MosaicAudio: Multi-Teacher On-Policy Distillation for Controllable Video-to-Audio Generation

Anonymous ACL submission

Abstract

Video-to-audio (V2A) generation aims to produce a single soundtrack that is simultaneously semantically faithful, temporally synchronized, aesthetically natural, and—for stereo output—spatially accurate. These perceptual axes are judged on the same waveform yet often pull generation in locally conflicting directions, while the dominant post-training recipe—reinforcement learning (RL) from an aggregated scalar reward—supervises them only at the final waveform with one entangled signal that gives little indication of which axis failed or where, so competing axes interfere through a single advantage value. We present MosaicAudio, a V2A framework that replaces terminal scalar rewards with dense, per-axis supervision distilled on-policy. MosaicAudio first trains four specialist flow-matching teachers, one per perceptual axis, then consolidates them into a single student: the student samples its own denoising trajectories and, at each visited state, receives a full teacher velocity field routed to one axis at a time, so that incompatible corrections are not averaged together. A perceptual manifold anchor keeps the student on a high-fidelity audio prior and curbs reward hacking during alignment. On the in-domain VGGSound test set and the out-of-domain AudioCanvas benchmark, MosaicAudio achieves state-of-the-art performance among compared V2A systems across objective metrics and human ratings, with especially consistent gains in complex multi-event scenes. Demo page: https://MosaicAudio-project.github.io.

1 Introduction

Video-to-Audio (V2A) generation, or video foley, synthesizes a soundtrack for a silent video, optionally guided by text. The difficulty is not producing plausible sound but satisfying four perceptual demands at once on the *same* waveform: naming the right events (*semantic* faithfulness), striking them at the right instants (*temporal* synchrony), sounding natural and usable (*aesthetic* quality), and placing sources in the stereo field (*spatial* accuracy). These axes routinely conflict at fine acoustic scale: rounding a transient for smoothness blurs the impact that synchrony needs, and widening the stereo image to localize a source perturbs the cues that shape its timbre. Optimizing one axis in isolation tends to regress another.

V2A systems have moved from visual-only synthesis (Luo et al., 2023; Wang et al., 2024a) to controllable multimodal generation with stronger text, video, and audio-language conditioning (Cheng et al., 2024a; Tian et al., 2025; Polyak et al., 2024; Wang et al., 2024b; Shan et al., 2025). ThinkSound (Liu et al., 2025a) makes generation explicit through a chain-of-thought (CoT) (Wei et al., 2022) plan, and PrismAudio (Liu et al., 2026) post-trains such a planner-conditioned generator with reinforcement learning (RL) against per-axis reward models. This line of work shows that explicit reasoning and perceptually grounded rewards both help.

Yet these systems share a structural limitation in how the reward reaches the generator. RL post-training scores a finished waveform and collapses the axes into one aggregated scalar advantage. That signal is *terminal*—only the final waveform is scored, so the same advantage is broadcast to every denoising step rather than telling each state how to move; *scalar*—one number cannot localize a failure to event identity, timing, texture, or placement; and *fixed*—a single set of weights governs every clip and every step. For coupled audio axes this is the wrong granularity: the corrections needed are local and often opposite—sharpen this onset, smooth that texture—so a terminal scalar both hides where the error lies and lets competing axes cancel through one advantage value.

We argue that multi-axis V2A alignment needs supervision that is dense, axis-specific, and applied on the trajectory the model actually follows. We

present MosaicAudio, which turns each axis into a specialist teacher and consolidates the teachers into one student by on-policy distillation. Four flow-matching teachers—semantic, temporal, aesthetic, and spatial—are each aligned to one reward; the student then samples its own denoising trajectories and, at every visited state, receives a full teacher velocity field as dense supervision, routed to a single axis at a time so incompatible corrections are not directly averaged together. Because teachers and student share one latent space and sampler, this guidance reduces to a tractable, timestep-weighted velocity-field discrepancy usable as a dense on-policy reward (Lu and Lab, 2025; Fang et al., 2026), and a frozen perceptual anchor keeps the student on a high-fidelity manifold. The name reflects the design: like a mosaic, each axis is reasoned about and supervised as a separate tile, then fused into one coherent scene by a single student that, at inference, runs as an ordinary generator with no teachers or reward models.

Our contributions are summarized as:

- We identify why aggregated scalar RL is mismatched to V2A's coupled perceptual axes—it is terminal, scalar, and applies a fixed trade-off—and argue for dense, axis-routed, on-policy supervision instead.
- We introduce MosaicAudio, which consolidates semantic, temporal, aesthetic, and spatial specialists through multi-teacher on-policy distillation, using axis-routed dense velocity rewards and a perceptual manifold anchor to jointly address reward sparsity, gradient interference, and audio-quality collapse.
- On the in-domain VGGSound test set and the out-of-domain AudioCanvas benchmark, MosaicAudio improves over strong V2A baselines across all four axes and in human ratings, with the largest gains in complex multi-event scenes where the axes most conflict.

2 Related Work

Video-to-Audio Generation. V2A generation has evolved from reconstructive visual-to-audio mappings to controllable multimodal generation. Diff-Foley (Luo et al., 2023) and V2A-Mapper (Wang et al., 2024a) rely primarily on visual representations, while later work adds stronger text, video, and audio-language conditioning (Mo et al., 2024; Tian et al., 2025; Polyak et al., 2024; Cheng et al., 2024a; Shan et al., 2025) and explicit foley control (Zhang et al., 2024; Xing et al., 2024; Mei et al., 2024b; Xie et al., 2024; Viertola et al., 2025). Most build on diffusion or flow-matching audio generators (Liu et al., 2023a; Copet et al., 2024). These systems improve controllability but typically operate as opaque generators, making it hard to diagnose which perceptual objective failed. ThinkSound (Liu et al., 2025a) introduces explicit CoT planning (Wei et al., 2022) for V2A, improving interpretability and semantic coherence, and PrismAudio (Liu et al., 2026) pairs a decomposed CoT plan with per-axis reward models to align the generator across multiple perceptual objectives via reinforcement learning. MosaicAudio builds on this decomposed-reasoning interface but supervises each axis with a specialist teacher, rather than aggregating per-axis rewards into a single scalar advantage.

RL and Distillation for Generative Models. Reinforcement learning aligns generative models with non-differentiable or preference objectives, from RLHF for language models (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024; Guo et al., 2025) to reward optimization for diffusion and flow models (Black et al., 2023; Fan et al., 2023; Wallace et al., 2024; Liu et al., 2025b; Xue et al., 2025; Li et al., 2025). Built on flow matching (Lipman et al., 2022), these methods are effective for a single target reward but become unstable when heterogeneous objectives are mixed into one scalar signal. On-policy distillation (OPD) (Lu and Lab, 2025) instead trains the student on samples from its own policy with dense teacher supervision, mitigating distribution shift; it has recently been formulated for flow-matching generators by turning a teacher's velocity field into a dense trajectory-level reward (Fang et al., 2026). We bring this dense, on-policy paradigm to controllable V2A generation, where a single output is aligned across several coupled perceptual axes at once.

3 Method

3.1 Overview

MosaicAudio post-trains a single V2A generator to satisfy all four perceptual axes at once, in three parts (Figure 1). (i) A *decomposed-CoT backbone* renders audio from video and a four-axis plan; we build on PrismAudio (Liu et al., 2026), a state-of-the-art open-source V2A system, as this backbone (Section 3.2). (ii) We construct four *specialist teachers*, one per perceptual axis, by single-reward

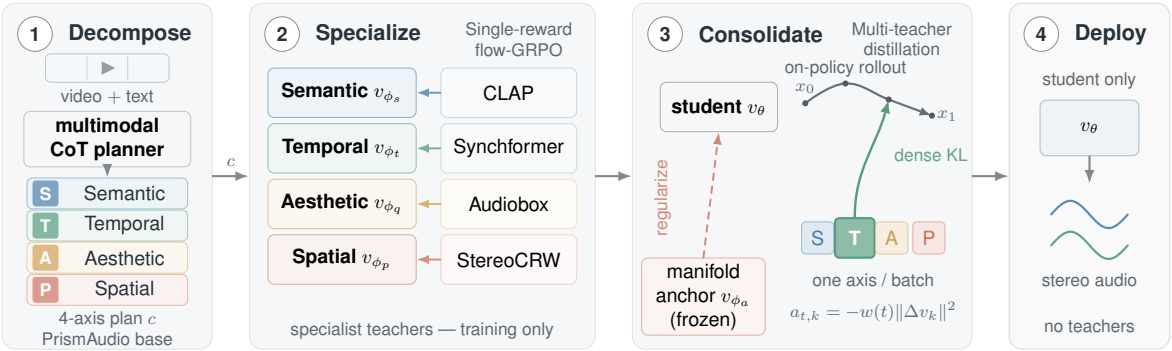

Figure 1: Overview of MosaicAudio, which replaces a terminal scalar reward with dense, per-axis, on-policy velocity supervision. **(1) Decompose** the soundscape into Semantic, Temporal, Aesthetic, and Spatial CoT blocks that condition the generator. **(2) Specialize** one flow-matching teacher per axis by single-reward alignment. **(3) Consolidate** the teachers into one student by multi-teacher on-policy distillation: at each visited denoising state the student receives a dense, axis-routed teacher velocity field as a velocity-KL reward, while a frozen perceptual anchor preserves fidelity; a balanced scheduler activates one axis (**S/T/A/P**, labeled in stage 1) per mini-batch. Teachers and reward models are used only in post-training; at inference the student alone maps video and plan to stereo audio.

alignment (Section 3.4). (iii) We consolidate them into a single *student* by multi-teacher on-policy distillation—dense, axis-routed velocity supervision on the student's own trajectories, with a frozen perceptual anchor (Sections 3.3–3.6). Our contributions are (ii)–(iii): per-axis specialist teachers and, centrally, the routed dense distillation that turns coupled perceptual axes into local velocity targets, together with the perceptual anchor.

3.2 Base Generator and Perceptual Decomposition

We build MosaicAudio on PrismAudio (Liu et al., 2026), a state-of-the-art open-source decomposed-CoT V2A system, as our backbone; we summarize it here. It is a flow-matching (Lipman et al., 2022) diffusion transformer (Peebles and Xie, 2023; Liu et al., 2023b) whose velocity field $v_\theta(x_t, t, c)$ at latent state x_t, timestep t, and condition c is trained by supervised flow matching,

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x_0,x_1,c}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2\right], \tag{1}$$

with $x_t = (1-t)x_0 + tx_1$ on the optimal-transport path. The condition c combines VideoPrism (Zhao et al., 2024) video features with structured text from a four-axis CoT plan encoded by T5-Gemma (Zhang et al., 2025). The plan decomposes the soundscape into **Semantic** (visible events, materials, actions), **Temporal** (onsets, ordering, synchronization), **Aesthetic** (fidelity, naturalness, production quality), and **Spatial** (direction, distance, motion) blocks generated by a multimodal planner (Comanici et al., 2025; Cheng et al., 2024b); at inference the blocks are concatenated in a fixed order (the annotation template is in Appendix B.1). The latent x_1 is decoded by a stereo autoencoder, so it carries the inter-channel cues the spatial axis depends on. Two properties make this backbone a clean substrate for our method: the planner is trained once and frozen, so post-training acts on the generator v_θ alone; and every teacher uses the same full-plan interface, differing only in its active-axis tag and reward used for alignment. Our contribution lies entirely in the post-training procedure below: per-axis teacher construction, routed dense distillation, and a perceptual anchor.

3.3 From Scalar Rewards to Dense Routed Supervision

The standard way to optimize the four axes is to sum their scalar rewards and run policy-gradient post-training. This is brittle for V2A because the reward vector is not merely multi-objective; its entries are measured at different perceptual levels. CLAP may reward a correct event label even when transients are late; an aesthetic score may prefer smooth texture that masks impact sounds; a stereo metric may improve localization while shifting timbre. Compressing these observations into one terminal advantage discards *which* part of the trajectory should change.

The failure is also visible at the gradient level. With a summed reward $R = \sum_k \lambda_k R_k$, the four axes share one advantage and one update direction; where two axes disagree, their gradients conflict

($\langle \nabla \mathcal{J}_i, \nabla \mathcal{J}_j \rangle < 0$) and the optimizer is free to trade one for another, with no signal that the sacrificed axis got worse (Fang et al., 2026). Re-weighting λ_k only moves the operating point along this trade-off; it does not remove the coupling.

MosaicAudio avoids this scalar mixing by never summing the axes, and separates two roles. Exploration stays on-policy: the student samples its own latent trajectory and exposes the states where it currently fails. Supervision becomes dense and routed: at each visited state, one teacher provides a full velocity target for a single axis. Each update then answers a local question—"how should this state move to satisfy this axis?"—rather than a global one—"was the final waveform better under a mixed scalar?" This matters for flow matching, where small intermediate errors accumulate into audible artifacts. The goal is thus not to add another reward term, but to raise the *bandwidth* (a full velocity field instead of one number) and *locality* (per-axis, per-state instead of terminal) of the signal the generator receives.

3.4 Specialist Teachers

We attach one reward function to each axis: semantic alignment from MS-CLAP (Elizalde et al., 2024), temporal synchrony from Synchformer (Iashin et al., 2024), audio aesthetics from Audiobox-Aesthetics (Tjandra et al., 2025), and stereo accuracy from StereoCRW (Chen et al., 2022). For each axis k we train one specialist teacher v_{ϕ_k} from the base with a single-reward, windowed flow-GRPO objective (Shao et al., 2024; Liu et al., 2025b) that confines stochastic exploration to a random window of denoising steps for efficiency (Li et al., 2025).

For each prompt, the old policy samples a group of G audio candidates, and rewards are normalized within the group:

$$A_k^i = \frac{R_k(x_1^i, c_k) - \bar{R}_k}{s_k + \epsilon}, \tag{2}$$

where $\bar{R}_k, s_k$ are the group mean/std of R_k (distinct from the noise scale σ_t), and $c_k = (c_{\text{full}}, e_k)$ is the full four-block plan with an axis-k tag e_k—all teachers share the plan and differ only in this tag. We optimize only a random window $\mathcal{W}(\ell) = \{\ell, \ldots, \ell + w - 1\}$ of denoising steps and integrate deterministically elsewhere. The teacher objective is

$$\mathcal{J}_k = \mathbb{E}\Big[\frac{1}{Gw} \sum_{i=1}^{G} \sum_{t \in \mathcal{W}(\ell)} \min\Big(r_t^i A_k^i, \operatorname{clip}(r_t^i, 1-\varepsilon, 1+\varepsilon) A_k^i\Big)\Big], \tag{3}$$

where $r_t^i = \pi_{\phi_k}(x_{t+\Delta t}^i \mid x_t^i, c_k) / \pi_{\phi_k^{\text{old}}}(x_{t+\Delta t}^i \mid x_t^i, c_k)$ is the transition-level policy ratio. Restricting exploration to a window of width $w \ll T$ cuts the policy-model evaluations per sample from T to w, which is what makes training four teachers affordable; the steps outside the window use the deterministic ODE, preserving the terminal distribution required for a valid reward. Each teacher is initialized from the cold-start checkpoint, so the four specialists and the student share one parameter space and differ only in the axis each was driven toward. These specialists are intentionally narrow: each exposes high-quality gradients for one axis, even when that specialization is not globally balanced—a property we exploit in the next step rather than deploy directly. Because all teachers also share the student's latent space and sampler, each defines a comparable local transition that the student can be supervised against.

3.5 Multi-Teacher On-Policy Distillation

The student integrates all specialists without collapsing them into a scalar mixture: it explores on-policy and is supervised densely, with the teacher–student velocity discrepancy serving as a dense on-policy reward, following on-policy distillation for flow models (Lu and Lab, 2025; Fang et al., 2026). We describe the four components in turn.

On-policy trajectory sampling. The current student generates G stochastic trajectories under the conditioning it will see at inference. Following the SDE view of flow sampling (Liu et al., 2025b), a transition inside the window $\mathcal{W}(\ell)$ is a Gaussian policy

$$\pi_\theta(x_{t+\Delta t} \mid x_t, c) = \mathcal{N}\left(\mu_\theta(x_t, t, c), \sigma_t^2 \Delta t I\right), \tag{4}$$

where μ_θ is set by the SDE drift parameterized by v_θ. We update the policy only through these in-window transition log-probabilities and stop gradients through the sampled states, so teachers are queried exactly where the student currently goes—supervision stays valid even as the student drifts from the supervised distribution.

Dimension routing. Because the four axes apply jointly to every clip rather than partitioning prompts into disjoint tasks, routing by prompt is not applicable here; instead a balanced cyclic scheduler assigns each mini-batch to a single axis k, cycling through all four. The routed teacher supplies a target transition policy

$$\pi_{\phi_k}(x_{t+\Delta t} \mid x_t, c_k) = \mathcal{N}\left(\mu_{\phi_k}(x_t, t, c_k), \sigma_t^2 \Delta t I\right). \tag{5}$$

Each teacher adds its active-axis tag to the full plan (c_k); the student keeps the tag-free plan c it uses at inference, so teacher differences come from reward alignment, not the interface, and the student has no train–test conditioning mismatch. Supervising one axis at a time avoids averaging teacher fields, which can cancel locally incompatible corrections—e.g., a temporal teacher sharpening a transient that an aesthetic teacher would smooth—while the balanced schedule ensures every axis is revisited.

Dense KL score. We score a transition by its student-to-teacher KL, $D_{\mathrm{KL}}(\pi_\theta \| \pi_{\phi_k})$, which directly penalizes the current student transition for deviating from the routed teacher transition. Since student and teacher share covariance, this KL reduces to a closed-form velocity-field distance. With $\Delta v_k(x_t, t) = v_\theta(x_t, t, c) - v_{\phi_k}(x_t, t, c_k)$,

$$\begin{aligned} D_{\mathrm{KL}}\left(\pi_\theta \| \pi_{\phi_k}\right) &= \frac{\|\mu_\theta - \mu_{\phi_k}\|_2^2}{2\sigma_t^2 \Delta t} \\ &= w(t)\, \|\Delta v_k(x_t, t)\|_2^2, \end{aligned} \tag{6}$$

and we take its negative as a dense, axis-specific score,

$$s_{t,k} = -w(t)\, \|\Delta v_k(x_t, t)\|_2^2. \tag{7}$$

Under the Euler–Maruyama discretization of our sampler, the weight is

$$w(t) = \frac{\Delta t}{2}\left(\frac{\sigma_t(1-t)}{2t} + \frac{1}{\sigma_t}\right)^2, \tag{8}$$

so the signal is larger at timesteps where a velocity mismatch more strongly changes the transition. To keep $w(t)$ finite we clip $t \in [\tau, 1]$ ($\tau > 0$) and floor $\sigma_t \leftarrow \max(\sigma_t, \sigma_{\min})$, so neither $1/t$ nor $1/\sigma_t$ diverges. Unlike regressing to offline teacher samples, this score corrects the student where it actually visits.

Clipped on-policy objective. The dense score $s_{t,k} \leq 0$ is non-positive, so we do not use it directly: we center it within each prompt's group of G rollouts: with the group mean $\bar{s}_{tj} = \frac{1}{G}\sum_i s_{tij}$, the advantage $\hat{A}_{tij} = s_{tij} - \bar{s}_{tj}$ removes the uniform negative offset and ranks the group's trajectories by teacher–student discrepancy. The student then maximizes a clipped surrogate over sampled states,

$$\begin{aligned} \mathcal{J}_{\mathrm{OPD}} = \mathbb{E}\Big[\frac{1}{BGw}\sum_{j=1}^{B}\sum_{i=1}^{G}\sum_{t\in\mathcal{W}(\ell)} \min\Big(\rho_{tij}\hat{A}_{tij}, \\ \mathrm{clip}(\rho_{tij}, 1-\varepsilon, 1+\varepsilon)\hat{A}_{tij}\Big)\Big], \end{aligned} \tag{9}$$

where the policy ratio $\rho_{tij} = \pi_\theta / \pi_{\theta_{\mathrm{old}}}$ is closed-form (shared-covariance Gaussians) and the clip bounds each on-policy update; Appendix A gives the ratio, the KL derivation, and the centered advantage. The objective delivers local, full-dimensional supervision from the routed specialist while keeping on-policy exploration. Because the student consolidation sees only a velocity-field discrepancy, heterogeneous raw reward scales do not enter this update, and each step queries one teacher (plus the anchor), so cost is linear in routed samples, not in the teacher bank.

3.6 Perceptual Manifold Anchor

Functional specialists can over-optimize their metrics and erode naturalness—a velocity-field analogue of reward hacking. We therefore keep a frozen anchor teacher v_{ϕ_a}, initialized from the cold-start base generator in all experiments, and regularize the student toward its perceptual manifold on *all* sampled states (not only the routed axis). Anchoring to a frozen high-quality teacher rather than a generic pretrained prior keeps the regularizer in the same conditioned velocity space as the functional teachers, so the two terms are directly comparable and reuse the same weight $w(t)$. With $\Delta v_a(x_t, t) = v_\theta(x_t, t, c) - v_{\phi_a}(x_t, t, c)$, the full objective is

$$\mathcal{L}_{\mathrm{total}} = -\mathcal{J}_{\mathrm{OPD}} + \lambda \mathbb{E}_{x_t,t,c}\big[w(t)\, \|\Delta v_a(x_t, t)\|_2^2\big]. \tag{10}$$

The weight λ behaves as an elastic anchor: too small and the functional teachers over-optimize their proxies (Section 4.3), too large and the student cannot move off the base. In summary, training proceeds in four stages: (1) instantiate a fixed decomposed-CoT backbone following PrismAudio as the cold start; (2) align four teachers from it, one

per axis; (3) consolidate a student with the routed dense-KL objective; and (4) regularize throughout with the anchor. At inference no teacher is used: a single student takes the video and concatenated plan and generates stereo audio with the standard flow sampler.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate in two regimes. *In-domain*, we use the VGGSound (Chen et al., 2020) test set, drawn from the same source distribution as our cold-start training and CoT annotations. *Out-of-domain*, we use AudioCanvas, the real-world V2A benchmark introduced by the CoT-RL system PrismAudio (Liu et al., 2026): it contains curated videos with strict off-screen-sound filtering, both single- and multi-event scenes, and structured four-axis CoT annotations. We use AudioCanvas strictly for evaluation—none of its clips or their source videos appear in our training or annotation data—so it measures generalization beyond the training distribution, which is where coupled-axis trade-offs are most visible.

Metrics. We use both objective and subjective metrics covering the four perceptual axes. For *objective* evaluation, semantic consistency uses CLAP (Elizalde et al., 2024); temporal alignment uses DeSync from Synchformer (Iashin et al., 2024) (lower is better); aesthetic quality uses the Audiobox-Aesthetics scores (Tjandra et al., 2025), including production quality (PQ) and content enjoyment (CE); spatial accuracy uses GCC (Knapp and Carter, 1976) and CRW (Chen et al., 2022) localization errors; and distributional similarity uses FD (Kilgour et al., 2018) and KL (Copet et al., 2024). For *subjective* evaluation, we run two 5-point mean-opinion-score studies on randomized, identity-hidden clips: MOS-Q rates standalone audio fidelity and naturalness, while MOS-C rates audio–video consistency in semantics, timing, and stereo placement. We sample 80 videos per test set, average ratings from multiple annotators, and report mean and standard deviation; the full protocol is in Appendix D.2.

Baselines. We compare with Frieren (Wang et al., 2024b), V2A-Mapper (Wang et al., 2024a), AudioX (Tian et al., 2025), HunyuanVideo-Foley (Shan et al., 2025), MMAudio (Cheng et al., 2024a), ThinkSound (Liu et al., 2025a), and the CoT-RL system PrismAudio (Liu et al., 2026), which post-trains a decomposed-CoT generator with multi-dimensional scalar rewards and is the strongest RL-aligned baseline in our setting. *MosaicAudio w/o OPD* denotes the CoT-aware cold-start model before multi-teacher consolidation.

Implementation. The student and all specialist teachers share one flow-transformer and a stereo latent autoencoder; each teacher is initialized from the cold-start model, trained on a single reward axis, and frozen during consolidation. We sample $G{=}12$ trajectories per prompt, confine stochastic exploration to a window of $w{=}3$ steps within the $T{=}25$-step sampler, clip the policy ratio at $\varepsilon{=}0.2$, and weight the anchor at $\lambda{=}0.5$; the cyclic router activates one axis per mini-batch. Full hyperparameters are in Appendix B.

4.2 Main Results

Table 1 reports the headline metrics on both benchmarks; the full objective and subjective metrics are in Appendix C.1 (Tables 5 and 6).

In-domain results (VGGSound). (1) MosaicAudio is the strongest generated model on the reported axes. Against the prior CoT baseline ThinkSound it improves semantics (CLAP $0.43 \rightarrow 0.50$), synchrony (DeSync $0.55 \rightarrow 0.38$), aesthetics (CE $3.95 \rightarrow 4.64$), and stereo placement (CRW $13.47 \rightarrow 7.11$). It also edges past the strongest RL-aligned baseline, PrismAudio (CLAP $0.47 \rightarrow 0.50$, DeSync $0.41 \rightarrow 0.38$, CRW $7.72 \rightarrow 7.11$, MOS-C $4.22 \rightarrow 4.25$), despite the two sharing the same decomposed-CoT interface—so the gain is attributable to the post-training rule, not the conditioning. **(2) The gain is driven by OPD, not the backbone.** The cold start (MosaicAudio w/o OPD) is the pre-consolidation foundation; multi-teacher OPD lifts it across the board (CLAP $0.42 \rightarrow 0.50$, DeSync $0.51 \rightarrow 0.38$, CE $3.94 \rightarrow 4.64$, CRW $10.29 \rightarrow 7.11$, MOS-C $4.11 \rightarrow 4.25$), so the routed dense supervision—not extra parameters—closes the gap to and past scalar CoT-RL. **(3) The improvement is free at inference.** Teachers and reward models are used only in post-training: the deployed student is the same 518M-parameter generator running a single $T{=}25$-step pass, whereas keeping the four teachers (or a reward model) in the loop would multiply per-step cost by the number of active models.

Table 1: Key results on VGGSound (in-domain) and AudioCanvas (out-of-domain). *MosaicAudio w/o OPD* is the shared decomposed-CoT backbone before any post-training—the same pre-RL cold-start that the scalar baseline starts from. "-" marks a method we do not evaluate on that benchmark, or a mono baseline without a stereo channel (CRW). Ground truth (GT) is shown for reference only: the automatic proxies are not calibrated upper bounds, so post-training can exceed GT on the metrics it directly optimizes (CLAP, DeSync, Audiobox scores), while GT stays highest on human MOS. Full metrics are in Appendix C.1.

	VGGSound (in-domain)					*AudioCanvas* (out-of-domain)				
Method	CLAP↑	DeSync↓	CE↑	CRW↓	MOS-C↑	CLAP↑	DeSync↓	CE↑	CRW↓	MOS-C↑
Ground truth	0.46	0.55	4.40	-	4.65	0.48	0.40	4.02	-	4.72
Frieren	0.32	0.85	3.57	-	3.51	-	-	-	-	-
V2A-Mapper	0.31	1.23	4.12	-	3.44	-	-	-	-	-
AudioX	0.41	1.24	3.86	19.25	3.65	-	-	-	-	-
Hunyuan-Foley	0.42	0.55	3.92	-	3.96	0.44	0.47	4.04	-	3.71
MMAudio	0.40	0.46	3.88	-	4.03	0.46	0.43	3.97	-	3.87
ThinkSound	0.43	0.55	3.95	13.47	4.18	0.48	0.80	4.10	22.82	3.80
PrismAudio	0.47	0.41	4.29	7.72	4.22	0.52	0.36	4.26	12.87	4.01
MosaicAudio w/o OPD	0.42	0.51	3.94	10.29	4.11	0.42	0.44	3.81	15.30	3.85
MosaicAudio	**0.50**	**0.38**	**4.64**	**7.11**	**4.25**	**0.55**	**0.35**	**4.36**	**12.20**	**4.13**

Table 2: Specialist analysis on AudioCanvas (cold start in Tables 1 and 3). **(a)** teachers vs. the unified student, best per column in **bold**; **(b)** leave-one-teacher-out, dropped axis underlined.

(a) Teachers vs. unified student

Model	CLAP↑	DeSync↓	PQ↑	CRW↓	FD↓
Semantic T.	**0.56**	0.58	6.62	11.89	1.84
Temporal T.	0.46	**0.35**	6.39	13.08	1.88
Aesthetic T.	0.46	0.42	**7.06**	13.51	4.50
Spatial T.	0.47	0.42	6.44	**11.88**	**1.77**
MosaicAudio	0.55	0.35	6.75	12.20	1.85

(b) Leave-one-teacher-out

Router	CLAP↑	DeSync↓	PQ↑	CRW↓	FD↓
– Semantic	<u>0.50</u>	0.36	6.73	12.35	1.88
– Temporal	0.54	<u>0.42</u>	6.72	12.45	1.90
– Aesthetic	0.54	0.36	<u>6.50</u>	12.40	1.92
– Spatial	0.54	0.36	6.73	<u>13.60</u>	1.95
MosaicAudio (all)	**0.55**	**0.35**	**6.75**	**12.20**	**1.85**

Out-of-domain results (AudioCanvas). (1) MosaicAudio generalizes to complex scenes. On the harder AudioCanvas the cold start sits well below PrismAudio, yet with no retuning OPD closes the gap and ends slightly ahead on every axis (CLAP $0.52 \rightarrow 0.55$, DeSync $0.36 \rightarrow 0.35$, CRW $12.87 \rightarrow 12.20$, MOS-C $4.01 \rightarrow 4.13$), so the teachers transfer a general per-axis correction rather than a benchmark-specific shortcut. **(2) Gains concentrate where the cold start is weakest.** Timing and spatial recover most (DeSync $0.44 \rightarrow 0.35$, CRW $15.30 \rightarrow 12.20$), as dense per-state guidance best repairs such cues at intermediate denoising steps rather than only at the final waveform.

4.3 Analysis

Are specialist teachers complementary? Panel (a) of Table 2 reports our four single-axis specialist policies—each fine-tuned from the shared cold start on one reward (Section 3.4)—and the distilled student on AudioCanvas. Each specialist reaches the best score on its target metric but regresses on another: the semantic teacher tops CLAP (0.56) yet has the worst synchrony (DeSync 0.58), and the aesthetic teacher tops PQ (7.06) yet more than doubles FD (4.50 vs. 1.85 for MosaicAudio). MosaicAudio is not a teacher average; it is a routed consolidation that stays close to each specialist on that specialist's metric while avoiding the off-axis regressions they incur, so among the models compared here it alone is strong on all axes at once.

Is every axis necessary? Panel (b) removes one teacher at a time and re-runs consolidation; the dropped axis then receives no routed supervision (the router cycles over the remaining three). Each removal degrades its own axis the most—semantic CLAP falls to 0.50, temporal DeSync to 0.42, aesthetic PQ to 6.50, and spatial CRW to 13.60—while the other three axes move only marginally. No teacher is redundant, and the off-axis stability is itself evidence that routing isolates each correction instead of letting it leak across axes.

Ablating the consolidation design. Table 3 ablates our post-training rule; all variants start from the same cold start and the same four axis rewards. *Scalar multi-reward RL*—summing the rewards and running flow-GRPO, the prior scalar CoT-RL recipe—is already strong (on par with that competitor on this set), but the seesaw between axes caps it below the full model. The lower block keeps the multi-teacher OPD setup and removes one design choice at a time: *offline distillation* (regressing to teacher samples instead of sampling on-policy) trails the full model because the student is corrected

Table 3: Ablating the post-training rule on AudioCanvas; all variants share the cold start and four axis rewards, and each lower-block row toggles one design choice of MosaicAudio. **Bold** = best per column; the w/o-anchor PQ is a reward-hacking artifact (note its FD).

Post-training	CLAP↑	DeSync↓	PQ↑	CRW↓	FD↓
Cold start (no PT)	0.42	0.44	6.45	15.30	2.10
Scalar multi-reward RL	0.52	0.36	6.68	12.87	1.92
Offline distillation	0.52	0.37	6.70	12.60	1.95
OPD, teacher average	0.51	0.38	6.72	12.50	2.05
OPD, routed, w/o anchor	0.50	0.39	**7.00**	13.10	3.50
MosaicAudio (full)	**0.55**	**0.35**	6.75	**12.20**	**1.85**

where the teacher data lies, not where it actually drifts; a *teacher average* improves over the cold start but trails hard routing, most on FD (2.05 vs. 1.85), since blending the four velocity fields dilutes locally incompatible corrections; and *removing the anchor* triggers reward hacking—PQ rises while FD collapses to 3.50. Only the full configuration—on-policy, hard-routed, and anchored—attains the best functional scores without the FD collapse, so each design choice is necessary.

Routing schedule. The balanced cyclic router visits each axis equally often, one axis per mini-batch. Two alternatives are worse on AudioCanvas. *Random* per-batch routing under the same teacher budget slightly destabilizes the under-visited axes (DeSync $0.35 \rightarrow 0.37$, CRW $12.20 \rightarrow 12.45$), since within a short window some axes may go unsupervised, and the uniform *teacher average* (Table 3) trails further still (DeSync 0.38, CRW 12.50). Equal cyclic coverage is thus a simple and effective default.

Scene complexity. We split AudioCanvas into multi- and single-event subsets. The OPD margin is larger in multi-event scenes, where several constraints apply at once: against PrismAudio, MosaicAudio improves on every axis (e.g., CLAP $0.50 \rightarrow 0.53$, DeSync $0.39 \rightarrow 0.37$, CRW $13.50 \rightarrow 12.60$), whereas the single-event gains, though consistent, are smaller. This mirrors our motivation: dense routed supervision helps most when several coupled axes need correcting together.

Consistent improvement across axes. Figure 2 plots the relative change from the cold start to MosaicAudio on AudioCanvas: every axis improves and none regresses. This is the intended behavior of multi-teacher OPD—it absorbs each specialist's correction without inheriting any one specialist's failure mode, in contrast to panel (a) of Table 2, where every single-axis teacher degrades at least one other metric.

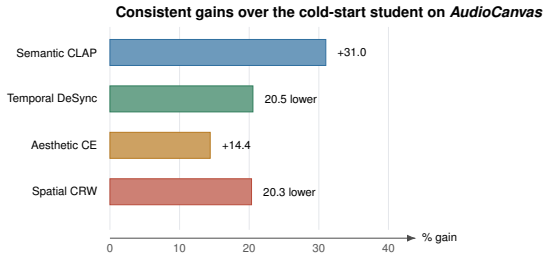

Figure 2: Relative improvement of MosaicAudio over its cold start (MosaicAudio w/o OPD) on AudioCanvas, by perceptual axis—semantic (CLAP), temporal (DeSync), aesthetic (CE), spatial (CRW). Each bar is the relative gain; for the DeSync and CRW *error* metrics it is the percentage reduction (so larger is better for all four). All four axes improve and none regresses. Absolute values are in Table 1.

Cross-domain generalization. The same routed teachers and student are used on both benchmarks with no per-benchmark tuning, yet the relative gain of OPD over the cold start persists across the distribution shift—CLAP rises 19% in-domain versus 31% on AudioCanvas, and CRW drops 31% versus 20%. The routed teachers thus encode a transferable per-axis correction rather than a benchmark-specific shortcut.

Human-study reliability. The subjective scores agree with the automatic metrics but emphasize a different aspect. The MOS-Q gap is modest ($4.21 \rightarrow 4.28$ on VGGSound; full scores in Tables 5, 6), as most recent models already produce plausible standalone audio. The MOS-C gap is larger and widens out of domain ($4.22 \rightarrow 4.25$ in-domain, $4.01 \rightarrow 4.13$ on AudioCanvas), because raters judge the audio together with the video and notice event-identity, timing, and stereo-placement errors. We therefore read MOS-C as the primary subjective signal for controllable V2A alignment.

5 Conclusion

We presented MosaicAudio, which treats multi-axis V2A alignment as a supervision-granularity problem: because the perceptual axes are judged on one waveform and conflict at fine acoustic scale, a terminal scalar reward cannot localize failures and lets axes cancel through a single advantage. MosaicAudio instead trains one specialist teacher per axis and consolidates them into a single student by on-policy distillation—dense, axis-routed velocity supervision on the student's own trajectories, under a perceptual anchor that preserves fidelity. Across VGGSound and AudioCanvas it improves over strong V2A baselines on all four axes.

Limitations

MosaicAudio depends on the quality and architectural compatibility of its specialist teachers. If a teacher learns a proxy-reward shortcut or fails on a rare event, dense supervision can transfer that error to the student. The current framework also uses fixed perceptual dimensions and a balanced routing schedule; future work should learn adaptive routing from scene complexity and listener intent. Finally, the AudioCanvas evaluation set broadens out-of-domain coverage but cannot span all real-world acoustic environments, languages, cultures, and production styles.

Ethics Statement

V2A generation can support media accessibility, film post-production, education, and assistive content creation. It can also be misused to fabricate convincing audio for silent videos or to alter the perceived context of real footage. We recommend provenance tracking, watermarking, and disclosure when generated audio is used in public media. Because reward models and training data may encode cultural and acoustic biases, deployments should evaluate performance across diverse scenes, environments, and user groups.

References

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

Ziyang Chen, David F Fouhey, and Andrew Owens. 2022. Sound localization by self-supervised time delay estimation. In *European Conference on Computer Vision*, pages 489–508. Springer.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. 2024a. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and 1 others. 2024b. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. 2024. Natural language supervision for general-purpose audio representations. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 336–340. IEEE.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. 2023. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885.

Zhen Fang, Wenxuan Huang, Yu Zeng, Yiming Zhao, Shuang Chen, Kaituo Feng, Yunlong Lin, Lin Chen, Zehui Chen, Shaosheng Cao, and Feng Zhao. 2026. Flow-OPD: On-policy distillation for flow matching models. *arXiv preprint arXiv:2605.08063*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Charles Knapp and G Clifford Carter. 1976. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327.

Junzhe Li, Yutao Cui, Tao Huang, Yinping Ma, Chun Fan, Miles Yang, and Zhao Zhong. 2025. Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. *arXiv preprint arXiv:2507.21802*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023a. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. 2023b. Vit-tts: visual text-to-speech with scalable diffusion transformer. *arXiv preprint arXiv:2305.12708*.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. 2026. PrismAudio: Decomposed chain-of-thoughts and multi-dimensional rewards for video-to-audio generation. In *International Conference on Learning Representations*.

Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. 2025a. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448*.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. 2025b. Flow-grpo: Training flow matching models via online rl. In *Advances in Neural Information Processing Systems*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. https://thinkingmachines.ai/blog/on-policy-distillation.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2023. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:48855–48876.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024a. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. 2024b. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Shentong Mo, Jing Shi, and Yapeng Tian. 2024. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, and 1 others. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. 2025. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, and 1 others. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. 2025. Temporally aligned audio for video with autoregression. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. 2024a. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15492–15501.

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. 2024b. Frieren: Efficient video-to-audio generation network with rectified flow matching. *arXiv preprint arXiv:2406.00320*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. 2024. Sonicvisionlm: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and 1 others. 2025. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*.

Biao Zhang, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder gemma: Improving the quality-efficiency trade-off via adaptation. *arXiv preprint arXiv:2504.06225*.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. 2024. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *arXiv preprint arXiv:2407.01494*.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, and 1 others. 2024. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217*.

A On-Policy Distillation: Derivation and Stability

Transition policy and ratio. Following the SDE view of flow sampling (Liu et al., 2025b), and using the same transition parameterization as Flow-GRPO/Flow-OPD (Fang et al., 2026), the probability-flow ODE is paired with an equivalent SDE $\mathrm{d}x_t = \mu_{\mathrm{SDE}}\,\mathrm{d}t + \sigma_t\,\mathrm{d}w$ whose drift is $\mu_{\mathrm{SDE}} = v_\theta + \frac{\sigma_t^2}{2t}(x_t + (1-t)v_\theta)$ (omitting the arguments of v_θ). An Euler–Maruyama step over Δt yields a Gaussian transition policy

$$\pi_\theta(x_{t+\Delta t} \mid x_t, c) = \mathcal{N}\big(\mu_\theta,\, \sigma_t^2 \Delta t\, I\big), \tag{11}$$

with mean $\mu_\theta = x_t + \mu_{\mathrm{SDE}}\Delta t$. As the old and current policies share the covariance $\sigma_t^2 \Delta t\, I$, the policy ratio is closed-form:

$$\rho_t(\theta) = \frac{\pi_\theta(x_{t+\Delta t} \mid x_t, c)}{\pi_{\theta_{\mathrm{old}}}(x_{t+\Delta t} \mid x_t, c)} = \exp\Big\{-\frac{\|x_{t+\Delta t} - \mu_\theta\|_2^2 - \|x_{t+\Delta t} - \mu_{\theta_{\mathrm{old}}}\|_2^2}{2\sigma_t^2 \Delta t}\Big\}. \tag{12}$$

Student-to-teacher KL as a velocity distance. For the routed teacher ϕ_k, the policy $\pi_{\phi_k} = \mathcal{N}(\mu_{\phi_k}, \sigma_t^2 \Delta t\, I)$ shares the covariance, so the student-to-teacher KL is a mean distance:

$$D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\phi_k}\big) = \frac{\|\mu_\theta - \mu_{\phi_k}\|_2^2}{2\sigma_t^2 \Delta t}. \tag{13}$$

Since the drift is affine in v, the state term x_t cancels in $\mu_\theta - \mu_{\phi_k}$, leaving

$$\mu_\theta - \mu_{\phi_k} = \Delta t\Big(1 + \tfrac{\sigma_t^2(1-t)}{2t}\Big)\,\Delta v_k, \tag{14}$$

with $\Delta v_k = v_\theta(x_t, t, c) - v_{\phi_k}(x_t, t, c_k)$. Substituting and factoring $1/\sigma_t$ gives the timestep weight of the main text,

$$D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\phi_k}\big) = \underbrace{\tfrac{\Delta t}{2}\Big(\tfrac{\sigma_t(1-t)}{2t} + \tfrac{1}{\sigma_t}\Big)^2}_{w(t)}\,\|\Delta v_k\|_2^2. \tag{15}$$

We use the student-to-teacher orientation because it directly scores how far the current student transition deviates from the routed teacher; under shared covariance this is simply a local, quadratic velocity-matching target. The two KL directions coincide in this mean-distance form, so we make no mode-seeking or mode-covering claim.

Note the deliberate conditioning asymmetry in Δv_k: the student velocity $v_\theta(x_t, t, c)$ uses the tag-free plan c, which is exactly its inference-time input, whereas the teacher $v_{\phi_k}(x_t, t, c_k)$ uses the axis-tagged c_k. The student is thus trained and deployed under identical tag-free conditioning—there is no train–test mismatch—and the target distills each teacher's axis-k behavior into that single tag-free student. The tag enters only on the teacher side, where it selects which axis-aligned velocity field serves as the local target, and is never required at inference.

Dense score, centered advantage, and stability. The per-step score $s_t = -w(t)\|\Delta v_k\|_2^2 \le 0$ is non-positive, so used directly it would discourage every sampled transition. We therefore subtract the per-timestep mean over the G rollouts of the same prompt, $\bar{s}_t = \frac{1}{G}\sum_{i=1}^{G} s_t^{(i)}$, and use the centered advantage $\hat{A}_t^{(i)} = s_t^{(i)} - \bar{s}_t$ in the surrogate. This removes the uniform negative offset and ranks sampled transitions within the group: transitions with smaller-than-average teacher–student discrepancy receive positive advantage and are reinforced. Gradients are stopped through the sampled states x_t, so the update is a policy gradient over transition log-probabilities rather than backpropagation through the sampler, and the PPO clip on ρ_t bounds each step.

B Implementation and Training Details

Architecture and data. MosaicAudio uses one student and four specialist teachers that share a flow-transformer backbone, a stereo latent audio autoencoder, and the conditioning interface of Section 3.2. The audio backbone is first pretrained for text-conditioned audio generation on standard audio–text corpora (AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024a)) together with VGGSound (Chen et al., 2020), and then given a CoT-aware cold start on VGGSound (≈190k video–audio clips) paired with four-axis CoT annotations produced by Gemini 2.5 Pro (Comanici et al., 2025) and a fine-tuned VideoLLaMA2 (Cheng et al., 2024b) planner. The planner is trained once and then *frozen* for all later stages, so teacher and student updates act only on the generator v_θ; the latent autoencoder is likewise frozen.

Reward models. The four axes use fixed, publicly available reward models, applied only during post-training: MS-CLAP (Elizalde et al., 2024) for semantic alignment, Synchformer (Iashin et al., 2024) for temporal synchrony, Meta Audiobox-Aesthetics (Tjandra et al., 2025) for aesthetic quality, and StereoCRW (Chen et al., 2022) for stereo accuracy. We use the official released checkpoints

and do not fine-tune them.

Teacher and student optimization. Each teacher is initialized from the cold-start checkpoint and trained on a single reward with the windowed flow-GRPO objective; within each group of G rollouts the reward is normalized by its z-score before computing the advantage. The student is then consolidated by OPD from the same checkpoint: a balanced cyclic router assigns exactly one axis to each mini-batch, the student samples stochastic trajectories inside a random denoising window and integrates deterministically elsewhere, and the frozen anchor is applied on all sampled states. The teacher bank and anchor are frozen throughout, so neither adds inference cost. Table 4 lists the hyperparameters used for all reported numbers; full configuration files will be released with the code. Timesteps are clipped to the sampler's nonzero range $t \in [\tau, 1]$ when evaluating the KL weight.

The *scalar multi-reward RL* comparison (Table 3) is a strict control: it starts from the same cold-start checkpoint, uses the same four reward models, the same group size G, rollout budget, update-step count, and learning-rate schedule as MosaicAudio, and received comparable tuning effort. The two differ only in the post-training rule—summed-scalar flow-GRPO versus routed multi-teacher OPD—so the gap between them isolates the effect of dense routed supervision rather than budget or reward differences.

Table 4: Post-training hyperparameters (teachers and student).

Setting	Value
Optimizer	AdamW (β_1=0.9, β_2=0.999)
Learning rate	1×10^{-5} (cosine decay)
Weight decay	0.01
Prompts per step	16
Group size G	12
Sampler steps T	25
SDE window width w	3
SDE noise level	0.7
PPO clip ε	0.2
Anchor weight λ	0.5
Teacher steps (per axis)	3,000
OPD consolidation steps	5,000
Reward normalization	per-group z-score
CFG dropout (per modality)	0.1
Hardware	8×A800 (80 GB)

The SDE noise level controls the stochasticity injected inside the sampled denoising window, while CFG dropout is the classifier-free guidance dropout applied to each conditioning modality during post-training.

Model size and computational budget. All reported variants share the same backbone: the student generator v_θ has 518M parameters and is the only network used at inference. Each of the four specialist teachers has the same 518M-parameter architecture (initialized from the cold start) and is discarded after consolidation, so the deployed model is a single 518M-parameter flow generator regardless of how many teachers were trained; the frozen latent autoencoder, the frozen multimodal planner, and the four reward models are not updated and are not part of the released generator. All post-training is run on one node of 8×A800 (80 GB) GPUs. Training the four specialist teachers (3,000 steps each) and the OPD consolidation (5,000 steps) together take about five days of wall-clock time on this 8×A800 node ($\approx$ 960 A800 GPU-hours in total); each single-axis teacher run is cheaper per step than the consolidation because it queries only one reward model. The CoT-aware cold start is trained once and reused across all post-training variants, so it is not repeated per variant. Inference for a 9-second stereo clip is a single T=25-step pass that takes $\approx$0.6 s on one A800—identical to the cold-start generator and to the scalar CoT-RL baseline we post-train, since the deployed architecture and step count are unchanged.

B.1 CoT Annotation Template

We use a fixed four-block annotation format for training the multimodal planner. The prompt asks the annotator model to ignore invisible or off-screen sound sources and to describe only sounds supported by visible evidence:

> **Semantic:** list visible sound-producing events, objects, and acoustic materials.
> **Temporal:** describe event order, onset timing, duration, and synchronization cues.
> **Aesthetic:** describe desired fidelity, naturalness, loudness balance, and production quality.
> **Spatial:** describe left-right location, distance, motion, and stereo movement.

The four blocks are kept separate during planner training and concatenated in the same order, with explicit section markers preserved, when used to condition the generator.

C Additional Quantitative Results

C.1 Full Main Results

Tables 5 and 6 provide the full metric set corresponding to the compact main-text tables.

Table 5: Full VGGSound results.

Method	Params	Semantic	Temporal	Aesthetic Quality				Spatial		Distribution		Subjective	
		CLAP↑	DeSync↓	PQ↑	PC↓	CE↑	CU↑	GCC↓	CRW↓	FD↓	KL↓	MOS-Q↑	MOS-C↑
ThinkSound	1.3B	0.43	0.55	6.15	3.53	3.95	5.48	4.65	13.47	1.17	1.35	4.05±0.55	4.18±0.51
PrismAudio	518M	0.47	0.41	6.38	3.24	4.29	5.68	3.77	7.72	1.08	1.23	4.21±0.35	4.22±0.29
MosaicAudio w/o OPD	518M	0.42	0.51	6.17	3.32	3.94	5.48	4.06	10.29	1.14	1.43	4.02±0.48	4.11±0.42
MosaicAudio	518M	**0.50**	**0.38**	**6.55**	**3.18**	**4.64**	**5.88**	**3.60**	**7.11**	**1.02**	**1.18**	**4.28±0.32**	**4.25±0.27**

Table 6: Full AudioCanvas results.

Method	Semantic	Temporal	Aesthetic Quality				Spatial		Distribution		Subjective	
	CLAP↑	DeSync↓	PQ↑	PC↓	CE↑	CU↑	GCC↓	CRW↓	FD↓	KL↓	MOS-Q↑	MOS-C↑
ThinkSound	0.48	0.80	6.48	3.50	4.10	5.94	4.43	22.82	1.95	2.54	3.79±0.58	3.80±0.54
PrismAudio	0.52	0.36	6.68	2.82	4.26	6.15	3.50	12.87	1.92	1.53	4.12±0.28	4.01±0.25
MosaicAudio w/o OPD	0.42	0.44	6.45	3.22	3.81	5.87	4.11	15.30	2.10	2.17	3.91±0.35	3.85±0.31
MosaicAudio	**0.55**	**0.35**	**6.75**	**2.78**	**4.36**	**6.22**	**3.38**	**12.20**	**1.85**	**1.45**	**4.18±0.26**	**4.13±0.23**

Table 7: Full-metric anchor ablation on AudioCanvas (cf. Table 3).

Method	Semantic	Temporal	Aesthetic Quality				Spatial		Distribution	
	CLAP↑	DeSync↓	PQ↑	PC↓	CE↑	CU↑	GCC↓	CRW↓	FD↓	KL↓
MosaicAudio w/o anchor	0.50	0.39	**7.00**	3.05	**4.40**	5.95	4.05	13.10	3.50	1.80
MosaicAudio	**0.55**	**0.35**	6.75	**2.78**	4.36	**6.22**	**3.38**	**12.20**	**1.85**	**1.45**

C.2 Anchor Ablation

Table 7 expands the anchor ablation from Table 3 to the full metric set. Removing the anchor improves the optimized aesthetic proxies but hurts distributional fidelity and spatial stability, which is the reward-hacking pattern the anchor is designed to prevent.

C.3 Statistical Reporting

For the human studies (MOS-Q, MOS-C) we report the mean and standard deviation over clips and raters; these are the ± values in Tables 5 and 6. Objective metrics are computed once per system with deterministic ODE decoding from a fixed post-trained checkpoint and a fixed evaluation seed, so each objective number is a single deterministic run rather than a max or an average over runs.

Significance testing protocol. Beyond means and standard deviations, we assess whether the MOS improvements over the strongest baseline are statistically reliable rather than rating noise. For each benchmark we form per-clip mean opinion scores (averaged over raters) on the 80 evaluated clips and compare MosaicAudio against PrismAudio with a two-sided Wilcoxon signed-rank test on the paired per-clip scores, together with a paired bootstrap (10,000 resamples over clips) for a 95% confidence interval on the mean per-clip difference. We release the per-clip ratings and the analysis scripts with the human-study materials, so that these tests and the per-axis significance can be reproduced and verified directly.

D Evaluation Details

D.1 Objective Metrics

CLAP measures semantic consistency between generated audio and text descriptions. DeSync is computed with a pretrained audio-visual synchronization model; lower values indicate better temporal alignment. PQ, PC, CE, and CU are the four outputs of the audio-aesthetic evaluator, where lower PC indicates less unnecessary production complexity and higher values are better for the other three scores. GCC and CRW measure stereo localization error. FD and KL measure distributional similarity between generated and reference audio. All metrics use official released checkpoints with default settings and no fine-tuning: semantic alignment uses the MS-CLAP audio–text model; temporal alignment uses the DeSync head of Synchformer; aesthetic scores are the four outputs of Meta Audiobox-Aesthetics; spatial accuracy uses GCC-PHAT and the StereoCRW cross-correlation localization error; and FD/KL are computed against the reference audio with the standard feature extractor used by prior V2A work. Audio is resampled to each model's expected sample rate and clips are loudness-normalized before scoring; every system

is evaluated with an identical pipeline and seed.

The four reward models (MS-CLAP, Synchformer, Audiobox-Aesthetics, StereoCRW) are also reported as the on-axis metrics they optimize, so improvements on CLAP, DeSync, the Audiobox scores, and CRW partly reflect direct optimization. We therefore additionally report FD, KL, GCC, and the two human studies (MOS-Q, MOS-C), none of which is used as a training reward, as held-out checks: a policy that merely games a reward model would gain on that reward's metric while degrading these. The anchor ablation (Table 7) is exactly this diagnostic—removing the anchor raises the optimized aesthetic scores yet collapses the held-out FD. Crucially, the held-out metrics move in the *same* direction as the rewards: from the cold start to MosaicAudio on AudioCanvas, FD improves $2.10 \rightarrow 1.85$, KL $2.17 \rightarrow 1.45$, and GCC (a spatial metric distinct from the StereoCRW reward) $4.11 \rightarrow 3.38$ (Tables 6, 7). Because none of these is a training reward, their consistent improvement indicates the gains are not solely reward exploitation.

Baseline conditioning. Each baseline is evaluated with its native conditioning on the same input video: the CoT-aware systems (ThinkSound, PrismAudio, MosaicAudio) receive the four-axis CoT plan, while systems that do not accept a structured plan (Frieren, V2A-Mapper, AudioX, HunyuanVideo-Foley, MMAudio) receive the text/video conditioning they were designed for. We neither withhold information from non-CoT models nor feed them inputs they cannot consume; every system sees the same video, and every CoT-capable system sees the same plan, so the comparison reflects each model's intended operating mode rather than an input handicap.

D.2 Human Evaluation Protocol

For each test set (VGGSound and AudioCanvas) we randomly sample 80 videos and render the 9-second output of every system. Clips are loudness-normalized, presented in randomized order with method identities hidden, and judged by three annotators per clip using headphones; calibration examples are used to familiarize raters with the scale before the study. We run two independent 5-point studies: **MOS-Q** (audio only) rates standalone fidelity and naturalness, and **MOS-C** (audio with video) rates audio–video consistency in semantics, timing, and stereo placement. Each integer point carries a written anchor (1 = bad/mismatched, 3 = acceptable, 5 = excellent/perfectly matched). We report the mean and standard deviation over clips and raters in Tables 5 and 6; the MOS-Q gap between systems is smaller than the MOS-C gap, consistent with most systems already producing plausible standalone audio.

Instructions given to raters. Each rater first reads a short briefing and three calibration examples per anchor level. The MOS-Q instruction is: *"Listen to the audio clip on headphones. Rate its standalone quality from 1 to 5, considering clarity, fidelity, absence of artifacts, and how natural it sounds (1 = bad/heavily degraded, 3 = acceptable, 5 = excellent). Judge only the audio; do not penalize it for matching or not matching any video."* The MOS-C instruction is: *"Watch the video with this audio track. Rate from 1 to 5 how well the audio matches the video in (i) what is making sound, (ii) when the sounds happen, and (iii) where they are placed left-to-right (1 = clearly mismatched, 3 = acceptable, 5 = perfectly matched)."* Raters may replay each item up to three times and rate at their own pace. The briefing also states that the clips are synthetic audio generated by research models, that the stimuli were selected to focus on non-speech sound effects and to avoid graphic content, and that participation is voluntary and may be stopped at any time without penalty; no audio is recorded from or about the raters.

Recruitment and payment. Raters were adult volunteers from an internal listener pool, including both audio-trained and general listeners to capture diverse perspectives, all reporting normal hearing. They were compensated for their time at a rate above the local minimum wage. Participation was voluntary: before starting, raters were told what the task involved, that the stimuli are synthetic non-speech audio, and that their ratings would be used only in aggregate (as mean-opinion scores) in this paper, and they could stop at any time. We recorded only the numerical ratings and no personal information about the raters.

E Scientific Artifacts, Data, and Documentation

Artifacts we use. MosaicAudio builds on a number of existing artifacts, all cited where first used in Section 4.1 and Appendix B: the VGGSound dataset (Chen et al., 2020) for cold-start training and in-domain evaluation; the AudioCanvas

benchmark for out-of-domain evaluation, introduced by the prior CoT-RL system PrismAudio (Liu et al., 2026); the VideoLLaMA2 planner (Cheng et al., 2024b) and Gemini 2.5 Pro (Comanici et al., 2025) for CoT annotation; and four fixed reward models—MS-CLAP (Elizalde et al., 2024), Synchformer (Iashin et al., 2024), Meta Audiobox-Aesthetics (Tjandra et al., 2025), and StereoCRW (Chen et al., 2022)—together with the GCC (Knapp and Carter, 1976), FD (Kilgour et al., 2018), and KL (Copet et al., 2024) evaluation metrics. The artifact we create is the MosaicAudio system itself (one 518M-parameter student generator plus the training recipe).

Licenses and terms of use. The encoders, planner, and reward/evaluation models we use are publicly released under open-source or research-oriented licenses; Gemini 2.5 Pro is accessed through its public API under the provider's terms of service. We use every artifact unmodified (official released checkpoints, no fine-tuning of the reward models) and only for non-commercial academic research, following their documented terms. VGGSound is distributed for research use as links to publicly available videos; AudioCanvas is likewise a research-only benchmark distributed under controlled access (Liu et al., 2026). We plan to release the MosaicAudio code and student checkpoint for non-commercial research use, with usage guidelines (Ethics Statement); because the cold-start data and CoT annotations were accessed for research, the model and any derivatives are intended for research contexts, consistent with the original access conditions.

Intended use. Our use of each existing artifact matches its intended purpose: the datasets are used for V2A training/evaluation, the planner and Gemini for caption/CoT generation, and the reward models for the perceptual measurements they were built to provide. The created MosaicAudio system is intended for research on controllable V2A generation and is released for that purpose only.

Personally identifying information and offensive content. Our released demos and human-evaluation stimuli focus on non-speech sound effects, which reduces the most common channel for identifying or sensitive content. The cold-start corpus and annotations are derived from VGGSound, and the out-of-domain benchmark applies strict filtering that removes off-screen narration, background speech, and clips with private or sensitive personal content (Liu et al., 2026); annotations describe only audio–visual and technical relationships and store no personal identifiers, and clips are referenced by anonymized identifiers. We did not collect or release any new raw user data.

Documentation and statistics of data. VGGSound covers 309 sound classes spanning everyday objects, animals, instruments, and environmental scenes; our cold-start generator and CoT annotations use its standard training split ($\approx$190k video–audio clips), and in-domain results are reported on its standard test split. AudioCanvas is used strictly for evaluation (no train/dev split is taken from it) and contains real-world videos across ≈ 300 V2A-relevant sound categories, including a curated multi-event subset alongside single-event scenes, with structured four-axis CoT annotations (Liu et al., 2026); none of its clips or source videos appear in our training or annotation data. For the subjective studies we draw 80 videos per benchmark and collect ratings from multiple annotators (Appendix D.2). The reward models cover the four perceptual axes (semantic, temporal, aesthetic, spatial) and are applied only at post-training and evaluation time. We report the full objective and subjective metric sets in Tables 5 and 6.

F Use of AI Assistants

We used AI assistants (general-purpose large language models) as writing aids for grammar correction, conciseness, consistency checks, and drafting support. The authors verified all technical statements, experiments, numerical results, and claims; the research ideas, method design, implementation, analysis, and conclusions are the authors' responsibility.