is not ambient (stratified: musical 21, mechanical 4, biological 3, speech 2), six catch prompts with ambient sound only, and three real-footage anchors (Creative-Commons clips) paired with copies whose audio is shifted by +400, −400 and +160 ms. File names are hash-blinded; the page shows only A/B; raters see no method names, hypotheses or instrument readings.

Session. 26 trials per rater: 10 OURS-vs-NS, 10 OURS-vs-B4 (disjoint prompt subsets rotated by rater number so that 15 raters give every prompt exactly 5 judgements per comparison), 3 catch, 3 anchors; order and left/right position from a PRNG seeded by the rater number; both clips must play to $\geq$ 90% before answering; headphones required. Questions: synchronization 2AFC (primary), confidence, overall preference, and a per-clip probe for clicks, pops or sudden jumps.

Analysis. H1 (primary): P(sync = OURS | vs. NS) $>$ 0.5, exact binomial one-sided and a rater-level sign test, both required; H2 vs. B4 two-sided; H3 overall preference vs. B4 two-sided; H4 glitch rate NS $>$ OURS (McNemar). Exclusion: fewer than two of the two ±400 ms anchors correct, or median trial time below 20 s; a sensitivity analysis under the looser rule of at least one anchor correct is pre-declared.

Table 9: Second backbone, sixteen prompts (four per class), four and eight windows of 5 s. Per-window readings are on each window in isolation; stitched readings slide the estimator over the full output. Paired rows are prompt-paired.

Windows	Continuation	window 0	later	later−first	stitched	hard cut
4 (20 s)	naive	0.204	0.369	+0.165 (t=2.76)	0.446	**23%**
	fixed noise	0.361	0.285	−0.077 (t=−1.22)	**0.347**	56%
	fixed−naive		−0.084 (t=−1.22)		−0.099 (t=−2.67)	+33 pt (t=2.83)
8 (40 s)	naive	0.204	0.340	+0.136 (t=3.78)	0.436	**12%**
	fixed noise	0.361	0.372	+0.011 (t=0.23)	**0.358**	33%
	fixed−naive		+0.032 (t=0.64)		−0.078 (t=−1.73)	+21 pt (t=2.23)

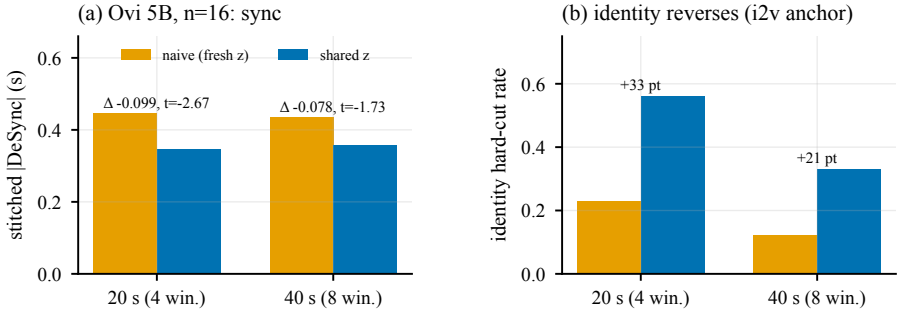


Figure 9: Second backbone: (a) shared noise lowers stitched |DeSync| at 20 and 40 s, with prompt-dependent effects and unequal estimator gains; (b) shared noise increases identity discontinuities.

Status at submission. Seven of fifteen planned raters had completed the study. Two pass the strict rule; four miss one ±400 ms anchor and one misses both. Median trial times range from 26 to 35 s across raters, all clips were played to completion, and raters selected the higher-confidence response on 85% of trials. These observations do not establish why anchor trials were missed. We report both pre-registered rules and treat the current results as descriptive. Table 11 reports these interim results. Under the looser rule, per-rater synchronization preference rates are $0.6, 0.5, 0.4, 0.9, 0.9, 0.7$ against NS and $0.7, 0.5, 0.4, 1.0, 0.9, 0.9$ against B4. Each of the -400, $+400$, and $+160$ ms anchor items was answered correctly by four of six raters.

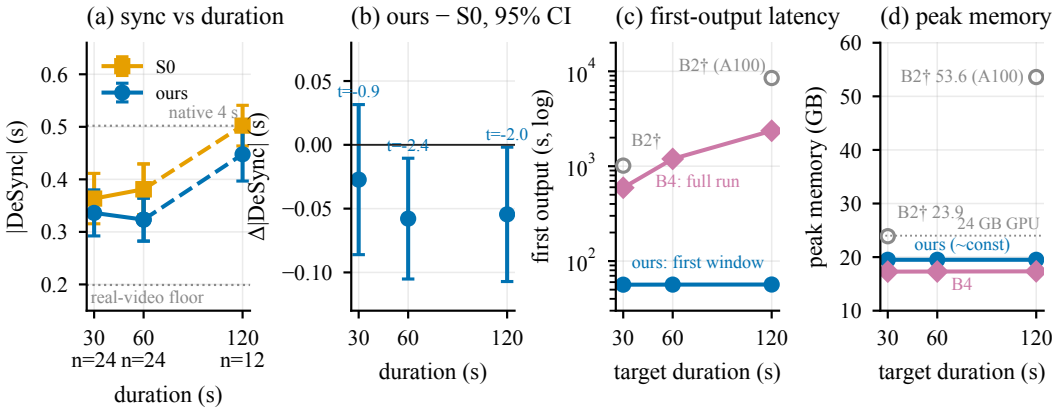


Figure 10: **Scaling to two minutes.** (a) |DeSync| vs. target duration on matched subsets (n=24 at 30/60 s, 12 at 120 s) and (b) paired OURS−NS with 95% CI. (c) First-output latency on an RTX 3090: streaming returns its first window after ≈56 s regardless of length, B4 returns nothing before the full run, and B2 needs 141 min and 53.6 GB on an A100 for 120 s. (d) Peak memory stays at ≈19.5 GB for streaming (Table 10, App. G).

Table 10: Cost versus target duration on one RTX 3090 (24 GB), seed-paired prompts, mean wall-clock seconds. First output is the time to the first emitted window for streaming systems and the full run for offline ones. ‡: 80 GB accelerator, listed for reference.

Method	Duration	Total wall (s)	First output (s)	Peak memory	Unbounded
OURS	30 / 60 / 120 s	571 / 1147 / 2293	56 / 57 / 57	19.5 GB	✓
B4 offline consensus	30 / 60 / 120 s	593 / 1193 / 2362	593 / 1193 / 2362	17.3 GB	–
B2 extrapolation	30 s	1014	1014	23.9 GB	–
B2 extrapolation	60 s	≈ 2460‡	≈ 2460‡	24 to 36 GB‡	–
B2 extrapolation	120 s	8475‡	8475‡	53.6 GB‡	–

Table 11: Human study, seven raters at submission (pre-registered; descriptive). Preference rows count choices of OURS; seam-artifact rows count reported events. Wilson 95% intervals summarize trials without adjusting for within-rater dependence. The strict rule requires both ±400 ms anchors correct; the looser rule requires at least one (pre-declared sensitivity analysis).

	strict (n=2 raters)		looser (n=6 raters)	
Read-out	vs. NS	vs. B4	vs. NS	vs. B4
better synchronized: OURS	13/20	13/20	40/60 [.54, .77]	44/60 [.61, .83]
rather watch: OURS	12/20	12/20	39/60 [.52, .76]	43/60 [.59, .81]
seam artifact heard: OURS / other	0 / 0	0 / 0	1 / 2	0 / 4
ambient catch trials choosing OURS	2/6		10/18	



Figure 11: Frames ± 0.3 s around three consecutive seams for three held-out prompts (30 s samples). NS changes camera, subject or layout at the illustrated seams; B4 preserves identity but exhibits appearance changes and blurring. OURS better preserves subject, clothing, and layout while pose changes.