AV-GRPO: MODALITY-ANCHORED DECOUPLING DIFFUSION REINFORCEMENT LEARNING FOR JOINT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent years have witnessed remarkable progress in joint audio-video generation. Nevertheless, existing models still suffer from limited per-modality fidelity, inadequate text-modality alignment, and weak cross-modal synchronization. Reinforcement-learning-based post-training offers a promising avenue for addressing these shortcomings. However, naively extending such approaches to joint audio-video generation is challenging. Heterogeneous multimodal rewards entangle the learning signals of the two modalities and obscure credit assignment. Jointly optimizing both modality towers also incurs substantial computational cost despite their distinct optimization dynamics. Furthermore, the difficulty of synchronization evaluation varies with the sampled counter-part modality, making fair reward comparison difficult. To address these challenges, we propose AV-GRPO, a modality-anchored online diffusion reinforcement learning framework, together with 5DAV, a fully decoupled and difficulty-controllable training dataset. AV-GRPO integrates three key components: (1) modality-anchored rollouts that disentangle learning signals while reducing anchor-induced difficulty variation; (2) trajectory-locked frozen-tower optimization that reduces training costs and redirects credit assignment, thereby simplifying optimization. and (3) adaptive objectives and perturbation strengths tailored to modality-specific dynamics and sample difficulty. Collectively, these designs transform coupled multi-modal preference learning into a set of conditional unimodal subproblems, enabling more accurate reward attribution and more effective synchronization optimization. We construct 5DAV, a dataset decoupled along five dimensions, to facilitate systematic training. Experiments on JavisBench and VABench demonstrate that AV-GRPO consistently outperforms LTX-2.3 in generation quality, semantic alignment, and cross-modal synchronization under both LoRA and full fine-tuning settings. Extensive ablation studies further validate the effectiveness of the proposed designs.

1 INTRODUCTION

Joint audio–video generation has advanced rapidly through unified and interacting dual-stream models (HaCohen et al., 2026; Low et al., 2025; Liu et al., 2026c; Team et al., 2026). Yet generated audio and video still struggle to achieve high quality together: either stream may contain perceptual artifacts, one or both may not reflect the text prompt, and otherwise plausible streams may depict mismatched events or drift out of sync. Scaling pretraining alone does not directly prioritize these failures. Reward-guided post-training instead turns perceptual and semantic evaluators into learning signals, with demonstrated benefits in language and visual generation (Rafailov et al., 2023; Shao et al., 2024; Liu et al., 2026a). Extending it to joint generation requires improving both streams while preserving their semantic and temporal dependence.

A natural baseline treats each generated audio–video pair as one policy output. For every prompt, it samples groups of paired rollouts, evaluates audio quality and alignment, video quality and alignment, and cross-modal synchronization, aggregates the heterogeneous rewards into a group-relative advantage, and updates both towers (Shao et al., 2024; Liu et al., 2026a; Zheng et al., 2026; Xue et al., 2025). This straightforward joint optimization has three difficulties.

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.

First, mixed rewards complicate model optimization and fair reward comparison. Audio quality, visual quality, text alignment, and synchronization may favor different candidates, making their joint optimization difficult. Moreover, synchronization rewards depend on the sampled counterpart: matching audio to a steady scene can be easier than frequent visual events. A higher score may therefore reflect an easier counterpart as well as better alignment, making comparisons across jointly varying pairs less controlled. **Second, joint updates incur high memory costs and can misassign credit.** Backpropagation and training states are required for both towers, yet a shared advantage updates both even when its improvement primarily comes from one modality. Their interactions throughout denoising further complicate assigning that improvement to the responsible tower. **Third, different modalities have different optimization dynamics.** Their pretrained capabilities, latent scales, reward sensitivities, and learning speeds differ. Shared objectives and noise strengths can favor one branch or destabilize the other, making a common configuration unsuitable.

We propose **AV-GRPO**, which alternates audio-anchored video optimization and video-anchored audio optimization (Figure 1). **First, modality-anchored rollouts disentangle rewards and control comparison conditions.** Each group shares one complete anchor trajectory while independently sampling the target modality. The anchor reward can be omitted, simplifying optimization to target-modality quality and alignment plus synchronization against a common counterpart. This controls anchor-induced difficulty variation within each group, while refreshing anchors across groups preserves diverse conditions. **Second, trajectory locking and tower freezing significantly reduce memory costs and direct credit to the target tower.** The anchor states remain fixed throughout each rollout and its policy update, and only the target tower is optimized. Applying the conditional advantage to this tower avoids updating both branches with the same signal; freezing the counterpart also reduces backward-pass and training-state costs. **Third, hyperparameter decoupling accommodates different branch dynamics.** We use modality-specific reward compositions, noise strengths, and loss weightings to support exploration and stable optimization in each tower. Alternating the target modality allows both towers to improve under controlled cross-modal conditions.

To support controlled post-training, we introduce **5DAV**, a training set of 5,760 prompts organized along five independently specified dimensions. We evaluate AV-GRPO on JavisBench and VABench against LTX-2.3 and GDPO under both LoRA and full fine-tuning. The results show broad gains in perceptual quality, text alignment, cross-modal coherence, and synchronization, while ablations test the dataset and the alternating schedule. The frozen-tower strategy also enables full-parameter post-training of the 22B LTX-2.3 model on eight NVIDIA A800 GPUs. Our contributions are:

- We propose **AV-GRPO**, a modality-anchored reinforcement learning framework that alternates controlled modality-wise updates for clearer reward attribution and balanced audio–video improvement.
- We develop three complementary mechanisms: (i) modality-anchored rollouts that disentangle reward objectives and enable controlled synchronization comparisons; (ii) trajectory locking and tower freezing that isolate modality-wise credit assignment while reducing memory overhead; and (iii) decoupled optimization hyperparameters that accommodate asymmetric audio–video learning dynamics.
- We introduce **5DAV**, a five-dimensionally decoupled training set. Experiments on JavisBench and VABench show broad gains in generation quality, text–modality alignment, and synchronization, supported by ablations.

2 METHOD

2.1 PRELIMINARIES

Joint audio–video generation models produce a video and its audio track from a text prompt c within a single denoising process. (HaCohen et al., 2026; Liu et al., 2026c)

The prevailing approach is to represent video and audio in separate latent spaces, denoted x^{V} and x^{A}, and noise them following rectified flow,

$$x_t^m = (1-t)\, x_0^m + t\, \epsilon^m, \qquad \epsilon^m \sim \mathcal{N}(0, I), \quad m \in \{\mathrm{A}, \mathrm{V}\}, \tag{1}$$

where the timestep t is shared by the two modalities and the noises are independent. The denoiser has one stream per modality, which we call the two towers and whose parameters we write as

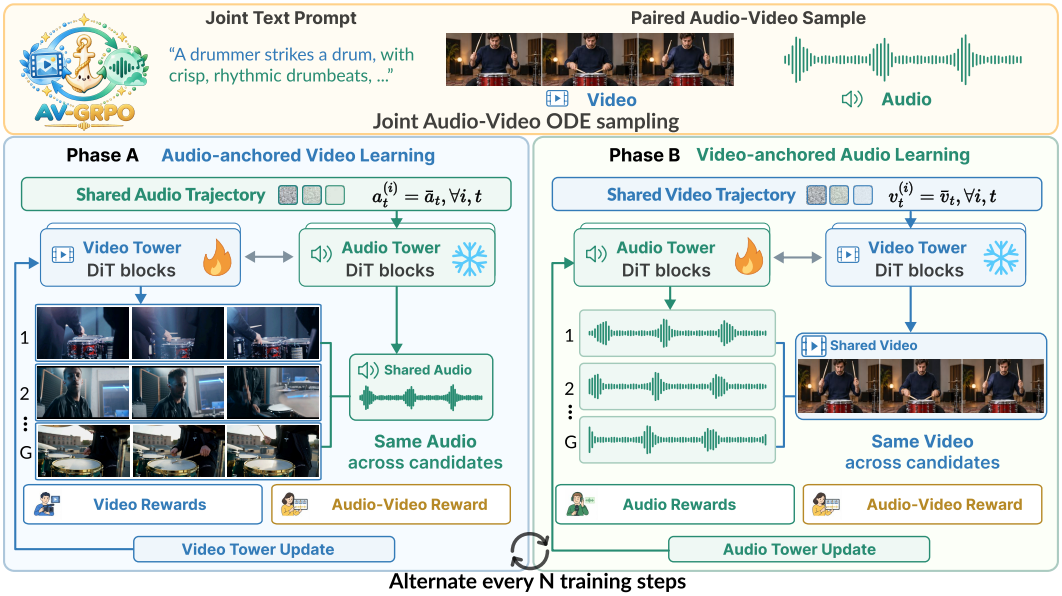


Figure 1: Overview of the AV-GRPO training pipeline.

$\theta = (\theta_{\mathrm{A}}, \theta_{\mathrm{V}})$; the towers exchange information through cross-modal attention or other mechanisms, so the velocity predicted by either tower depends on the current latents of both modalities:

$$\mathrm{d}x_t^m = v_\theta^m\big(x_t^{\mathrm{A}}, x_t^{\mathrm{V}}, t, c\big)\,\mathrm{d}t, \qquad m \in \{\mathrm{A}, \mathrm{V}\}. \tag{2}$$

At sampling time, both modalities integrate Eq. (2) in parallel from $t = 1$ to $t = 0$.

Flow-GRPO. Online reinforcement learning with GRPO (Shao et al., 2024) optimizes the step-by-step sampler as a policy: each denoising step is an action, and the reward is assigned to the final sample. This requires the sampling process to be stochastic, so that different samples can be drawn for the same prompt, and the per-step transition probability to be computable, so that there is a probability ratio to adjust. The deterministic ODE of Eq. (2) satisfies neither: the same initial noise always yields the same sample, and computing its transition probability is computationally expensive due to divergence estimation. Flow-GRPO (Liu et al., 2026a) therefore replaces the ODE with an SDE that has the same marginal distribution at every timestep, introducing stochasticity without changing the distribution the pretrained model generates. Discretized over T steps, the sample advances by

$$x_{t+\Delta t} = x_t + \Big[v_\theta + \frac{\sigma_t^2}{2t}\big(x_t + (1-t)\,v_\theta\big)\Big]\Delta t + \sigma_t\sqrt{|\Delta t|}\,\epsilon, \qquad \sigma_t = a\sqrt{\tfrac{t}{1-t}}, \tag{3}$$

where v_θ abbreviates $v_\theta(x_t, t, c)$, $\epsilon \sim \mathcal{N}(0, I)$ controls how strongly the sampling is perturbed. Each step of Eq. (3) is a Gaussian transition $\pi_\theta(x_{t+\Delta t} \mid x_t, c)$. With this sampler, Flow-GRPO draws G trajectories per prompt and maximizes

$$J(\theta) = \mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{T}\sum_{t}\Big(\min\big(r_t^i\hat{A}^i,\ \mathrm{clip}(r_t^i, 1-\varepsilon, 1+\varepsilon)\hat{A}^i\big) - \beta\, D_{\mathrm{KL}}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)\Big)\Big], \tag{4}$$

where $\hat{A}^i = (R^i - \mathrm{mean}_i R^i)/\mathrm{std}_i R^i$ is the advantage of trajectory i, obtained by standardizing the reward of its final sample, $R^i = R(x_0^i, c)$, within the group, and indicates whether the sample is better or worse than the others in its group; r_t^i is the probability ratio between the current policy and the sampling policy at this step, i.e., a ratio of two Gaussian densities; ε is the clip range; β is the KL coefficient and π_{ref} the pretrained model, for which the KL term has a closed form proportional to $\|v_\theta - v_{\mathrm{ref}}\|^2$.

2.2 AV-GRPO

AV-GRPO applies the Flow-GRPO update of Eq. (4) to one tower at a time. In an *audio-anchored* phase, it fixes one audio trajectory, samples a group of video trajectories, and updates only the video

tower; the *video-anchored* phase is symmetric (Figure 1). The phases alternate every n training steps. Let $m \in \{\mathrm{A}, \mathrm{V}\}$ denote the anchor modality, $\bar{m}$ the target, and $\tau^m = (x_1^m, \ldots, x_0^m)$ a complete denoising trajectory.

Modality-Anchored Rollouts. For prompt c, we first generate one audio–video pair by sampling both towers with Eq. (3). We retain the full trajectory τ_{anc}^m of modality m and independently sample G target trajectories. At each step, its anchor state $x_{\mathrm{anc},t}^m$ replaces the state of modality m in cross-modal attention. The i-th target advances by

$$x_{t+\Delta t}^{\bar{m},i} = x_t^{\bar{m},i} + \left[v_\theta^{\bar{m}} + \frac{\sigma_t^2}{2t}\left(x_t^{\bar{m},i} + (1-t)\, v_\theta^{\bar{m}}\right)\right]\Delta t + \sigma_t \sqrt{|\Delta t|}\, \epsilon^{\bar{m},i}, \qquad \epsilon^{\bar{m},i} \sim \mathcal{N}(0, I), \quad (5)$$

where $v_\theta^{\bar{m}}$ is evaluated at $(x_t^{\bar{m},i}, x_{\mathrm{anc},t}^m, t, c)$. This gives a conditional Gaussian transition $\pi_\theta^{\bar{m}}(x_{t+\Delta t}^{\bar{m}} \mid x_t^{\bar{m}}, x_{\mathrm{anc},t}^m, c)$ and G final pairs $(x_{\mathrm{anc},0}^m, x_0^{\bar{m},i})$. The anchor is common to the group, whereas the target sample varies.

Reward Disentanglement. Each pair receives a target-modality quality and text-alignment reward $R_{\bar{m}}$ and a synchronization reward R_{sync} (Section 3.2.1). In a joint group, changes in either modality can alter the synchronization score. Here, the anchor reward is constant across all candidates and disappears under group centering, leaving

$$R^i = R_{\bar{m}}(x_0^{\bar{m},i}) + R_{\mathrm{sync}}(x_0^{\bar{m},i}, x_{\mathrm{anc},0}^m). \quad (6)$$

Meanwhile, this controls the variation in synchronization difficulty of the anchor modality: within a group, all video (or audio) samples share the same audio (or the same video), enabling fair comparison and more targeted learning of audio–video synchronization.

This grouping changes the conditional policy being optimized, rather than merely removing a reward term. For a fixed anchor trajectory, the corresponding KL-regularized subproblem can be written as

$$\mathcal{L}_{\bar{m}}(\pi \mid \tau_{\mathrm{anc}}^m, c) = \mathbb{E}_{\tau^{\bar{m}} \sim \pi^{\bar{m}}(\cdot \mid \tau_{\mathrm{anc}}^m, c)}[R_{\bar{m}} + R_{\mathrm{sync}}] - \beta D_{\mathrm{KL}}\big(\pi^{\bar{m}}(\cdot \mid \tau_{\mathrm{anc}}^m, c) \| \pi_{\mathrm{ref}}^{\bar{m}}(\cdot \mid \tau_{\mathrm{anc}}^m, c)\big)\,. \quad (7)$$

The video and audio phases solve the two symmetric conditional subproblems. Holding the anchor fixed makes their reward comparisons controlled, while refreshing it between groups prevents training on a single counterpart. The formal relation between these conditional objectives and a joint KL-regularized objective is developed in Appendix A.

For comparison, the joint formulation optimizes a distribution over paired trajectories, $\mathcal{L}_{\mathrm{joint}}(P) = \mathbb{E}_P[r(\tau^{\mathrm{V}}, \tau^{\mathrm{A}}, c)] - \beta D_{\mathrm{KL}}(P \| P_0)$. Its samples change both audio and video at once, whereas each AV-GRPO phase conditions on one realized trajectory and changes only the other. Fixing the *complete* trajectory matters because the two towers interact at every denoising step: a fixed final audio or video sample alone would leave the intermediate cross-modal states uncontrolled. The conditional view explains why the within-group reward has a clearer interpretation without assuming that the two generation streams are independent.

Trajectory Locking and Tower Freezing. The same anchor states are reused throughout each rollout and during its policy update. We freeze θ_m and update only $\theta_{\bar{m}}$, retaining gradient and optimizer states for the active tower while the anchor supplies a fixed cross-modal condition. Resampling the anchor for each group exposes the target to varied counterpart trajectories; switching towers every n steps lets both branches improve. For the fixed anchor, the conditional GRPO objective is

$$\mathcal{J}_{\bar{m}}(\theta) = \mathbb{E}_{c, \tau_{\mathrm{anc}}^m, \{\tau^{\bar{m},i}\}_{i=1}^G \sim \pi_{\theta_{\mathrm{old}}}^{\bar{m}}(\cdot \mid \tau_{\mathrm{anc}}^m, c)} \Bigg[\frac{1}{GT} \sum_{i=1}^{G} \sum_t \Big(\min\left(r_t^{\bar{m},i} \hat{A}^{\bar{m},i},\ \mathrm{clip}(r_t^{\bar{m},i}, 1-\varepsilon, 1+\varepsilon)\hat{A}^{\bar{m},i}\right) - \beta D_{\mathrm{KL}}\big(\pi_\theta^{\bar{m}} \| \pi_{\mathrm{ref}}^{\bar{m}}\big) \Big) \Bigg], \quad (8)$$

where $r_t^{\bar{m},i}$ is the current-to-old conditional transition-probability ratio; both policies and the reference are conditioned on the anchor. Appendix A analyzes an idealized counterpart: under a common reference joint law and contraction assumptions, alternating exact conditional Gibbs kernels converge to the joint KL-regularized optimum.

Freezing the anchor also separates the costs of generating a condition from those of learning under it. An anchor trajectory must still be sampled for each group, but its tower does not require backward

activations or optimizer updates in that phase. The target tower sees the same anchor at all G comparisons, allowing reward variation to reflect its sampled trajectories rather than changes in the counterpart. Alternating phases then updates each tower under the other's latest distribution instead of permanently fixing either one.

2.3 Adaptive Noise Clipping for Robust Sampling

The diffusion coefficient $\sigma_t = a\sqrt{t/(1-t)}$ in Eq. (3) grows near $t = 1$. On our coarse grid, the resulting stochastic increment can overwhelm the latent signal during noising and destabilize training. **Adaptive Noise Clipping (ANC)** bounds its standard deviation using

$$s_t = \sigma_t \sqrt{|\Delta t|}, \qquad \lambda_t = \min\left(1, \frac{\tau}{s_t + \delta}\right), \qquad \tilde{\sigma}_t = \lambda_t \sigma_t, \tag{9}$$

where $\tau > 0$ is a clipping threshold and $\delta > 0$ prevents division by zero. We substitute $\tilde{\sigma}_t$ for σ_t in both the stochastic term and the drift correction of Eq. (3):

$$x_{t+\Delta t} = x_t + \left[v_\theta + \frac{\tilde{\sigma}_t^2}{2t}\big(x_t + (1-t)v_\theta\big)\right]\Delta t + \tilde{\sigma}_t \sqrt{|\Delta t|}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{10}$$

When clipping is inactive, Eq. (10) reduces to the original update; otherwise the stochastic term scales by λ_t and the drift correction by λ_t^2. We use 15 sampling steps and find ANC stabilizes training at larger noise strengths.

2.4 Hyperparameter Decoupling and Training Stability

A shared configuration can limit exploration in one tower while destabilizing the other. AV-GRPO naturally supports independent optimization configurations through its alternating frozen-tower updates, allowing noise strengths and objectives to be tailored to each modality. For LTX-2.3, we use $a_{\mathrm{V}} = 0.02$ and $a_{\mathrm{A}} = 0.8$ in $\sigma_t = a\sqrt{t/(1-t)}$: too little noise limits exploration, whereas too much can break denoising.

With the original Flow-GRPO objective, the video tower becomes over-exposed after a few iterations. The late denoising stages, which control brightness and fine detail, receive weak gradients; changing only the scalar KL coefficient does not resolve this imbalance. Following LongCat-Video (Team et al., 2025), we reweight the video policy and KL terms at each step:

$$\mathcal{J}_{\text{video}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\Big(\lambda_{\text{policy}}(t,\Delta t)\cdot\mathcal{L}_{\text{policy}}(\theta) - \beta\,\lambda_{\text{KL}}(t,\Delta t)\cdot D_{\text{KL}}(\pi_\theta^{\text{video}}\|\pi_{\text{ref}}^{\text{video}})\Big)\right], \tag{11}$$

The weighting functions are defined as:

$$\lambda_{\text{policy}}(t,\Delta t) = \sqrt{\frac{t}{\Delta t(1-t)}}, \qquad \lambda_{\text{KL}}(t,\Delta t) = \frac{t}{\Delta t(1-t)}. \tag{12}$$

The audio tower retains the original weighting. Appendix C gives the full objective and gradient derivation.

3 Experiment

3.1 5DAV Dataset

To support GRPO optimization for joint audio-video generation models, we construct a **5D decoupled training set (5DAV)** that is fully disentangled across five dimensions: semantic hierarchy, sound source type, synchronization difficulty, temporal complexity, and instruction granularity. Through Cartesian product of these dimensions, our dataset achieves comprehensive capability coverage with controllable difficulty levels. The five dimensions are defined as follows:

- **Semantic Hierarchy (W)**: Entity-level (single object/person/animal and its inherent sound), Scene-level (static scene with background atmosphere), Event-level (dynamic interactions and causal events among multiple entities).

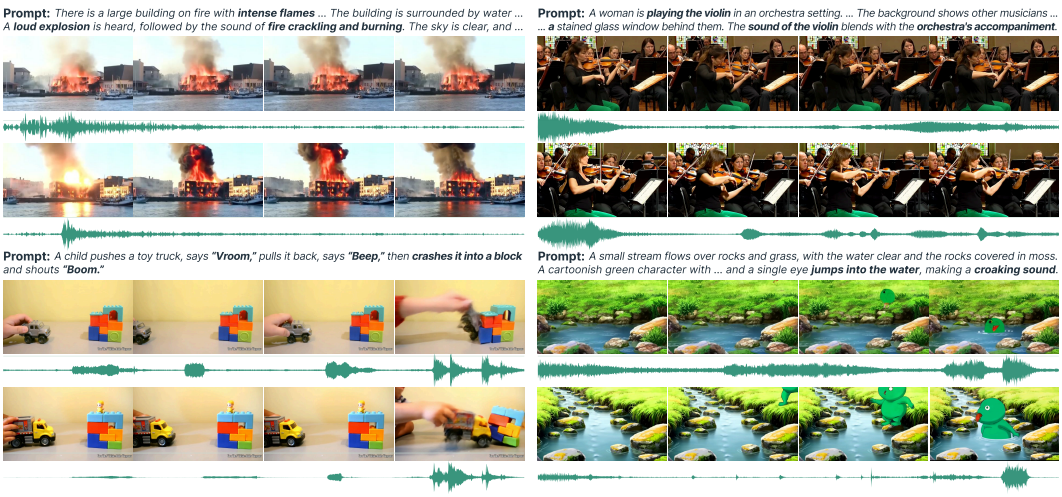


Figure 2: Qualitative comparison of LTX-2.3 and AV-GRPO (full). For each example, rows 2–3 show the video frames and corresponding audio waveform generated by the original LTX-2.3 model; rows 4–5 show those generated by AV-GRPO (full). Bold text in the prompt highlights the visual events and sound cues that require audio–video synchronization.

- **Sound Source Type (H)**: Human voice, Environmental and object sounds, Music, Mixed sources (two or more types).
- **Synchronization Difficulty (D)**: Irrelevant (audio and video independent), Category matching (audio category matches the scene), Frame-level synchronization (audio and video fully aligned), Physical causal synchronization (e.g., pitch rising as a train approaches).
- **Temporal Complexity (T)**: Single-event (single scene/event, no scene transition), Multi-event (multiple scenes/causal events, with scene transitions).
- **Instruction Granularity (C)**: Weak constraint (core semantics only), Medium constraint (explicit sound source/scene/event), Strong constraint (fine-grained details including precise timing, spatial position, pitch/volume variations).

The five-dimensional decomposition yields $3 \times 4 \times 4 \times 2 \times 3 = 288$ orthogonal categories, with 20 prompts generated per category, resulting in a total of 5,760 data samples. This design promotes comprehensive coverage of audio-video generation application scenarios. Moreover, the sampling ratio of categories within each dimension can be flexibly adjusted according to the requirements of different training stages, enabling the dataset to adapt to diverse optimization objectives and model iteration paces. This design makes the training process traceable and model weaknesses localizable, providing clear guidance for data selection and reward function design.

3.2 Experimental Setup

Training configuration. All LoRA and full-parameter runs use a single node of eight NVIDIA A800 80 GB GPUs. LTX-2.3 22B generates 544×960 clips of 97 frames at 24 FPS with 15 denoising steps. The video tower is updated first, and the active tower switches every four steps. Appendix B provides the remaining implementation details.

3.2.1 Reward Models and Reward Composition

For video-tower optimization, we use VideoAlign (Liu et al., 2026b), CLIP (Radford et al., 2021), and DeSync (Iashin et al., 2024) to assess video quality, video–prompt alignment, and audio–video synchronization, respectively. For audio-tower optimization, we use Audiobox Aesthetics (Tjandra et al., 2025), CLAP (Wu et al., 2023), and DeSync to assess audio quality, audio–prompt alignment, and audio–video synchronization, respectively.

Following GDPO (Liu et al., 2026e), the video score sums standardized VideoAlign, CLIP, and negative DeSync, whereas the audio score sums standardized audio aesthetics, CLAP, and negative DeSync (Since lower DeSync is better, its sign is reversed). We standardize the resulting score again within the group to obtain the trajectory advantage. For audio, a CLAP guardrail uses only the alignment advantage when the group's average CLAP falls below a threshold; this prevents quality and synchronization gains from masking a loss of prompt fidelity. Appendix D gives the equations.

3.2.2 Evaluation Benchmarks

We adopt JavisBench (Liu et al., 2026c) and VABench (Hua et al., 2026) as our evaluation benchmarks to comprehensively assess the generated audio-video content.

JavisBench evaluates generated samples across four complementary dimensions: (1) **Unimodal Generation Fidelity**, measured by Visual Quality (VQ) and Audio Quality (AQ), assessing the perceptual quality of each modality independently; (2) **Text-Modal Alignment**, where ImageBind similarity, CLIP score, and CLAP score are employed to measure the semantic consistency between the textual prompt and each generated modality; (3) **Audio-Video Semantic Coherence**, evaluated via Audio-Video ImageBind similarity (AV-IB) and AVH Score to quantify cross-modal semantic alignment; (4) **Audio-Video Synchronization**, assessed by JavisScore and DeSync to measure temporal correspondence between audio and video streams.

VABench organizes its evaluation into two paradigms. The first paradigm relies on expert models to provide objective perceptual quality assessments, including speech quality and naturalness (SpeechQual&Nat), audio aesthetics (AudioAesthetic), and lip synchronization accuracy (Lip-Sync). The second paradigm leverages Multimodal Large Language Models (MLLMs) to simulate human judgments on complex audio-video semantics, covering high-level criteria such as global Alignment, Artistic quality (Artistry), and Expressiveness.

3.3 Main Results and Analysis

3.3.1 Quantitative Analysis

Table 1 and 2 report the results of AV-GRPO under both LoRA and full fine-tuning on JavisBench and VABench. Compared with LTX-2.3 (22B) and GDPO, AV-GRPO consistently improves nearly all metrics under both settings, with full fine-tuned model attaining the best overall performance.

In terms of perceptual quality, the full fine-tuned model improves AQ on JavisBench from 5.097 to 5.798 and Audio Aesthetics on VABench from 3.319 to 3.631 over LTX-2.3. For semantic alignment, it boosts CLIP on JavisBench from 0.318 to 0.327 and CLAP from 0.408 to 0.468. For cross-modality consistency and synchronization, substantial gains are observed across the board: AVH Score, DeSync, and AV-Align on JavisBench, as well as Lip Sync and DeSync on VABench, all exhibit significant improvements. In all these aspects, AV-GRPO consistently outperforms GDPO by a considerable margin.

The two training regimes show different strengths. Full fine-tuning yields the highest JavisBench AQ and AV-IB scores (5.798 and 0.247), while LoRA attains a slightly higher CLIP score (0.330 versus 0.327) and a lower DeSync value (0.554 versus 0.607). The improvements are therefore broad rather than uniform: LoRA's VQ is 5.816, below the base model's 5.855, and the full model's VABench visual-realism score is 4.395 versus 4.399 for the base model. These exceptions do not change the gains in most quality and alignment measures, but they show that different modalities and training regimes retain distinct tradeoffs.

Compared with GDPO, AV-GRPO (full) raises JavisBench AV-IB from 0.224 to 0.247 and JavisScore from 0.202 to 0.222, reducing DeSync from 0.708 to 0.607. On VABench, its Lip Sync increases from 1.439 to 1.646 and DeSync decreases from 0.726 to 0.542. These comparisons focus on cross-modal measures, where controlling the ref-modality is most relevant to the learning signal.

3.3.2 Qualitative Analysis

Figure 2 showcases several representative cases. As shown in the figure, compared with the original model, AV-GRPO-trained LTX-2.3 achieves improvements across different aspects: per-modality fidelity, modality-text alignment, and cross-modal synchronization.

Table 1: Main results on JavisBench. Best results are in **bold**, second-best are underlined. VQ: Visual Quality, AQ: Audio Quality. (↑: higher is better; ↓: lower is better).

Model	size	AV-Quality		Text-Consistency				AV-Consistency		AV-Synchrony		
		VQ↑	AQ↑	TV-IB↑	TA-IB↑	CLIP↑	CLAP↑	AV-IB↑	AVHScore↑	JavisScore↑	DeSync↓	AV-align↑
-T2A+A2V												
TempoTKn	1.3B	-	-	0.084	-	0.205	-	0.139	0.122	0.103	1.532	-
TPoS	1.0B	-	-	0.201	-	0.229	-	0.124	0.129	0.095	1.493	-
-T2V+V2A												
ReWaS	0.6B	-	-	-	0.123	-	0.280	0.110	0.104	0.079	1.071	-
See&Hear	0.4B	-	-	-	0.129	-	0.263	0.160	0.143	0.112	1.099	-
FolleyCrafter	1.2B	-	-	-	0.149	-	0.383	0.193	0.186	0.151	0.952	-
MMAudio	0.1B	-	-	-	0.160	-	0.407	0.198	0.182	0.150	0.849	-
-T2AV												
JavisDiT	3.1B	1.291	4.478	0.263	0.143	0.302	0.391	0.197	0.179	0.154	1.039	-
UniVerse-1	6.4B	1.357	4.839	0.272	0.111	0.309	0.245	0.104	0.098	0.077	0.929	-
JavisDiT++	2.1B	1.462	5.049	0.282	0.164	0.316	0.424	0.198	0.184	0.159	0.832	-
LTX-2.3	22B	5.855	5.097	0.284	0.143	0.318	0.408	0.212	0.201	0.183	0.757	0.354
LTX-2.3+GDPO	22B	5.870	5.406	0.283	0.165	0.313	0.431	0.224	0.217	0.202	0.708	0.360
LTX-2.3+AV-GRPO(LoRA)	22B	5.816	5.633	**0.289**	0.174	**0.330**	0.449	0.236	0.235	0.213	**0.554**	0.382
LTX-2.3+AV-GRPO(full)	22B	**5.902**	**5.798**	0.288	**0.185**	0.327	**0.468**	**0.247**	**0.241**	**0.222**	0.607	**0.398**

Table 2: Main results on VABench. Best results are in **bold**, second-best are underlined. For all metrics except DeSync, higher is better; for DeSync, lower is better.

Model	Speech Q&N	Audio Aes	T-V Align	T-A Align	A-V Align	Lip Sync	DeSync↓	Alignment	Express-siveness	Visual Realism	Audio Realism	Artistry
LTX-2.3	1.383	3.319	0.211	0.398	0.227	1.351	0.800	4.455	4.371	4.399	3.830	3.766
LTX-2.3+GDPO	1.465	3.452	0.210	0.424	0.243	1.439	0.726	4.481	4.409	**4.413**	3.849	3.784
LTX-2.3+AV-GRPO(lora)	1.487	3.553	**0.217**	0.446	0.261	1.585	0.594	**4.512**	**4.450**	4.402	3.865	3.792
LTX-2.3+AV-GRPO(full)	**1.510**	**3.631**	0.215	**0.468**	**0.272**	**1.646**	**0.542**	4.510	4.434	4.395	**3.878**	**3.798**

The examples span event-driven effects, music performance, object sounds, and natural ambience. In each row, the video frames and corresponding waveform are shown together so that visual changes can be compared with the timing and character of the generated audio. These cases complement the aggregate benchmark scores by illustrating the kinds of cross-modal relationships optimized through the shared-anchor comparisons.

3.4 Ablation Study

3.4.1 Dataset

To demonstrate the effectiveness of our proposed dataset, we conduct a comparative experiment against the well-known audio-video dataset VGGSound (Chen et al., 2020) under full fine-tuning settings. All training configurations are kept identical across experiments. The results on JavisBench-mini are reported in Figure 3.

We conduct ablation studies on the difficulty levels of Synchronization Difficulty (**D**) and Instruction Granularity (**C**). Specifically, we compare the easiest and most difficult categories within each of these two dimensions (denoted as **D/C easiest/hardest**) while keeping all other settings identical. The experimental results are reported in Figure 3. The results demonstrate that the decoupling and categorical partitioning of dimensions and difficulty levels contribute substantially to the model's performance.

The left panel compares the two training sets across quality, text alignment, cross-modal consistency, and synchronization metrics; the middle panel reports the relative change when an easy or hard level of **C** or **D** is removed. Together, these views test both the value of the complete category grid and whether its difficulty levels supply distinct training signals. In particular, changes in the synchronization measures under the **D** removals show why a single aggregate benchmark score is insufficient to assess the data design.

3.4.2 Alternating Modality-wise Optimization

To verify the effectiveness of alternating optimization, we conduct ablation studies on whether to apply alternating optimization, and provide experimental results with varying switch intervals (i.e., the number of training steps per modality before switching to the other). All experiments are conducted on JavisBench-mini, as shown in Figure 3. The results demonstrate that alternating optimization substantially benefits the performance of both towers. Under our current configuration, switching every four steps achieves relatively optimal performance.

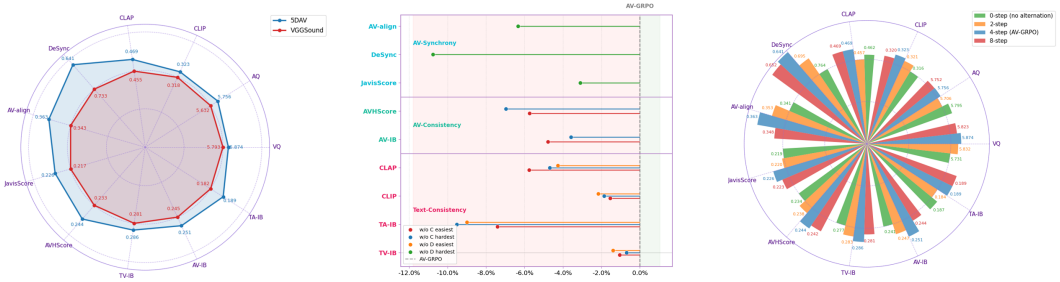


Figure 3: **Left**: Radar chart comparing performance on the 5DAV and VGGSound datasets. **Middle**: Metrics are reported as relative percentage changes with respect to the baseline. **Right**: Comparison of model performance with different alternating step intervals.

The right panel includes no alternation and intervals of two, four, and eight steps. The four-step schedule gives the strongest overall balance across the plotted measures, although individual metrics need not peak at the same interval. This comparison motivates the switching interval used for the main LoRA and full-parameter runs.

4 Related Work

4.1 Joint Audio–Video Generation

Early audio–visual generation methods typically adopt asymmetric pipelines that synthesize one modality conditioned on the other. For example, FoleyCrafter (Zhang et al., 2026b) and MMAudio (Cheng et al., 2025) generate temporally aligned audio from video. While such methods benefit from strong unimodal priors, their asymmetric formulations limit direct bidirectional co-generation. Recent works instead generate audio and video within a unified process. JavisDiT (Liu et al., 2026c), UniVerse-1 (Wang et al., 2025), Ovi (Low et al., 2025) and LTX-2 (HaCohen et al., 2026) employ interacting dual-stream architectures with cross-modal fusion. These advances primarily focus on model architecture, data construction, and large-scale pretraining. In contrast, our work studies reward-guided post-training of joint audio–video models, aiming to improve both modalities without entangling their learning signals.

4.2 Reinforcement Learning-based Post-training

Reward-guided post-training has been widely explored for diffusion and flow-based generation. DDPO (Black et al., 2023) and DPOK (Fan et al., 2023) formulate reverse denoising as a sequential decision process and optimize diffusion models using policy gradients. More recently, Flow-GRPO (Liu et al., 2026a) converts deterministic flow sampling into an equivalent stochastic process for online GRPO. DiffusionNFT (Zheng et al., 2026) instead performs online reward optimization through the forward process. For multiple objectives, GDPO (Liu et al., 2026e) independently normalizes individual rewards before aggregation, but does not resolve cross-modal credit assignment when both modalities vary simultaneously.

5 Conclusion

We presented AV-GRPO, a modality-anchored reinforcement learning framework for joint audio–video post-training. By combining anchored rollouts with alternating frozen-tower optimization, AV-GRPO disentangles rewards and controls comparison conditions, reduces training overhead, redirects credit assignment, and adapts to modality-specific dynamics and sample difficulty. We also introduced 5DAV, a five-dimensionally decoupled dataset for controllable post-training. Experiments on JavisBench and VABench show consistent improvements over LTX-2.3 and GDPO in perceptual quality, text alignment, and audio–video coherence and synchronization under both LoRA and full fine-tuning. These results demonstrate an effective and scalable approach to improving coupled audio–video generators.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript polishing and grammar checking. We have not used generative AI tools for conducting experiments or deriving mathematical results, and all other tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for no additional tasks. We have reviewed all AI-assisted work. The authors checked all AI-assisted edits to ensure grammatical correctness and preserve the intended technical meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The main paper describes the AV-GRPO algorithm, the construction of the 5DAV dataset, and the experimental setup, including evaluation benchmarks and metrics. Appendix B provides additional training configurations and hyperparameters, while Appendix D details reward computation and normalization. The assumptions and proofs underlying our theoretical analysis are provided in Appendix A.

REFERENCES

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

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

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 *CVPR*, 2025.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, P. Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *ArXiv*, abs/2305.16381, 2023. URL `https://api.semanticscholar.org/CorpusID:258947323`.

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.

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

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

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026a.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026b.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *ICLR*, 2026c.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *The Fourteenth International Conference on Learning Representations*, 2026d.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcák, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *ArXiv*, abs/2601.05242, 2026e. URL https://api.semanticscholar.org/CorpusID:284543897.

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

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. URL https://api.semanticscholar.org/CorpusID:269149104.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

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

Meituan LongCat Team, Bairui Wang, Bin Xiao, Bo Zhang, Bolin Rong, Borun Chen, Chang Wan, Chao Zhang, Chen Huang, Chen Chen, et al. Longcat-flash-omni technical report. *arXiv preprint arXiv:2511.00279*, 2025.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

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

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

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

Guohui Zhang, Xiaoxiao Ma, Jie Huang, Hang Xu, Hu Yu, Siming Fu, Yuming Li, Zeyue Xue, Lin Song, Haoyang Huang, Nan Duan, and Feng Zhao. Omninft: Modality-wise omni diffusion reinforcement for joint audio-video generation. *ArXiv*, abs/2605.12480, 2026a. URL https://api.semanticscholar.org/CorpusID:288257189.

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, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. In *International Conference on Learning Representations*, volume 2026, pp. 134129–134150, 2026.

A Proof of the Convergence and Joint Optimality of AV-GRPO

This section studies KL-regularized reinforcement learning fine-tuning of audio-video two-tower diffusion models under the protocol of "alternately fixing a single-modality complete denoising trajectory". We introduce a common reference joint law P_0, and specify the two single-tower operation kernels as the regular conditional kernels of P_0; the reward r is bounded measurable, and $\beta > 0$. We define the exponentially tilted joint law $\mathrm{d}P^*/\,\mathrm{d}P_0 \propto e^{r/\beta}$, and construct the two conditional Gibbs optimal kernels K_v^*, K_a^*. Under the Dobrushin contraction condition $\delta_v\delta_a < 1$, we prove: P^* is the unique global optimum of the joint KL-regularized objective; K_v^*, K_a^* are the unique optimal kernels of the conditional subproblems, respectively, and are automatically compatible; the systematic-scan Gibbs chain with kernels K_v^*, K_a^* has a unique stationary distribution P^*; starting from any initial joint law, the joint laws of both sampling phases converge geometrically in total variation to P^*, with an explicit convergence rate. If the model implements K_v^*, K_a^* via a non-interfering exact kernel oracle, then both the training rollouts and the inference distribution using the same alternating full-trajectory protocol possess the above convergence properties.

A.1 Problem Setup and Basic Assumptions

A.1.1 Spaces and the Reference Joint Law

Fix the text condition c, which is omitted in the following. Let $\mathcal{X}$ be the space of complete video denoising trajectories, and $\mathcal{Y}$ the space of complete audio denoising trajectories. Assume that $\mathcal{X}, \mathcal{Y}$ are spaces equipped with fixed Polish topologies, with their Borel σ-algebras denoted by $\mathcal{B}_\mathcal{X}, \mathcal{B}_\mathcal{Y}$, respectively. Thus $\mathcal{X} \times \mathcal{Y}$ is also a Polish space, and regular conditional probability kernels exist. Let $\mathcal{P}(\mathcal{X}), \mathcal{P}(\mathcal{Y}), \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ denote the corresponding sets of probability measures.

Assumption 1 (Common reference joint law H0)**.** There exists a reference joint probability measure $P_0 \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$. The actual reference kernel for "running the video tower with the audio trajectory fixed", $K_v^0 : \mathcal{Y} \times \mathcal{B}_\mathcal{X} \to [0, 1]$, and the reference kernel for "running the audio tower with the video trajectory fixed", $K_a^0 : \mathcal{X} \times \mathcal{B}_\mathcal{Y} \to [0, 1]$, are two versions of the regular conditional distributions of P_0:

$$K_v^0(\cdot \mid y) = P_0(\,\mathrm{d}x \mid y), \quad P_{0,Y}\text{-a.s.}, \qquad K_a^0(\cdot \mid x) = P_0(\,\mathrm{d}y \mid x), \quad P_{0,X}\text{-a.s.},$$

and we select a jointly measurable version of them that is defined on all of $\mathcal{X}, \mathcal{Y}$. Hence

$$P_0(A \times B) = \int_B K_v^0(A \mid y)\, P_{0,Y}(\,\mathrm{d}y) = \int_A K_a^0(B \mid x)\, P_{0,X}(\,\mathrm{d}x).$$

A.1.2 Reward and the Joint Objective

Let $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a bounded Borel measurable function, $\beta > 0$, and let $M = \|r\|_\infty$. Define the joint objective

$$J(P) = \mathbb{E}_P[r] - \beta\, D_{\mathrm{KL}}(P \parallel P_0), \qquad P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}),$$

with the convention that $D_{\mathrm{KL}}(P \parallel P_0) = \infty$ when $P \not\ll P_0$, so that $J(P) = -\infty$.

A.1.3 Conditional Subproblem Objectives

Fix $y \in \mathcal{Y}$; for any $q \in \mathcal{P}(\mathcal{X})$, define the video subproblem

$$L_v(q; y) = \int_\mathcal{X} r(x, y)\, q(\,\mathrm{d}x) - \beta\, D_{\mathrm{KL}}\big(q \parallel K_v^0(\cdot \mid y)\big).$$

Fix $x \in \mathcal{X}$; for any $q \in \mathcal{P}(\mathcal{Y})$, define the audio subproblem

$$L_a(q; x) = \int_{\mathcal{Y}} r(x, y)\, q(\,\mathrm{d}y) - \beta\, D_{\mathrm{KL}}\big(q \parallel K_a^0(\cdot \mid x)\big).$$

A.2 Exponential Tilting and Conditional Gibbs Optimal Kernels

A.2.1 Joint Exponential Tilting

Define the normalizing constant and the probability measure

$$Z = \int_{\mathcal{X}\times\mathcal{Y}} e^{r(x,y)/\beta}\, P_0(\,\mathrm{d}x,\, \mathrm{d}y), \qquad \frac{\mathrm{d}P^*}{\mathrm{d}P_0}(x, y) = Z^{-1} e^{r(x,y)/\beta}.$$

Since r is bounded, $e^{-M/\beta} \le Z \le e^{M/\beta}$, hence $0 < Z < \infty$, and P^* and P_0 are mutually absolutely continuous. Their marginals are denoted by P_X^* and P_Y^*, respectively.

A.2.2 Conditional Normalizing Constants and Optimal Kernels

Define

$$Z_v(y) = \int_{\mathcal{X}} e^{r(x,y)/\beta}\, K_v^0(\,\mathrm{d}x \mid y), \qquad Z_a(x) = \int_{\mathcal{Y}} e^{r(x,y)/\beta}\, K_a^0(\,\mathrm{d}y \mid x).$$

By the measurability theorem for kernel integrals, Z_v, Z_a are measurable, and $e^{-M/\beta} \le Z_v(y), Z_a(x) \le e^{M/\beta}$, so the denominators are everywhere positive and finite. Define the optimal conditional kernels

$$K_v^*(\,\mathrm{d}x \mid y) = Z_v(y)^{-1} e^{r(x,y)/\beta} K_v^0(\,\mathrm{d}x \mid y),$$

$$K_a^*(\,\mathrm{d}y \mid x) = Z_a(x)^{-1} e^{r(x,y)/\beta} K_a^0(\,\mathrm{d}y \mid x).$$

The integrands are jointly measurable, and the normalizing constants are measurable and everywhere positive; therefore K_v^*, K_a^* are probability kernels defined on all conditional states.

A.3 Joint KL Variational Identity

Lemma 1 (Joint Gibbs variational identity)**.** *For any $P \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, in the extended real-valued sense,*

$$J(P) = \beta \log Z - \beta\, D_{\mathrm{KL}}(P \parallel P^*).$$

Therefore P^* is the unique global optimum of J over all joint probability measures, with optimal value $\beta \log Z$.

Proof. Since $\mathrm{d}P^* / \mathrm{d}P_0 = e^{r/\beta}/Z$ has strictly positive finite upper and lower bounds, P^* and P_0 are equivalent. Therefore $P \ll P_0$ if and only if $P \ll P^*$. If this condition does not hold, both sides are $-\infty$, and the identity holds. Assume $P \ll P_0$. By the Radon–Nikodym chain rule, P-a.s. we have

$$\log \frac{\mathrm{d}P}{\mathrm{d}P^*} = \log \frac{\mathrm{d}P}{\mathrm{d}P_0} - \frac{r}{\beta} + \log Z.$$

Integrating both sides against P and multiplying by β, we obtain

$$\beta\, D_{\mathrm{KL}}(P \parallel P^*) = \beta\, D_{\mathrm{KL}}(P \parallel P_0) - \mathbb{E}_P[r] + \beta \log Z.$$

Because r is bounded, the two KL terms are either simultaneously finite or simultaneously ∞, so the above identity is unambiguous in the extended real-valued sense. Rearranging yields

$$J(P) = \beta \log Z - \beta\, D_{\mathrm{KL}}(P \parallel P^*) \le \beta \log Z.$$

Finally, $D_{\mathrm{KL}}(P \parallel P^*) = 0$ if and only if $P = P^*$. Hence P^* is the unique optimum. □

A.4 Optimal Kernels of the Conditional Subproblems

Lemma 2 (Conditional Gibbs optimality)**.** *For each $y \in \mathcal{Y}$, $K_v^*(\cdot \mid y)$ is the unique global optimum of $L_v(\cdot; y)$ over $\mathcal{P}(\mathcal{X})$, with optimal value $\beta \log Z_v(y)$. For each $x \in \mathcal{X}$, $K_a^*(\cdot \mid x)$ is the unique global optimum of the audio subproblem, with optimal value $\beta \log Z_a(x)$.*

Proof. Fix y, and replace P_0, P^*, Z, r by $K_v^0(\cdot \mid y), K_v^*(\cdot \mid y), Z_v(y), r(\cdot, y)$, respectively. The Radon–Nikodym computation of Lemma 1 applies verbatim, giving

$$L_v(q; y) = \beta \log Z_v(y) - \beta\, D_{\mathrm{KL}}\big(q \parallel K_v^*(\cdot \mid y)\big).$$

KL is nonnegative and vanishes only when $q = K_v^*(\cdot \mid y)$, so the conclusion holds. The audio side is entirely symmetric. □

A.5 Automatic Compatibility of the Two Optimal Conditional Kernels

Lemma 3 (Marginals and regular conditional kernels of P^*)**.**

$$P_Y^*(\,\mathrm{d}y) = \frac{Z_v(y)}{Z}\, P_{0,Y}(\,\mathrm{d}y), \qquad P_X^*(\,\mathrm{d}x) = \frac{Z_a(x)}{Z}\, P_{0,X}(\,\mathrm{d}x).$$

Hence P_Y^* is equivalent to $P_{0,Y}$, and P_X^* is equivalent to $P_{0,X}$. The selected K_v^*, K_a^* are, respectively, versions of the video and audio regular conditional distributions of P^*.

Proof. For any $B \in \mathcal{B}_{\mathcal{Y}}$, using the decomposition of P_0 via K_v^0:

$$P_Y^*(B) = Z^{-1} \int_B \int_{\mathcal{X}} e^{r/\beta} K_v^0(\,\mathrm{d}x \mid y) P_{0,Y}(\,\mathrm{d}y) = Z^{-1} \int_B Z_v(y) P_{0,Y}(\,\mathrm{d}y).$$

Thus $P_Y^*(\,\mathrm{d}y) = Z^{-1} Z_v(y) P_{0,Y}(\,\mathrm{d}y)$. Z_v has strictly positive finite upper and lower bounds, so the two audio marginals are equivalent. The video marginal formula follows analogously.

Next, taking $A \in \mathcal{B}_{\mathcal{X}}, B \in \mathcal{B}_{\mathcal{Y}}$, we have

$$\int_B K_v^*(A \mid y) P_Y^*(\,\mathrm{d}y) = Z^{-1} \int_B \int_A e^{r/\beta} K_v^0(\,\mathrm{d}x \mid y) P_{0,Y}(\,\mathrm{d}y) = P^*(A \times B).$$

This is precisely the definition of K_v^* being a regular conditional distribution of P^*. K_a^* is symmetric. □

A.6 Alternating Kernels, Marginal Operators, and the Systematic-Scan Chain

A.6.1 Kernel–Measure Mixing

For $\mu \in \mathcal{P}(\mathcal{Y})$, $\nu \in \mathcal{P}(\mathcal{X})$, define the joint distributions of the two phases

$$M_v\mu(\,\mathrm{d}x,\, \mathrm{d}y) = K_v^*(\,\mathrm{d}x \mid y)\mu(\,\mathrm{d}y), \qquad M_a\nu(\,\mathrm{d}x,\, \mathrm{d}y) = \nu(\,\mathrm{d}x) K_a^*(\,\mathrm{d}y \mid x).$$

Define the corresponding marginal operators

$$V\mu(\,\mathrm{d}x) = \int_{\mathcal{Y}} K_v^*(\,\mathrm{d}x \mid y)\mu(\,\mathrm{d}y), \qquad W\nu(\,\mathrm{d}y) = \int_{\mathcal{X}} K_a^*(\,\mathrm{d}y \mid x)\nu(\,\mathrm{d}x).$$

Here the Y-marginal of $M_v\mu$ is μ, and its X-marginal is $V\mu$; the X-marginal of $M_a\nu$ is ν, and its Y-marginal is $W\nu$. Let $F = WV : \mathcal{P}(\mathcal{Y}) \to \mathcal{P}(\mathcal{Y})$.

A.6.2 One Complete Gibbs Sweep

Starting from any joint law $Q \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$, first keep y and resample $x' \sim K_v^*(\cdot \mid y)$, then keep x' and resample $y' \sim K_a^*(\cdot \mid x')$. Denote one round of transition by G; then

$$QG = M_a(VQ_Y), \qquad (QG)_Y = WVQ_Y = FQ_Y.$$

In particular, the joint law after one round depends only on the Y-marginal of the input joint law.

A.6.3 INVARIANCE OF P^*

By Lemma 3, $P^* = M_v P_Y^* = M_a P_X^*$, and moreover $P_X^* = V P_Y^*$, $P_Y^* = W P_X^*$. Therefore

$$P^* G = M_a(V P_Y^*) = M_a P_X^* = P^*.$$

So P^* is a stationary distribution of the alternating Gibbs chain. This step requires no ergodicity or convergence assumptions.

A.7 DOBRUSHIN CONTRACTION AND THE BASIC CONTRACTION LEMMA

A.7.1 TOTAL VARIATION AND DOBRUSHIN COEFFICIENTS

Define

$$\|\mu - \mu'\|_{\mathrm{TV}} = \sup_A |\mu(A) - \mu'(A)| \in [0,1],$$

$$\delta_v = \sup_{y,y'} \|K_v^*(\cdot \mid y) - K_v^*(\cdot \mid y')\|_{\mathrm{TV}}, \qquad \delta_a = \sup_{x,x'} \|K_a^*(\cdot \mid x) - K_a^*(\cdot \mid x')\|_{\mathrm{TV}}.$$

Assumption 2 (Verifiable weak coupling/contraction condition H1)**.**

$$\rho = \delta_v \delta_a < 1.$$

This condition is used to derive uniqueness and convergence. The supremum here is taken with respect to the full-domain conditional kernel versions selected above; H1 is an assumption on these versions.

A.7.2 MARKOV KERNEL CONTRACTION LEMMA

Lemma 4 (Dobrushin contraction)**.** *Let $K : S \times \mathcal{B}_T \to [0,1]$ be a probability kernel, and $\delta(K) = \sup_{z,z'} \|K(\cdot \mid z) - K(\cdot \mid z')\|_{\mathrm{TV}}$. Then for any $\eta, \eta' \in \mathcal{P}(S)$,*

$$\|\eta K - \eta' K\|_{\mathrm{TV}} \le \delta(K) \|\eta - \eta'\|_{\mathrm{TV}}.$$

Proof. Let $\sigma = \eta - \eta'$. If $t = \|\eta - \eta'\|_{\mathrm{TV}} = 0$, the conclusion is obvious. Otherwise the Jordan decomposition of σ satisfies $\sigma^+ = tp, \sigma^- = tq$, where p, q are probability measures. For any $C \in \mathcal{B}_T$, let $f_C(z) = K(C \mid z)$; then

$$|(\eta K - \eta' K)(C)| = t \left| \int f_C \, \mathrm{d}p - \int f_C \, \mathrm{d}q \right| \le t(\sup_z f_C(z) - \inf_z f_C(z)) \le t\delta(K).$$

Taking the supremum over C yields the conclusion. □

A.7.3 APPLICATION TO V, W

$$\|V\mu - V\mu'\|_{\mathrm{TV}} \le \delta_v \|\mu - \mu'\|_{\mathrm{TV}}, \qquad \|W\nu - W\nu'\|_{\mathrm{TV}} \le \delta_a \|\nu - \nu'\|_{\mathrm{TV}},$$
$$\|F\mu - F\mu'\|_{\mathrm{TV}} \le \rho \|\mu - \mu'\|_{\mathrm{TV}}, \qquad \rho = \delta_v \delta_a < 1.$$

A.8 EXISTENCE, UNIQUENESS, AND IDENTIFICATION OF THE SELF-CONSISTENT MARGINALS

Lemma 5 (Unique self-consistent marginals)**.** *Under H1, $F = WV$ is a contraction mapping on the complete metric space $(\mathcal{P}(\mathcal{Y}), \|\cdot\|_{\mathrm{TV}})$. Hence there exists a unique $\mu^* \in \mathcal{P}(\mathcal{Y})$ satisfying $F\mu^* = \mu^*$. Let $\nu^* = V\mu^*$; then (μ^*, ν^*) is the unique solution of the equations $\mu = W\nu$, $\nu = V\mu$.*

Proof. Finite signed measures form a Banach space under the total variation norm, and $\mathcal{P}(\mathcal{Y})$ is a closed subset thereof, so $(\mathcal{P}(\mathcal{Y}), \mathrm{TV})$ is complete. By Lemma 4, the Lipschitz constant of F is at most $\rho < 1$. The Banach fixed point theorem yields the unique fixed point μ^*; and for any $\mu_0 \in \mathcal{P}(\mathcal{Y})$,

$$\|F^n \mu_0 - \mu^*\|_{\mathrm{TV}} \le \rho^n \|\mu_0 - \mu^*\|_{\mathrm{TV}}.$$

Let $\nu^* = V\mu^*$; then $W\nu^* = WV\mu^* = F\mu^* = \mu^*$, so (μ^*, ν^*) is self-consistent. Conversely, any self-consistent pair (μ, ν) satisfies $\mu = W\nu = WV\mu = F\mu$, hence $\mu = \mu^*$; and then $\nu = V\mu = V\mu^* = \nu^*$. □

Remark 1 (Identification with the marginals of the jointly optimal distribution)**.** By Lemma 3, the two marginals of P^* satisfy

$$P_X^* = VP_Y^*, \qquad P_Y^* = WP_X^* = WVP_Y^* = FP_Y^*.$$

So P_Y^* is a fixed point of F. By uniqueness, $\mu^* = P_Y^*$, $\nu^* = P_X^*$. Thus the self-consistent marginals not only exist and are unique, but are exactly the marginals of the jointly KL-optimal distribution.

A.9 The Unique Stationary Joint Distribution

Lemma 6 (Unique stationary distribution of the alternating chain)**.** *Under H0, bounded reward, and H1, the stationary distribution of the systematic-scan Gibbs chain G exists and is unique, and the unique stationary distribution is P^*.*

Proof. Existence has already been proved in A.6.3: $P^*G = P^*$.

We now prove uniqueness. Let Q be any stationary distribution, i.e., $Q = QG$. By the structure of one round of updates,

$$Q = M_a(VQ_Y), \qquad Q_Y = WVQ_Y = FQ_Y.$$

So Q_Y is a fixed point of F. Lemma 5 gives $Q_Y = \mu^* = P_Y^*$. Substituting back into the joint distribution formula:

$$Q = M_a(V\mu^*) = M_a\nu^* = M_aP_X^* = P^*.$$

Therefore P^* is the unique stationary distribution. □

Remark 2 (Strict distinction among optimality, stationarity, and convergence)**.** That P^* is the unique joint optimum follows from the KL variational identity and does not require $\rho < 1$; the conditional compatibility of K_v^*, K_a^* follows from the common P_0 and the common reward tilting, and does not require $\rho < 1$; that P^* is a stationary distribution follows from the two conditional kernels of P^*, and does not require $\rho < 1$; uniqueness of the stationary distribution follows from the strict contraction of $F = WV$, and requires $\rho < 1$; geometric convergence from any initial value follows from Banach iteration and the TV isometry, and requires $\rho < 1$.

A.10 Geometric Rollout Convergence from Arbitrary Initial Laws

A.10.1 The Two Sampling Phases

Take any initial joint law Q_0, and let $\mu_0 = Q_{0,Y}$. For $n \geq 1$, recursively define

$$Q_n^v = M_v\mu_{n-1}, \qquad \nu_n = V\mu_{n-1},$$
$$Q_n^a = M_a\nu_n, \qquad \mu_n = W\nu_n = F\mu_{n-1}.$$

Q_n^v is the joint law after the video resampling; Q_n^a is the joint law after the subsequent audio resampling, and is also the chain distribution after one complete sweep.

A.10.2 TV Isometric Embedding of the Mixing Operators

M_v keeps y as is and only randomly generates x. Hence the non-expansiveness of Markov kernels gives $\|M_v\mu - M_v\mu'\|_{\mathrm{TV}} \leq \|\mu - \mu'\|_{\mathrm{TV}}$, while taking the Y-marginal gives the reverse inequality. Thus

$$\|M_v\mu - M_v\mu'\|_{\mathrm{TV}} = \|\mu - \mu'\|_{\mathrm{TV}}.$$

Similarly, $\|M_a\nu - M_a\nu'\|_{\mathrm{TV}} = \|\nu - \nu'\|_{\mathrm{TV}}$.

Theorem 1 (Geometric TV convergence of the two phases)**.** *Let $\rho = \delta_v\delta_a < 1$. Then for any Q_0 and $n \geq 1$:*

$$\|\mu_n - P_Y^*\|_{\mathrm{TV}} \leq \rho^n\|\mu_0 - P_Y^*\|_{\mathrm{TV}},$$
$$\|\nu_n - P_X^*\|_{\mathrm{TV}} \leq \delta_v\rho^{n-1}\|\mu_0 - P_Y^*\|_{\mathrm{TV}},$$
$$\|Q_n^v - P^*\|_{\mathrm{TV}} \leq \rho^{n-1}\|\mu_0 - P_Y^*\|_{\mathrm{TV}},$$
$$\|Q_n^a - P^*\|_{\mathrm{TV}} \leq \delta_v\rho^{n-1}\|\mu_0 - P_Y^*\|_{\mathrm{TV}}.$$

When $\rho = 0$ and $n = 1$, we take $\rho^0 = 1$ by convention.

Proof. The first inequality is the Banach iteration. The second follows from $\nu_n = V\mu_{n-1}$ and the δ_v-contraction of V. Since $P^* = M_v P_Y^* = M_a P_X^*$, combining the two isometry identities yields the third and fourth inequalities. □

A.11 Main Theorem

Theorem 2 (Idealized AV-GRPO Gibbs theorem)**.** *Fix c. Let $\mathcal{X}, \mathcal{Y}$ be Polish trajectory spaces, and suppose H0 holds; r is bounded measurable, and $\beta > 0$. Define P^*, K_v^*, K_a^* from P_0 and r. Assume the selected full-domain versions satisfy $\rho = \delta_v \delta_a < 1$. Then:*

1. *P^* is the unique global optimum of the joint objective $J(P) = \mathbb{E}_P[r] - \beta D_{\mathrm{KL}}(P \parallel P_0)$, with optimal value $\beta \log Z$;*
2. *K_v^*, K_a^* are, respectively, the unique optimal kernels of the conditional KL subproblems when the counterpart trajectory is fixed, and they are two versions of the regular conditional distributions of P^*;*
3. *the alternating Gibbs chain has a unique stationary distribution P^*;*
4. *the self-consistent marginal equations $\mu = W\nu,\ \nu = V\mu$ have a unique solution (P_Y^*, P_X^*);*
5. *starting from any initial joint law, the joint laws of the video phase and the audio phase both converge to P^* in total variation at the explicit rates of Theorem 1.*

Proof. Conclusion (1) is Lemma 1. Conclusion (2) follows from Lemma 2 and Lemma 3. The invariance of P^* follows from A.6.3; Lemma 6 then uses Dobrushin contraction to prove its uniqueness, giving (3). Lemma 5 identifies the unique self-consistent marginals, giving (4). Finally, Theorem 1 gives the geometric TV convergence of both phases, giving (5). All conclusions have been proved by the preceding lemmas. □

A.12 Connection with Alternating Training

Assumption 3 (exact-kernel non-interference oracle H2)**.** The video tower policy class contains the selected full-domain kernel K_v^*, and the audio tower policy class contains K_a^*. The video update oracle returns, for each y, the unique optimal kernel $K_v^*(\cdot \mid y)$ of Lemma 2; the audio update oracle returns, for each x, $K_a^*(\cdot \mid x)$. Furthermore, we assume independently that updating the parameters of either tower does not change the conditional kernel already realized by the other tower.

Corollary 1 (Idealized training–inference consistency)**.** *Under the main theorem and H2, after the video tower and the audio tower each complete one exact update, the two policy kernels are fixed as K_v^*, K_a^*, respectively. Thereafter, the joint distributions of the video phase and the audio phase of the training rollouts, as well as the inference distribution adopting the same full-trajectory alternating resampling protocol, all converge geometrically in TV to P^* from any initial joint law. The limiting joint distribution attains the global optimal value $\beta \log Z$ of J.*

Proof. By Lemma 2, the unique optimal solutions of the pointwise conditional subproblems are K_v^* and K_a^*, respectively. H2 guarantees that the two towers realize these full-domain kernels after updating, and that subsequent updates do not change the other already-realized kernel. Therefore, after each tower has been updated once, all subsequent sampling is described exactly by the fixed operators V, W, M_v, M_a. Theorem 2(5) directly gives the geometric TV convergence of both phases and the inference chain to P^*; Theorem 2(1) gives the objective value of P^*. □

A.13 Conclusion

Under the common reference joint law H0, bounded reward, $\beta > 0$, and the Dobrushin contraction condition $\rho = \delta_v \delta_a < 1$, this section has proved:

1. the exponentially tilted joint law P^* is the unique global optimum of the joint KL-regularized objective, with optimal value $\beta \log Z$;

2. the two conditional Gibbs optimal kernels K_v^*, K_a^* are the unique optimal kernels of the conditional subproblems, respectively, are automatically compatible, and are the regular conditional distributions of P^*;
3. the systematic-scan Gibbs chain with kernels K_v^*, K_a^* has a unique stationary distribution P^*;
4. the self-consistent marginal equations have a unique solution, which is exactly the two marginals of P^*;
5. starting from any initial joint law, the joint laws of the two sampling phases converge geometrically in total variation to P^*, with explicit rates;
6. if the model implements K_v^*, K_a^* via a non-interfering exact-kernel oracle, then both the training rollouts and the inference distribution possess the same geometric convergence properties.

The proof is closed under the idealized assumptions: optimality comes from the KL variational identity, compatibility comes from the common reference joint law, and uniqueness and convergence come from Dobrushin contraction.

B Training Details

All training runs, including LoRA and full fine-tuning, are conducted on a single 8-GPU A800 (80 GB) node. All prompts in the training dataset are fully randomly shuffled (rather than sampled in category-ordered sequences). Given a sampled prompt, the LTX-2.3 model generates audio-video clips with resolution 544×960, 97 frames, and 24 FPS via 15-step denoising inference using the default random seed (42), producing 8 samples per prompt. The composition of the reward used to evaluate each audio-video sample is described in Appendix D. For LoRA training, the learning rate starts from 3×10^{-6} and follows a cosine decay schedule over 1440 steps in total (i.e., all data is seen once by each of the two towers: 5760 × 2/8 = 1440), with training stopped early at 480 steps. The noise level for sampling of the video tower is set to 0.02 with a KL coefficient of 0.01; for the audio tower, the sampling noise level is 0.8 and the KL coefficient is 0.002. GDPO is trained for 480 steps under identical hyperparameters and the same data ordering. Since GDPO directly generates paired audio–video samples and updates both towers simultaneously, the same data is used to train each tower twice. For full fine-tuning, the learning rate starts at 1×10^{-6} and follows a cosine decay schedule defined over 1440 steps (keeping the same total step count as LoRA), with training stopped early at 480 steps. The noise levels and KL coefficients for both towers are kept identical to those used in LoRA training. Alternating optimization is adopted: training switches between the two towers every 4 steps, with 8 prompts consumed per step and the video tower updated first. Within each round, the video tower is first trained on $4 \times 8 = 32$ prompts. Training then switches to the audio tower, which is updated using the same set of 32 prompts, before proceeding to the next round with fresh data.

C Flow-GRPO and LongCat-Video Framework

In this section, we present the complete details of the Flow-GRPO (Liu et al., 2026a) and the LongCat-Video (Team et al., 2025) framework.

C.1 Flow-GRPO

Flow-GRPO converts the deterministic flow ODE into an equivalent SDE that preserves identical marginal distributions with the original model at all timesteps. The reversed SDE for rectified flow is formulated as:

$$d\boldsymbol{x}_t = \left[\boldsymbol{v}_t(\boldsymbol{x}_t) + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_t(\boldsymbol{x}_t)\big)\right] dt + \sigma_t d\boldsymbol{w}. \tag{13}$$

where $d\boldsymbol{w}$ denotes the Wiener process increment, and σ_t is the time-dependent noise schedule. Flow-GRPO sets $\sigma_t = a\sqrt{\frac{t}{1-t}}$, where scalar hyperparameter a controls the stochastic intensity. Applying Euler-Maruyama discretization to the SDE yields the iterative update formula:

$$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \left[\boldsymbol{v}_\theta(\boldsymbol{x}_t, t) + \frac{\sigma_t^2}{2t}\big(\boldsymbol{x}_t + (1-t)\boldsymbol{v}_\theta(\boldsymbol{x}_t, t)\big)\right] \Delta t + \sigma_t \sqrt{\Delta t}\, \boldsymbol{\epsilon}, \tag{14}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ introducing Gaussian noise. From Equation (2), the transition distribution $\pi_\theta(\boldsymbol{x}_{t+\Delta t} \mid \boldsymbol{x}_t, c)$ is an isotropic Gaussian distribution. Thus, the KL divergence between current policy and reference policy admits a closed-form solution:

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\Delta t}{2} \left(\frac{\sigma_t (1-t)}{2t} + \frac{1}{\sigma_t} \right)^2 \| \boldsymbol{v}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{v}_{\text{ref}}(\boldsymbol{x}_t, t) \|^2. \tag{15}$$

Flow-GRPO adopts the group-relative formulation for advantage estimation. Given condition c, the model samples a group of G reverse trajectories. All timesteps of one trajectory share the identical advantage computed via intra-group reward normalization:

$$\hat{A}_t^i = \frac{R(\boldsymbol{x}_0^i, c) - \mathrm{mean}(\{R(\boldsymbol{x}_0^i, c)\}_{i=1}^G)}{\mathrm{std}(\{R(\boldsymbol{x}_0^i, c)\}_{i=1}^G)}. \tag{16}$$

The importance sampling ratio characterizes the probability ratio between old and new policy:

$$r_t^i(\theta) = \frac{p_\theta(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, c)}{p_{\theta_{\text{old}}}(\boldsymbol{x}_{t-1}^i \mid \boldsymbol{x}_t^i, c)}. \tag{17}$$

Flow-GRPO optimizes the policy by maximizing the following objective:

$$\mathcal{J}_{\text{Flow-GRPO}}(\theta) = \mathbb{E}_{c, \pi^{\text{old}}, \{\boldsymbol{x}^i\}_{i=1}^G} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{T} \sum_{t=0}^{T-1} \Big(\min \left(r_t^i \hat{A}_t^i,\ \mathrm{clip}(r_t^i, 1-\varepsilon, 1+\varepsilon) \hat{A}_t^i \right) - \beta D_{\mathrm{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Big) \Bigg]. \tag{18}$$

C.2 LongCat-Video

The gradient of the policy loss $\mathcal{L}_{\text{policy}}(\theta) = r_t^i(\theta) \hat{A}_t^i$ with respect to parameters θ can be derived. The gradient computation proceeds as follows:

$$\nabla_\theta \mathcal{L}_{\text{policy}}(\theta) = \hat{A}_t^i \nabla_\theta r_t^i(\theta). \tag{19}$$

$$\nabla_\theta r_t^i(\theta) = \frac{p_\theta \left(x_{t-1}^i \mid x_t^i, c \right)}{p_{\theta_{\text{old}}} \left(x_{t-1}^i \mid x_t^i, c \right)} \nabla_\theta \log p_\theta \left(x_{t-1}^i \mid x_t^i, c \right) = \nabla_\theta \log p_\theta \left(x_{t-1}^i \mid x_t^i, c \right). \tag{20}$$

Combining these results yields the policy gradient:

$$\nabla_\theta \mathcal{L}_{\text{policy}}(\theta) = \hat{A}_t^i r_t^i(\theta) \nabla_\theta \log p_\theta \left(x_{t-1}^i \mid x_t^i, c \right). \tag{21}$$

The score function $\nabla_\theta \log p_\theta \left(x_{t-1} \mid x_t, c \right)$ is computed next. The conditional distribution is Gaussian:

$$p_\theta \left(x_{t-1} \mid x_t, c \right) = \mathcal{N} \left(x_{t-1}; \mu_\theta \left(x_t, t, c \right), \sigma_t^2 \Delta t \mathbf{I} \right). \tag{22}$$

$$\nabla_\theta \log p_\theta = \frac{1}{\sigma_t^2 \Delta t} \left(x_{t-1} - \mu_\theta \right) \cdot \nabla_\theta \mu_\theta. \tag{23}$$

From the SDE sampling process, the following reparameterization holds:

$$x_{t-1} = \mu_\theta + \sigma_t \sqrt{\Delta t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{24}$$

Substituting:

$$\nabla_\theta \log p_\theta = \frac{1}{\sigma_t^2 \Delta t} \left(\sigma_t \sqrt{\Delta t} \epsilon \right) \cdot \nabla_\theta \mu_\theta = \frac{1}{\sigma_t \sqrt{\Delta t}} \epsilon \cdot \nabla_\theta \mu_\theta. \tag{25}$$

$$\mu_\theta = x_t + \left[v_\theta \left(x_t, t, c \right) + \frac{\sigma_t^2}{2t} \left(x_t + (1-t) v_\theta \left(x_t, t, c \right) \right) \right] (-\Delta t) \tag{26}$$

Simplifying the drift term:

$$\begin{aligned}\text{drift} &= v_\theta + \frac{\sigma_t^2}{2t}x_t + \frac{\sigma_t^2}{2t}(1-t)v_\theta \\ &= v_\theta\left(1 + \frac{\sigma_t^2(1-t)}{2t}\right) + \frac{\sigma_t^2}{2t}x_t\end{aligned} \tag{27}$$

Thus:

$$\mu_\theta = x_t - \Delta t \cdot \text{drift} \tag{28}$$

Taking the gradient with respect to θ (noting that x_t is constant):

$$\nabla_\theta \mu_\theta = -\Delta t \cdot \nabla_\theta \text{drift} = -\Delta t \cdot \left(1 + \frac{\sigma_t^2(1-t)}{2t}\right)\nabla_\theta v_\theta \tag{29}$$

Substituting into $\nabla_\theta \log p_\theta$:

$$\begin{aligned}\nabla_\theta \log p_\theta &= \frac{1}{\sigma_t\sqrt{\Delta t}}\epsilon \cdot \left[-\Delta t \cdot \left(1 + \frac{\sigma_t^2(1-t)}{2t}\right)\nabla_\theta v_\theta\right] \\ &= -\frac{\sqrt{\Delta t}}{\sigma_t}\left(1 + \frac{\sigma_t^2(1-t)}{2t}\right)\epsilon \cdot \nabla_\theta v_\theta\end{aligned} \tag{30}$$

Therefore, the gradient of the policy loss is:

$$\nabla_\theta \mathcal{L}_{\text{policy}}(\theta) = \hat{A}_t^i r_t^i(\theta) \cdot \left[-\frac{\sqrt{\Delta t}}{\sigma_t}\left(1 + \frac{\sigma_t^2(1-t)}{2t}\right)\epsilon \cdot \nabla_\theta v_\theta\right] \tag{31}$$

Substituting $a = 1$ and $\sigma_t = \sqrt{\frac{t}{1-t}}$ (so $\sigma_t^2 = \frac{t}{1-t}$). The coefficient term is computed as:

$$1 + \frac{\sigma_t^2(1-t)}{2t} = 1 + \frac{\frac{t}{1-t}\cdot(1-t)}{2t} = 1 + \frac{1}{2} = \frac{3}{2} \tag{32}$$

And the scaling term:

$$\frac{\sqrt{\Delta t}}{\sigma_t} = \frac{\sqrt{\Delta t}}{\sqrt{\frac{t}{1-t}}} = \sqrt{\Delta t}\cdot\sqrt{\frac{1-t}{t}} = \sqrt{\frac{\Delta t(1-t)}{t}} \tag{33}$$

Substituting these simplifications gives the final policy gradient expression:

$$\nabla_\theta \mathcal{L}_{\text{policy}}(\theta) = -\frac{3}{2}\hat{A}_t^i\sqrt{\frac{\Delta t(1-t)}{t}}\epsilon \cdot \nabla_\theta v_\theta \tag{34}$$

A reweighting coefficient is introduced, defined as:

$$\lambda_{\text{policy}}(t, \Delta t) = \kappa(t, \Delta t)^{-1} = \sqrt{\frac{t}{\Delta t(1-t)}} \tag{35}$$

The reweighted policy loss becomes:

$$\mathcal{L}_{\text{policy, reweighted}}(\theta) = \lambda_{\text{policy}}(t, \Delta t) \cdot \mathcal{L}_{\text{policy}}(\theta) \tag{36}$$

This yields the modified gradient:

$$\nabla_\theta \mathcal{L}_{\text{policy, reweighted}}(\theta) = -\frac{3}{2}\hat{A}_t^i \cdot \epsilon \cdot \nabla_\theta v_\theta \tag{37}$$

Similarly, the gradient of the KL divergence term can be derived as:

$$\nabla_\theta D_{\text{KL}}(\theta) = \Delta t \cdot \frac{9}{4} \cdot \frac{1-t}{t} \cdot (v_\theta - v_{\text{ref}}) \cdot \nabla_\theta v_\theta \tag{38}$$

This expression reveals that the KL loss gradient suffers from the same scaling issues as the policy loss gradient. To address this, a KL reweighting coefficient is also introduced:

$$\lambda_{\text{KL}}(t, \Delta t) = k_{\text{KL}}(t, \Delta t)^{-1} = \frac{t}{\Delta t(1-t)} \tag{39}$$

The reweighted KL loss becomes:

$$\mathcal{L}_{\text{KL, reweighted}}(\theta) = \lambda_{\text{KL}}(t, \Delta t) \cdot D_{\text{KL}}(\theta) \tag{40}$$

yielding the simplified gradient:

$$\nabla_\theta \mathcal{L}_{\text{KL, reweighted}}(\theta) = \frac{9}{4} \cdot (v_\theta - v_{\text{ref}}) \cdot \nabla_\theta v_\theta \tag{41}$$

Based on the reweighting coefficients for the policy loss and KL loss, the revised GRPO objective function is as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{c\sim\mathcal{C},\, t'\sim\mathcal{U}(0,T'-1),\\ \{\boldsymbol{x}^i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|c,t')}} \left[\frac{1}{G} \sum_{i=1}^{G} \left(\lambda_{\text{policy}} \left(\frac{t'}{T}, \Delta \frac{t'}{T} \right) \cdot \mathcal{L}_{\text{policy}}(\theta) \right.\right.$$
$$\left.\left. -\beta \lambda_{\text{KL}} \left(\frac{t'}{T}, \Delta \frac{t'}{T} \right) \cdot D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) \right) \right] \tag{42}$$

D Reward Models and Reward Composition

For video-tower optimization, we use VideoAlign (Liu et al., 2026b), CLIP (Radford et al., 2021), and DeSync (Iashin et al., 2024) to assess video quality, video–prompt alignment, and audio–video synchronization, respectively. For audio-tower optimization, we use Audiobox Aesthetics (Tjandra et al., 2025), CLAP (Wu et al., 2023), and DeSync to assess audio quality, audio–prompt alignment, and audio–video synchronization, respectively.

Following GDPO, we first compute group-wise normalized advantages for each individual metric, then sum them to obtain a cumulative score per sample, and finally perform a second group-wise normalization to obtain the final advantage $\hat{A}_t^i$.

Specifically, when training the video tower with audio anchored, the cumulative score for video sample i is:

$$S^i_{\text{video}} = \text{norm}(\text{VQ}^i) + \text{norm}(\text{CLIP}^i) + \text{norm}(\text{-Desync}^i),$$

where $\text{norm}(r^i) = \frac{r^i - \text{mean}(\{r^i\}_{i=1}^G)}{\text{std}(\{r^i\}_{i=1}^G)}$ denotes group-wise standardization. **Note that for the DeSync metric, the lower the better.** The final advantage is:

$$\hat{A}^i_{\text{video}} = \text{norm}\big(\{S^i_{\text{video}}\}_{i=1}^G\big).$$

Symmetrically, when training the audio tower with video anchored as τ_v, the cumulative score for audio sample i is:

$$S^i_{\text{audio}} = \text{norm}(\text{AQ}^i) + \text{norm}(\text{CLAP}^i) + \text{norm}(\text{-Desync}^i),$$

with the final advantage:

$$\hat{A}^i_{\text{audio}} = \text{norm}\big(\{S^i_{\text{audio}}\}_{i=1}^G\big).$$

In practice, however, we observe that when training the audio tower, the model is prone to reward hacking: it easily finds a shortcut that sacrifices the CLAP score while substantially boosting AQ and Desync. This behavior severely undermines audio-prompt semantic alignment and is unacceptable.

Inspired by GDPO, we introduce a **semantic alignment guardrail** for CLAP. Let the group-wise average CLAP score be $\bar{r}_{\text{CLAP}} = \frac{1}{G}\sum_{i=1}^{G} \text{CLAP}^i$, with a predefined threshold η_{CLAP}. Only when $\bar{r}_{\text{CLAP}} \geq \eta_{\text{CLAP}}$ are AQ and Desync incorporated into the advantage; otherwise, the CLAP advantage serves directly as the final advantage. Formally:

$$\hat{A}^i_{\text{audio}} = \begin{cases} \text{norm}\big(\{S^i_{\text{audio}}\}_{i=1}^G\big), & \bar{r}_{\text{CLAP}} \geq \eta_{\text{CLAP}}, \\ \text{norm}\big(\{\text{CLAP}^i\}_{i=1}^G\big), & \bar{r}_{\text{CLAP}} < \eta_{\text{CLAP}}. \end{cases}$$

When the group-average CLAP falls below the threshold, the model receives learning signals exclusively from CLAP, guiding it to prioritize audio-prompt semantic alignment before attending to other metrics.

E More Qualitative Results

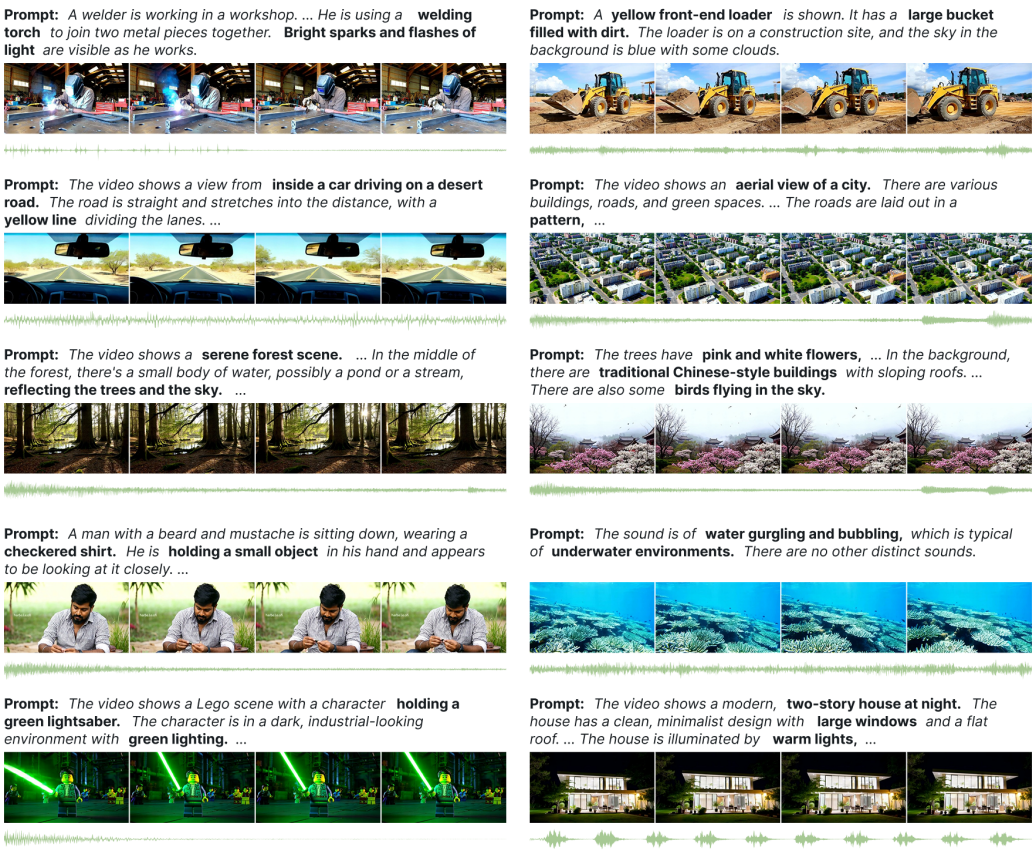


Figure 4: More qualitative results of AV-GRPO (full).