HORSES FOR COURSES: BREAKING MODALITY SYMMETRY IN FEW-STEP AUDIO-VIDEO DIFFUSION DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified audio-video diffusion transformers generate synchronized video and audio, but the prohibitive cost of iterative sampling necessitates aggressive few-step distillation. Prevailing distillation practice inherits a modality-symmetric recipe, indiscriminately applying identical distribution matching (DM) objectives and synchronous step schedules across both modalities. In this work, we reveal that this symmetric recipe systematically undermines the relational temporal structure of audio. Through a Fisher information analysis of Gaussian data along the diffusion path, we prove that re-noising suppresses information about cross-time correlations at a quadratic rate $\mathcal{O}(\rho_\tau^2)$ in the signal fraction ρ_τ, whereas information about first-order statistics decays only linearly, $\mathcal{O}(\rho_\tau)$. Consequently, weak long-range acoustic dependencies receive an attenuated distributional gradient that is readily overwhelmed by the bias of the learned fake score. Teacher re-noising diagnostics confirm that audio loses disproportionately more structure than video at the same signal-to-noise ratio. To address this modality asymmetry, we propose Modality-Aware Asynchronous Distillation (MAAD), which breaks distillation homogeneity by decoupling optimization objectives and step schedules within a single shared student. Exploiting the substantial token-count disparity between modalities, asynchronous denoising inserts audio-only sub-steps that reuse the visual key-value cache, affording dense acoustic refinement at merely 6% additional computation. Student-context trajectory alignment (SCTA) evaluates the teacher's audio velocity directly under the student's detached visual cache, removing the cross-modal conditioning mismatch between the visual context of the target and that of the student. On MiniMax-H3 (33B), MAAD outperforms compared methods on all nine singing metrics, reducing MuQ FAD from 11.37 to 4.58, and matches or exceeds the 50-step teacher on four of six song-quality scores.

1 INTRODUCTION

Joint audio-video generation has become a core capability of recent video generators (Google DeepMind, 2025; OpenAI, 2025), including unified single-stream diffusion transformers that denoise concatenated video and audio tokens with one set of parameters (SII-GAIR et al., 2026; MiniMax, 2026). Deploying these large models requires aggressive few-step distillation (FastVideo contributors, 2026). Few-step distillation follows two strategies: *trajectory-based* methods train the student to follow the teacher's sampling trajectory from each noise sample (Salimans & Ho, 2022; Song et al., 2023), whereas *distribution matching* (DM) aligns the student's output distribution with the teacher's through the gap between the teacher score and a learned fake score on re-noised samples (Yin et al., 2024b;a; Luo et al., 2025). Current joint audio-video distillation follows a *modality-symmetric recipe*: it adopts DM-based objectives developed for images and video and applies the same objective and step schedule to both modalities (FastVideo contributors, 2026; Su et al., 2026; Yang et al., 2026). While DM produces sharp video (Yin et al., 2025; Huang et al., 2025; Zheng et al., 2026), whether this symmetric recipe serves audio remains unclear.

In this work, we demonstrate that this symmetric recipe erodes the temporal structure of audio. Because DM observes the student solely through re-noised marginals, its supervisory gradient depends entirely on how far the re-noised student and teacher distributions differ. Through a Fisher

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.

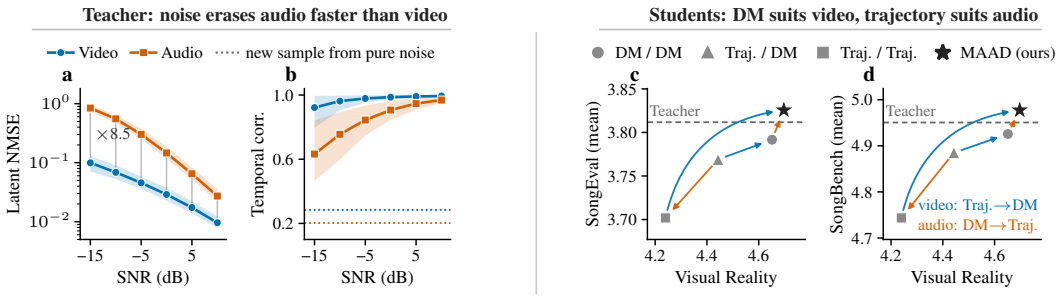

Figure 1: Modality asymmetry. (a, b) Teacher samples re-noised to the same SNR and regenerated (35 clips; mean and 10th-90th percentile): (a) latent NMSE; (b) correlation of the frame-change profile (video) or energy envelope (audio) with the original; dotted: a new sample from pure noise. (c, d) Students with each objective per modality (video/audio) on the Sing350 singing prompts; each arrow changes the objective of one modality.

information analysis of Gaussian data along the diffusion path, we prove that re-noising suppresses information about cross-time correlations at a quadratic rate $\mathcal{O}(\rho_\tau^2)$ in the signal fraction ρ_τ, whereas information about first-order statistics decays only linearly, $\mathcal{O}(\rho_\tau)$. Since audio quality hinges on long-range relations across time, the DM signal on this *relational temporal structure* is intrinsically weak and readily dominated by the systematic bias of the learned fake score. Consistent with this analysis, when teacher samples of MiniMax-H3 (MiniMax, 2026) are re-noised to the same signal-to-noise ratio and regenerated, audio drifts markedly further from the original than video, both in latent NMSE and in its temporal profile (Fig. 1a,b). In distillation, DM improves video but degrades the musical quality of audio, whereas trajectory supervision faithfully preserves it (Fig. 1c,d).

These findings motivate pairing DM for video with trajectory supervision for audio, yet realizing this division within one joint student exposes two coupled challenges. Trajectory alignment matches the teacher's instantaneous velocity field and thus demands fine integration steps to keep its discretization error small, but the step budget of a joint model is dictated by video tokens: a five-second 768×1344 clip in H3 comprises 37,296 video tokens but only 414 audio tokens. Moreover, the teacher's audio velocity depends on its visual context, and DM lets the student's video depart from the teacher's trajectory. A target evaluated under the teacher's video, or with the student's video re-encoded by a full teacher pass, therefore induces a *cross-modal conditioning mismatch* with the visual context that the student's audio actually attends to at inference.

To resolve these coupled difficulties, we propose *Modality-Aware Asynchronous Distillation* (MAAD), which breaks distillation homogeneity within a single student (Fig. 2; Sec. 4). Exploiting the token disparity, *asynchronous denoising* inserts audio-only sub-steps that reuse the visual key-value cache of sparse joint steps, granting audio 20 updates alongside 4 video updates at about 6% extra computation. *Student-context trajectory alignment* (SCTA) removes the conditioning mismatch by evaluating the teacher's audio velocity directly under the student's detached visual cache, at the cost of the audio tokens alone, while video is distilled with trajectory distribution matching (TDM) (Luo et al., 2025) on the same rollouts. Our main contributions are as follows.

- We identify a modality asymmetry in few-step distillation of joint audio-video generators: re-noising suppresses information about cross-time correlations faster than about first-order statistics, so DM benefits video but erodes the relational temporal structure of audio. We derive this in a Gaussian model and verify it on the teacher and in distillation (Secs. 3 and 5.4).
- We propose MAAD, an asynchronous distillation framework that couples visual key-value reuse for dense audio sub-steps with SCTA for on-policy cross-modal trajectory alignment (Sec. 4).
- On the 33B MiniMax-H3, MAAD outperforms released four-step distillations on all nine singing metrics, reducing MuQ FAD from 11.37 to 4.58 at 6% extra computation, and ranks first among them on five of nine general-generation metrics; ablations show that both asynchronous denoising and SCTA are needed (Sec. 5).

2 Related Work

Joint audio-video generation. Cascaded pipelines synthesize video first and then generate audio conditioned on the frames (Iashin & Rahtu, 2021; Luo et al., 2023b; Zhang et al., 2026; Cheng et al., 2025; Polyak et al., 2024), so the video cannot adapt to the audio. Joint generation instead denoises both modalities together. Multi-branch architectures couple modality-specific backbones through cross-modal or joint attention (Ruan et al., 2023; Wang et al., 2025; Low et al., 2025; HaCohen et al., 2026; Huang et al., 2026), whereas single-stream models process video and audio tokens as one sequence with a shared transformer (SII-GAIR et al., 2026; MiniMax, 2026). In this single-stream setting, which we study, one set of weights governs both modalities and every step processes all tokens, tying the audio step budget to the cost of video.

Few-step distillation across modalities. Few-step distillation either follows the teacher's sampling trajectory (Salimans & Ho, 2022; Song et al., 2023; Luo et al., 2023a; Lu & Song, 2025) or matches its output distribution with DM (Yin et al., 2024b;a). Video distillation now relies mainly on DM (Luo et al., 2025; Yin et al., 2025; Huang et al., 2025; Fan et al., 2026), although consistency-based students exist (Wang et al., 2023; 2024b;a). Few-step audio generators, by contrast, have largely been trajectory-based (Ye et al., 2023; Bai et al., 2024; Liu et al., 2024; 2025), and those that use DM add adversarial or metric losses to it (Novack et al., 2025; Li et al., 2025). Joint audio-video distillation has followed the video practice: FastH3 and daVinci-MagiHuman use DMD2 (FastVideo contributors, 2026; SII-GAIR et al., 2026), OmniForcing combines DMD with ODE regression and Self Forcing (Su et al., 2026), and TurboT2VA combines consistency distillation with a DM-style score regularizer and per-modality loss weighting (Yang et al., 2026). All of them share the objective and the step schedule across modalities. Mutual Forcing instead self-distills a fast autoregressive audio-video character generator from a weight-shared multi-step mode (Zhou et al., 2026).

Modality-specific and on-policy distillation. Unlike them, MAAD decouples the objective and step budget of the two modalities within one student. Heterogeneous step allocation (Chu & Patel, 2026) also varies steps across tokens and keeps the cached keys and values of skipped tokens, but as a training-free scheme that skips steps for slowly evolving tokens of a multi-step video model; MAAD instead gives the few audio tokens more updates than video in a trained few-step student. Trajectory targets in prior joint distillation are computed on teacher trajectories or re-noised data (Su et al., 2026; Yang et al., 2026), whereas on-policy distillation evaluates the teacher on the student's own samples, first for language models (Agarwal et al., 2024) and recently for diffusion models (Li et al., 2026). SCTA extends on-policy distillation across modalities by evaluating the teacher's audio velocity under the student's visual cache, removing the cross-modal conditioning mismatch.

3 Diagnosing the Modality-Symmetric Recipe

3.1 Preliminaries

Let x^V and x^A be the clean video and audio latents and c the text condition; H3 uses the flow-matching path

$$z_t^m = t\, x^m + (1-t)\, \epsilon^m, \qquad \epsilon^m \sim \mathcal{N}(0, I), \qquad m \in \{V, A\}, \tag{1}$$

with $t = 0$ at noise and $t = 1$ at data. The model takes a separate time for each modality and predicts the velocities $(v^V, v^A) = F(z^V_{t_V}, z^A_{t_A}, t_V, t_A; c)$, each of which defines a clean prediction $f^m(z^m_{t_m}, t_m) = z^m_{t_m} + (1-t_m)\, v^m$.

We distill a frozen teacher F_ϕ into a student F_θ. DM trains an auxiliary fake-score model F_ψ on student samples and, for a student sample re-noised to a probe time τ, moves the clean prediction against

$$g^m(z^m_\tau, \tau) = f^m_\psi(z^m_\tau, \tau) - f^m_\phi(z^m_\tau, \tau), \tag{2}$$

which is proportional to the gradient of $\mathrm{KL}(p^\theta_\tau \,\|\, p^\phi_\tau)$ between the re-noised student and teacher distributions (Yin et al., 2024b; Luo et al., 2025). Trajectory-based supervision instead regresses the student velocity onto the teacher velocity at the same state,

$$\mathcal{L}_{\mathrm{traj}} = \left\| v^m_\theta - \mathrm{sg}(v^m_\phi) \right\|^2. \tag{3}$$

Sharing one objective and one step schedule across modalities is the modality-symmetric recipe.

3.2 Noise Suppresses Relational Structure

The direction in Eq. (2) is informative only to the extent that the re-noised student and teacher distributions differ. Near the teacher, $\mathrm{KL}(p_\tau^\theta \,\|\, p_\tau^\phi) = \frac{1}{2}\,\delta^\top \mathcal{I}_\tau\,\delta + o(\|\delta\|^2)$, where δ parameterizes the student-teacher discrepancy and $\mathcal{I}_\tau$ is the Fisher information of the re-noised distribution. We compute $\mathcal{I}_\tau$ for two kinds of discrepancy in a Gaussian model $x \sim \mathcal{N}(\mu, \Sigma)$, using the signal fraction $\rho_\tau = \tau^2/(\tau^2 + (1-\tau)^2) = \mathrm{SNR}_\tau/(1+\mathrm{SNR}_\tau)$ (proof in Appendix A).

Proposition 1 (Noise suppresses relational information)**.** *Under Eq. (1): (i) for $\Sigma = I$, a mean discrepancy $\Delta\mu$ changes the re-noised distribution by $\mathrm{KL} = \frac{\rho_\tau}{2}\|\Delta\mu\|^2$; (ii) for two unit-variance positions with correlation r, the re-noised marginals do not depend on r, and the Fisher information about r is*

$$\mathcal{I}_\tau(r) = \rho_\tau^2\,\frac{1+\rho_\tau^2 r^2}{(1-\rho_\tau^2 r^2)^2}. \tag{4}$$

At high noise ($\rho_\tau \to 0$), the information about first-order statistics thus decays linearly in the signal fraction ρ_τ, whereas the information about correlations decays quadratically, $\mathcal{O}(\rho_\tau^2)$. These orders do not rely on the Gaussian model: for any data distributions with bounded support, a Hermite expansion shows that at high noise mean differences enter the KL divergence at $\mathcal{O}(\rho_\tau)$ and, at matched means, second-moment differences at $\mathcal{O}(\rho_\tau^2)$ (Proposition 2 in Appendix A). The factor $(1+\rho_\tau^2 r^2)/(1-\rho_\tau^2 r^2)^2 \geq 1$ grows only for strong correlations, so weak correlations, typical of long-range relations, are suppressed the most. Intuitively, audio quality hinges on such relations: rhythmic regularity, syllable durations, and the alignment of vocals with accompaniment and lips govern how acoustic events are arranged over time, and each segment can sound plausible while the arrangement is wrong. DM therefore receives a weak signal on this *relational temporal structure*, consistent with few-step audio generators rarely relying on DM alone.

This weakness matters because Eq. (2) is estimated through a learned fake score. Along one relational direction with discrepancy δ, DM is locally $\alpha\delta^2/2$ with curvature $\alpha = \mathbb{E}_\tau[w(\tau)\,\mathcal{I}_\tau]$ for loss weight w, so a systematic fake-score bias b, which shifts the gradient to $\alpha\delta + b$, leaves an equilibrium error $\delta^* = -b/\alpha$ that grows as α shrinks. Adding a trajectory term $\kappa(\delta - d)^2/2$, whose per-sample target is offset by d from the distributional optimum, moves the equilibrium to $\delta^* = (\kappa d - b)/(\alpha+\kappa)$. Where α is small relative to $|b|$, this term anchors the student against the bias; where α is large, it only adds an offset $\kappa d/(\alpha+\kappa)$ toward individual teacher samples, consistent with the preference for DM in video. The analysis thus predicts that if audio relies on such structure more than video does, re-noising to the same SNR erases more of audio, increasingly so as the SNR drops.

3.3 Empirical Check on H3

We test this prediction on the H3 teacher by re-noising its samples, regenerating them from the noisy state, and measuring how much of each sample survives (protocol in Appendix B). As shown in Fig. 1a, at the same SNR, which fixes ρ_τ, audio drifts further than video in every clip: its latent NMSE is $2.8\times$ that of video at 10 dB and $8.5\times$ at -15 dB. Fig. 1b shows that the loss extends to temporal structure: at -15 dB, the energy envelope of audio correlates with the original at 0.63, against 0.92 for the frame-change profile of video. The teacher's schedule does not close this gap, since at the same sampling step it already gives audio a 6.2 dB higher SNR than video (Table 3). The asymmetry persists after distillation: DM improves video but lowers the musical quality of audio, whereas trajectory supervision preserves it (Fig. 1c,d).

3.4 From Diagnosis to Design

These findings suggest supervising audio with Eq. (3) while keeping DM for video. Eq. (3) matches the teacher's instantaneous velocity, and first-order integration with step size Δt follows the teacher trajectory only up to an $O(\Delta t)$ global error, so the audio schedule must be fine. The step schedule of a joint model, however, is set by video, which dominates the cost: at 768×1344, one clip has 37,296 video tokens but only 414 audio tokens, so extra audio steps add little computation. The joint model also ties the audio target to the video: the teacher's audio velocity depends on the video, and DM, which does not preserve the teacher's noise-to-sample mapping, lets the student's video depart from the teacher's, which causes the cross-modal conditioning mismatch. MAAD thus adds audio-only steps at small cost and computes the audio target under the student's video.

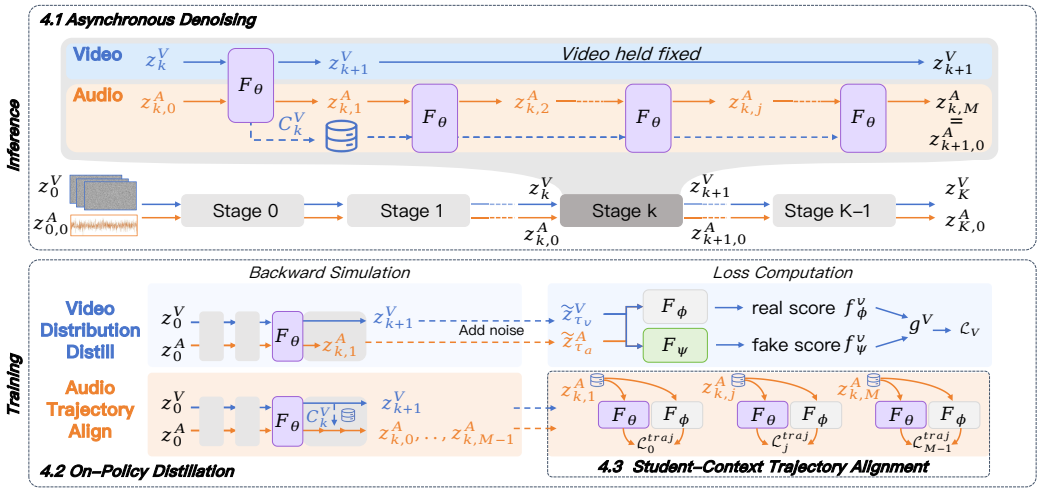

Figure 2: Overview of MAAD (K=4, M=5). **Top** (inference): each stage runs one joint step and $M-1$ audio-only steps that reuse the visual key-value cache C_k^V. **Bottom** (training): the student's rollout to a sampled stage is saved; its video endpoint is re-noised for TDM, and SCTA supervises every audio node under the student's visual cache.

4 Modality-Aware Asynchronous Distillation

MAAD departs from the modality-symmetric recipe along three axes: the step schedule, the objective of each modality, and the visual context of the audio target (Fig. 2). Asynchronous denoising gives audio a finer schedule than video at small cost (Sec. 4.1). On this schedule, on-policy distillation supervises video with TDM and audio with trajectory alignment (Sec. 4.2), and student-context trajectory alignment (SCTA) removes the cross-modal conditioning mismatch by computing the audio target under the student's own visual cache (Sec. 4.3).

4.1 Asynchronous Denoising

Trajectory alignment needs a fine audio schedule, and audio tokens are few, so audio can take extra steps between joint steps at a fraction of the joint cost. We divide the raw time interval into K stages with boundaries $s_k = k/K$, split each stage into M audio intervals, and map raw times to model times with the teacher's shifts,

$$s_{k,j} = \frac{k + j/M}{K}, \qquad t_k^V = h_V(s_k), \qquad t_{k,j}^A = h_A(s_{k,j}), \tag{5}$$

for $k < K$ and $j \le M$, with $t_K^V = 1$, where $h_m(s) = s/\big(1 + (\gamma_m - 1)(1 - s)\big)$. Let z_k^V be the video state at the start of stage k and $z_{k,j}^A$ the audio state at node j. Each stage starts with a joint step $(u_k^V, u_{k,0}^A, C_k^V) = F_\theta(z_k^V, z_{k,0}^A, t_k^V, t_{k,0}^A; c)$, where u denotes student velocities and C_k^V collects the layer-wise visual keys and values. Video then takes an Euler step to t_{k+1}^V, whereas audio takes one only to $t_{k,1}^A$. For $j = 1, \dots, M-1$, the video remains frozen, and only the audio and text tokens are recomputed: in self-attention, their queries attend to their own keys and values concatenated with the cached visual ones,

$$u_{k,j}^A = F_{\theta,A}^{\text{cache}}\big(z_{k,j}^A, t_{k,j}^A;\, t_k^V, C_k^V, c\big), \tag{6}$$

each followed by an Euler step to $t_{k,j+1}^A$. The video is frozen in denoising time, not in clip time: the cache covers the whole clip, so each audio-only step attends to the full stage-entry video. The next joint step refreshes the cache and pairs the updated video z_{k+1}^V with the refined audio $z_{k+1,0}^A = z_{k,M}^A$, so the acoustic refinement of one stage conditions the video updates of the next. With K=4 and M=5, video is updated N_V=K=4 times and audio N_A=KM=20 times, at the cost of K joint passes and $K(M-1)$ audio-only passes of about 1.6% of a joint pass each, i.e., 4.25 token-normalized function evaluations (NFE) at inference, 6% more than four joint steps (Algorithm 1; Appendix E).

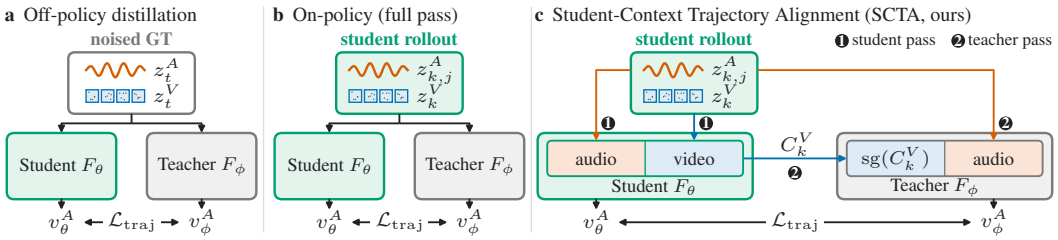

Figure 3: Where the audio target is computed. (a) Off-policy: noised data. (b) On-policy with a full teacher pass on the student's rollout. (c) SCTA: the student pass (❶) caches the visual keys and values C_k^V at the stage-entry joint step, and the teacher pass (❷) reads $\text{sg}(C_k^V)$ in place of its own video.

4.2 On-Policy Distillation

Both objectives supervise the states that the student itself visits at inference, so that it learns to recover from its own errors rather than compound them. At each iteration, we sample a stage k uniformly, run Algorithm 1 from noise (ϵ^V, ϵ^A) to the joint step of stage k without gradients, and record its states, velocities, and cache C_k^V. We write f_θ^m for the student's clean predictions at this joint step and z_{out}^m for the states after its Euler updates, i.e., $z_{\text{out}}^V = z_{k+1}^V$ and $z_{\text{out}}^A = z_{k,1}^A$. The student θ and the fake-score model ψ are separate LoRA adapters (Hu et al., 2022) on the frozen H3 backbone.

Video: TDM on joint steps. DM remains effective for video, so we supervise the video of each joint step with TDM (Luo et al., 2025), which applies DM along the student's trajectory with a separate target for each step. The outgoing video state z_{out}^V is re-noised to a time τ_V within the current step, and f_θ^V is moved against the direction g^V of Eq. (2), which the teacher and the fake model evaluate on this probe. Audio enters the probe only as context, and the loss acts only on the video output. The fake model is trained on the same probes (Appendix C.1).

Audio: trajectory alignment at every node. Audio is supervised at every node of stage k, since $M-1$ of its M updates in a stage are audio-only steps. Node 0 is the incoming audio state of the joint step, nodes $j \geq 1$ follow from z_{out}^A through Eq. (6), and at each node $z_{k,j}^A$ the loss $\mathcal{L}_{A,k,j}$ is Eq. (3) on the velocities of Sec. 4.3. Each iteration updates ψ, and every R iterations the student takes one gradient step on the video loss and M sequential ones on the audio losses (Algorithm 2).

4.3 Student-Context Trajectory Alignment

Trajectory alignment in a joint model raises a question that does not arise for audio alone: under which video should the teacher's audio velocity be evaluated? The answer matters because the audio output receives no DM loss: in the analysis of Sec. 3.2 ($\alpha = b = 0$), the audio student then converges to the trajectory target itself, $\delta^* = d$, and inherits any mismatch between this target and the video that the student generates. Since the teacher's audio velocity depends on the video, which differs between student and teacher, the target should be computed on the inputs that the student's audio receives at inference, namely its own audio state and the visual context of its stage.

Existing choices fall short of this requirement (Fig. 3). Off-policy targets are computed on noised data or teacher trajectories (Song et al., 2023; Su et al., 2026; Yang et al., 2026) and differ from the student in both the audio state and the video. On-policy targets evaluate the teacher on the student's own rollout (Agarwal et al., 2024; Li et al., 2026) and thus match the audio state. In a joint model, however, a full teacher pass re-encodes the video with the teacher's weights, whereas the student's audio-only steps attend to the cache C_k^V, and it costs a joint pass for every audio node. Both choices thus leave a cross-modal conditioning mismatch, which SCTA removes by evaluating both models on the audio and text tokens alone, attending to the detached student cache,

$$v_\omega^A = F_{\omega,A}^{\text{cache}}\big(z_{k,j}^A, t_{k,j}^A;\, t_k^V, \text{sg}(C_k^V), c\big), \qquad \omega \in \{\theta, \phi\}, \tag{7}$$

so the student and the teacher receive identical inputs, and the teacher pass costs only the audio and text tokens. As the student is a LoRA adapter atop the frozen teacher backbone, its cache

preserves the teacher's layer-wise format, which the teacher reads without modification. The cached keys and values differ from the teacher's own encoding of the same video, which is intended: as Self Forcing (Huang et al., 2025) conditions each frame on the model's own history, SCTA grounds the teacher's target in the student's actual inference context. The stop-gradient keeps trajectory gradients out of the video tokens, so the student cannot trivially lower the audio loss by altering its visual context; the shared adapter still couples the two modalities.

5 EXPERIMENTS

We compare MAAD with released four-step distillations of H3 (Sec. 5.2), ablate asynchronous denoising and SCTA (Sec. 5.3), and test the objective assignment against its alternatives (Sec. 5.4). Generated videos with audio are provided in the supplementary material.

5.1 EXPERIMENTAL SETUP

Benchmarks. General text-to-audio-video generation is evaluated on the 778 prompts of the VABench T2AV split (Hua et al., 2026), with 398 easy and 380 hard prompts. For singing, we build Sing350, 350 prompts composed from MusicCaps descriptions (Agostinelli et al., 2023), English lyric excerpts from SongBench (Wu et al., 2026), and descriptions of the scene, performers, and camera; they span nine genre groups and target five-second clips with singing and accompaniment from the start (Appendix E).

Compared methods. We compare with three released four-step distillations of H3, the FP8 H3 Turbo by larryvrh (Larry Turbo) (LarryVRH, 2026), Sol-H3 with the FastH3 adapter (Sana-Sol contributors, 2026; FastVideo contributors, 2026), and LightX2V Turbo (LightX2V contributors, 2026), and report the teacher with 50 joint steps and no classifier-free guidance as a reference. For controlled comparisons, we train variants in our framework that share data, initialization, and training budget, differ only in the stated factor, and are evaluated at the same training step; DM denotes TDM, and Traj. denotes trajectory supervision with Eq. (3) on the student's rollouts.

Metrics. We group metrics by what they measure: video quality (VABench Visual Reality, Expressiveness, and Visual QA), audio-video and text alignment (VABench Alignment, DeSync, ViCLIP, and CLAP), local audio quality (AudioBox and DNSMOS), and song quality on Sing350 (SongBench (Wu et al., 2026) and SongEval (Yao et al., 2025), whose dimensions include melody and coherence). We also measure fidelity to the teacher on the same prompt with MuQ and CLAP FAD and with Onset F1 at a ± 50 ms tolerance, and video motion with the Dynamic Degree of VBench (Huang et al., 2024).

Implementation. MAAD uses $K{=}4$ stages and $M{=}5$ audio updates per stage, i.e., 4 video and 20 audio updates, and generates 124 frames of 768×1344 video at 24 fps with 32 kHz stereo audio. Each model generates one sample per prompt with seed 0; released models use their own samplers. Cost is reported in NFE (Appendix E), and training hyperparameters are in Appendix C.2.

5.2 MAIN RESULTS

General generation. On VABench (Table 1, top), MAAD ranks first among the four-step models on five of nine metrics (Visual QA, Alignment, ViCLIP, CLAP, and AudioBox) and second on three (Visual Reality, DeSync, and DNSMOS). It also exceeds the teacher in Visual QA (0.729 vs. 0.691), ViCLIP (0.223 vs. 0.217), and CLAP (0.361 vs. 0.352) and has lower DeSync (0.501 vs. 0.569 s). Its video quality is on par with the released models.

Singing. On Sing350 (Table 1, bottom), MAAD is the best four-step model on all nine metrics. It matches or exceeds the teacher on four of the six song-quality scores (Melody 4.769 vs. 4.700, Musicality 3.909 vs. 3.891, Coherence 3.875 vs. 3.851, Clarity 3.791 vs. 3.791) and falls slightly below it on Vocal (5.091 vs. 5.149) and Instrumental (5.094 vs. 5.117), whereas every released model falls below the teacher on every score. The margin is largest in fidelity to the teacher, consistent with audio converging to the teacher's trajectory target: compared with the best released model, MuQ FAD drops from 11.37 to 4.58 and Onset F1 rises from 0.727 to 0.803.

Table 1: Comparison with released four-step distillations of H3. Top: VABench T2AV (778 prompts). Bottom: Sing350 (350 singing prompts); FAD and Onset F1 are measured against the teacher. NFE: token-normalized function evaluations. Bold and underline: best and second best among four-step models; the teacher is a reference.

VABench T2AV	*NFE*	Video			Alignment				Local audio	
		Visual Reality↑	Expressiveness↑	Visual QA↑	Alignment↑	DeSync↓	ViCLIP↑	CLAP↑	AudioBox↑	DNSMOS↑
Teacher	50	4.528	4.275	0.691	4.478	0.569	0.217	0.352	3.586	1.454
Larry Turbo	4	**4.495**	<u>4.294</u>	<u>0.716</u>	<u>4.431</u>	0.550	<u>0.219</u>	0.312	<u>3.444</u>	1.346
Sol-H3	4	4.404	4.242	0.647	4.404	0.601	0.199	0.296	3.328	1.393
LightX2V Turbo	4	4.386	**4.323**	0.714	4.400	**0.469**	0.214	<u>0.318</u>	3.433	**1.471**
MAAD	4.25	<u>4.476</u>	4.293	**0.729**	**4.478**	<u>0.501</u>	**0.223**	**0.361**	**3.544**	<u>1.418</u>

Sing350		Fidelity to teacher			SongBench↑			SongEval↑		
		MuQ FAD↓	CLAP FAD↓	Onset F1↑	Vocal	Melody	Instrumental	Musicality	Coherence	Clarity
Teacher	50	–	–	–	5.149	4.700	5.117	3.891	3.851	3.791
Larry Turbo	4	<u>11.37</u>	0.070	<u>0.727</u>	<u>4.822</u>	<u>4.565</u>	4.825	<u>3.799</u>	<u>3.744</u>	<u>3.678</u>
Sol-H3	4	11.64	<u>0.059</u>	0.715	4.703	4.501	<u>4.833</u>	3.773	3.716	3.615
LightX2V Turbo	4	15.77	0.113	0.723	4.377	4.237	4.751	3.496	3.446	3.344
MAAD	4.25	**4.58**	**0.019**	**0.803**	**5.091**	**4.769**	**5.094**	**3.909**	**3.875**	**3.791**

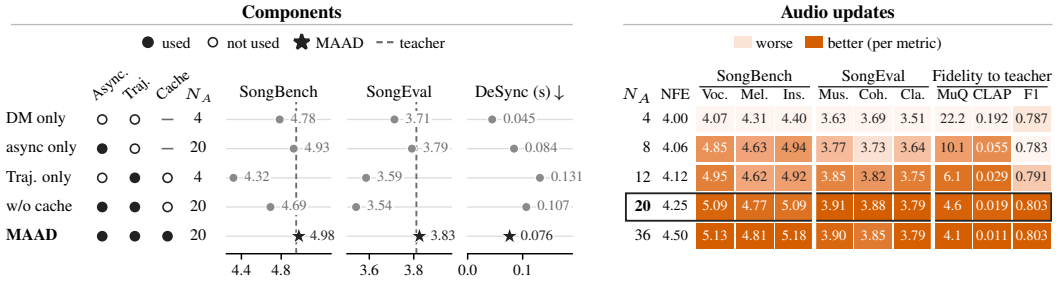

Figure 4: Component ablation on Sing350. Left: configurations with their SongBench, SongEval, and DeSync; Cache: the teacher reads the student's visual cache (SCTA) instead of running a full joint pass. Right: audio updates N_A with four video updates; MuQ and CLAP are FAD against the teacher. Voc./Mel./Ins.: SongBench Vocal, Melody, Instrumental; Mus./Coh./Cla.: SongEval Musicality, Coherence, Clarity.

5.3 Component Ablation

We ablate asynchronous denoising and SCTA on Sing350 (Fig. 4; Tables 4 and 5). The baseline, DM only, is the modality-symmetric recipe: four joint steps with TDM on both modalities.

Asynchronous denoising. Extra audio updates help even under DM: 16 audio-only steps raise SongBench from 4.78 to 4.93 and SongEval from 3.71 to 3.79, but DeSync grows from 0.045 to 0.084 s. Trajectory alignment, in turn, needs these steps. On the four joint steps alone, it lowers SongBench and SongEval to 4.32 and 3.59, as expected from its $O(\Delta t)$ error, whereas MAAD, which adds 16 audio-only steps and SCTA, reaches 4.98 and 3.83, with DeSync 0.076 s, between the 0.084 s of async only and the 0.045 s of DM only. With MAAD's objectives and a separately trained student for each N_A, every audio metric is better at N_A=20 than at N_A=4, e.g., MuQ FAD against the teacher drops from 22.2 to 4.6 and SongEval Coherence rises from 3.69 to 3.88, while NFE grows only from 4 to 4.25. At N_A=36, SongBench and FAD improve further and Onset F1 is unchanged, but all three SongEval scores drop slightly, so we use N_A=20.

SCTA versus a full teacher pass. With asynchronous steps, where the audio target is computed decides whether trajectory alignment helps. If the teacher computes it with a full joint pass on the student's rollout (w/o cache; Fig. 3b), SongEval falls to 3.54, below the DM-only baseline, with SongBench 4.69 and DeSync 0.107 s. With SCTA (MAAD), these become 3.83, 4.98, and 0.076 s. This is expected: without a DM loss, the audio inherits the cross-modal conditioning mismatch between the visual context of its target and the cache it attends to. At the target level (12 replayed

Table 2: Objective assignment on Sing350 (N_A=20). DM: distribution matching (TDM); Traj.: trajectory supervision. SongBench and SongEval: means over their dimensions; Coh.: SongEval Coherence; Mel.: SongBench Melody; Onset F1 is measured against the teacher; DD: Dynamic Degree. Bold: best.

Objective		Video			Song quality		Musical structure			
Video	Audio	VR↑	Expr.↑	DeSync↓	SongBench↑	SongEval↑	Coh.↑	Mel.↑	Onset F1↑	DD (%)
DM	DM	4.65	4.11	0.084	4.93	3.79	3.84	4.69	0.758	8.9
Traj.	DM	4.44	4.07	0.352	4.88	3.77	3.80	4.61	0.777	4.9
Traj.	Traj.	4.24	4.04	0.707	4.74	3.70	3.75	4.50	0.783	2.6
DM	Traj. (MAAD)	**4.70**	**4.13**	**0.076**	**4.98**	**3.83**	**3.88**	**4.77**	**0.803**	**51.1**

Sing350 prompts, relative ℓ_2), the full-pass target lies 8.4% from the SCTA target, comparable to the student's own distance of 7.7%, and a target under the teacher's own video, as in off-policy distillation, lies 14.0% away (Appendix E); we do not train an off-policy variant, so this comparison stays at the target level.

5.4 Which Objective for Which Modality

We distill four students that assign DM or Traj. to each modality, all with N_A=20 (Table 2, Fig. 1c,d, and Appendix Fig. 8). Video favors DM: switching video from Traj. to DM raises Visual Reality from 4.44 to 4.65 with audio on DM and from 4.24 to 4.70 with audio on Traj., and lowers DeSync from 0.352 to 0.084 s and from 0.707 to 0.076 s. Audio favors trajectory supervision once video uses DM, as in MAAD: switching audio to Traj. raises SongEval Coherence from 3.84 to 3.88, SongBench Melody from 4.69 to 4.77, and Onset F1 against the teacher from 0.758 to 0.803, while video also improves slightly. These metrics depend on relations across time, which DM observes weakly (Sec. 3.2). With video on Traj., audio trajectory supervision does not help, and Coherence and Melody drop. Trajectory supervision on both modalities is the weakest configuration on every video and song metric except Onset F1, so the gain comes from assigning each objective to the modality it suits rather than from trajectory supervision alone.

The audio objective affects video. With both modalities on DM, visual and audio quality remain good, but few videos are dynamic: Dynamic Degree is 8.9%, against 43.7% for the teacher. Switching only the audio to SCTA raises it to 51.1% without degrading Visual Reality or DeSync, so in the joint model the audio objective also shapes the video (Appendix E).

5.5 Limitations

MAAD can inherit errors from its teacher. In multi-person scenes, several people may move their mouths while only one voice is audible. Our distillation objectives do not explicitly correct speaker-to-person correspondence, so such errors may persist. MAAD focuses on preserving audio quality during distillation. Although refined audio influences subsequent video updates, we do not explicitly optimize this interaction to improve video. Future work could exploit audio timing and rhythm to guide lip movements, facial expressions, and body motion.

6 Conclusion

We showed that the prevailing modality-symmetric recipe of few-step audio-video distillation is ill-suited to audio. At high noise, the information DM receives about correlations decays quadratically in the signal fraction, whereas that about first-order statistics decays linearly, and re-noising the teacher's samples erases more of the relational temporal structure of audio than of video. MAAD therefore distills video with TDM and audio with trajectory alignment, gives audio extra steps reusing the visual key-value cache, and computes the audio target under the student's visual cache. On H3, it outperforms released four-step distillations on all nine singing metrics at 6% extra computation. Modality-specific objectives and step budgets may also benefit other joint generators.

AI USE STATEMENT

In this work, we used generative AI tools to complete missing textual descriptions during the construction of our evaluation benchmark. We also used these tools to check grammar and improve the readability of the manuscript, assist with code debugging, and organize experimental results. All AI-assisted work was reviewed by the authors, including manuscript edits, suggested code changes, compiled results, and generated text used in the benchmark. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We constructed a test set of 350 samples using information from two publicly available datasets: MusicCaps (Agostinelli et al., 2023), which provides musical descriptions, and SongBench (Wu et al., 2026), which provides English lyric excerpts. Large language models were used to complete missing textual descriptions, and the resulting content was reviewed by the authors. Public availability does not remove the rights associated with the source material; any release of the test set will respect applicable licenses and redistribution restrictions, with appropriate attribution to the original sources. Our user study involved participants watching generated videos and providing subjective evaluations. It did not involve physiological measurements or biomedical interventions.

REPRODUCIBILITY STATEMENT

The method and experimental sections of the main paper, together with the appendix, describe the proposed approach, experimental settings, benchmark construction, and evaluation protocols. To facilitate reproducibility and further research, we will publicly release the model weights, inference and evaluation code, and the 350-sample test set, subject to the applicable terms governing the source data. The release will include instructions for running inference and evaluation with the released resources.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, and Somayeh Sojoudi. ConsistencyTTA: Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation. In *Interspeech 2024*, pp. 3285–3289, 2024. doi: 10.21437/Interspeech.2024-1333.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, June 2025.

Ernie Chu and Vishal M. Patel. Not all tokens need 40 steps: Heterogeneous step allocation in diffusion transformers for efficient video generation, 2026. URL https://arxiv.org/abs/2605.06892.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased DMD: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41667–41676, 2026. URL https://arxiv.org/abs/2510.27684.

FastVideo contributors. FastVideo-FastH3-4-step-Preview-v1-LoRA. https://huggingface.co/FastVideo/FastVideo-FastH3-4-step-Preview-v1-LoRA, 2026. Dense-datafree adapter; accessed September 2026.

Google DeepMind. Veo 3 technical report. https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient Joint Audio-Visual Foundation Model, 2026. URL https://arxiv.org/abs/2601.03233.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. VABench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Xiaohu Huang, Haoyang He, Hao Zhou, Qiangpeng Yang, Shilei Wen, and Kai Han. JoVA: Unified multimodal learning for joint video-audio generation and editing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 346–365. Springer, 2026. doi: 10.1007/978-3-032-37324-3_19.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.08009.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2021. doi: 10.5244/C.35.336.

LarryVRH. MiniMax-H3-Turbo-LoRA: Turbo v4 release. https://huggingface.co/larryvrh/MiniMax-H3-Turbo-Lora, 2026. Accessed September 2026.

Quanhao Li, Junqiu Yu, Kaixun Jiang, Yujie Wei, Zhen Xing, Pandeng Li, Ruihang Chu, Shiwei Zhang, Yu Liu, and Zuxuan Wu. DiffusionOPD: A unified perspective of on-policy distillation in diffusion models, 2026. arXiv:2605.15055. To appear in SIGGRAPH Asia 2026.

Yinghao Aaron Li, Rithesh Kumar, and Zeyu Jin. DMOSpeech: Direct metric optimization via distilled diffusion model in zero-shot speech synthesis. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35186–35208. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25ay.html.

LightX2V contributors. Minimax-H3-Turbo: Four-step v1.0 release. https://huggingface.co/lightx2v/Minimax-h3-Turbo, 2026. Accessed September 2026.

Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. AudioLCM: Efficient and High-Quality Text-to-Audio Generation with Minimal Inference Steps. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7008–7017, 2024. doi: 10.1145/3664647.3681072. URL https://doi.org/10.1145/3664647.3681072.

Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Zhou Zhao, and Wei Xue. FlashAudio: Rectified flow for fast and high-fidelity text-to-audio generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13694–13710, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.673. URL https://aclanthology.org/2025.acl-long.673/.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin Backbone Cross-Modal Fusion for Audio-Video Generation, 2025. URL https://arxiv.org/abs/2510.01284.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.11081.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023a. URL https://arxiv.org/abs/2310.04378.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48855–48876. Curran Associates, Inc., 2023b. doi: 10.52202/075280-2121. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728, 2025. doi: 10.1109/ICCV51701.2025.01646.

MiniMax. MiniMax-H3. Official model repository and model card, 2026. URL https://huggingface.co/MiniMaxAI/MiniMax-H3. Accessed September 23, 2026.

Zachary Novack, Ge Zhu, Jonah Casebeer, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Presto! distilling steps and layers for accelerating music generation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Gj5JTAwdoy.

OpenAI. Sora 2 system card. https://openai.com/index/sora-2-system-card/, September 2025.

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, Ishan Misra, Ji Hou, Jialiang Wang, Kiran 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 Tsai, Samaneh 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 Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie Gen: A Cast of Media Foundation Models, 2024. URL https://arxiv.org/abs/2410.13720.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10219–10228, 2023. doi: 10.1109/CVPR52729.2023.00985.

Tim Salimans and Jonathan Ho. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Sana-Sol contributors. Sana-Sol: Sol-h3 inference backend. https://github.com/NVlabs/Sana/tree/sol-engine/models/minimax_h3/Sol-H3, 2026. Sana sol-engine branch; accessed September 2026.

SII-GAIR, Sand.ai, Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, Lingzhi Li, Lyumanshan Ye, Min Hu, Qiangang Wang, Quanwei Qi, Steffi Chern, Tao Bu, Taoran Wang, Teren Xu, Tianning Zhang, Tiantian Mi, Weixian Xu, Wenqiang Zhang, Wentai Zhang, Xianping Yi, Xiaojie Cai, Xiaoyang Kang, Yan Ma, Yixiu Liu, Yunbo Zhang, Yunpeng Huang, Yutong Lin, Zewei Tao, Zhaoliang Liu, Zheng Zhang, Zhiyao Cen, Zhixuan Yu, Zhongshu Wang, Zhulin Hu, Zijin Zhou, Zinan Guo, Yue Cao, and Pengfei Liu. Speed by Simplicity: A Single-Stream Architecture for Fast Audio-Video Generative Foundation Model, 2026. URL https://arxiv.org/abs/2603.21986.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Haoyang Huang, and Nan Duan. OmniForcing: Unleashing real-time joint audio-visual generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 566–584. Springer, 2026. doi: 10.1007/978-3-032-37271-0_31.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. UniVerse-1: Unified audio-video generation via stitching of experts, 2025. URL https://arxiv.org/abs/2509.06155.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 83951–84009. Curran Associates, Inc., 2024a. doi: 10.52202/079017-2668. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/98a29475083c502c34949f9baa1aa2ef-Paper-Conference.pdf.

Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. AnimateLCM: Computation-efficient personalized style video generation without personalized video data. In *SIGGRAPH Asia 2024 Technical Communications*, pp. 1–5. ACM, 2024b. doi: 10.1145/3681758.3698013.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. VideoLCM: Video Latent Consistency Model, 2023. URL https://arxiv.org/abs/2312.09109.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. SongBench: A fine-grained multi-aspect benchmark for song quality assessment. *arXiv preprint arXiv:2604.25937*, 2026.

Xiaoda Yang, Yuxiang Liu, Kaiwen Zheng, Yuan Liu, Yibo Lai, Shengpeng Ji, Kai Jiang, Jianfei Chen, Shan Yang, Sen Liang, Xiaobin Hu, Shuicheng Yan, Jintao Zhang, Jun Zhu, and Zhou Zhao. TurboT2VA: Fast large-scale text-to-video-audio generation via score-regularized consistency distillation, 2026. URL https://arxiv.org/abs/2608.24674.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, et al. SongEval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

Zhen Ye, Wei Xue, Xu Tan, Jie Chen, Qifeng Liu, and Yike Guo. CoMoSpeech: One-step speech and singing voice synthesis via consistency model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1831–1839. ACM, 2023. doi: 10.1145/3581783.3612061.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487. Curran Associates, Inc., 2024a. doi: 10.52202/079017-1505. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/54dcf25318f9de5a7a01f0a4125c541e-Paper-Conference.pdf.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623, June 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.html.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, June 2025.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. FoleyCrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026. doi: 10.1007/s11263-025-02649-3.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.08431.

Yupeng Zhou, Lianghua Huang, Zhifan Wu, Jiabao 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, 2026. URL https://arxiv.org/abs/2604.25819.

A Proofs for Section 3

Proof of Proposition 1. Under Eq. (1), $z_\tau = \tau x + (1-\tau)\epsilon \sim \mathcal{N}\big(\tau\mu,\ \tau^2\Sigma + (1-\tau)^2 I\big)$. Dividing by $\sigma_\tau = (\tau^2 + (1-\tau)^2)^{1/2}$ is a bijection applied to both the student and the teacher distributions, so it leaves the KL divergence unchanged and gives $\bar{z}_\tau = z_\tau/\sigma_\tau \sim \mathcal{N}(\sqrt{\rho_\tau}\,\mu,\ \Sigma_\tau)$ with $\Sigma_\tau = \rho_\tau\Sigma + (1-\rho_\tau)I$.

(i) A mean discrepancy $\Delta\mu$ shifts the mean of $\bar{z}_\tau$ by $\sqrt{\rho_\tau}\,\Delta\mu$ and leaves its covariance Σ_τ unchanged. Since two Gaussians with a shared covariance have half the squared Mahalanobis distance of their means as KL divergence, this gives $\mathrm{KL} = \frac{\rho_\tau}{2}\,\Delta\mu^\top \Sigma_\tau^{-1}\Delta\mu$ for general Σ, which reduces to $\frac{\rho_\tau}{2}\|\Delta\mu\|^2$ for $\Sigma = I$.

(ii) We first treat a general covariance discrepancy $\Delta\Sigma$, which perturbs Σ_τ by $E = \rho_\tau\Delta\Sigma$. With $P = \Sigma_\tau^{-1}E$,

$$\mathrm{KL}\big(\mathcal{N}(0, \Sigma_\tau + E)\,\|\,\mathcal{N}(0,\Sigma_\tau)\big) = \tfrac{1}{2}\big[\operatorname{tr} P - \log\det(I+P)\big] = \tfrac{1}{4}\operatorname{tr}(P^2) + O(\|P\|^3),$$

using $\log\det(I+P) = \operatorname{tr} P - \frac{1}{2}\operatorname{tr} P^2 + O(\|P\|^3)$. Substituting $P = \rho_\tau\Sigma_\tau^{-1}\Delta\Sigma$ gives $\mathrm{KL} = \frac{\rho_\tau^2}{4}\operatorname{tr}\big[(\Sigma_\tau^{-1}\Delta\Sigma)^2\big] + O(\|\Delta\Sigma\|^3)$, i.e., covariance discrepancies are suppressed quadratically in ρ_τ.

For two unit-variance positions with correlation r, Σ_τ has unit diagonal and off-diagonal entry $r_\tau = \rho_\tau r$, so the re-noised marginals are $\mathcal{N}(0,1)$ regardless of r. The Fisher information of a standard bivariate normal with respect to its correlation r_τ is $(1+r_\tau^2)/(1-r_\tau^2)^2$, and the chain

rule with $\mathrm{d}r_\tau/\mathrm{d}r = \rho_\tau$ yields Eq. (4). As $\rho_\tau \to 0$, $\Sigma_\tau \to I$ and both expressions reduce to $\mathrm{KL} \approx \frac{\rho_\tau}{2}\|\Delta\mu\|^2$ and $\mathrm{KL} \approx \frac{\rho_\tau^2}{4}\|\Delta\Sigma\|_F^2$. □

Beyond the Gaussian model. The orders in Proposition 1 do not depend on Gaussian data at high noise.

Proposition 2 (High-noise expansion)**.** *Let the student and teacher data distributions p_1^θ and p_1^ϕ on $\mathbb{R}^D$ have bounded support, with differences $\Delta\mu$ in the mean and ΔS in the second moment $\mathbb{E}[xx^\top]$. For $\rho \in [0,1)$, let q_ρ and p_ρ be the laws of $\sqrt{\rho}\,x + \sqrt{1-\rho}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ under the student and the teacher, so that $\mathrm{KL}(q_{\rho_\tau} \,\|\, p_{\rho_\tau}) = \mathrm{KL}(p_\tau^\theta \,\|\, p_\tau^\phi)$ by the rescaling above. As $\rho \to 0$, $\mathrm{KL}(q_\rho \,\|\, p_\rho) = \frac{\rho}{2}\|\Delta\mu\|^2 + O(\rho^{3/2})$, and if $\Delta\mu = 0$, then $\Delta S = \Delta\Sigma$ and $\mathrm{KL}(q_\rho \,\|\, p_\rho) = \frac{\rho^2}{4}\|\Delta\Sigma\|_F^2 + O(\rho^{5/2})$.*

Both leading terms are the high-noise limits of the Gaussian expressions above; for two unit-variance positions whose correlation differs by Δr, the second gives $\frac{\rho^2}{2}\Delta r^2$, which matches $\frac{1}{2}\mathcal{I}_\tau(r)\,\Delta r^2$ from Eq. (4) as $\rho_\tau \to 0$.

Proof of Proposition 2. Let φ_s denote the density of $\mathcal{N}(0, sI)$, $\varphi = \varphi_1$, and let $\mathrm{He}_\nu(z) = \prod_i \mathrm{He}_{\nu_i}(z_i)$, for multi-indices $\nu \in \mathbb{N}^D$, be the probabilists' Hermite polynomials, which are orthogonal under φ with $\mathbb{E}_\varphi[\mathrm{He}_\nu^2] = \nu!$. Mehler's formula, applied to each coordinate, gives for all x, z and $\rho < 1$

$$\frac{\varphi_{1-\rho}(z - \sqrt{\rho}\,x)}{\varphi(z)} = \sum_\nu \frac{\rho^{|\nu|/2}}{\nu!}\,\mathrm{He}_\nu(x)\,\mathrm{He}_\nu(z).$$

By Cramér's inequality, $|\mathrm{He}_\nu(x)| \le 1.09^D\sqrt{\nu!}\,e^{\|x\|^2/4}$, so for $\|x\| \le B$ the series converges absolutely and can be integrated term by term. Writing $m_\nu^\omega = \mathbb{E}_{p_1^\omega}[\mathrm{He}_\nu(x)]$ for $\omega \in \{\theta, \phi\}$ and $\Delta m_\nu = m_\nu^\theta - m_\nu^\phi$,

$$\frac{p_\rho}{\varphi} = \sum_\nu \frac{\rho^{|\nu|/2}}{\nu!}\,m_\nu^\phi\,\mathrm{He}_\nu, \qquad \chi_\rho := \frac{q_\rho - p_\rho}{\varphi} = \sum_{|\nu|\ge 1} \frac{\rho^{|\nu|/2}}{\nu!}\,\Delta m_\nu\,\mathrm{He}_\nu,$$

where $\Delta m_\nu^2/\nu!$ is bounded uniformly in ν. By orthogonality, $\mathbb{E}_\varphi[\chi_\rho^2] = \sum_{|\nu|\ge 1}\rho^{|\nu|}\Delta m_\nu^2/\nu!$, and since the number of multi-indices with $|\nu| = n$ grows polynomially in n, the terms with $|\nu| \ge k$ sum to $O(\rho^k)$. The first-order Hermite moments are the means, and the second-order ones are $\mathbb{E}[x_i^2] - 1$ and $\mathbb{E}[x_i x_j]$ for $i \ne j$. Hence $\mathbb{E}_\varphi[\chi_\rho^2] = \rho\|\Delta\mu\|^2 + O(\rho^2)$, and if $\Delta\mu = 0$, $\mathbb{E}_\varphi[\chi_\rho^2] = \rho^2\big(\frac{1}{2}\sum_i \Delta S_{ii}^2 + \sum_{i<j}\Delta S_{ij}^2\big) + O(\rho^3) = \frac{\rho^2}{2}\|\Delta\Sigma\|_F^2 + O(\rho^3)$.

It remains to pass from χ_ρ to the KL divergence. Let $\eta = q_\rho/p_\rho - 1 = \chi_\rho\,\varphi/p_\rho$. Since $\mathbb{E}_{p_\rho}[\eta] = 0$ and $|(1+\eta)\log(1+\eta) - \eta - \frac{1}{2}\eta^2| \le |\eta|^3$ for $\eta \ge -1$,

$$\mathrm{KL}(q_\rho \,\|\, p_\rho) = \tfrac{1}{2}\,\mathbb{E}_\varphi\big[\chi_\rho^2\,\varphi/p_\rho\big] + O\big(\mathbb{E}_\varphi\big[|\chi_\rho|^3(\varphi/p_\rho)^2\big]\big).$$

For $\rho \le \frac{1}{4}$, the Gaussian kernel gives $|\log(p_\rho(z)/\varphi(z))| \le C\sqrt{\rho}\,(1+\|z\|)^2$ with C depending only on B and D, so $|\varphi/p_\rho - 1| \le C\sqrt{\rho}\,(1+\|z\|)^2 e^{C\sqrt{\rho}(1+\|z\|)^2}$. Hence, for small ρ, $\mathbb{E}_\varphi[(\varphi/p_\rho)^4]$ is bounded and $\mathbb{E}_\varphi[(\varphi/p_\rho - 1)^2] = O(\rho)$. Moreover, χ_ρ is the Ornstein-Uhlenbeck semigroup applied to $\chi_{5\rho}$, which multiplies the terms of degree n by $5^{-n/2}$, so Gaussian hypercontractivity gives $\mathbb{E}_\varphi[\chi_\rho^4]^{1/4} \le \mathbb{E}_\varphi[|\chi_\rho|^6]^{1/6} \le \mathbb{E}_\varphi[\chi_{5\rho}^2]^{1/2}$. By Cauchy-Schwarz, the first term differs from $\frac{1}{2}\mathbb{E}_\varphi[\chi_\rho^2]$ by $O(\sqrt{\rho}\,\mathbb{E}_\varphi[\chi_{5\rho}^2])$, and the remainder is $O(\mathbb{E}_\varphi[\chi_{5\rho}^2]^{3/2})$. Since $\mathbb{E}_\varphi[\chi_{5\rho}^2]$ is $O(\rho)$ in general and $O(\rho^2)$ if $\Delta\mu = 0$, the two expansions of $\mathbb{E}_\varphi[\chi_\rho^2]$ prove the claim. □

B Noise-Recovery Protocol

Clips and re-generation. We use 35 teacher samples at 864×480, each with 124 video frames at 24 fps and 5.175 s of stereo audio at 32 kHz; the reference latents x^m are the VAE encodings of the saved media. For each raw time in $\{0.98, 0.96, 0.9, 0.8, 0.6, 0.4\}$ on the teacher's 50-step grid, we map it to t_m with the shift h_m of Eq. (5), form $z_{t_m}^m$ by Eq. (1) with noise drawn from seed 0, and run the remaining teacher steps. Re-generating from pure noise with the same prompt gives the level of each metric at which nothing of the original is kept.

Metrics. The latent NMSE is $\mathrm{mean}\big((\hat{x}^m - x^m)^2\big)/\mathrm{mean}\big((x^m)^2\big)$, where $\hat{x}^m$ is the re-generated latent before decoding. The measured SNR of a clip is $10\log_{10}\big[\mathrm{mean}\big((t_m x^m)^2\big)/\mathrm{mean}\big(((1-t_m)\epsilon^m)^2\big)\big]$. The frame-change profile averages the absolute RGB difference between consecutive frames downscaled to 216×120 (123 points). The energy envelope is the RMS over both channels in non-overlapping 10 ms windows (517 points). Both profiles are compared with those of the original clip by Pearson correlation.

Matched SNR. For each clip and modality, we interpolate log NMSE and the Pearson correlation linearly in the measured SNR between the two bracketing levels, at $-15, -10, \ldots, 10$ dB. Every clip covers this range, so no extrapolation is needed. Fig. 1a,b reports the mean and the 10th-90th percentile over clips; 95% intervals of the mean from 5,000 bootstrap resamples of clips lie within $\pm8\%$ for NMSE and ±0.05 for the Pearson correlation. The ordering holds clip by clip: audio NMSE exceeds video NMSE in all 35 clips at every SNR, by 1.7-4.3$\times$ at 10 dB and 5.1-13.5$\times$ at -15 dB, and the video profile correlates better with the original than the audio envelope in at least 34 of 35 clips at every SNR.

Shared raw time. Table 3 lists the measurements at each raw time. Audio SNR exceeds video SNR by 6.2 dB at every level: the smaller audio shift contributes 12.0 dB, and the lower power of audio latents removes 5.8 dB. Audio NMSE is nevertheless higher at all but the smallest noise level.

Table 3: Noise recovery at a shared raw time on the teacher's grid (mean over 35 clips). At every raw time, the teacher's schedule already gives audio about 6.2 dB higher SNR than video, yet audio NMSE is higher at all but the smallest noise level, so the schedule does not close the gap between the modalities.

	Noise level $1-t_m$		SNR (dB)		Latent NMSE	
Raw time	Video	Audio	Video	Audio	Video	Audio
0.98	0.197	0.058	12.5	18.7	0.0070	0.0063
0.96	0.333	0.111	6.3	12.5	0.0153	0.0177
0.90	0.571	0.250	−2.2	4.0	0.0365	0.0775
0.80	0.750	0.429	−9.3	−3.0	0.0654	0.2384
0.60	0.889	0.667	−17.8	−11.5	0.1234	0.6679
0.40	0.947	0.818	−24.8	−18.6	0.2149	1.0763

C Method Details

C.1 Video TDM Objective

The student's clean video prediction at the joint step of stage k is $f_\theta^V = z_k^V + (1-t_k^V)\, v_\theta^V$, where the video velocity v_θ^V is recomputed with gradients from the saved incoming states. For each modality, we sample a raw time uniformly within the current joint step, map it through Eq. (5) to a time τ_m between the start of that step and its end time e_m, and re-noise the outgoing state z_{out}^m from e_m to τ_m. Under Eq. (1), and dropping the modality superscript, a state at time e is $z_e = e\,x + (1-e)\,\epsilon$, and for $\tau \le e$ the forward transition is

$$\tilde{z}_\tau = \frac{\tau}{e}\,\mathrm{sg}(z_e) + \sqrt{(1-\tau)^2 - \left(\frac{\tau}{e}\right)^2 (1-e)^2}\;\eta, \qquad \eta \sim \mathcal{N}(0, I), \tag{8}$$

which gives $\tilde{z}_\tau = \tau x + (1-\tau)\epsilon'$ with $\epsilon' \sim \mathcal{N}(0, I)$, so $\tilde{z}_\tau$ has the marginal of time τ. The square root is real because $\tau \le e$ implies $(1-\tau) \ge (\tau/e)(1-e)$, and at $\tau = e$ the transition returns z_e. The teacher and the fake model process this joint probe with full forward passes, and the student loss uses their video outputs through g^V of Eq. (2):

$$\mathcal{L}_V = \tfrac{1}{2}\left\| f_\theta^V - \mathrm{sg}\left(f_\theta^V - \mathrm{clip}_G\left(g^V(\tilde{z}_{\tau_V}^V, \tau_V)\right)\right)\right\|_{\mathrm{mse}}^2, \tag{9}$$

where clip_G bounds the per-sample norm by G and $\|\cdot\|_{\mathrm{mse}}^2$ is the squared norm averaged over latent entries, which we also use for Eq. (3). The fake model regresses the video velocity toward the

Algorithm 1 Asynchronous sampling with MAAD

Require: student F_θ, condition c, stages K, audio updates per stage M

$z_0^V \sim \mathcal{N}(0, I)$, $z_{0,0}^A \sim \mathcal{N}(0, I)$
for $k = 0, \dots, K-1$ **do**
  $(u_k^V, u_{k,0}^A, C_k^V) \leftarrow F_\theta(z_k^V, z_{k,0}^A, t_k^V, t_{k,0}^A; c)$ ▷ joint step
  $z_{k+1}^V \leftarrow z_k^V + (t_{k+1}^V - t_k^V)\, u_k^V$
  $z_{k,1}^A \leftarrow z_{k,0}^A + (t_{k,1}^A - t_{k,0}^A)\, u_{k,0}^A$
  for $j = 1, \dots, M-1$ **do**
    $u_{k,j}^A \leftarrow F_{\theta,A}^{\text{cache}}(z_{k,j}^A, t_{k,j}^A;\, t_k^V, C_k^V, c)$ ▷ audio-only step
    $z_{k,j+1}^A \leftarrow z_{k,j}^A + (t_{k,j+1}^A - t_{k,j}^A)\, u_{k,j}^A$
  end for
  $z_{k+1,0}^A \leftarrow z_{k,M}^A$ ▷ cache rebuilt at stage $k{+}1$
end for
return $(z_K^V,\, z_{K,0}^A)$

detached student prediction,

$$\mathcal{L}_{\text{fake}}(\psi) = \left\| v_\psi^V - \frac{\text{sg}(f_\theta^V) - \tilde{z}_{\tau_V}^V}{1 - \tau_V} \right\|_{\text{mse}}^2. \tag{10}$$

In the implementation, small lower bounds stabilize the variance in Eq. (8) and the denominator $1 - \tau_V$ in Eq. (10).

C.2 Algorithms and Hyperparameters

Algorithm 1 gives the asynchronous sampler of Sec. 4.1, which is the only procedure used at inference. Algorithm 2 gives the training loop of Secs. 4.2 and 4.3. Times t_k^V and $t_{k,j}^A$ follow Eq. (5), and all state updates are Euler steps.

Optimization. Each iteration updates ψ with $\mathcal{L}_{\text{fake}}$. Every R iterations, the student takes one step on $\lambda_V \mathcal{L}_V$ and then M sequential steps on $(\lambda_A/M)\, \mathcal{L}_{A,k,j}$, one per audio node, continuing the detached rollout with the updated student; afterwards, $\psi \leftarrow \beta\psi + (1-\beta)\theta$. Gradients are truncated across rollout nodes and optimizer steps.

The teacher's shifts are γ_V=12 and γ_A=3. Default hyperparameters: K=4, M=5, R=5, λ_V=1, λ_A=0.5, β=0.97, G=5, LoRA rank 128 for both adapters, learning rates 10^{-4} (student) and 3×10^{-4} (fake), and teacher guidance scale 1.

D User Study

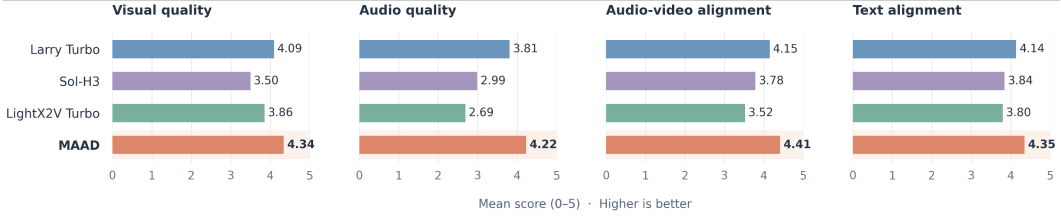

Figure 5: User study of four distillation models on four evaluation dimensions.

We collected ratings from 46 evaluators for 10 video clips under each method (on a scale of 0–5), resulting in a total of 1,840 ratings across four dimensions. MAAD achieved average scores of 4.34, 4.22, 4.41, and 4.35 for visual quality, audio quality, audio-visual alignment, and text alignment, respectively. Larry Turbo ranked second across all four dimensions, while LightX2V Turbo outperformed Sol-H3 in visual quality (3.86 vs. 3.50) and had a text alignment score close to that of Sol-H3 (3.80 vs. 3.84).

E Experimental Details

Algorithm 2 MAAD training

Require: teacher F_ϕ (frozen), student adapter θ, fake adapter $\psi \leftarrow \theta$
Require: K, M, R, λ_V, λ_A, β, G
for $n = 1, 2, \ldots$ **do**
  sample prompt c, stage $k \sim \mathcal{U}\{0, \ldots, K-1\}$, noise (ϵ^V, ϵ^A)
  run Alg. 1 without gradients to the joint step of stage k
  record z_k^V, $z_{k,0}^A$, u_k^V, $u_{k,0}^A$, z_{out}^V, z_{out}^A, C_k^V
  sample τ_V, τ_A and build the probe $(\tilde{z}_{\tau_V}^V, \tilde{z}_{\tau_A}^A)$ ▷ Eq. (8)
  update ψ on $\mathcal{L}_{\text{fake}}$ ▷ Eq. (10)
  if $n \bmod R = 0$ **then**
    recompute f_θ^V from $(z_k^V, z_{k,0}^A)$ with gradients
    update θ on $\lambda_V \mathcal{L}_V$ ▷ Eq. (9)
    $z \leftarrow z_{k,0}^A$ ▷ audio node 0
    for $j = 0, \ldots, M-1$ **do**
      update θ on $(\lambda_A / M)\, \mathcal{L}_{A,k,j}$ ▷ Eqs. (3) and (7)
      if $j < M-1$ **then** ▷ detached
        if $j = 0$ **then**
          $z \leftarrow z_{\text{out}}^A$
        else
          $z \leftarrow z + (t_{k,j+1}^A - t_{k,j}^A)\, u$
        end if
        $u \leftarrow F_{\theta,A}^{\text{cache}}(z, t_{k,j+1}^A; t_k^V, C_k^V, c)$
      end if
    end for
    $\psi \leftarrow \beta\psi + (1-\beta)\theta$
  end if
end for

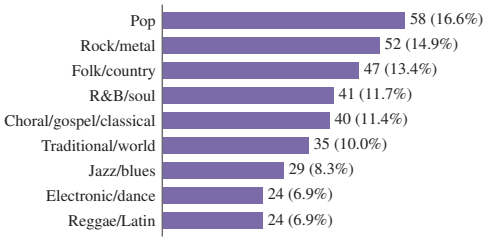

Figure 6: Sing350 prompts per music-style category (count and share).

Singing benchmark Sing350. Evaluating audiovisual singing requires prompts that jointly specify a visual performance, musical characteristics, and the lyrics to be sung. Existing benchmarks focused on scene and environmental-sound generation or lyric-conditioned music generation do not directly provide this combination: scene descriptions need not specify singable lyrics, while lyric-conditioned music prompts need not describe a visible singing performance. We therefore construct a unified benchmark of 350 prompts that combines visual performance descriptions, musical specifications, and explicit English lyrics for joint video and audio generation of singing with instrumental accompaniment. The benchmark covers nine broad music-style categories (Fig. 6) and four vocal configurations: lead vocal, lead vocal with backing vocals, group or harmony vocals, and duet. Each prompt draws its musical description from a distinct MusicCaps record (Agostinelli et al., 2023), including caption-supported instrumentation, vocal characteristics, tempo, and mood. English lyrics are taken verbatim from 50 SongBench records (Wu et al., 2026), producing 350 unique excerpts. Each excerpt consists of one or two consecutive lines containing 7-12 words. All prompts specify a five-second performance with melodic singing and instrumental accompaniment already in progress.

Metrics. DeSync is in seconds. AudioBox is $(\text{CE} + \text{CU} - \text{PC} + \text{PQ})/4$, and DNSMOS reports OVRL; both are speech-oriented predictors of local quality. FAD is the finite-sample FAD against the teacher outputs, computed with native CLAP window embeddings and time-pooled MuQ clip embeddings without extra normalization; values are not comparable across embedding spaces. Onset F1 matches onsets with the teacher output of the same prompt at a ± 50 ms tolerance.

Token-normalized function evaluations. Following H3's officially recommended generation configuration, we profile 124-frame videos at 24 frames per second with a resolution of 768×1344

and stereo audio sampled at 32 kHz. Under this configuration, H3's latent representation contains L_V=37,296 video tokens and L_A=414 audio tokens, while the profiled prompt contributes L_T=177 text tokens.

We measure the denoising workload in token-normalized function evaluations (NFE), counting one full joint audio-video forward pass as one. The sampler of Sec. 4.1 runs K joint steps and $K(M-1)$ audio-only steps, and each audio-only step reuses the visual keys and values and updates only the audio and text tokens, so

$$\mathrm{NFE} = K + K(M-1)\,\frac{L_A + L_T}{L_V + L_A + L_T}. \tag{11}$$

For K=4 and M=5,

$$\mathrm{NFE} = 4 + 16 \times \frac{414 + 177}{37{,}296 + 414 + 177} \approx 4.25. \tag{12}$$

Under this definition, the teacher uses 50 NFE and the four-step baselines use 4. MAAD adds approximately 6.24% token-normalized computation relative to four joint steps for this input. This metric approximates the dominant Transformer workload; realized latency also depends on cache transfers, communication, numerical precision, sparsity, and kernel efficiency.

SCTA diagnostics. The diagnostics of Sec. 5.3 (Fig. 7) replay the rollouts of 12 Sing350 prompts with seed 0 and evaluate, at each of the 20 audio nodes, the student and the teacher on the same audio state. The teacher's audio velocity is computed under the student's cached keys and values, which is the SCTA target, and under fresh joint passes on the student's stage-entry video, on the teacher's own stage-entry video, and on the stage-entry video of another prompt, all at the stage-entry time t_k^V. Prompts are the unit of pairing and resampling, with 20,000 bootstrap resamples; brackets give 95% intervals. At stages 1-3, the teacher-video target differs from the SCTA target by 14.0% [12.4, 15.6], $1.81\times$ [1.60, 2.08] the student's distance of 7.7% [7.1, 8.4], and exceeds it in 12 of 12 prompts and 98% of nodes. Per stage, these distances are 22.1%, 13.2%, and 6.7%, against student distances of 11.5%, 7.6%, and 4.1%. The fresh joint pass on the student's video differs from the SCTA target by 8.4% [7.7, 9.0], $1.09\times$ [1.01, 1.17] the student's distance (9 of 12 prompts), and by 9.8% at the first node of stage 0, where the video, its time, and the audio coincide. This difference combines the network that computed the cached keys and values, the audio state present when they were computed, and the cached computation path; separating these requires a cache built by the teacher itself, so we do not attribute it to any one of them. With the path, time, and audio fixed, replacing the student's stage-entry video by the teacher's changes the teacher's audio velocity by 12.4% [10.8, 14.2] (18.8%, 12.7%, and 5.8% at stages 1-3), 0.71 [0.62, 0.80] times the change caused by the video of an unrelated prompt. Because the student is trained on the SCTA target, its smaller error to this target (7.9% against 10.8% for the fresh joint pass, over all stages) is expected and is not evidence for SCTA. Excluding prompt 11, whose teacher replay did not reproduce the stored media, and, for the unrelated-video comparison, prompt 2, whose unrelated video comes from prompt 11, changes these values by at most 0.9 percentage points.

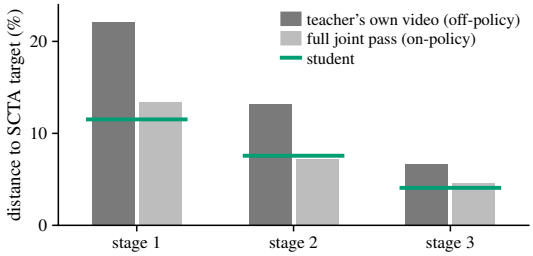

Figure 7: Distance of alternative audio targets from the SCTA target in relative ℓ_2 (12 Sing350 prompts, prompt means). The green line is the student's distance to the SCTA target.

Audio objective and video motion. Table 10 lists Dynamic Degree for the four objective assignments of Sec. 5.4. With video on DM, switching the audio objective from DM to SCTA raises it

from 8.9% to 51.1%; with video on Traj., it stays below 5% for either audio objective. Two routes connect the audio objective to the video, and we do not separate them. In the single-stream model, video tokens attend to audio tokens, and the refined audio state enters the next joint step (Sec. 4.1), so the audio trajectory steers the denoising of the video. In training, the stop-gradient in Eq. (7) blocks only the path through the cached video tokens, while the audio losses still update the adapter weights that the video tokens share.

Full results. Tables 4 and 5 list the component ablation of Sec. 5.3. Table 6 lists all VABench metrics. Tables 7-10 list all results of the objective-assignment comparison in Sec. 5.4. Fig. 8 plots the main metrics of this comparison against the video objective.

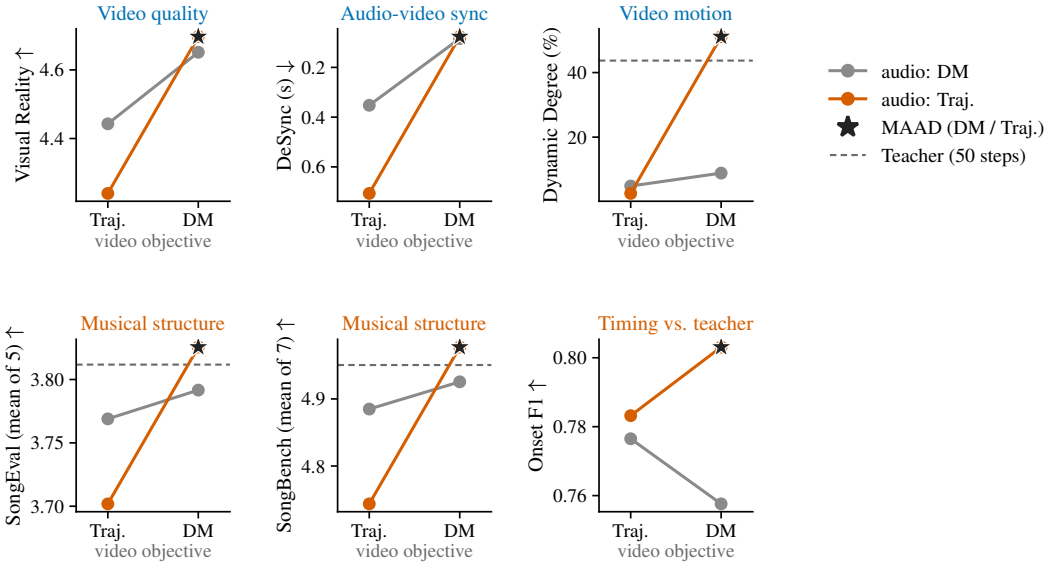

Figure 8: Objective assignment on Sing350. Each panel plots one metric against the video objective; line color gives the audio objective.

Table 4: Component ablation on Sing350 (Fig. 4, left). Async.: audio-only steps; Traj.: audio trajectory alignment; Cache: the teacher reads the student's visual cache (SCTA) instead of running a full joint pass. SongBench and SongEval are means over their dimensions.

Configuration	Async.	Traj.	Cache	N_V/N_A	NFE	SongBench↑	SongEval↑	DeSync↓
DM only	×	×	–	4/4	4.00	4.784	3.713	**0.045**
async only	✓	×	–	4/20	4.25	4.926	3.792	0.084
Traj. only	×	✓	×	4/4	4.00	4.320	3.586	0.131
w/o cache	✓	✓	×	4/20	4.25	4.692	3.541	0.107
MAAD	✓	✓	✓	4/20	4.25	**4.977**	**3.826**	0.076

Table 5: Audio updates on Sing350 (Fig. 4, right). N_V/N_A: video and audio updates, with a separately trained student for each N_A. NFE: token-normalized function evaluations; fidelity is measured against the teacher.

	SongBench			SongEval			Fidelity to teacher			
N_V/N_A	Vocal↑	Melody↑	Instrumental↑	Musicality↑	Coherence↑	Clarity↑	MuQ FAD↓	CLAP FAD↓	Onset F1↑	NFE
4/4	4.0699	4.3148	4.4011	3.6289	3.6859	3.5149	22.2141	0.1920	0.7869	4
4/8	4.8498	4.6305	4.9422	3.7669	3.7342	3.6440	10.0641	0.0551	0.7830	4.06
4/12	4.9530	4.6250	4.9172	3.8533	3.8204	3.7462	6.0801	0.0286	0.7911	4.12
4/20 (MAAD)	5.0907	4.7688	5.0940	3.9090	3.8755	3.7912	4.5799	0.0187	0.8031	4.25
4/36	5.1326	4.8125	5.1752	3.8955	3.8477	3.7897	4.1263	0.0115	0.8031	4.50

Table 6: VABench T2AV results on 778 prompts with seed 0. All scores are reported on their original scales. ↑/↓ denotes higher/lower is better. Among the four accelerated models, best results are bold and second-best results are underlined, including ties. The teacher is a reference and is excluded from ranking. LSA is not ranked (see below).

Model	DNSMOS ↑	NISQA ↑	AudioBox ↑	ViCLIP ↑	CLAP ↑	DeSync ↓	ImageBind ↑	LSA† ↑
Teacher (50 steps)	1.4542	1.5614	3.5858	0.2171	0.3518	0.5689	0.2561	1.4283
Larry Turbo	1.3458	**1.5868**	3.4440	0.2189	0.3119	0.5496	**0.2490**	1.8148
Sol-H3	1.3930	1.5810	3.3282	0.1992	0.2962	0.6012	0.2032	1.3159
LightX2V Turbo	**1.4712**	1.5467	3.4327	0.2139	0.3184	**0.4685**	0.2342	1.9003
MAAD	1.4181	1.4991	**3.5438**	**0.2232**	**0.3612**	0.5006	0.2382	1.6788

Model	Alignment ↑	Audio real. ↑	Visual real. ↑	Express. ↑	Artistry ↑	Audio QA ↑	Visual QA ↑
Teacher (50 steps)	4.4781	4.0373	4.5283	4.2751	3.6799	0.6973	0.6913
Larry Turbo	4.4306	**4.0103**	**4.4949**	4.2943	**3.6941**	**0.7668**	0.7156
Sol-H3	4.4036	3.9357	4.4036	4.2416	3.6645	0.6722	0.6474
LightX2V Turbo	4.3997	3.9267	4.3856	**4.3226**	**3.6941**	0.7263	0.7138
MAAD	**4.4781**	3.9602	4.4756	4.2931	3.6902	0.7441	**0.7290**

†LSA is averaged over each model's valid subset: teacher: 192/778, Larry Turbo: 216/778, Sol-H3: 156/778, LightX2V Turbo: 286/778, MAAD: 183/778. These subsets differ across models, so LSA scores are not ranked. All other metrics cover all 778 samples.

Table 7: SongBench results for different video/audio supervision assignments.

	Supervision		SongBench						
Variant	Video	Audio	Vocal ↑	Instrumental ↑	Melody ↑	Structure ↑	Arrangement ↑	Mixing ↑	Musicality ↑
Teacher (50 steps)	–	–	5.1493	5.1167	4.7004	4.5330	4.6828	5.2157	5.2559
DM / DM	DM	DM	5.1263	5.1321	4.6869	4.4569	4.6739	5.1548	5.2477
MAAD	DM	Traj.	5.0907	5.0940	**4.7688**	**4.6235**	**4.7342**	**5.2143**	**5.3158**
Traj. / DM	Traj.	DM	**5.1616**	**5.1538**	4.6078	4.3455	4.6725	5.0758	5.1774
Traj. / Traj.	Traj.	Traj.	4.8903	4.8795	4.4994	4.3742	4.5211	4.9855	5.0547

Means over Sing350 with seed 0. Bold denotes the best mean among the four supervision variants; the teacher is a reference and is excluded from ranking.

Table 8: SongEval results for different video/audio supervision assignments.

	Supervision		SongEval				
Variant	Video	Audio	Coherence ↑	Musicality ↑	Memorability ↑	Clarity ↑	Naturalness ↑
Teacher (50 steps)	–	–	3.8508	3.8908	3.9563	3.7909	3.5699
DM / DM	DM	DM	3.8386	3.8602	3.9233	3.7713	3.5647
MAAD	DM	Traj.	**3.8755**	**3.9090**	**3.9617**	**3.7912**	**3.5905**
Traj. / DM	Traj.	DM	3.7991	3.8526	3.9140	3.7550	3.5240
Traj. / Traj.	Traj.	Traj.	3.7511	3.7795	3.8434	3.6814	3.4539

Means over Sing350 with seed 0. Bold denotes the best mean among the four supervision variants; the teacher is a reference and is excluded from ranking.

Table 9: Paired audio similarity and distribution distance to the teacher.

	Supervision		Paired agreement			Distribution distance	
Variant	Video	Audio	CLAP cosine ↑	MuQ cosine ↑	Onset F1 ↑	CLAP FAD ↓	MuQ FAD ↓
Teacher (50 steps)	–	–	1.0000	1.0000	1.0000	0.0000	0.0000
DM / DM	DM	DM	0.8870	0.8940	0.7576	0.0404	6.6519
MAAD	DM	Traj.	**0.9182**	**0.9149**	**0.8031**	**0.0187**	**4.5799**
Traj. / DM	Traj.	DM	0.8899	0.8987	0.7765	0.0436	6.2400
Traj. / Traj.	Traj.	Traj.	0.8939	0.8975	0.7832	0.0269	5.3721

The cosine similarities and Onset F1 are averaged over the 350 paired Sing350 clips (seed 0); FAD compares each 350-clip distribution with the teacher. Bold denotes the best result among the four variants; the teacher is excluded from ranking. CLAP and MuQ cosines compare audio with audio. Onset F1 uses a ± 50 ms tolerance. Finite-sample FAD uses native CLAP window embeddings and time-pooled MuQ clip embeddings, without extra L2 normalization. These metrics measure agreement with the teacher, not perceptual quality.

Table 10: Dynamic Degree on Sing350. Each model is evaluated with seed 0. Δ denotes the difference from the teacher in percentage points.

Model	Video sup.	Audio sup.	Dynamic / Total	DD (%)	Δ (pp)
Teacher (50 steps)	–	–	153/350	43.71	–
DM / DM	DM	DM	31/350	8.86	-34.86
MAAD	DM	Traj.	**179/350**	**51.14**	$\mathbf{+7.43}$
Traj. / DM	Traj.	DM	17/350	4.86	-38.86
Traj. / Traj.	Traj.	Traj.	9/350	2.57	-41.14