MODALITY-DECOUPLED DISTRIBUTION MATCHING DISTILLATION FOR FEW-STEP AUDIO-VISUAL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint audio-visual diffusion transformers produce a video and its soundtrack from one denoising network, but iterative sampling makes them costly to serve. Distribution matching distillation (DMD) is a natural route to a few-step joint student, and its direct extension treats the two modalities as one tensor: they are renoised on a shared schedule, and one fake-score objective is fitted to both generated marginals at once. We show that both couplings limit the student. A shared diffusion time does not place video and audio at comparable denoising difficulty, and a summed fake-score loss does not constrain the per-modality estimates that the generator update actually consumes. Decoupling them is not simply a matter of giving each modality its own noise level: a jointly pretrained audio-visual model has only been trained on synchronized noise levels, so a mixed-noise pair evaluates it outside its training distribution. We introduce AV-DMD, which decouples the two DMD estimates while keeping every network evaluation synchronized. From one timestep and one Gaussian draw it builds two internally synchronized joint states, reads the video target from the first and the audio target from the second, and supervises the shared fake score on those same two states with branch-specific losses. On a four-step LTX-2.3 audio-visual student this raises the normalized 16-dimension VBench aggregate from 0.642 to 0.674 and the JavisBench-mini CLAP score from 0.372 to 0.416, with the gain concentrated in semantic alignment and a small decrease in the VBench quality aggregate.

1 INTRODUCTION

A joint audio-visual diffusion transformer synthesizes a video and its soundtrack with a single denoising network, which lets the two modalities attend to each other at every layer instead of being aligned after the fact. Latent diffusion transformers now drive most high-resolution video generators (Rombach et al., 2022; Peebles & Xie, 2023; Blattmann et al., 2023a; Yang et al., 2024), and recent systems extend the same joint backbone to sound, replacing cascades that generate audio from a finished video (Iashin & Rahtu, 2021; Luo et al., 2023b; Cheng et al., 2025). The cost is iterative sampling through a large joint model: the published 40-step LTX-2 audio-visual teacher spends roughly twenty times the sampling time of a four-step student of the same backbone (Yang et al., 2026b). Few-step distillation is therefore the operative question for deploying joint audio-visual generators, and the one we study here.

Distribution matching distillation trains a student by comparing a frozen real score against a fake score fitted online to the student's own samples (Yin et al., 2024b;a). Unlike trajectory-based accelerators that compress the sampling chain (Salimans & Ho, 2022; Song et al., 2023), it constrains the generated distribution directly, which has made it the basis for recent few-step video systems (Yin et al., 2025). Extending it to a joint generator admits an implementation so direct that it is rarely questioned: concatenate the video and audio latents, renoise them at one sampled diffusion time, read both modality gradients from that single noisy state, and fit the fake score by summing the two denoising losses before one backward pass. This inherits two couplings from the single-modality formulation. The two modalities share one noise level, and they share one fake-score objective.

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.

Neither coupling survives inspection. Video latents are spatially structured and temporally redundant, whereas audio latents carry dense transients, pitch, and broadband texture at a different token density (Kreuk et al., 2023; Liu et al., 2023a). A noise level that gives video a global correction can erase the temporal detail that the audio branch must recover, so a shared diffusion time does not place the two modalities at comparable denoising difficulty. That a schedule is a design variable in its own right is well established for single-modality models (Karras et al., 2022; Chen, 2023; Esser et al., 2024); what is not established is what to do when one network serves two modalities at once. The shared fake-score objective fails for a separate reason. The generator update consumes a *per-modality* score difference, but the summed loss $\ell^v + \ell^a$ only constrains the total; a small sum is compatible with a poorly fitted audio marginal, and that error passes directly into the generator gradient. Section 4.1 makes both effects measurable without appealing to end-task scores: the normalized denoising error of the two modalities separates under matched noise, and training the two fake-score branches separately reduces the audio fake-loss variance from 0.362 to 0.191 at a comparable video-loss range.

The obstacle is that the obvious repair is unavailable. Assigning video a noise level σ_{V} and audio a lower σ_{A}, then feeding the mixed pair $(\boldsymbol{x}^v_{\sigma_{\mathrm{V}}}, \boldsymbol{x}^a_{\sigma_{\mathrm{A}}})$ to the score networks, keeps a shared time index but evaluates a jointly pretrained model on an input it has never seen: its real score, which DMD relies on as a fixed target, was trained only on synchronized noise levels. Decoupling the noise *inside* a model evaluation therefore trades a schedule mismatch for an out-of-distribution one.

AV-DMD resolves this by decoupling the two DMD *estimates* while leaving every individual model evaluation synchronized. From one sampled timestep and one Gaussian draw we construct two joint states that are each internally synchronized, one at the video noise level and one at the audio level, and we read the video target only from the first and the audio target only from the second. The shared fake score is then supervised on those same two states with branch-specific losses, so it is fitted at exactly the noise scales at which the generator queries it. Reusing one timestep and one noise draw makes the schedule the only stochastic difference between the two states, and no component introduces a second score network, a discriminator, or a reward model.

On a four-step LTX-2.3 audio-visual student distilled from 63k text prompts, this raises the normalized 16-dimension VBench aggregate from 0.642 to 0.674 and JavisBench-mini CLAP from 0.372 to 0.416, with the gain concentrated in semantic alignment ($+0.063$) and a small decrease in the VBench quality aggregate that we report in full.

Our contributions are as follows:

- We show that the two couplings inherited by joint audio-visual DMD are distinct failure modes rather than one, and we measure each directly: a shared noise level does not equalize denoising difficulty across modalities, and a summed fake-score loss leaves the per-modality estimates that the generator consumes unconstrained.
- We identify why naive noise decoupling cannot be used to repair either one, and give a construction that avoids it: two internally synchronized joint states built from one timestep and one noise draw, from which the two DMD targets and the two fake-score losses are read. The pretrained joint model is never evaluated off its training distribution, and the fake score remains a single set of parameters.
- We evaluate on the full 16-dimension VBench suite (Huang et al., 2024), the 1,000-prompt JavisBench-mini split with the official CLAP (Wu et al., 2023) and Audiobox Aesthetics (Tjandra et al., 2025) protocols, all 470 AVBench prompts (Yang et al., 2026a), and a 100-clip text-conditioned digital-human protocol, and report dimension-level results including the cases that regress.

2 Related Work

2.1 Few-Step Distillation of Diffusion Models

Diffusion models generate by reversing a noising process (Ho et al., 2020; Song et al., 2021b), and their cost is the length of that reverse chain. *Trajectory* methods keep the teacher's path and traverse it in fewer steps, through larger deterministic steps (Song et al., 2021a), repeated halving of the step count (Salimans & Ho, 2022), or networks mapping any trajectory point to its endpoint (Song et al., 2023; Luo et al., 2023a; Kim et al., 2024). *Distribution* methods discard the path: score distillation

backpropagates a frozen score into a generator (Poole et al., 2023), variational score distillation corrects its mode-seeking bias with a second score fitted to the generator's own outputs (Wang et al., 2023), and DMD makes that difference the generator objective (Yin et al., 2024b), with DMD2 adding online fake-score tracking and multi-step backward simulation (Yin et al., 2024a). Variants alter the two scores (Zhou et al., 2024), add an adversarial term (Sauer et al., 2023), or move to autoregressive video (Yin et al., 2025). We build on DMD2 and change neither score nor objective, but how both are *evaluated*.

Two design variables bear on our method. Noise schedules determine where a model is supervised (Karras et al., 2022) and must be retuned with resolution (Chen, 2023; Hoogeboom et al., 2023); flow-matching backbones (Lipman et al., 2023; Liu et al., 2023b) express this as a shift set by token count (Esser et al., 2024). That work varies the schedule across *samples*; we vary it across *modalities within a sample*, which Section 4.2 shows a jointly pretrained score model constrains. For two simultaneous objectives, standard tools reweight by gradient norm (Chen et al., 2018) or project away conflicts (Yu et al., 2020). Our method instead changes the joint states used for fake-score estimation and modality-specific supervision.

2.2 Joint Audio-Visual Generation

Video generators progressed from pixel-space diffusion (Ho et al., 2022) through latent diffusion (Rombach et al., 2022; Blattmann et al., 2023b) to large transformer backbones (Peebles & Xie, 2023; Blattmann et al., 2023a; Yang et al., 2024), with audio generators following a parallel path for speech, sound, and music (Liu et al., 2023a; Kreuk et al., 2023; Copet et al., 2023). Combining them admits three arrangements. *Cascaded* systems generate video first and add conditioned sound (Iashin & Rahtu, 2021; Luo et al., 2023b; Cheng et al., 2025), so audio cannot influence the image. *Composed* systems link separately trained experts through shared latents (Tang et al., 2023; Ruan et al., 2023). *Joint* systems, our target, denoise both streams in one network, and work here adds architectural synchronization structure: spatio-temporal priors (Liu et al., 2025), modality-specific experts with aligned position encoding (Liu et al., 2026), or causal streaming that exposes temporal and token-density asymmetry between modalities (Su et al., 2026). LTX-Video supplies the backbone we distill (HaCohen et al., 2024), and TurboT2VA distills it to four steps by combining consistency learning with distribution matching (Yang et al., 2026b). These change the backbone or the objective; our changes leave both intact.

Audio-driven human video is a separate line, scoring lip sync with a pretrained expert (Chung & Zisserman, 2016; Prajwal et al., 2020) and synthesizing portraits with diffusion (Tian et al., 2024; Ji et al., 2024). It conditions on a reference identity and real audio, whereas our generator receives only text; we borrow its measurements as a synchronization probe, not as a comparison.

2.3 Evaluating Audio-Visual Generation

No single score covers joint generation. Visual quality and semantics use VBench (Huang et al., 2024); text-audio alignment uses CLAP (Wu et al., 2023) and perceptual audio quality the four Audiobox Aesthetics axes (Tjandra et al., 2025). Cross-modal consistency uses AVBench (Yang et al., 2026a) and JavisBench (Liu et al., 2025), whose paired similarities live in a shared multimodal embedding space (Girdhar et al., 2023). The digital-human protocol reports distributional distances (Heusel et al., 2017; Unterthiner et al., 2018) with SyncNet scores (Chung & Zisserman, 2016) over S3FD tracks (Zhang et al., 2017) on HDTF clips (Zhang et al., 2021). We keep these families separate, because a higher semantic score need not imply cleaner audio.

3 Background

Notation. An italic superscript $m \in \{v, a\}$ indexes a modality, video or audio; an upright subscript V or A names a *noise schedule* rather than a modality. A clean joint latent is $\boldsymbol{z}_0 = (\boldsymbol{x}_0^v, \boldsymbol{x}_0^a)$, a noisy joint state is $\boldsymbol{z}_t$, and $\epsilon = (\epsilon^v, \epsilon^a)$ is a Gaussian draw with one tensor per modality. A noisy state interpolates as $(1-\sigma)\boldsymbol{x}_0 + \sigma\epsilon$, and we write $\alpha = 1 - \sigma$. Let $G_\theta(\epsilon, c)$ be a few-step audio-visual generator conditioned on text c that produces $\boldsymbol{z}_0$.

Distribution matching distillation. DMD fits a fake score s_ϕ to noisy samples from the current generator and uses a frozen real score s_{real} from the pretrained model. Both are joint networks: one evaluation on a joint state returns both modality components, $s_\phi(\boldsymbol{z}_t, t, c) = (s^v_\phi, s^a_\phi)$. At diffusion time t the generator update is induced by the modality-wise score difference

$$\boldsymbol{g}^m(\boldsymbol{z}_t, t) = w^m_t \left(s^m_\phi(\boldsymbol{z}_t, t, c) - s^m_{\text{real}}(\boldsymbol{z}_t, t, c) \right), \qquad m \in \{v, a\}, \tag{1}$$

with w^m_t the usual DMD normalization. Writing $\boldsymbol{g}^m$ as a function of the joint state records that the modality-m target depends on how *both* modalities were noised. The fake score is trained online by denoising generated samples, and we write

$$\ell^m(\boldsymbol{z}_t, t) = \mathcal{D}^m(s_\phi(\boldsymbol{z}_t, t, c), \boldsymbol{x}^m_0, \epsilon^m) \tag{2}$$

for the modality-m branch of the pretrained model's native denoising objective, evaluated from one joint forward pass at $\boldsymbol{z}_t$. The direct extension of DMD to a joint generator builds one $\boldsymbol{z}_t$ in which both modalities carry the same noise level, reads $\boldsymbol{g}^v$ and $\boldsymbol{g}^a$ from that single state, and minimizes $\ell^v + \ell^a$ in one backward pass; Section 4 replaces both decisions.

Multi-step distillation. We use backward simulation as in DMD2: the student constructs its inference-time trajectory and one trajectory state is selected for the DMD update (Yin et al., 2024a). Our four transitions use the unshifted time points $[1.0, 0.75, 0.5, 0.25, 0]$, with the zero-noise endpoint excluded from generator-step sampling.

4 Modality-Decoupled Audio-Visual DMD

Section 4.1 measures the two couplings that the joint baseline of Section 3 inherits. Section 4.2 states the condition that rules out the direct repair, and Sections 4.3 and 4.4 give a construction that satisfies it. Figure 2 summarizes the generator and fake-score updates.

4.1 Two Measured Couplings

A shared time is not a shared difficulty. Video latents are spatially structured and temporally redundant, whereas audio latents carry dense temporal variation. Under the joint baseline both are renoised to the same level at a sampled t, which fixes the noise scale but not the optimization problem it induces. Renoising paired samples to matched levels and measuring each modality's normalized denoising error yields two separated curves, with the gap widest at high noise (Figure 1(a)). A scale that gives video a global correction removes temporal detail that audio must still recover.

A summed loss does not constrain both branches. The generator update reads $\boldsymbol{g}^v$ and $\boldsymbol{g}^a$ separately (Eq. 1), so the fake score has to be accurate on each generated marginal; minimizing $\ell^v + \ell^a$ constrains only their sum. Figure 1(b) shows the consequence: the per-sample audio errors disperse over an order of magnitude more than the video errors, so the aggregate can be small while one branch is loosely fitted. The audio fake-loss variance across training reflects the same asymmetry, 0.362 against 0.191 once the objectives are separated (Figure 1(c)).

4.2 The Synchronization Constraint

Suppose we accept the diagnosis and give each modality its own noise level at the same time index, $\sigma_{\text{V}}(t)$ for video and $\sigma_{\text{A}}(t)$ for audio. The direct implementation forms the mixed-noise state

$$\boldsymbol{z}^{\text{mix}}_t = (\alpha_{\text{V}} \boldsymbol{x}^v_0 + \sigma_{\text{V}} \epsilon^v,\ \alpha_{\text{A}} \boldsymbol{x}^a_0 + \sigma_{\text{A}} \epsilon^a) \tag{3}$$

and evaluates the joint networks on it. This preserves a shared time index but violates the condition under which those networks were trained. Joint audio-visual pretraining draws one noise level per sample and applies it to both modalities, so every state the pretrained model has seen satisfies $\sigma^v = \sigma^a$; the frozen real score has this property by construction and the fake score inherits it from initialization. Evaluating either on $\boldsymbol{z}^{\text{mix}}_t$ queries the model off its training distribution, and the resulting score difference is then unreliable for *either* modality: a schedule mismatch has been traded for an out-of-distribution one.

We therefore impose the following condition on any modality-decoupled update.

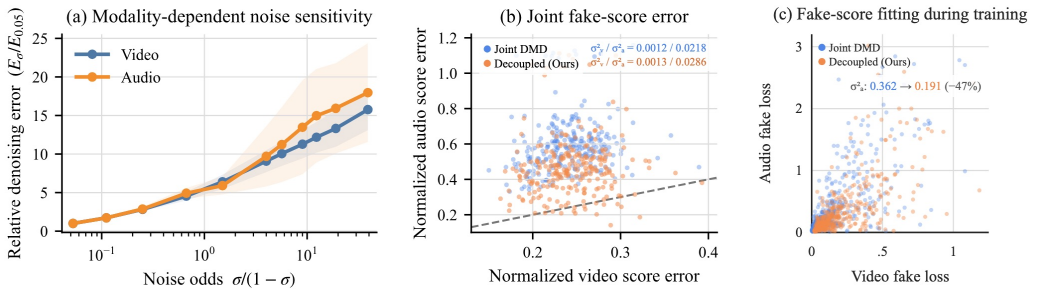


Figure 1: **Audio and video induce different DMD optimization dynamics.** (a) Normalized denoising error of each modality under matched noise levels. (b) Per-sample video and audio fake-score errors under joint DMD and under decoupled training; each point is one sample. (c) Video and audio fake losses sampled during training; each point is one training step. Variances are quoted in the text.

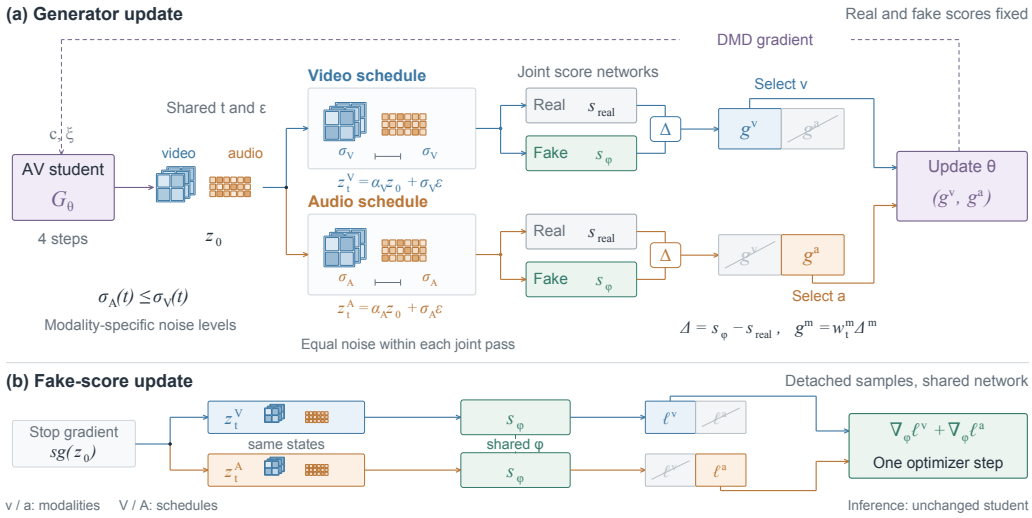


Figure 2: Overview of AV-DMD. (a) From one generated audio-video pair, a shared timestep and Gaussian draw produce two internally synchronized states. The joint real and fake scores are evaluated on both states; only the video component from schedule V and the audio component from schedule A contribute to the generator update. (b) Detached samples use the same two-state construction to train the shared fake score with modality-specific losses. Gradients are accumulated into one optimizer step. Crossed-out components are discarded; inference uses the unchanged four-step student.

> *Synchronization constraint.* Every evaluation of s_ϕ or s_{real} is made on a joint state whose two modalities carry the same noise level.

This is a design requirement imposed by the pretrained model, not an experimental finding: we do not evaluate Eq. 3 and do not quantify how much it degrades the target. The constraint forbids modality-specific noise levels *within one forward pass*, not across passes. Two consequences follow: the two DMD targets must be read at two different noise levels and no single admissible state carries both, so a decoupled update needs two evaluations per score network; and since both evaluations still contain both modalities, audio-video interaction is preserved in each.

4.3 MODALITY-SPECIFIC RENOISING

We sample one time index t and one Gaussian pair $\epsilon = (\epsilon^v, \epsilon^a)$. Two schedules map t to two noise levels. Schedule V is the backbone's own schedule. Schedule A differs from it in a single scalar: the maximum shift of the backbone's logit-space time shift is lowered from its default 2.05 to 1.2, with the base shift, terminal value, and terminal rescaling left unchanged. Because that shift is monotone in its maximum, schedule A assigns audio a noise level no higher than video's at the same t, the two coinciding at the terminal point. Shifted schedules are standard for high-resolution flow models, where the shift is set by token count (Esser et al., 2024); here it is set per modality instead. From these schedules we form two admissible joint states,

$$\boldsymbol{z}_t^{\mathrm{V}} = (\alpha_{\mathrm{V}}\boldsymbol{x}_0^v + \sigma_{\mathrm{V}}\epsilon^v,\ \alpha_{\mathrm{V}}\boldsymbol{x}_0^a + \sigma_{\mathrm{V}}\epsilon^a)\,, \tag{4}$$

$$\boldsymbol{z}_t^{\mathrm{A}} = (\alpha_{\mathrm{A}}\boldsymbol{x}_0^v + \sigma_{\mathrm{A}}\epsilon^v,\ \alpha_{\mathrm{A}}\boldsymbol{x}_0^a + \sigma_{\mathrm{A}}\epsilon^a)\,, \tag{5}$$

each of which satisfies the synchronization constraint: within $\boldsymbol{z}_t^{\mathrm{V}}$ both modalities carry σ_{V}, and within $\boldsymbol{z}_t^{\mathrm{A}}$ both carry σ_{A}. We then read each DMD target from the state built at its own noise level,

$$\boldsymbol{g}_t = \big(\boldsymbol{g}^v\big(\boldsymbol{z}_t^{\mathrm{V}}, t\big)\,,\ \boldsymbol{g}^a\big(\boldsymbol{z}_t^{\mathrm{A}}, t\big)\big)\,, \tag{6}$$

and discard the off-schedule component of each pass. Reusing ϵ across the two states makes the schedule the only stochastic difference between them, so the two targets differ by noise level and not by noise realization.

4.4 MODALITY-DECOUPLED FAKE-SCORE TRAINING

Eq. 6 reads $\boldsymbol{g}^v$ at σ_{V} and $\boldsymbol{g}^a$ at σ_{A}. The fake score must therefore be accurate on the video marginal at the first level and on the audio marginal at the second, and a single joint loss evaluated at one level supplies neither guarantee. We reuse one generated sample and the same ϵ, and train the fake score on the same two states with one branch active in each:

$$\mathcal{L}_{\mathrm{fake}}^v = \ell^v\big(\boldsymbol{z}_t^{\mathrm{V}}, t\big)\,, \tag{7}$$

$$\mathcal{L}_{\mathrm{fake}}^a = \ell^a\big(\boldsymbol{z}_t^{\mathrm{A}}, t\big)\,. \tag{8}$$

Each pass contributes loss only for the modality whose target is read from that state, which aligns the fake score's training distribution with the states the generator queries. The two gradients are accumulated and applied in one fake-score optimizer step. The branches share one network: Modality-Decoupled Fake-Score Training decouples the training states and the losses, not the parameters.

Appendix C tests whether this produces the intended effect, by measuring the trained fake score's error against each generated marginal across noise levels: video accuracy is essentially unchanged, while audio accuracy improves at every level, with the largest margin in the mid-noise range where the four student transitions fall.

4.5 TRAINING OBJECTIVE AND COST

The generator loss keeps equal modality weights,

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{DMD}}^v + \mathcal{L}_{\mathrm{DMD}}^a. \tag{9}$$

Satisfying the synchronization constraint costs one additional forward pass per score network per update, since Eqs. 4 and 5 must both be evaluated. Parameter count, pretrained weights, and the inference-time student are unchanged, so the cost is confined to distillation. To isolate the effect of modality-aware score estimation, the main text-to-audio-video experiments add no GAN, reward, or loudness objective, no RMS normalization, and no memory or image-conditioned branch.

5 EXPERIMENTS

5.1 SETUP

We evaluate a four-step LTX-2.3 student on JavisBench, VBench and text-conditioned digital-human generation. Baseline and Ours use matched training and evaluation settings. Main results use the full benchmark splits; audio ablations use a common subset. Implementation details, evaluation protocols and extended analyses are provided in Appendices A–B.

5.2 JAVISBENCH RESULTS

Table 1: JavisBench comparison. Higher is better except FVD, FAD and DeSync. Results from different protocols are not directly comparable.

Method	Protocol	CLAP	AQ	CE	CU	PC	PQ	AVA	AV-IB	Javis	DeSync	FVD	FAD
MM-Diffusion	Full-10K	0.079	–	–	–	–	–	–	0.119	0.070	0.875	2311.9	27.5
JavisDiT	Full-10K	0.391	–	–	–	–	–	–	0.197	0.154	1.039	204.1	7.2
UniVerse-1	Full-10K	0.245	–	–	–	–	–	–	0.104	0.077	0.929	194.2	8.7
JavisDiT++	Full-10K	**0.424**	–	–	–	–	–	–	**0.198**	**0.159**	**0.832**	**141.5**	**5.5**
LTX-2 teacher	200	0.340	**4.902**	–	–	–	–	–	0.199	0.165	0.617	–	–
TurboT2VA	200	**0.347**	4.486	–	–	–	–	–	**0.229**	**0.196**	**0.388**	–	–
Joint DMD	Mini-1K	0.372	5.585	5.207	6.789	**3.363**	6.981	0.2420	–	–	–	–	–
Ours	Mini-1K	**0.416**	**5.686**	**5.387**	**6.948**	3.354	**7.055**	**0.2431**	–	–	–	–	–
OmniForcing (reported)	Published	0.401	–	–	–	–	–	–	–	–	–	–	–
OmniForcing V1 repro.	Internal-64	0.2717	–	–	–	–	–	–	–	–	–	–	–
Ours + OmniForcing	Internal-64	**0.3223**	–	–	–	–	–	–	–	–	–	–	–

Full-10K: 10,140 prompts (Liu et al., 2026); 200: TurboT2VA protocol (Yang et al., 2026b); Mini-1K: 1,000 prompts. OmniForcing's reported CLAP is from Su et al. (2026); Internal-64 compares the strict V1 reproduction and our method at step 5600 on the same 64 prompts. AQ: AudioBox quality; CE/CU/PC/PQ: content enjoyment/usefulness and production complexity/quality; AVA: AV-Align. –: unavailable.

Ours improves text–audio alignment and perceptual audio quality over the matched baseline, with an 11.9% gain in CLAP (Table 1). AudioBox gains come from content enjoyment, usefulness and production quality rather than greater complexity. The much smaller change in AV-Align suggests that improved semantic alignment should not be conflated with better event timing. Published results provide context, but their different evaluation protocols prevent a direct cross-paper ranking.

5.3 VBENCH RESULTS

Table 2: Representative VBench dimensions. Scores are multiplied by 100; higher is better. Bold marks the better value within each matched pair.

Method	Aesthetic	Imaging	Subject cons.	Motion smooth.	Temp. flicker
LTX-2 teacher†	56.9	57.4	94.5	99.3	98.8
OmniForcing†	**59.5**	**59.4**	**95.5**	**99.5**	**98.9**
Joint DMD	61.5	64.0	93.0	**99.2**	98.8
Ours	**62.4**	**70.4**	**95.2**	**99.2**	**99.3**

† Reported by OmniForcing (Su et al., 2026) under its own per-video protocol, not our evaluation split. Complete dimensions, including dynamic degree and color, are in Table 5.

Table 2 compares representative dimensions with published LTX-2 and OmniForcing results. Within our matched pair, the improvement is concentrated in semantic composition: Overall rises by 3.23 points and Semantic by 6.30 points, with the largest gains in object class, multiple objects and spatial relations. Imaging quality also improves, while dynamic degree and color regress (Table 5). This indicates stronger prompt grounding rather than uniform visual dominance. The separate AVBench evaluation shows a similar pattern: AV and VT consistency improve, whereas AT is slightly lower (Appendix B).

5.4 DIGITAL-HUMAN RESULTS

Table 3: Digital-human results on HDTF. Published portrait-driven methods and our text-only models use different conditioning and evaluation splits; their scores do not constitute a task-equivalent ranking.

Method	Conditioning	FID ↓	FVD ↓	Sync-C ↑	Sync-D ↓	Smooth ↑
SadTalker	Image + audio	61.672	397.114	1.755	10.695	0.9961
AniPortrait	Image + audio	36.965	471.452	1.095	12.461	0.9921
V-Express	Image + audio	47.396	758.023	1.256	12.394	0.9968
Hallo	Image + audio	30.176	347.358	4.060	9.551	0.9941
Hallo2	Image + audio	38.673	328.540	4.136	9.465	0.9942
EchoMimic	Image + audio	33.207	384.304	2.514	10.743	0.9934
Sonic	Image + audio	29.104	301.173	4.197	9.371	0.9970
Joint DMD baseline	Text only	82.924	476.616	6.936	8.333	0.9939
Ours	Text only	80.000	478.936	**7.841**	**7.560**	0.9935

Published results follow Sonic (Ji et al., 2024). Our FVD uses 16 frames per clip. Protocol and paired-sample analysis: Appendix B.

Ours improves lip synchronization while increasing face-track success from 81% to 93%. The improvement persists when both methods are evaluated on the same successfully tracked clips, ruling out changed detection coverage as its sole explanation. FID improves, while FVD is nearly unchanged with a small regression. We therefore interpret these results as a synchronization gain, not a general improvement in human-video fidelity. Published portrait-driven methods use different conditioning and are not directly ranked against Ours.

5.5 ABLATION STUDIES

Table 4: Component ablation. VBench uses the full set; audio metrics use a common subset. Higher is better.

Configuration	Overall	Quality	Semantic	CLAP	AQ	CE	CU	PC	PQ
Baseline	0.6417	**0.8419**	0.4861	0.268	4.978	4.267	**6.358**	2.480	**6.807**
DecShift only	0.6283	0.8225	0.4774	0.304	**5.010**	**4.434**	6.328	**2.601**	6.677
Ours (DecShift + DFake)	**0.6740**	0.8345	**0.5491**	**0.306**	4.930	4.264	6.251	2.534	6.671

AQ averages CE/CU/PC/PQ. Bold marks the best value in each column. Full settings: Appendix B.

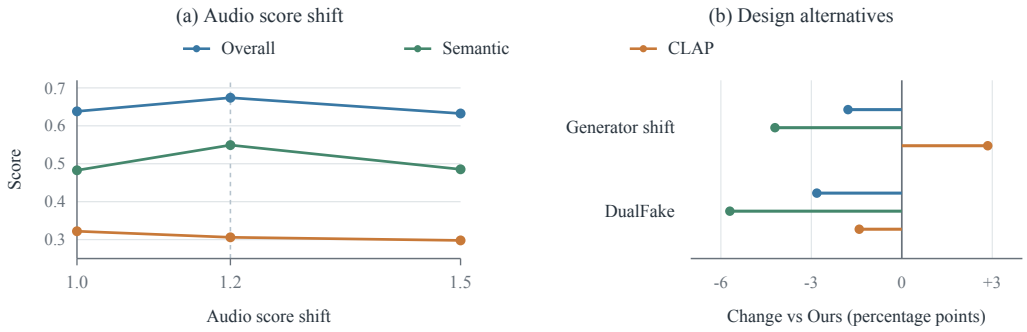


Figure 3: Design analysis. (a) Audio score-shift sweep. (b) Changes relative to Ours for generator shifting and independent DualFake parameters; positive values indicate improvements. CLAP uses the common audio subset.

Audio score shift. In Figure 3(a), shift 1.2 gives the strongest Overall and Semantic scores. Although 1.0 achieves higher CLAP, it sacrifices visual semantics. The selected shift balances the joint task; simply lowering the audio noise level does not improve every metric.

Generator schedule. Changing the generator trajectory is different from changing the noise level at which its correction is estimated. Applying an audio-specific shift to the generator improves CLAP but weakens Overall and Semantic (Figure 3(b)). Keeping the native joint trajectory preserves the stronger semantic result and confines our intervention to score estimation and supervision.

Fake-score parameter sharing. The shared fake score is trained on the same synchronized states used to construct the generator targets. Adding this training scheme recovers the visual-semantic loss of DecShift-only, while independent DualFake parameters perform worse on Overall and CLAP (Figure 3(b)). Some perceptual audio metrics favor DualFake, so the advantage concerns joint semantics rather than every quality axis. These controls support the shared, state-matched design; independent component attribution still requires the standalone DFake control. Further comparisons and the gradient-accumulation implementation are in Appendix B.

5.6 Qualitative Comparison

Figure 4 illustrates the bidirectional and digital-human results using matched video timestamps and aligned waveforms. Across all three rows, prompts and displayed timestamps are fixed. The examples span stylized motion, multi-person performance, object interaction, and speech; they complement the aggregate results by showing improved audio-visual agreement without claiming uniformly better appearance or motion fidelity.

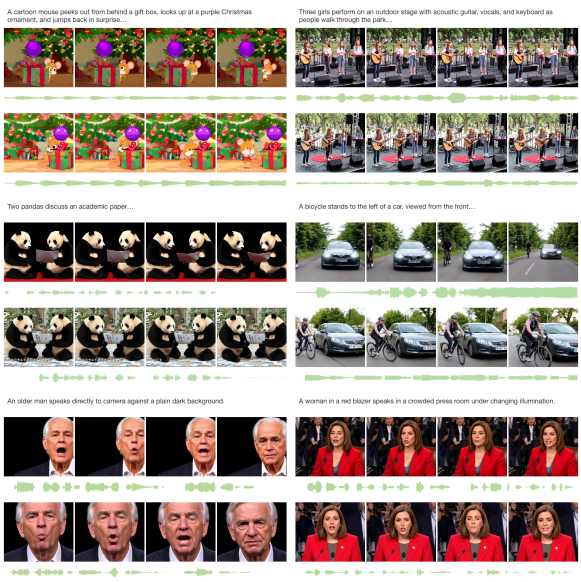


Figure 4: Qualitative audio-visual comparison in three representative rows, with matched prompts, video timestamps, and aligned audio waveforms. Within each case, the baseline is shown above AV-DMD.

6 Conclusion

AV-DMD decouples audio-visual score estimation across synchronized joint states and trains a shared fake score on those same states, without changing student inference. The resulting gains are strongest in semantic alignment and lip synchronization, with mixed changes in motion and perceptual quality. The ablations favor this joint design, while independent component attribution remains open.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023b.

Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. In *CVPR*, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *ACCV Workshops*, 2016.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *NeurIPS*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

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 *CVPR*, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *NeurIPS*, 2022.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *ICML*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024.

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *BMVC*, 2021.

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, Qinglin Lu, and Chengjie Wang. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *ICLR*, 2024.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *ICLR*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *ICML*, 2023a.

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*, 2025.

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*, 2026.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023b.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023a.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

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 *CVPR*, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021b.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023.

Yaofeng Su, Yuming Li, Zeyue Xue, Jie Huang, Siming Fu, Haoran Li, Ying Li, Zezhong Qian, Haoyang Huang, and Nan Duan. Omniforcing: Unleashing real-time joint audio-visual generation. *arXiv preprint arXiv:2603.11647*, 2026.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. In *NeurIPS*, 2023.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. EMO: Emote portrait alive – generating expressive portrait videos with audio2video diffusion model under weak conditions. In *ECCV*, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

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

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *NeurIPS*, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023.

Jialiang Yang, Bin Xia, Ruihang Chu, Dingdong Wang, Wanke Xia, Zhun Mou, Tianyang Zhong, Yiting Zhao, and Wenming Yang. AVBench: Human-aligned and automated evaluation benchmark for audio-video generative models. *arXiv preprint arXiv:2605.24652*, 2026a.

Xiaoda Yang, Yuxiang Liu, Kaiwen Zheng, Yuan Liu, Yibo Lai, Shengpeng Ji, Kai Jiang, Jianfei Chen, 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. *arXiv preprint arXiv:2608.24674*, 2026b.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.

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 *CVPR*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. S^3FD: Single shot scale-invariant face detector. In *ICCV*, 2017.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *CVPR*, 2021.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. In *ICML*, 2024.

AI USE STATEMENT

Generative AI tools were used only for language editing and polishing of the manuscript. They were not used to formulate the research ideas, design or run experiments, analyze data, or determine the scientific claims and conclusions. The authors reviewed and take full responsibility for all content in the paper.

A IMPLEMENTATION DETAILS

Both score networks and the generator are initialized from the same pretrained LTX-2.3 audio-visual checkpoint. The real score is frozen; the generator and the fake score are optimized over all parameters with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.01), learning rate 2×10^{-5}, and global gradient-norm clipping at 3.0. One generator update follows every five fake-score updates. Video and audio carry equal weights in both the generator and the fake-score objective. Classifier-free guidance (Ho & Salimans, 2022) for the frozen real score uses scale 3.0 for video and 5.0 for audio. Training uses seed 42 and all benchmarks use seed 12345. Runs stop at 4000 steps and the matched comparison checkpoint is step 3600.

The audio schedule of Section 4.3 sets the maximum shift of the backbone's time shift to 1.2 in place of its default 2.05; base shift 0.95, terminal value 0.1, and terminal rescaling are unchanged. Both schedules are applied only inside the synchronized score passes of Eqs. 4 and 5; the generator's own sampling trajectory stays on a single joint schedule.

B EXTENDED EXPERIMENTAL ANALYSIS

B.1 SETUP

Bidirectional audio-visual distillation. We distill the LTX-2.3 22B audio-visual checkpoint into a four-step text-to-audio-video generator. Training uses 63k text prompts without paired ground-truth videos or audio. Videos contain 121 frames at 512×768 and 25 fps. The student time points are $[1.0, 0.75, 0.5, 0.25, 0]$ before the native LTX scheduler transform. The real-score guidance scales are 3.0 for video and 5.0 for audio. All three networks start from the same pretrained checkpoint; the real score remains frozen, while the generator and fake score are fully optimized with learning rate 2×10^{-5}. We use one generator update per five fake-score updates, global gradient clipping at 3.0, and train for 4k steps. Unless stated otherwise, results use checkpoint 3600 and evaluation seed 12345.

Controlled comparisons. The matched baseline and our model use the same pretrained initialization, prompt collection, training horizon, student step count and evaluation seed. The main benchmarks and the audio ablation subset answer different questions: the former measure full-set generation performance, while the latter compares design choices on a common development set. We therefore compare CLAP within each set, rather than interpreting its different absolute values across sets as a training improvement. The ablations change the indicated schedule, parameter-sharing or update-order choice without introducing a reward model or an additional generator objective.

Evaluation benchmarks. We report the full 16-dimension VBench suite on 944 prompts. Following the official normalization, we show the aggregate score and its quality and semantic groups. For audio, we generate the official 1,000-prompt JavisBench-mini split (Liu et al., 2025), whose captions describe both visual and acoustic content. We use the official unfused CLAP model with the provided audio captions and report the official AudioBox aggregate together with its four axes. We also report AV-Align, but treat it as a lightweight onset–motion diagnostic rather than a complete synchronization measure. We additionally use the official AVBench evaluators (Yang et al., 2026a) on all 470 prompts (350 normal and 120 hard), reporting audio–video (AV), video–text (VT) and audio–text (AT) consistency. All 470 samples are valid for both runs; full-set means weight the two splits by their sample counts.

Digital-human protocol. We additionally evaluate text-conditioned talking-human generation on a fixed 100-case manifest derived from four-second HDTF clips. Our generator receives only text

captions and does not condition on a reference identity or source frame. We report S3FD face-track success, SyncNet confidence and distance, absolute temporal offset, FID, FVD16, and temporal VBench diagnostics. This matches the set size and clip duration used by Sonic, but its exact random sample list is not public; Table 3 therefore reports a matched protocol rather than the exact Sonic split.

B.2 JAVISBENCH RESULTS

Published few-step and joint-generation references. Table 1 places existing systems and our models in one table under a common metric header, with the protocol recorded for each row. On the complete JavisBench-10K protocol, JavisDiT++ reports the strongest published joint-model results among the listed methods. The published TurboT2VA block provides a few-step reference on its separate 200-prompt protocol. These public values provide context; our controlled Mini-1K comparison is interpreted only against its matched baseline.

The 200-prompt comparison illustrates why both modalities must be inspected when accelerating a joint model: TurboT2VA improves CLAP from 0.340 to 0.347 and AV-IB from 0.199 to 0.229, while reducing DeSync from 0.617 to 0.388. Its reported sampling time falls from 50.52 to 2.51 seconds when moving from the 40-step teacher to the four-step student (Yang et al., 2026b). This is a published within-protocol efficiency reference, not a runtime measurement of our implementation. Our own comparison holds the four-step student fixed and asks whether its distillation targets can be improved. The absence of matched joint-distance and synchronization results for our Mini-1K runs prevents a claim of superiority over these external systems.

Text–audio alignment and perceptual quality. On the full JavisBench-mini split, CLAP improves from 0.372 to 0.416 ($+11.9\%$), while the official AudioBox aggregate rises from 5.585 to 5.686. AudioBox content enjoyment, usefulness, and production quality increase from $5.207/6.789/6.981$ to $5.387/6.948/7.055$, while production complexity changes from 3.363 to 3.354. AV-Align changes from 0.2420 to 0.2431, a small positive difference rather than a decisive synchronization gain.

The largest relative change is in text–audio alignment. The AudioBox aggregate improves by 0.101 (about 1.8%), with gains in content enjoyment ($+0.180$), usefulness ($+0.159$) and production quality ($+0.074$), but essentially no increase in production complexity. Thus the measured audio benefit is more consistent with improved relevance and content than with uniformly richer or more complex sound. AV-Align improves by only about 0.42% relative, so the CLAP gain should not be interpreted as proof of better event-level audio–video timing. Listening tests and the missing joint synchronization metrics remain necessary for that stronger claim.

B.3 VBENCH RESULTS

Table 5: Complete VBench results for the matched joint DMD comparison. Scores are multiplied by 100; higher is better.

Method	Ovl.	Qual.	Sem.	SC	BC	TF	MS	DD	AQ	IQ	OC	MO	HA	Col	SR	Scn	AS	TS	OCons
Baseline	64.17	**84.19**	48.61	93.0	95.1	98.8	**99.2**	**77.8**	61.5	64.0	67.2	47.8	**89.0**	**91.4**	63.6	10.1	19.8	24.2	24.4
Ours	**67.40**	83.45	**54.91**	**95.2**	**95.3**	**99.3**	**99.2**	62.5	**62.4**	**70.4**	**90.9**	**63.6**	**89.0**	89.0	**78.4**	**12.4**	**20.7**	**24.6**	**25.6**

Ovl./Qual./Sem.: Overall/Quality/Semantic. SC/BC: subject/background consistency; TF: temporal flickering; MS: motion smoothness; DD: dynamic degree; AQ/IQ: aesthetic/imaging quality; OC/MO: object class/multiple objects; HA: human action; Col: color; SR: spatial relationship; Scn: scene; AS/TS: appearance/temporal style; OCons: overall consistency.

Semantic composition rather than a uniform quality gain. Table 5 collects the aggregate scores and all VBench dimensions. In the bidirectional setting, AV-DMD improves Overall from 0.642 to 0.674. The gain is concentrated in semantic alignment, which rises by 0.063, with gains in object class ($+0.237$), multiple objects ($+0.158$), spatial relationships ($+0.148$), and imaging quality ($+0.064$). The visual improvements are not uniform: the quality aggregate decreases slightly, chiefly due to dynamic degree and color. The same table includes these regressions alongside the improved dimensions.

On the percentage scale used in the table, Overall rises by 3.23 points and Semantic by 6.30 points, whereas Quality decreases by 0.74 points. Object class, multiple objects and spatial relationships account for the largest displayed improvements (+23.7, +15.8 and +14.8 points). These dimensions test whether the requested entities and relations appear, so they locate the improvement in prompt composition rather than merely smoother frames. Imaging quality also rises by 6.4 points, while subject and background consistency remain high.

Motion and appearance trade-offs. Dynamic degree decreases from 77.8 to 62.5, and color from 91.4 to 89.0. Motion smoothness is unchanged at the displayed precision even as dynamic degree falls: a smooth sequence need not contain enough motion. Consequently, the higher Overall score is not evidence that every temporal aspect improves. The combination of stronger composition and weaker dynamic degree motivates checking both prompt satisfaction and motion in the qualitative examples, rather than selecting cases solely by a single aggregate score. All dimensions remain visible in the same comparison table.

Cross-modal consistency on AVBench. On this separate benchmark, full-set AV consistency increases from 0.8845 to 0.8991 and VT consistency from 0.7832 to 0.7947, while AT changes from 0.8694 to 0.8671. The VT gain is larger on the hard split (0.7491 to 0.7718) than on the normal split (0.7949 to 0.8026), consistent with the stronger visual-semantic composition seen in VBench. AV consistency improves on both splits, from 0.8701 to 0.8875 on normal and 0.9268 to 0.9327 on hard. The slight AT regression is important: the higher CLAP score on JavisBench-mini does not imply an improvement under every audio–text evaluator or prompt distribution. Taken together, the measurements indicate stronger visual grounding and joint consistency, with mixed audio-side changes. These learned consistency scores do not directly measure fine-grained lip timing, which we examine next on HDTF.

B.4 Digital-Human Results

Synchronization under text-only conditioning. Table 3 places published reference-conditioned methods and our text-only model in one table. The task difference prevents a direct ranking against Sonic and related systems. Within our matched pair, AV-DMD raises face-track success from 81% to 93%, increases Sync-C from 6.936 to 7.841, and reduces Sync-D from 8.333 to 7.560. On the stricter paired subset of 80 clips with valid tracks for both methods, Sync-C improves for 59 clips, Sync-D improves for 56, and both improve for 54. FID improves by 3.53%, while FVD16 is nearly unchanged, with a marginal regression, and motion smoothness changes only slightly.

Detection coverage and paired evaluation. The success-rate change means that the unpaired SyncNet means are computed on different valid-track sets, so their difference alone is not a clean paired estimate. Restricting the analysis to the 80 clips tracked for both models retains a mean Sync-C gain of 1.435 and a Sync-D reduction of 1.022. Sync-C improves for 73.75% of these clips, Sync-D for 70%, and both improve for 67.5%. Absolute temporal offset also decreases from 1.778 to 1.387 frames over valid tracks, with a 0.488-frame reduction on the paired set. This shows that the synchronization improvement is not solely an artifact of the changed detection coverage, without claiming that every individual clip improves.

Synchronization is not distributional fidelity. FID decreases from 82.924 to 80.000, while FVD changes only marginally from 476.616 to 478.936 (about 0.49% worse). Motion smoothness remains close (0.993913 versus 0.993522). These metrics probe different properties: the SyncNet gains support a narrower lip-synchronization benefit, whereas the FVD regression warns against claiming a general improvement in human-video dynamics or fidelity. Moreover, the reference-image and driving-audio inputs available to Sonic and related methods are absent from our text-only task; their rows are contextual measurements, not a task-equivalent leaderboard.

B.5 Ablation and Analysis

Table 6: Component and design ablations at step 3600. VBench uses the full standard set; audio metrics use a shared evaluation subset. Only evaluated configurations are listed. Higher is better.

Configuration	Overall	Quality	Semantic	CLAP	AQ	CE	CU	PC	PQ
Baseline	0.6417	0.8419	0.4861	0.268	4.978	4.267	6.358	2.480	6.807
DecShift only	0.6283	0.8225	0.4774	0.304	5.010	4.434	6.328	2.601	6.677
Ours (score shift 1.2)	**0.6740**	0.8345	**0.5491**	0.306	4.930	4.264	6.251	2.534	6.671
Score shift 1.0	0.6381	0.8379	0.4827	0.322	4.925	4.262	6.245	2.520	6.675
Score shift 1.5	0.6325	0.8216	0.4853	0.298	4.953	4.234	6.161	**2.768**	6.648
Generator audio shift 1.5	0.6562	0.8479	0.5071	**0.335**	–	–	–	–	–
Independent DualFake	0.6458	0.8435	0.4921	0.292	**5.058**	**4.482**	**6.502**	2.392	**6.858**
Sequential AV updates	0.6651	**0.8523**	0.5195	0.310	4.943	4.338	6.254	2.530	6.650

Score-shift variants retain shared fake-score parameters and the native generator schedule. Generator shift, DualFake and sequential updates retain score shift 1.2. DualFake separates video/audio fake-score parameters; sequential AV uses separate generator optimizer steps. AQ is the mean of CE/CU/PC/PQ. – denotes unevaluated AudioBox metrics. Bold marks the best observed value per column.

Table 6 reports only measured configurations. DecShift-only improves CLAP and several AudioBox axes on the fixed audio subset, but lowers VBench Overall from 0.6417 to 0.6283. The complete method reaches 0.6740. A standalone DFake control has not yet been evaluated, so these results do not establish independent component effects or complementarity.

The conditional comparison is nevertheless informative. Starting from DecShift-only, adding the shared fake-score training construction recovers 0.0457 Overall and 0.0717 Semantic, while subset CLAP changes by only 0.002 and AudioBox AQ decreases by 0.080. The substantial recovery is therefore in the joint model's visual-semantic behavior, not in a uniform audio-quality increase. This is consistent with the need to fit the fake score on the same states used for DMD, but does not isolate state matching from the additional score evaluations or establish a standalone DFake gain.

Choosing the audio score shift. Table 6 varies the maximum audio score shift over 1.0, 1.2 and 1.5, while retaining the shared fake score and the same generator schedule. Overall VBench is 0.6381, 0.6740 and 0.6325, respectively, with semantic scores 0.4827, 0.5491 and 0.4853. The choice is not monotone and does not maximize every metric: shift 1.0 has slightly higher visual quality and higher subset CLAP (0.322 versus 0.306), whereas 1.2 improves semantic composition and imaging quality. We choose 1.2 for this joint trade-off, not as a universal optimum or as evidence that lower audio noise is always better.

In particular, selecting the shift by subset CLAP alone would prefer 1.0 and lose 0.0359 Overall and 0.0664 Semantic relative to 1.2. The 1.5 setting is also worse on these two aggregates. The three-point sweep supports an interior operating choice among the tested settings; it does not establish the shape of the response between them. Since the runs share one training seed, the choice still requires multi-seed confirmation.

Why not shift the generator as well? The score shift selects the noise levels at which the DMD correction is estimated; applying separate video/audio shifts inside the generator instead changes its denoising trajectory. We tested this additional change while retaining score shift 1.2 and decoupled fake-score training. With generator video/audio shifts $2.05/1.5$, Overall and Semantic fall from $0.6740/0.5491$ to $0.6562/0.5071$, although Quality rises from 0.8345 to 0.8479 and subset CLAP rises from 0.306 to 0.335. A generator audio shift of 1.2 gives subset CLAP 0.302; its available video diagnostic is a subset evaluation and is not substituted for the full-set scores in the table. These results show a change in the audio–visual trade-off, not a uniform failure of generator shifting. We keep the native joint generator trajectory to preserve the stronger semantic result and confine the proposed change to score estimation and supervision.

Why shared parameters and branch-specific fake-score losses? The generator queries video at $\boldsymbol{z}_t^{\mathrm{V}}$ and audio at $\boldsymbol{z}_t^{\mathrm{A}}$, so each fake-score branch is supervised on the same state from which its DMD target is read. This matches the training inputs to the score queries while retaining joint audio–video

context in each pass. Stronger parameter separation is not better in this comparison: Independent DualFake reaches Overall 0.6458, Semantic 0.4921 and subset CLAP 0.292, versus 0.6740, 0.5491 and 0.306 for the shared model. DualFake improves some quality-related axes, so the advantage is in the joint semantic result rather than every perceptual metric. Separating the generator's video/audio optimizer steps also changes the trade-off: Overall is 0.6651, Semantic 0.5195 and CLAP 0.310. Our two fake-score backward calls are a memory-saving implementation of gradient accumulation, not a distinct objective by themselves: for fixed states, shared parameters and no intervening optimizer step, $\nabla_\phi(\ell^v + \ell^a) = \nabla_\phi \ell^v + \nabla_\phi \ell^a$. The design distinction is the two synchronized training states and the modality selected on each, not the number of backward calls.

What the diagnostic figures establish. Figure 1 separates matched-noise denoising difficulty from the dispersion of the training losses, while Figure 5 reports fake-score errors across noise levels. These diagnostics describe the behavior of the evaluated configurations; a reduction in loss variance is not itself a quality metric, and a difference between the baseline and full method cannot assign the effect to one component. The design ablations in Table 6 are therefore needed alongside those plots. The evidence supports the tested shared-network, state-matched construction as a joint semantic trade-off, not a general claim that independent experts or sequential updates must fail.

Training cost and inference scope. The proposed construction evaluates each joint score network at two synchronized states rather than one. Our fixed-step comparisons thus hold the generator architecture and sampling budget constant, but are not matched in training FLOPs. A forward-pass-matched control is needed to separate the effect of the construction from extra score computation. At inference none of these score-query branches is used: the same four-step joint student produces the audio-video pair. We therefore claim unchanged inference structure, not a measured hardware speedup over the matched baseline.

C Fake-Score Accuracy

Section 4.4 argues that a summed fake-score objective leaves the audio branch less well fitted, and that branch-specific losses correct it. We test this directly by measuring the normalized score error of the trained fake score against each generated marginal, over a grid of noise levels.

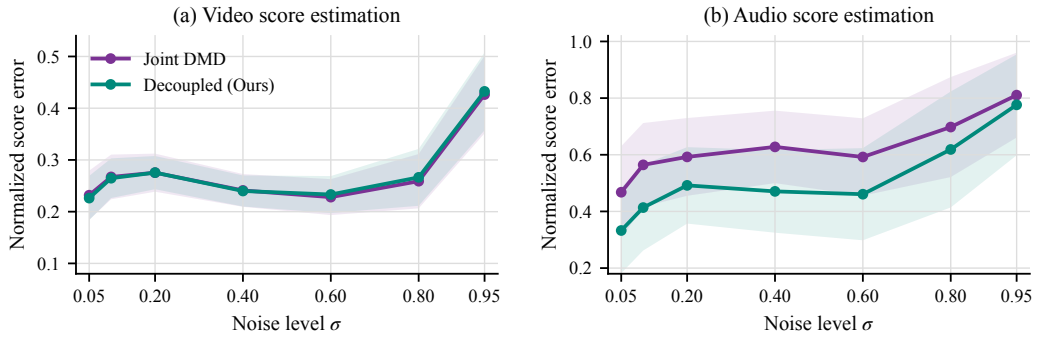


Figure 5: **Fake-score accuracy on each generated marginal.** Normalized score error of the trained fake score against (a) the video and (b) the audio marginal, as a function of noise level; shaded bands are across evaluation samples. Decoupling the objective leaves video accuracy essentially unchanged and improves audio accuracy at every noise level.

On video the two schemes are close, with decoupled training slightly ahead at low noise. On audio the gap is present at every noise level: the joint objective leaves an error between 0.47 and 0.81, while decoupled training holds it between 0.33 and 0.78 and widens the margin through the mid-noise range where the four student transitions fall. The summed objective was therefore not merely reporting two losses together; it left the audio branch measurably less accurate at the noise levels the generator reads.

D CAUSAL SELF-FORCING TRANSFER

We also probe whether the two components transfer outside the bidirectional student, using an OmniForcing-style causal self-forcing stage (Su et al., 2026). Both variants start from the same step-4000 ODE initialization, use a five-block prefix rerun, and apply the loss to the last generated block. The reproduced baseline of this stage is not converged, so Table 7 is a diagnostic and is not evidence for the claims made in the main text.

Table 7: Provisional causal self-forcing reproduction at step 5200. The reproduced baseline is qualitatively underconverged; these numbers are retained for diagnosis and are not used to support the main claim. Higher is better.

Method	VBench	Quality	Semantic	CLAP	AQ	CE	CU	PC	PQ
Reproduced baseline	0.5293	0.7352	0.3692	**0.377**	5.089	4.191	6.186	**3.672**	6.308
Ours	**0.6045**	**0.7891**	**0.4610**	0.348	**5.282**	**4.494**	**6.525**	3.483	**6.626**