TURBOT2VA: FAST LARGE-SCALE TEXT-TO-VIDEO-AUDIO GENERATION VIA SCORE-REGULARIZED CONSISTENCY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint text-to-video-audio generation produces synchronized visual and acoustic content, but the long sampling trajectories and heterogeneous multimodal computation of large models make inference prohibitively expensive. We present TurboT2VA, a distillation and inference framework for accelerating a 19B-parameter joint video-audio model. Compared with discrete-time consistency, continuous-time consistency avoids finite-step discretization errors in the training target and provides more faithful supervision along the teacher trajectory, making it particularly attractive for few-step distillation. However, scaling continuous-time consistency to a 19B joint T2VA model is challenged by modality-imbalanced optimization, numerically fragile trajectory-tangent estimation, and the quality–diversity trade-off. TurboT2VA addresses these issues with per-modality normalization and a progressive dCM→sCM→sCM+DMD curriculum, where dCM establishes a stable initialization, sCM more faithfully learns the continuous teacher trajectory while preserving diversity, and DMD improves perceptual quality. On LTX-2, four-step distillation reduces generator latency from 50.52s to 2.51s at the standard evaluation resolution of 512×768, achieving a 20.1× speedup while maintaining strong visual quality, audio fidelity, diversity, and video-audio synchronization. We further develop an architecture-aware inference stack that combines guarded W8A8 and fused operators, padded-text compaction, and modality-aware sparse attention while preserving dense cross-modal and text-conditioning paths without retraining. Under the high-resolution deployment setting at 1024×1792, the complete stack reduces generator latency from 318.74s to 5.83s on one NVIDIA H20, achieving a 54.67× generator-only speedup over the original teacher.

1 INTRODUCTION

Motivation. Text-to-video-audio (T2VA) generation has advanced toward realistic video synthesis with synchronized audio, with models such as LTX-2 (HaCohen et al., 2026) demonstrating unified audio-visual generation. However, these models rely on long diffusion or flow sampling trajectories, making inference prohibitively expensive. Existing acceleration methods mainly follow consistency distillation and distribution matching. Within consistency distillation, dCM enforces consistency between finitely separated timesteps and is affected by timestep discretization, whereas sCM operates in continuous time and provides faithful supervision along the teacher trajectory (Song et al., 2023; Lu & Song, 2025). Such continuous-trajectory supervision is attractive for T2VA, where visual motion, acoustic evolution, and their temporal correspondence must be preserved jointly. Distribution Matching Distillation (DMD) instead aligns student and teacher distributions through score discrepancies (Yin et al., 2024b;a). Score-regularized continuous-time consistency (rCM) combines sCM with distribution-level score supervision to improve the quality–diversity trade-off (Zheng et al., 2026), but scaling this paradigm to large joint T2VA models remains underexplored.

Challenge. Directly extending existing T2I/T2V acceleration methods to T2VA introduces three challenges. (C1) **Modality imbalance.** Video and audio differ substantially in scale and temporal resolution, causing naive joint objectives to be dominated by one modality. (C2) **Stable continuous-time trajectory learning.** Trajectory-tangent estimation through JVPs becomes numerically fragile

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.

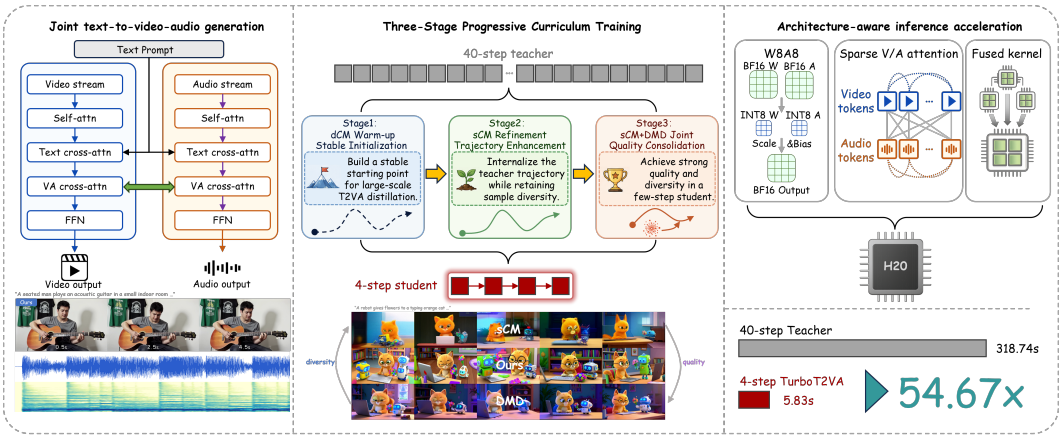

Figure 1: Overview. A joint video-audio Transformer is distilled from a 40-step teacher into a four-step student through a progressive dCM→sCM→sCM+DMD curriculum, improving the quality–diversity trade-off for synchronized video-audio generation. The distilled student is further accelerated by an architecture-aware inference stack consisting of modality-aware sparse attention dispatch, guarded W8A8 linear operators, and fused Transformer kernels, yielding a 54.67× generator-only speedup under the 1024×1792 high-resolution deployment setting on one NVIDIA H20.

when scaled to a 19B joint model, particularly for audio. (C3) **Quality–diversity trade-off.** Consistency learning favors trajectory coverage and diversity, whereas distribution matching improves realism but can reduce diversity.

Approach. We present **TurboT2VA**, a joint distillation and inference framework for large-scale T2VA generation (Figure 1). (A1) We decouple video and audio losses with per-modality normalization while retaining end-to-end optimization through the shared cross-modal generator. (A2) We adopt a progressive dCM→sCM strategy, using dCM for stable initialization and sCM for continuous teacher-trajectory refinement. (A3) After establishing a stable trajectory prior, we jointly optimize sCM and DMD to balance trajectory fidelity, perceptual quality, and sample diversity.

Performance. We instantiate TurboT2VA on LTX-2, a 19B joint T2VA model with a 14B video backbone and a 5B audio backbone. At 512×768, four-step distillation reduces generator latency from 50.52s to 2.51s, achieving a 20.1× speedup while maintaining strong visual quality, audio fidelity, diversity, and video-audio alignment. At 1024×1792, the complete architecture-aware inference stack further reduces generator latency from 318.74s to 5.83s on one NVIDIA H20, yielding a 54.67× generator-only speedup. Across representative cascaded, open-source, and closed-source T2VA baselines and extensive ablations, TurboT2VA remains competitive across multimodal quality metrics while substantially improving efficiency.

Contributions. Our contributions are as follows:

- We propose **TurboT2VA**, a unified acceleration framework for large-scale joint T2VA generation that combines few-step distillation with architecture-aware inference acceleration.
- We develop modality-aware joint T2VA distillation with decoupled video and audio objectives and per-modality normalization, mitigating optimization imbalance while preserving coupled video-audio generation through the shared model.
- We combine continuous-time consistency distillation with distribution matching in a progressive dCM→sCM→sCM+DMD curriculum, stabilizing large-scale training while balancing teacher-trajectory fidelity, diversity, and perceptual quality.
- We develop an architecture-aware inference stack with modality-aware sparse attention, guarded W8A8 operators, and fused multimodal kernels, achieving a 54.67× generator-only speedup at 1024×1792 on one NVIDIA H20.

2 BACKGROUND

2.1 FROM VIDEO GENERATION TO JOINT VIDEO-AUDIO GENERATION

Recent text-to-video models have achieved strong visual fidelity, motion coherence, and temporal consistency (Singer et al., 2022; Yang et al., 2025; HaCohen et al., 2024; Zheng et al., 2024). Yet most generate silent videos, while realistic multimedia requires semantically relevant and temporally aligned audio. Cascaded pipelines generate audio from video (Jeong et al., 2025; Zhang et al., 2026b; Cheng et al., 2025) or video from audio (Yariv et al., 2024; Jeong et al., 2023), but sequential modeling may introduce semantic and temporal mismatch. Joint T2VA systems such as JavisDiT, OVI, LTX-2, and DaVinci-MagiHuman instead synthesize both modalities within unified or coupled architectures (Liu et al., 2026; Low et al., 2025; HaCohen et al., 2026; Chern et al., 2026), enabling direct cross-modal interaction and improved video-audio coherence. However, generating both modalities through long diffusion or flow trajectories incurs substantial inference cost, motivating distillation of multi-step T2VA teachers into few-step students.

2.2 CONSISTENCY AND DISTRIBUTION-MATCHING DISTILLATION

Distillation for diffusion and flow models mainly follows consistency learning and distribution matching. For clarity, we use dCM to denote discrete-time consistency training, which enforces consistency between predictions at finitely separated noise levels. Since dCM constructs targets through numerical PF-ODE transitions on a discretized timestep grid, its supervision is affected by grid design and finite-step discretization errors. In contrast, sCM follows the infinitesimal tangent of the teacher trajectory and provides more faithful continuous-time supervision (Song et al., 2023; Lu & Song, 2025), but requires Jacobian–vector products (JVPs), making large-scale training more numerically demanding (Lu & Song, 2025; Zheng et al., 2026). Distribution-matching methods align student and teacher distributions through score discrepancies, improving few-step synthesis quality but potentially reducing diversity (Yin et al., 2024b;a; Zheng et al., 2026). Score-regularized continuous-time consistency combines trajectory-level consistency with distribution matching to improve the quality–diversity trade-off (Zheng et al., 2026). TurboT2VA extends this principle to large-scale joint T2VA generation with modality-balanced optimization and progressive distillation.

3 CROSS-MODAL JOINT DISTILLATION

TurboT2VA extends score-regularized continuous-time consistency (Zheng et al., 2026) to joint text-to-video-audio generation, as illustrated in Figure 2. Let $\mathcal{M} = \{v, a\}$ denote the video and audio modalities. Given a text condition c, a clean sample is represented as $z_0 = (z_0^v, z_0^a)$. We sample a shared timestep t and independent noises ϵ^m to construct

$$z_t^m = \cos(t) z_0^m + \sin(t)\epsilon^m, \quad m \in \mathcal{M}. \tag{1}$$

The student predicts both modalities in one forward pass:

$$(\hat{z}_{0,\theta}^v, \hat{z}_{0,\theta}^a) = G_\theta(z_t^v, z_t^a, t, c). \tag{2}$$

Although the objectives are computed per modality, both predictions share the same timestep, text condition, and cross-modal Transformer. Video and audio are therefore optimized through a coupled forward graph rather than two independent distillation branches.

3.1 JOINT CONTINUOUS-TIME CONSISTENCY

For each modality $m \in \mathcal{M}$, teacher and student predictions are converted into TrigFlow fields:

$$F_q^m = \frac{\cos(t) z_t^m - \hat{z}_{0,q}^m}{\sin(t)}, \quad q \in \{T, \theta\}. \tag{3}$$

To capture coupled video-audio dynamics, we compute the JVP of each student field along the joint teacher direction:

$$J_\theta^m = c_t \left(\frac{\partial F_\theta^m}{\partial t} + \frac{\partial F_\theta^m}{\partial z_t^v} F_T^v + \frac{\partial F_\theta^m}{\partial z_t^a} F_T^a \right), \quad c_t = \cos(t)\sin(t). \tag{4}$$

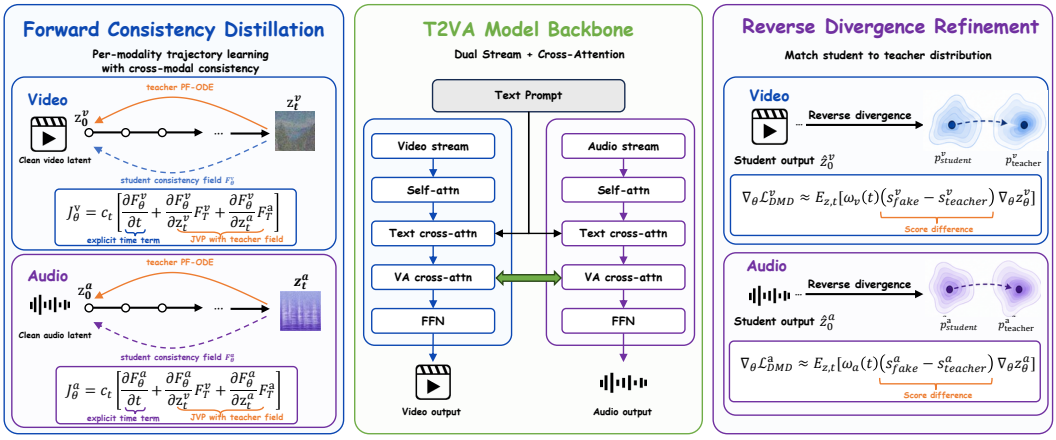

Figure 2: Dual-Divergence Distillation for T2VA. Left: cross-modal consistency distillation; middle: joint T2VA backbone; right: distribution-matching refinement.

Because F_θ^m depends on both modalities through the shared cross-modal Transformer, the video direction depends on audio tokens and vice versa. This distinguishes joint T2VA consistency distillation from simply summing two independent modality objectives.

Following sCM, we substitute this joint JVP into the continuous-time consistency direction g_{sCM}^m. Further details on the sCM direction and cross-modal JVP are provided in Appendix A.1. Since video and audio differ substantially in latent scale and dimensionality, we normalize this direction separately for each modality:

$$\bar{g}_{\text{sCM}}^m = \frac{g_{\text{sCM}}^m}{\|g_{\text{sCM}}^m\|_2 + \epsilon_{\text{sCM}}}, \quad m \in \mathcal{M}. \tag{5}$$

The norm is computed per sample over all latent elements of modality m. The sCM objective is

$$\mathcal{L}_{\text{sCM}}^{\text{joint}} = \sum_{m \in \mathcal{M}} w_m \left\| F_\theta^m - \text{sg}(F_\theta^m) - \bar{g}_{\text{sCM}}^m \right\|_2^2, \tag{6}$$

where w_m controls the contribution of each modality. This objective transfers the coupled teacher trajectory while balancing video and audio optimization.

3.2 Joint Distribution Matching

While sCM emphasizes trajectory fidelity and diversity, few-step generation may still lose perceptual quality. We therefore complement it with DMD using paired on-policy samples generated by the same student. Given $\hat{z}_0 = (\hat{z}_0^v, \hat{z}_0^a)$ and its perturbed sample $\hat{z}_t$, the modality-wise DMD direction is

$$g_{\text{DMD}}^m = \hat{z}_{0,\text{fake}}^m(\hat{z}_t, t, c) - \hat{z}_{0,T}^m(\hat{z}_t, t, c), \quad m \in \mathcal{M}, \tag{7}$$

where $\hat{z}_{0,T}$ is the teacher prediction and $\hat{z}_{0,\text{fake}}$ is produced by the fake-score network trained on student-generated samples. Under this parameterization, the prediction discrepancy corresponds to the score discrepancy used in distribution matching.

Because video and audio differ in residual scale, DMD uses a modality-wise residual normalizer:

$$c_{\text{DMD}}^m = \frac{1}{N_m} \sum_{i \in m} \left| \hat{z}_{0,i}^m - \hat{z}_{0,T,i}^m \right|, \qquad \bar{g}_{\text{DMD}}^m = \frac{g_{\text{DMD}}^m}{c_{\text{DMD}}^m + \epsilon_{\text{DMD}}}, \tag{8}$$

where N_m is the number of latent elements in modality m.

We aggregate the DMD surrogate with a sum reduction:

$$\mathcal{L}_{\text{DMD}}^{\text{joint}} = \sum_{m \in \mathcal{M}} w_m \sum_{i \in m} \left[\hat{z}_{0,i}^m - \text{sg}\left(\hat{z}_{0,i}^m - \bar{g}_{\text{DMD},i}^m \right) \right]^2. \tag{9}$$

The sum reduction retains the accumulated distribution-matching signal over the paired video-audio latent, while modality-wise normalization controls the relative scales of the two modalities. Since both modalities are generated by the same student under shared conditioning, DMD refines perceptual quality without separating the coupled video-audio generation process. Further analysis of the DMD surrogate and modality-wise normalization is provided in Appendix A.2.

3.3 Modality-Balanced Joint Optimization

TurboT2VA therefore applies modality-wise balancing to both objectives: sCM uses an L2 direction normalizer, whereas DMD uses a teacher-residual scale normalizer. The balancing operates only at the objective level; the student remains a single coupled T2VA generator.

During joint optimization, sCM and DMD operate on the same paired batch and text condition. The final objective is

$$\mathcal{L}_{\text{stage3}} = \lambda_{\text{sCM}}\mathcal{L}_{\text{sCM}}^{\text{joint}} + \lambda_{\text{DMD}}\mathcal{L}_{\text{DMD}}^{\text{joint}}. \tag{10}$$

The sCM term maintains teacher-trajectory fidelity and diversity, whereas DMD improves perceptual quality through distribution matching.

LTX-2 uses a rectified-flow parameterization, whereas sCM is formulated under TrigFlow. We therefore use a lightweight interface wrapper without modifying the underlying backbone; the mapping and implementation details are provided in Appendix A.3.

4 Progressive Curriculum Distillation

Direct sCM+DMD optimization is feasible, but the optimization path substantially affects the final quality–diversity trade-off. As shown in Figure 5(a), the staged schedule overtakes direct joint training after entering the final stage and maintains stronger late-training performance.

The two objectives play complementary roles. sCM transfers the teacher trajectory and promotes trajectory coverage and sample diversity, whereas DMD provides distribution-level supervision that improves perceptual quality. Joint optimization from the base initialization therefore exposes early trajectory learning to the DMD gradient before the student has acquired a sufficiently structured trajectory prior. We consequently delay DMD until after dCM warm-up and sCM refinement, yielding the progressive dCM→sCM→sCM+DMD curriculum illustrated in the middle panel of Figure 1. A gradient-level interpretation of this optimization path is provided in Appendix B.

Stage 1: dCM warm-up. We first train with dCM for 2,000 steps to establish a coarse joint denoising prior. Since dCM avoids JVP-based trajectory-tangent estimation, it provides a stable initialization for subsequent continuous-time consistency learning.

Stage 2: sCM refinement. Starting from the dCM checkpoint, we optimize the sCM objective for 500 steps. This stage replaces discrete-time consistency with continuous trajectory supervision, improving trajectory fidelity and preserving diversity before distribution matching is introduced.

Stage 3: sCM+DMD joint optimization. We then train for 4,500 steps with the joint objective in Section 3. sCM continues to preserve teacher-trajectory fidelity and diversity, while DMD refines the student distribution for improved perceptual quality.

The selected four-step student is trained for 7,000 steps in total. Figure 5(b) further shows that, after the same dCM+sCM prefix, joint sCM+DMD optimization achieves stronger late-training performance than continuing either objective alone. Together with Table 2, these results show that the staged curriculum achieves a better quality–diversity trade-off.

5 Architecture-Aware Inference Acceleration

Few-step distillation reduces the number of Transformer evaluations but not their per-step cost, which becomes dominant at high resolution. Joint T2VA models further contain heterogeneous video, audio, cross-modal, and text-conditioning paths with different sequence lengths, masks, and

matrix shapes. Building on TurboDiffusion (Zhang et al., 2025b), we therefore specialize sparse attention, W8A8 linear operators, and fused kernels to the architecture of LTX-2. The resulting stack is applied after distillation and requires no retraining.

5.1 Modality-Aware Sparse Attention

The joint Transformer contains video self-attention, audio self-attention, bidirectional video-audio attention, and masked text cross-attention. We apply SageSLA (Zhang et al., 2025a; 2026a) only to the unmasked video and audio self-attention paths, while keeping cross-modal and text-conditioning attention dense. This role-aware dispatch concentrates approximation on the long modality-specific sequences without altering explicit video-audio exchange or text-conditioning connectivity.

For each selected layer, SageSLA retains a fraction ρ_ℓ of key blocks using a query-dependent block map. To avoid degenerate sparsity on short audio sequences, we enforce at least one retained key block. Our default high-resolution configuration uses $\rho_\ell = 0.3$ for all selected layers.

5.2 Guarded Low-Precision and Fused Execution

Guarded W8A8. For compatible linear layers, we use post-scale W8A8 with per-output-channel weight and per-row activation quantization. INT8 products are accumulated in INT32 across the full reduction dimension before rescaling. Because matrix shapes differ across video, audio, and cross-modal branches, unsupported cases fall back to native BF16 computation.

Fused execution. We further reduce kernel-launch and memory overhead by fusing recurring normalization, adaptive modulation, and gated residual operations.

Padded-text compaction. For batch-size-one inference, we compact padded text context before cross-attention using the same valid-token set for video and audio. This optimization is disabled for larger batches, where valid-token patterns may differ across prompts.

6 Experiments

6.1 Experimental Setup

We evaluate TurboT2VA on LTX-2, a 19B joint T2VA model comprising a 14B video backbone and a 5B audio backbone. Training uses 100K paired text-video-audio samples at 512×768 with 121 frames on 8 NVIDIA H20 GPUs. The four-step student is trained for 7K steps following the curriculum in Section 4, requiring 21 hours. Additional training details are provided in Appendix C.

Standard-resolution evaluation. All main model comparisons, ablations, and sampling-step experiments use the same 200-prompt split at 512×768 resolution with 121 frames. We report average per-sample inference time, excluding model loading and evaluation overhead. We report Javis-Bench (Liu et al., 2026) for multimodal generation quality and alignment, VBench (Huang et al., 2024) for video fidelity, TTA-Bench (Wang et al., 2026) for audio quality, MS-CLAP (Elizalde et al., 2023) for text-audio alignment, and ImageBind (Girdhar et al., 2023) for within-prompt diversity.

High-resolution deployment. We separately profile the architecture-aware inference stack on one NVIDIA H20 at batch size one, 1024×1792 resolution, and 121 frames. Unless otherwise specified, latency is generator-only and excludes checkpoint loading, decoding, video-audio muxing, and disk I/O. This setting is used only for the high-resolution acceleration experiments, while the standard-resolution protocol is used for model-level comparisons and ablations.

6.2 Main Results

6.2.1 Generation Quality

Table 1 compares TurboT2VA with representative cascaded, open-source, and closed-source T2VA systems. Relative to the 40-step teacher, the four-step student reduces inference time from 50.52s

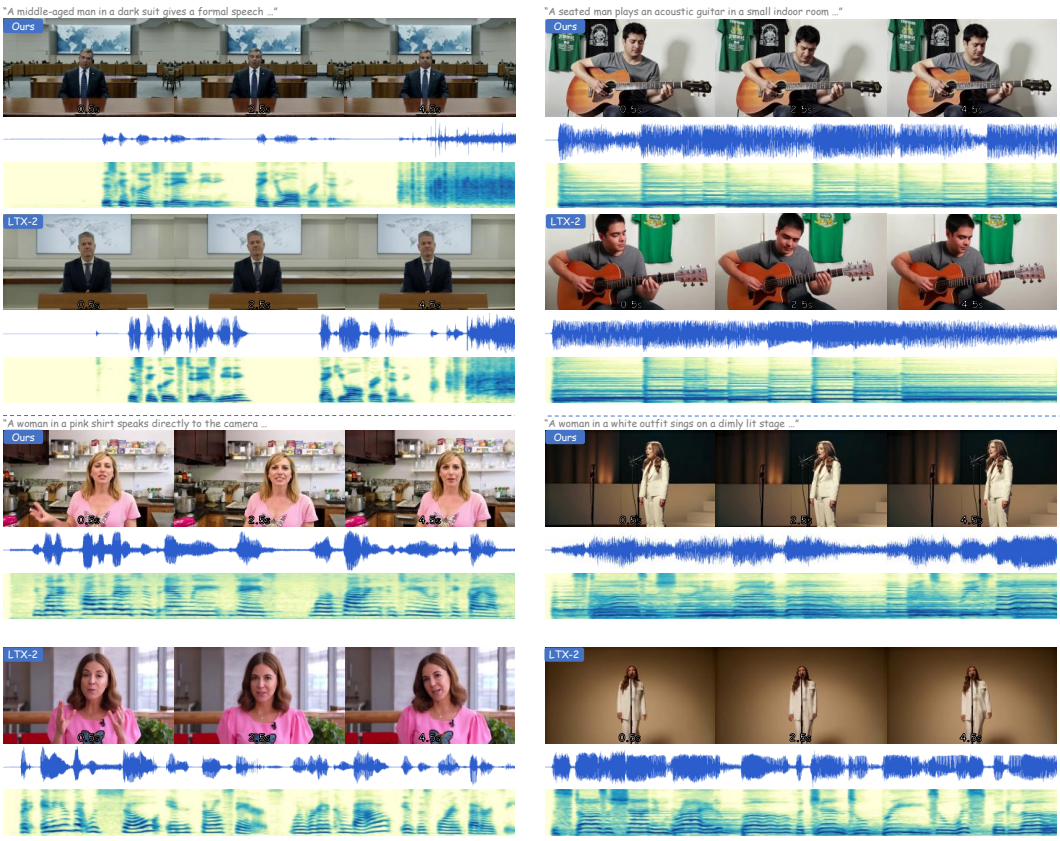

Figure 3: Qualitative comparison between the 40-step teacher and four-step TurboT2VA under matched prompts, showing representative video frames with waveforms and mel spectrograms.

Table 1: JavisBench comparison with representative T2VA systems under the standard-resolution setting. Best and second-best open-source results are **bold** and underlined.

Model	Time (s) ↓	Perceptual Quality ↑			Text Consistency ↑				V-A Consistency ↑			V-A Alignment	
		Visual	Motion	Audio	IB-TV	IB-TA	CLIP	CLAP	IB-AV	CAVP	AVH	Javis ↑	Desync ↓
Closed-source T2VA Models													
Kling v3	–	3.478	1.104	5.499	0.246	0.154	0.309	0.421	0.231	0.793	0.210	0.177	0.907
Sora 2	–	2.836	0.185	4.802	0.252	0.161	0.308	0.434	0.240	0.792	0.217	0.187	0.414
Veo 3	–	4.221	0.735	5.534	0.219	0.178	0.310	0.469	0.351	0.798	0.326	0.285	0.336
Cascaded T2A + A2V													
TempoToken	7.76	0.885	-0.671	–	**0.298**	–	0.292	–	0.133	0.796	0.111	0.071	1.021
TPoS	18.01	1.592	-0.330	–	0.240	–	0.261	–	0.145	0.793	0.138	0.098	1.112
Cascaded T2V + V2A													
ReWaS	40.49	–	–	4.124	–	0.075	–	0.309	0.113	0.791	0.107	0.083	0.772
FoleyCrafter	36.74	–	–	4.225	–	0.079	–	0.212	0.182	0.798	0.168	0.134	0.934
MMAudio	15.16	–	–	4.308	–	**0.164**	–	0.401	0.190	0.794	0.180	0.142	0.524
Open-source T2VA Models													
JavisDiT	49.73	1.488	0.591	4.405	0.250	0.145	0.297	**0.413**	0.212	0.798	0.183	0.152	0.848
OVI	76.79	2.145	0.294	5.024	0.244	0.132	**0.307**	0.407	0.209	0.795	0.206	0.175	0.638
DaVinci-MagiHuman	6.70	0.854	0.382	**5.245**	0.227	0.124	0.284	0.362	0.201	0.800	0.181	0.145	0.730
Teacher and TurboT2VA													
Teacher (40 steps)	50.52	1.968	0.783	4.902	0.268	0.139	0.291	0.340	0.199	0.796	0.195	0.165	0.617
Ours (4 steps)	**2.51**	**2.389**	**0.849**	4.486	0.218	0.145	0.292	0.347	**0.229**	**0.803**	**0.224**	**0.196**	**0.388**

to 2.51s, achieving a 20.1× speedup, while improving Javis from 0.165 to 0.196 and Desync from 0.617 to 0.388. Several modality-specific metrics improve while others decrease moderately; among systems with reported runtimes, TurboT2VA achieves the lowest latency. Detailed modality-specific results are provided in Appendix C.2. Figure 3 further shows that the four-step student preserves coherent visual evolution and audio structure under matched prompts.



Figure 4: Quality–diversity comparison under matched prompts and random seeds. Rows correspond to sCM-only, Ours (staged), and DMD-only; columns use different random seeds.

Table 2: Quality–diversity comparison under different distillation objectives. Best and second-best results are **bold** and underlined.

Model	Quality	Diversity		
	Javis ↑	Video ↑	Audio ↑	V-A Avg. ↑
sCM-only	0.1131	**0.3026**	**0.5037**	**0.4032**
DMD-only	<u>0.1812</u>	0.1821	0.3561	0.2691
Ours (staged)	**0.1963**	<u>0.2322</u>	<u>0.4197</u>	<u>0.3259</u>

6.2.2 Quality–Diversity Trade-off

We evaluate within-prompt diversity using eight samples per prompt with different random seeds and compute pairwise ImageBind distances separately for video and audio; their arithmetic mean is reported as V-A Avg. Detailed evaluation procedures are provided in Appendix C.3. As shown in Table 2, sCM-only yields the highest diversity but substantially lower Javis quality, whereas DMD-only improves quality at the cost of diversity. The staged model achieves the highest Javis score while retaining substantially more video and audio diversity than DMD-only, demonstrating a more favorable quality–diversity trade-off. Figure 4 further provides a qualitative comparison under matched prompts and random seeds. sCM-only shows broader variation but weaker prompt fidelity, whereas DMD-only is more repetitive; the staged model retains cross-seed variation while better preserving the requested interaction.

6.3 High-Resolution Inference Acceleration

We evaluate the inference stack at 1024×1792. The dense 40-step teacher requires 318.74s, whereas the complete TurboT2VA stack with $\rho = 0.3$ requires 5.83s, yielding a $54.67\times$ generator-only speedup. Relative to the unoptimized four-step student at 16.50s, the inference stack provides an additional $2.83\times$ acceleration. The cumulative acceleration path is provided in Appendix C.4.

Component-wise acceleration. Table 3 isolates the contribution of individual inference components. Fused operations and W8A8 reduce dense four-step latency from 16.50s to 13.77s and 14.73s, respectively, while their combination reaches 12.16s. Sparse attention provides the largest single-component gain at 6.24s, and the complete stack further reduces latency to 5.83s.

Speed–quality trade-off. We vary the SageSLA retention ratio while fixing the four-step student, prompts, sampling schedule, and random seeds. As shown in Table 4, all sparse settings substantially reduce latency with metrics close to the dense reference. We choose $\rho = 0.3$ as default because it achieves the best CAVP and lower Desync than the faster $\rho = 0.2$ setting.

6.4 Ablation Study

Progressive curriculum. Figure 5(a) compares the staged curriculum with direct sCM+DMD optimization from the base initialization. Direct joint training is feasible, but the staged route overtakes

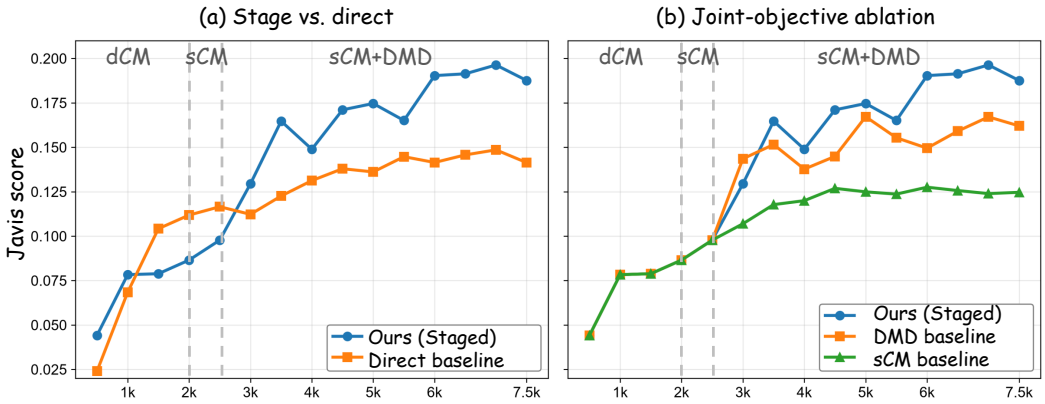

Figure 5: Training-curve ablations. (a) Staged dCM→sCM→sCM+DMD training versus direct sCM+DMD optimization. (b) Final-stage objectives starting from the same dCM+sCM prefix.

Table 3: High-resolution generator latency breakdown at 1024×1792 on NVIDIA H20. Speedups are relative to the dense four-step student. Best and second-best are **bold** and underlined.

Metric	Teacher	Dense	Fused Only	W8A8 Only	W8A8+Fused	Sparse Only	Full	ρ=0.5	ρ=0.4	ρ=0.3	ρ=0.2
Gen. (s) ↓	318.74	16.50	13.77	14.73	12.16	6.24	6.44	6.13	5.83	**5.57**
Speedup ↑	–	1.00×	1.20×	1.12×	1.36×	2.65×	2.56×	2.69×	2.83×	**2.97**×

Table 4: High-resolution speed–quality trade-off of four-step TurboT2VA. Best and second-best results are **bold** and underlined; rankings use unrounded values.

Config.	Gen. (s)↓	Javis↑	AVH↑	Desync↓	CAVP↑	IB-AV↑
Dense	16.50	**0.1948**	**0.2330**	0.4275	0.7964	0.2370
Full, ρ=0.5	6.44	0.1939	0.2287	**0.3990**	0.7974	0.2375
Full, ρ=0.4	6.13	0.1901	0.2247	0.4080	0.7974	0.2345
Full, ρ=0.3	5.83	0.1914	0.2269	0.4085	**0.7985**	0.2348
Full, ρ=0.2	**5.57**	0.1948	0.2305	0.4145	0.7971	**0.2380**

it after entering the joint stage and maintains stronger late-training Javis performance at the common 7,500-step endpoint. This empirically supports the optimization schedule introduced in Section 4.

Joint sCM+DMD objective. Figure 5(b) compares final-stage objectives after the same dCM+sCM prefix. Joint sCM+DMD optimization outperforms both sCM-only continuation and DMD refinement in late-training Javis score, reaching 0.1963 at the selected 7,000-step checkpoint and remaining ahead at 7,500 steps. Together with the diversity results in Table 2, this supports retaining both trajectory- and distribution-level objectives in the final stage.

Sampling steps. We additionally compare one-, two-, and four-step sampling in Appendix C.5. Four-step inference performs best, while two-step sampling already provides reasonable quality.

7 Conclusion

We presented TurboT2VA, a unified framework for accelerating large-scale joint text-to-video-audio generation through progressive consistency distillation and architecture-aware inference optimization. The proposed dCM→sCM→sCM+DMD curriculum improves the quality–diversity trade-off while preserving coupled video-audio generation. On LTX-2, TurboT2VA achieves a 20.1× speedup at 512×768 with four-step sampling, and the complete inference stack reaches a 54.67× generator-only speedup at 1024×1792 on one NVIDIA H20. These results demonstrate the complementary benefits of few-step distillation and system-level acceleration for efficient joint T2VA generation.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and wording refinement. All technical content, experimental results, and conclusions were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REFERENCES

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Ethan Chern, Hansi Teng, Hanwen Sun, Hao Wang, Hong Pan, Hongyu Jia, Jiadi Su, Jin Li, Junjie Yu, Lijie Liu, et al. Speed by simplicity: A single-stream architecture for fast audio-video generative foundation model. *arXiv preprint arXiv:2603.21986*, 2026.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190. IEEE, 2023.

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

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7788–7798. IEEE, 2023.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Scofield Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *International Conference on Learning Representations*, volume 2026, pp. 139160–139194, 2026.

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

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, volume 2025, pp. 50611–50649, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. Tta-bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33512–33520, 2026.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6639–6647, 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. *Advances in neural information processing systems*, 37:47455–47487, 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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 2024b.

Jintao Zhang, Pengle Zhang, Jun Zhu, Jianfei Chen, et al. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. In *International Conference on Learning Representations*, volume 2025, pp. 71566–71585, 2025a.

Jintao Zhang, Kaiwen Zheng, Kai Jiang, Haoxu Wang, Ion Stoica, Joseph E Gonzalez, Jianfei Chen, and Jun Zhu. Turbodiffusion: Accelerating video diffusion models by 100-200 times. *arXiv preprint arXiv:2512.16093*, 2025b.

Jintao Zhang, Haoxu Wang, Kai Jiang, Shuo Yang, Kaiwen Zheng, Haocheng Xi, Ziteng Wang, Hongzhou Zhu, Min Zhao, Ion Stoica, et al. Sla: Beyond sparsity in diffusion transformers via fine-tunable sparse–linear attention. In *International Conference on Learning Representations*, volume 2026, pp. 138288–138305, 2026a.

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, 2026b.

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*, volume 2026, pp. 2582–2603, 2026.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

A Additional Distillation Details

This appendix provides additional formulations and implementation details for the joint distillation method in Section 3. The main paper contains the complete cross-modal training objectives; the following sections provide additional objective details, theoretical interpretations, and backbone-specific interface adaptations.

A.1 Full sCM Direction and Cross-Modal JVP Analysis

Following score-regularized continuous-time consistency, the modality-wise sCM direction used in the main text is

$$\begin{aligned} g^m_{\mathrm{sCM}} = &- \beta \cos(t) \sqrt{1 - \gamma^2 \sin^2(t)}\, (\mathrm{sg}(F^m_\theta) - F^m_T) \\ &- \gamma \cos(t) \sin(t) z^m_t - J^m_\theta, \end{aligned} \tag{11}$$

where β is the consistency boost factor, γ is the tangent warm-up ratio, and $\mathrm{sg}(\cdot)$ denotes stop-gradient. Unlike the standard single-modality formulation, J^m_θ in TurboT2VA is the joint video-audio JVP defined in Section 3.

Cross-modal JVP analysis. To make the coupling in the JVP explicit, define the joint latent and teacher direction as

$$z_t = \begin{bmatrix} z^v_t \\ z^a_t \end{bmatrix}, \qquad F_T = \begin{bmatrix} F^v_T \\ F^a_T \end{bmatrix}. \tag{12}$$

The directional derivative of the modality-m student field along the joint teacher direction is

$$\frac{\partial F^m_\theta}{\partial t} + \nabla_{z_t} F^m_\theta F_T = \frac{\partial F^m_\theta}{\partial t} + \frac{\partial F^m_\theta}{\partial z^v_t} F^v_T + \frac{\partial F^m_\theta}{\partial z^a_t} F^a_T. \tag{13}$$

Multiplying this expression by the TrigFlow factor $c_t = \cos(t)\sin(t)$ gives the joint JVP used in the main text. Because the video and audio streams interact through the cross-modal Transformer, F^v_θ can depend on z^a_t and F^a_θ can depend on z^v_t. Therefore, a modality-separable JVP that differentiates only with respect to its own modality would omit the corresponding cross-modal Jacobian term. The joint JVP retains these dependencies when transferring the teacher trajectory to the student.

After modality-wise normalization,

$$\bar{g}^m_{\mathrm{sCM}} = \frac{g^m_{\mathrm{sCM}}}{\|g^m_{\mathrm{sCM}}\|_2 + \epsilon_{\mathrm{sCM}}}, \tag{14}$$

the surrogate residual can be written as

$$r^m_{\mathrm{sCM}} = F^m_\theta - \mathrm{sg}(F^m_\theta) - \bar{g}^m_{\mathrm{sCM}}. \tag{15}$$

Hence,

$$\mathcal{L}^{\mathrm{joint}}_{\mathrm{sCM}} = \sum_{m \in \mathcal{M}} w_m \, \|r^m_{\mathrm{sCM}}\|_2^2 . \tag{16}$$

The normalization is performed independently for each sample and modality over all corresponding latent elements, reducing sensitivity to differences in the magnitudes of the video and audio consistency directions while retaining backpropagation through the shared cross-modal Transformer.

A.2 DMD Surrogate Form and Gradient Interpretation

For an on-policy paired student sample $\hat{z}_0 = (\hat{z}^v_0, \hat{z}^a_0)$, we perturb it to $\hat{z}_t$ and evaluate the teacher and fake-score networks under the same timestep and text condition. The modality-wise DMD direction is

$$g^m_{\mathrm{DMD}} = \hat{z}^m_{0,\mathrm{fake}}(\hat{z}_t, t, c) - \hat{z}^m_{0,T}(\hat{z}_t, t, c), \qquad m \in \mathcal{M}. \tag{17}$$

Relation to score discrepancy. Let

$$\alpha_t = \cos(t), \qquad \sigma_t = \sin(t).$$

Under the clean-prediction parameterization associated with $\hat{z}_t^m = \alpha_t \hat{z}_0^m + \sigma_t \epsilon^m$, the score induced by a clean predictor $\hat{z}_{0,q}^m$, $q \in \{\text{fake}, T\}$, can be written as

$$s_q^m(\hat{z}_t, t, c) = \frac{\alpha_t \hat{z}_{0,q}^m(\hat{z}_t, t, c) - \hat{z}_t^m}{\sigma_t^2}. \tag{18}$$

Therefore,

$$s_{\text{fake}}^m - s_T^m = \frac{\alpha_t}{\sigma_t^2}\left(\hat{z}_{0,\text{fake}}^m - \hat{z}_{0,T}^m\right) = \frac{\cos(t)}{\sin^2(t)} g_{\text{DMD}}^m. \tag{19}$$

Hence, for interior timesteps, the clean-prediction discrepancy used in our implementation is proportional to the corresponding score discrepancy by a positive time-dependent factor. The two therefore define the same update direction up to temporal scaling.

Stop-gradient surrogate. Using the modality-wise residual scale

$$c_{\text{DMD}}^m = \frac{1}{N_m} \sum_{i \in m} \left| \hat{z}_{0,i}^m - \hat{z}_{0,T,i}^m \right|, \qquad \bar{g}_{\text{DMD}}^m = \frac{g_{\text{DMD}}^m}{c_{\text{DMD}}^m + \epsilon_{\text{DMD}}}, \tag{20}$$

we define the surrogate residual

$$\Delta_{\text{DMD}}^m = \hat{z}_0^m - \text{sg}\left(\hat{z}_0^m - \bar{g}_{\text{DMD}}^m\right). \tag{21}$$

The joint DMD objective is then

$$\mathcal{L}_{\text{DMD}}^{\text{joint}} = \sum_{m \in \mathcal{M}} w_m \sum_{i \in m} \left(\Delta_{\text{DMD},i}^m\right)^2. \tag{22}$$

Although $\Delta_{\text{DMD}}^m = \bar{g}_{\text{DMD}}^m$ in the forward pass, the stop-gradient operator blocks gradients through the target term. Consequently,

$$\nabla_\theta \mathcal{L}_{\text{DMD}}^{\text{joint}} = 2 \sum_{m \in \mathcal{M}} w_m \left(\frac{\partial \hat{z}_0^m}{\partial \theta}\right)^\top \text{sg}\left(\bar{g}_{\text{DMD}}^m\right), \tag{23}$$

where the vector notation absorbs the summation over latent elements. Thus, the squared surrogate does not train toward a fixed regression target; instead, it realizes the normalized distribution-matching direction while avoiding backpropagation through the teacher and fake-score discrepancy.

Effect of modality-wise normalization. The normalization also reduces sensitivity to heterogeneous latent scales across video and audio. Consider a uniform rescaling of modality m, $z^m \mapsto a_m z^m$ with $a_m > 0$, and assume that the associated clean predictions transform equivariantly. Then

$$g_{\text{DMD}}^m \mapsto a_m g_{\text{DMD}}^m, \qquad c_{\text{DMD}}^m \mapsto a_m c_{\text{DMD}}^m,$$

so that

$$\bar{g}_{\text{DMD}}^m \mapsto \frac{a_m g_{\text{DMD}}^m}{a_m c_{\text{DMD}}^m + \epsilon_{\text{DMD}}} = \frac{g_{\text{DMD}}^m}{c_{\text{DMD}}^m + \epsilon_{\text{DMD}}/a_m}. \tag{24}$$

For a small stabilizing constant ϵ_{DMD}, the normalized direction is therefore approximately invariant to a modality-specific change of latent scale. This property is useful for joint T2VA training, where video and audio latents can have markedly different numerical magnitudes and dimensionalities.

Finally, although the objective is written as a sum of modality-wise terms, the two terms are not optimized independently. Both $\hat{z}_0^v$ and $\hat{z}_0^a$ are produced by the same cross-modal generator $G_\theta(\hat{z}_t^v, \hat{z}_t^a, t, c)$ and therefore share parameters and cross-modal representations. The resulting gradient jointly updates the shared generator using both video and audio distribution-matching signals.

A.3 LTX-2 TrigFlow Adaptation

LTX-2 follows a rectified-flow velocity parameterization, whereas sCM is formulated under TrigFlow. We map TrigFlow latents and timesteps to the native LTX-2 interface as

$$z_{\mathrm{rf}} = \frac{z_{\mathrm{trig}}}{\cos(t) + \sin(t)}, \qquad t_{\mathrm{rf}} = \frac{\sin(t)}{\cos(t) + \sin(t)}. \tag{25}$$

The wrapper leaves the underlying LTX-2 backbone unchanged while exposing both video and audio predictions through the TrigFlow training interface. Gradients and JVPs remain differentiable through timestep embeddings, video-audio tokenization, positional embeddings, and the cross-modal Transformer. This allows the joint sCM and DMD objectives to operate on both modalities within one student model without modifying the original architecture.

B Additional Analysis of Curriculum Optimization

This section provides a gradient-level interpretation of the progressive curriculum described in Section 4. The analysis is intended to explain the optimization behavior observed empirically rather than to claim that sCM and DMD gradients are always conflicting.

B.1 Gradient Interaction in Direct Joint Optimization

Let

$$G_{\mathrm{sCM}} = \nabla_\theta \mathcal{L}_{\mathrm{sCM}}^{\mathrm{joint}}, \qquad G_{\mathrm{DMD}} = \nabla_\theta \mathcal{L}_{\mathrm{DMD}}^{\mathrm{joint}} \tag{26}$$

denote the parameter gradients of the trajectory-level and distribution-level objectives. For a local first-order interpretation, we consider an idealized gradient-descent update:

$$\Delta\theta_{\mathrm{direct}} = -\eta \left(\lambda_{\mathrm{sCM}} G_{\mathrm{sCM}} + \lambda_{\mathrm{DMD}} G_{\mathrm{DMD}}\right), \tag{27}$$

where η is the step size.

The component of this update along the sCM gradient is

$$\langle G_{\mathrm{sCM}}, \Delta\theta_{\mathrm{direct}} \rangle = -\eta \left(\lambda_{\mathrm{sCM}} \|G_{\mathrm{sCM}}\|_2^2 + \lambda_{\mathrm{DMD}} \langle G_{\mathrm{sCM}}, G_{\mathrm{DMD}} \rangle\right). \tag{28}$$

The cross term

$$\langle G_{\mathrm{sCM}}, G_{\mathrm{DMD}} \rangle \tag{29}$$

therefore determines whether distribution matching locally reinforces or interferes with trajectory learning. Their normalized alignment can be characterized by

$$A(\theta) = \frac{\langle G_{\mathrm{sCM}}, G_{\mathrm{DMD}} \rangle}{\|G_{\mathrm{sCM}}\|_2 \, \|G_{\mathrm{DMD}}\|_2}. \tag{30}$$

When $\langle G_{\mathrm{sCM}}, G_{\mathrm{DMD}} \rangle < 0$, the DMD term reduces the first-order descent contribution toward the sCM objective relative to an sCM-only update. Conversely, positive alignment can reinforce this descent. We therefore do not assume that sCM and DMD are inherently conflicting; rather, this decomposition shows why their relative alignment can matter during early optimization.

B.2 Motivation for the Staged Schedule

sCM and DMD emphasize different aspects of the teacher distribution. sCM transfers continuous teacher-trajectory information and supports broad trajectory coverage, while DMD provides distribution-level supervision that favors teacher-like final samples. Optimizing both objectives from the base initialization exposes trajectory acquisition to the DMD gradient before the student has learned a sufficiently structured few-step trajectory.

TurboT2VA instead uses the staged objective

$$\mathcal{L}(k) = \begin{cases} \mathcal{L}_{\mathrm{dCM}}, & 1 \le k \le K_1, \\ \mathcal{L}_{\mathrm{sCM}}^{\mathrm{joint}}, & K_1 < k \le K_2, \\ \lambda_{\mathrm{sCM}} \mathcal{L}_{\mathrm{sCM}}^{\mathrm{joint}} + \lambda_{\mathrm{DMD}} \mathcal{L}_{\mathrm{DMD}}^{\mathrm{joint}}, & k > K_2, \end{cases} \tag{31}$$

with $K_1 = 2{,}000$ and $K_2 = 2{,}500$ in the main training schedule.

The first stage establishes a coarse denoising prior without JVP-based trajectory-tangent estimation. The second stage introduces continuous-time trajectory supervision before distribution matching is active. Only in the final stage is the DMD gradient added, after the student has already acquired a structured trajectory prior. This schedule separates trajectory acquisition from distribution refinement during early training while retaining both objectives in the final stage.

B.3 Empirical Interpretation

Figure 5(a) shows that direct sCM+DMD training from the base initialization is feasible, confirming that joint optimization itself is not unstable. However, the staged schedule overtakes the direct route after entering the final stage and maintains stronger late-training Javis scores at the common 7,500-step endpoint.

Figure 5(b) evaluates a complementary question by fixing the same dCM+sCM prefix and varying only the final-stage objective. Joint sCM+DMD optimization achieves stronger late-training performance than either sCM-only continuation or DMD refinement. Together, these observations support the use of staged initialization followed by joint trajectory- and distribution-level optimization.

The gradient analysis above should be interpreted as a local optimization view of these empirical results rather than as a direct measurement of gradient conflict. In particular, we do not claim that $A(\theta) < 0$ throughout early training, since gradient-alignment statistics are not explicitly measured in our current experiments.

B.4 Training Endpoints

The selected TurboT2VA checkpoint is trained for 7,000 global steps: 2,000 dCM warm-up steps, 500 sCM refinement steps, and 4,500 joint sCM+DMD steps. The training-curve ablations in Figure 5 are extended to a common 7,500-step endpoint for controlled late-training comparison. For the staged schedule, the 7,500-step endpoint therefore contains 5,000 joint-stage steps.

The 7,500-step endpoint is used only for comparing optimization trajectories, whereas the main quantitative results report the selected 7,000-step student.

C Additional Experimental Details

This section provides additional training details and complementary analyses for the experiments in Section 6.

C.1 Training Configuration

TurboT2VA is trained on 8 NVIDIA H20 GPUs with per-GPU batch size 1, giving an effective batch size of 8. The selected four-step student follows the progressive curriculum in Section 4: 2K dCM warm-up steps, 0.5K sCM refinement steps, and 4.5K joint sCM+DMD optimization steps, for 7K optimization steps in total.

The joint stage uses AdamW with a learning rate of 2×10^{-5}, $\beta_1 = 0.0$, $\beta_2 = 0.999$, weight decay 0.01, and equal sCM and DMD coefficients. Teacher predictions use video and audio guidance scales of 3.0 and 5.0, respectively. Training uses BF16 mixed precision, gradient checkpointing, and FSDP. The full 7K training run takes approximately 21 hours on 8 H20 GPUs, excluding evaluation and checkpoint-sweep overhead.

C.2 Modality-Specific Quality Analysis

We further compare the 40-step teacher and four-step TurboT2VA using modality-specific fidelity metrics. As shown in Table 5, TurboT2VA improves video aesthetic quality, imaging quality, and subject consistency while closely preserving motion smoothness and temporal consistency. Audio results remain broadly comparable: content usefulness and production quality improve slightly, while content enjoyment, production complexity, and CLAP decrease.

Table 5: Modality-specific fidelity comparison between the 40-step teacher and four-step TurboT2VA. Best results are underlined.

Model	Video Fidelity					Audio Fidelity				
	Aes.↑	Img.↑	Mot.↑	Subj.↑	Temp.↑	Enj.↑	Use.↑	Comp.↑	Qual↑	CLAP↑
Teacher (40 steps)	0.529	0.588	0.992	0.959	0.984	4.765	6.307	3.275	6.577	0.353
Ours (4 steps)	0.552	0.652	0.991	0.961	0.977	4.612	6.334	2.964	6.584	0.347

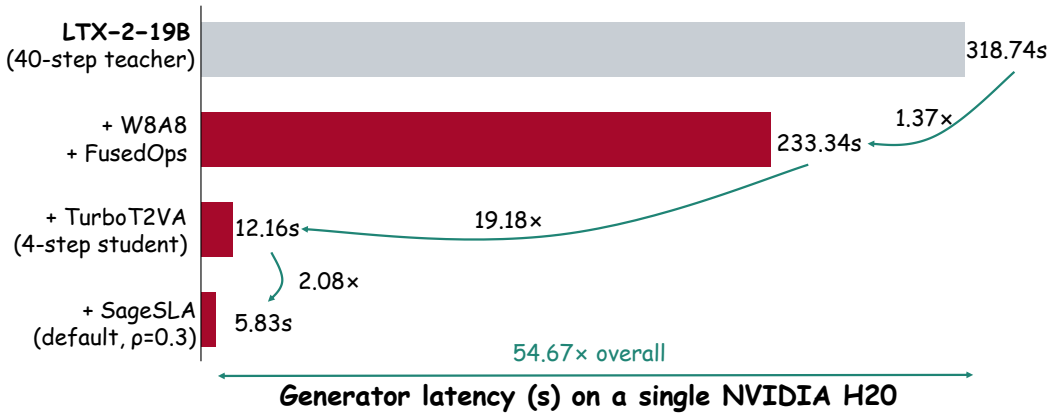

Figure 6: Cumulative high-resolution inference acceleration. Starting from the dense 40-step teacher, W8A8 and fused operations reduce generator latency from 318.74s to 233.34s. Replacing the teacher trajectory with the four-step TurboT2VA student under the same operator stack reduces latency to 12.16s, while modality-aware sparse attention and padded-text compaction further reduce it to 5.83s.

C.3 Diversity Evaluation Details

For each of the $P = 200$ evaluation prompts, we generate $K = 8$ video-audio samples using different random seeds. Video and audio are encoded separately using a fixed pretrained ImageBind encoder. Pairwise distances are computed only among samples generated from the same prompt.

Let $e^m_{p,i}$ denote the ImageBind embedding of sample i for prompt p and modality $m \in \{v, a\}$. We define the modality-specific diversity score as the average within-prompt cosine distance:

$$D_m = \frac{1}{P}\sum_{p=1}^{P} \frac{2}{K(K-1)} \sum_{1\le i<j\le K} \left(1 - \frac{\langle e^m_{p,i}, e^m_{p,j}\rangle}{\left\|e^m_{p,i}\right\|_2 \left\|e^m_{p,j}\right\|_2}\right). \tag{32}$$

With $K = 8$, this gives $\binom{8}{2} = 28$ pairs per prompt and 5,600 pairs per modality over the full evaluation set.

We report video and audio diversity separately and define their arithmetic mean as

$$D_{\text{VA}} = \frac{D_{\text{video}} + D_{\text{audio}}}{2}. \tag{33}$$

A larger within-prompt distance indicates greater cross-seed variation under the same text condition, but does not by itself imply higher generation quality. We therefore interpret diversity jointly with the Javis score in Table 2.

C.4 Cumulative High-Resolution Acceleration

Figure 6 shows the cumulative acceleration path at 1024×1792 on one NVIDIA H20. All stages are cumulative and use generator-only latency, excluding checkpoint loading, VAE decoding, video-audio muxing, and disk I/O.

Table 6: Sampling-step ablation for TurboT2VA. Best and second-best results are shown in **bold** and underlined, respectively.

Steps	Visual ↑	Motion ↑	Audio ↑	IB-AV ↑	CLIP ↑	CLAP ↑	CAVP ↑	AVH ↑	Javis ↑	Desync ↓
1	2.0576	0.6680	4.0746	0.0959	0.2915	0.2367	0.7955	0.0847	0.0663	0.6220
2	2.2328	0.7269	4.3416	0.1915	**0.2926**	0.3198	0.7959	0.1845	0.1576	**0.3780**
4	**2.3892**	**0.8493**	**4.4860**	**0.2290**	0.2921	**0.3472**	**0.8033**	**0.2241**	**0.1963**	0.3880

The complete stack achieves a 54.67× generator-only speedup relative to the dense 40-step teacher. For comparison, the unoptimized four-step student requires 16.50s, so architecture-aware inference optimization contributes an additional 2.83× acceleration beyond sampling-step reduction alone.

C.5 Sampling-Step Ablation

We evaluate the same TurboT2VA student using one, two, and four sampling steps. Table 6 shows that four-step inference achieves the strongest results on most quality and cross-modal metrics, while two-step inference already provides competitive performance and obtains slightly better CLIP and Desync scores.