CORD: Bridging the Audio–Text Reasoning Gap via Weighted On-policy Cross-modal Distillation

Jing Hu[1,2*], Danxiang Zhu[1*], Xianlong Luo[1], Dan Zhang[1], Shuwei He[1,3], Yishu Lei[1], Shikun Feng[1†‡], Hai-Tao Zheng[2†], Jingzhou He[1], Yu Sun[1], Hua Wu[1], Haifeng Wang[1]
[1]ERNIE Team, Baidu, [2]Shenzhen International Graduate School, Tsinghua University,
[3]College of Computer Science, Inner Mongolia University,
cminuser@gmail.com, zhudanxiang@baidu.com,
fengshikun01@baidu.com, zheng.haitao@sz.tsinghua.edu.cn

Abstract

Large Audio Language Models (LALMs) have garnered significant research interest. Despite being built upon text-based large language models (LLMs), LALMs frequently exhibit a degradation in knowledge and reasoning capabilities. We hypothesize that this limitation stems from the failure of current training paradigms to effectively bridge the acoustic-semantic gap within the feature representation space. To address this challenge, we propose CORD, a unified alignment framework that performs online cross-modal self-distillation. Specifically, it aligns audio-conditioned reasoning with its text-conditioned counterpart within a unified model. Leveraging the text modality as an internal teacher, CORD performs multi-granularity alignment throughout the audio rollout process. At the token level, it employs on-policy reverse KL divergence with importance-aware weighting to prioritize early and semantically critical tokens. At the sequence level, CORD introduces a judge-based global reward to optimize complete reasoning trajectories via Group Relative Policy Optimization (GRPO). Empirical results across multiple benchmarks demonstrate that CORD consistently enhances audio-conditioned reasoning and substantially bridges the audio–text performance gap with only 80k synthetic training samples, validating the efficacy and data efficiency of our on-policy, multi-level cross-modal alignment approach.

1 Introduction

Large Language Models (LLMs) (Qwen et al., 2025; DeepSeek-AI et al., 2025; OpenAI, 2025) have demonstrated exceptional semantic understanding capabilities, sparking research to extend this intelligence to multimodal domains through end-to-end processing. In the audio domain, most state-of-the-art Large Audio-Language Models (LALMs) (Chu et al., 2023, 2024; Wu et al., 2025; Ding et al., 2025; Zeng et al., 2024; Fang et al., 2024) are built upon pretrained text-based LLMs by incorporating an audio encoder (Radford et al., 2022) and a modality alignment module. Raw audio signals are first encoded into acoustic representations, which are then projected into the LLM's embedding space through audio-text paired supervision. This paradigm implicitly assumes that training on audio-text interleaved data is sufficient to align audio and text into a unified semantic space.

However, recent studies (Wang et al., 2024; Cuervo et al., 2025) have revealed a persistent performance disparity between the two modalities. Despite receiving semantically equivalent inputs, LALMs often exhibit markedly inferior performance on audio-conditioned tasks compared to their text-conditioned counterparts. This gap is particularly pronounced in data-constrained regimes (Chu et al., 2024; Wu et al., 2025; Xu et al., 2025), where limited training samples expose the limitations of existing alignment mechanisms and highlight the need for more data-efficient cross-modal alignment.

Several recent approaches have attempted to improve audio-conditioned reasoning performance, primarily through supervised fine-tuning with labeled speech data or knowledge distillation from external text-based teachers. Despite their progress, these methods suffer from three fundamental limitations. **(1) Limited scalability.** Supervised fine-tuning (He et al., 2024; Minixhofer et al., 2025) relies on large-scale, high-quality annotated speech data, which is expensive to collect and difficult to scale across diverse tasks and domains. **(2) Off-policy distillation and distribution mismatch.** Teacher-based distillation methods (Cuervo et al., 2025; Wang et al., 2024; Tseng et al., 2025) typically provide supervision along the teacher's text-generation trajectories, rather than the student's actual audio-conditioned inference states. This off-

*Equal contribution.
†Corresponding author.
‡Project leader.

Findings of the Association for Computational Linguistics: ACL 2026, pages 31602–31612

policy supervision leads to distribution mismatch and limits the ability to correct accumulated audio-specific reasoning errors. **(3) Uniform token-level supervision.** Conventional KL-based distillation treats all tokens equally (Wang et al., 2025; Li et al., 2025; Tseng et al., 2025), failing to emphasize semantically critical tokens that drive cross-modal misalignment, and lacking sequence-level constraints to explicitly regulate global reasoning trajectories.

To address these challenges, we propose **CORD** (**C**ross-modal Weighted **O**n-policy **R**eward-guided **D**istillation), a unified alignment framework that performs online cross-modal self-distillation without relying on any external teacher. CORD leverages the model's internal text modality as an in-model teacher and conducts multi-granularity distillation along the student model's real audio-conditioned reasoning trajectories. By constructing on-policy alignment objectives based on audio modality rollouts, CORD directly aligns audio-conditioned reasoning behavior with its text-conditioned counterpart within a unified model architecture.

At the token level, CORD introduces a fine-grained weighting mechanism that prioritizes tokens exhibiting high cross-modal divergence as well as those appearing at early stages of reasoning, where semantic deviations are more likely to propagate and dominate the final outcome. This weighted reverse KL objective enables targeted correction of modality-specific reasoning errors. At the sequence level, CORD formulates a cross-modal reward function and employs Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to align complete reasoning trajectories, encouraging the audio modality to follow reasoning policies consistent with the text modality.

Extensive experiments on multiple reasoning benchmarks demonstrate that CORD substantially improves audio-conditioned reasoning performance. In particular, CORD reduces the audio–text performance gap by an average of **41.6%** on Qwen2-Audio-7B-Instruct and **44.8%** on Step-Audio2-mini, significantly outperforming conventional distillation baselines. On several tasks, CORD nearly eliminates the modality gap altogether, indicating that on-policy cross-modal self-distillation with an internal teacher provides an effective and scalable solution for mitigating reasoning degradation in LALMs.

Our contributions can be summarized as follows:

- **Internal on-policy cross-modal self-distillation.** We propose **CORD**, a fully in-model, weighted on-policy cross-modal self-distillation framework that aligns audio-conditioned reasoning with text-conditioned behavior without relying on any external teacher, avoiding off-policy mismatch and architecture-induced noise.

- **Multi-granularity alignment of reasoning trajectories.** CORD jointly enforces token-level and sequence-level alignment by emphasizing semantically critical and early reasoning tokens with weighted reverse KL, while regulating global reasoning trajectories via a reward-guided GRPO objective.

- **Effective reduction of the audio-text reasoning gap.** We demonstrate that CORD consistently reduces the audio-text gap by over 40% across various backbones, nearly reaching parity with text-conditioned performance in several reasoning tasks.

2 Related Work

2.1 Audio-Text Alignment in LALMs

While cascaded ASR–LLM pipelines largely preserve text-domain performance, they discard speaker and paralinguistic cues essential for speech interaction (Maimon et al., 2025) , prompting recent work on end-to-end LALMs (Tang et al., 2024; Chu et al., 2024; Xie and Wu, 2024). However, despite architectural advances, such models consistently underperform text-based LLMs on language understanding and reasoning benchmarks, revealing a persistent *text–speech understanding gap* (Cui et al., 2025) . Existing approaches mainly attempt to reduce this gap via representation-level cross-modal alignment or large-scale synthetic speech data augmentation (Held et al., 2025), but often show limited gains on broad reasoning tasks or rely on massive proprietary datasets that hinder reproducibility (Zeng et al., 2025) .

2.2 On Policy Distillation

On-policy distillation (OPD) trains a student using supervision computed on the *student's own rollouts*, rather than on teacher-generated trajectories. Agarwal et al. (2024) introduces this paradigm for language models, showing that supervising at the student's prefixes can correct error cascades

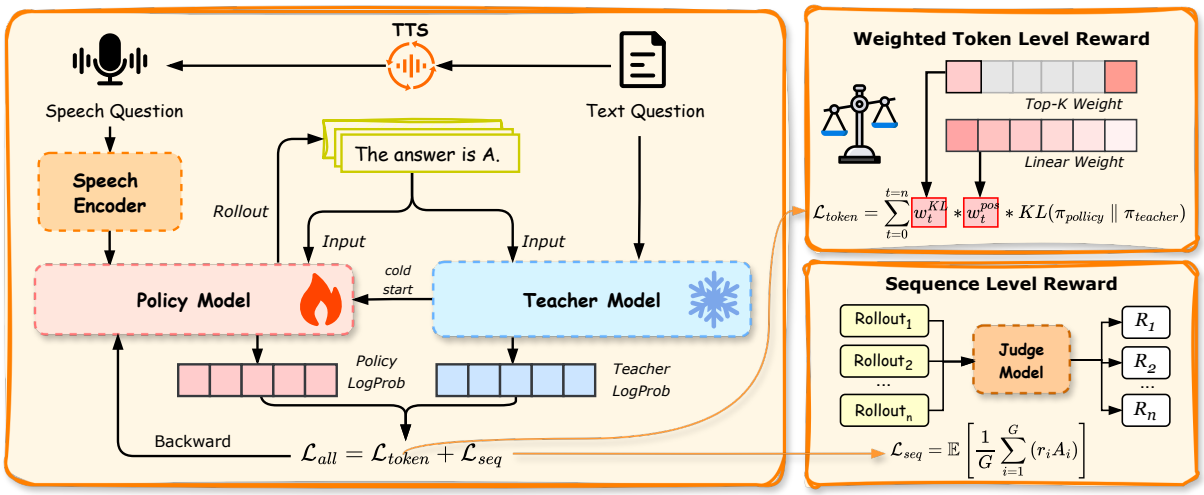


Figure 1: The overall framework of **CORD**. Given semantically equivalent audio and text inputs, CORD performs on-policy cross-modal self-distillation within a single model. Audio-conditioned trajectories are aligned to text-conditioned behaviors at two levels: (i) a token-level objective that applies importance-aware and position-aware reverse KL weighting along audio rollouts, and (ii) a sequence-level objective that uses a judge-based reward optimized via GRPO to enforce global reasoning consistency.

that would not be exposed under teacher trajectories. MiniLLM (Gu et al., 2025) further develops on-policy knowledge distillation for compressing LLMs, emphasizing that matching behaviors on student trajectories can yield strong performance with a smaller student. Recently, some works (Yang et al., 2025; Lu and Lab, 2025; Chen et al., 2025) reports that on-policy distillation can reach comparable reasoning performance to RL-style alignment at a fraction of the training cost under their training recipe. Beyond white-box settings, Ye et al. (2025) study *black-box* on-policy distillation where only teacher API access is available, proposing mechanisms to stabilize on-policy supervision without full teacher internals.

3 Method

In this section, we present CORD (**C**ross-modal Weighted **O**n-policy **R**eward-guided **D**istillation), a framework designed to bridge the performance gap in LALMs by aligning audio-conditioned reasoning trajectories with their text-conditioned counterparts. Unlike static or off-policy supervision, CORD enforces alignment along the model's actual inference paths using an in-model teacher strategy. As shown in Figure 1, CORD employs a multi-granularity alignment approach consisting of two complementary objectives:

- **Token-level alignment** corrects fine-grained semantic deviations at each decoding step using importance-weighted KL.
- **Sequence-level alignment** enforces global consistency across entire trajectories via GRPO.

By combining local step-wise correction with global trajectory regulation, CORD provides a unified framework for cross-modal alignment. In the following, we detail the problem formulation, the specific alignment objectives, and provide a comprehensive analysis of the underlying mechanisms that ensure training stability and semantic fidelity.

3.1 Problem Formulation

Given semantically equivalent audio and text inputs (x^a, x^t), a LALM p_θ induces autoregressive distributions over output sequences y:

$$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, x), \quad x \in \{x^a, x^t\}. \tag{1}$$

Empirically, LALMs often exhibit divergent behaviors across modalities despite identical input semantics. CORD aims to minimize this discrepancy by explicitly constraining the *audio-conditioned inference trajectories* $y \sim p_\theta(\cdot \mid x^a)$, rather than merely matching marginal output distributions.

3.2 On-policy Cross-modal Distillation

As shown in Figure 1, CORD aligns the model's cross-modal behavior along trajectories sampled from its current policy. For each x^a, we sample $y \sim p_\theta(\cdot \mid x^a)$ to obtain an on-policy decoding

trajectory. At each step t, the model induces two distributions over the vocabulary $\mathcal{V}$ conditioned on the same prefix $y_{<t}$: $p_\theta(\cdot \mid y_{<t}, x^a)$ and $p_\theta(\cdot \mid y_{<t}, x^t)$.

To quantify the discrepancy at each state, we employ **Reverse Kullback–Leibler (KL) Divergence**:

$$D_t = \mathrm{KL}\big(p_\theta(\cdot \mid y_{<t}, x^a) \,\big\|\, p_\theta(\cdot \mid y_{<t}, x^t)\big) = \sum_{v \in \mathcal{V}} p_\theta(v \mid y_{<t}, x^a) \log \frac{p_\theta(v \mid y_{<t}, x^a)}{p_\theta(v \mid y_{<t}, x^t)}. \tag{2}$$

Compared to forward KL, reverse KL places stronger emphasis on high-probability tokens under the text-conditioned distribution, encouraging the audio-conditioned policy to recover critical reasoning decisions made by the text modality. When applied *on-policy* along audio-conditioned trajectories, this formulation enables targeted correction of semantic deviations that arise during audio inference, rather than enforcing global distributional matching.

3.3 Analysis of Cross-modal Discrepancy

To motivate the design of CORD, we first investigate the reasoning behavior of LALMs under audio and text modalities. Let $y = (y_1, \ldots, y_T)$ denote a sequence sampled from the audio-conditioned policy $p_\theta(\cdot \mid x^a)$. At each decoding step t, given the same prefix $y_{<t}$, the model induces two conditional token distributions over the vocabulary $\mathcal{V}$: one conditioned on audio input x^a, and the other on text input x^t. We quantify their discrepancy using the token-level Kullback–Leibler divergence D_t.

A common practice is to uniformly average token-level KL divergences,

$$\frac{1}{T} \sum_{t=1}^{T} D_t. \tag{3}$$

This approach implicitly assumes that all tokens contribute equally to cross-modal alignment. However, our empirical analysis on the MMSU benchmark reveals that the distribution of D_t is highly skewed and non-uniform.

As illustrated in Figure 2, the distribution of D_t is highly skewed: most tokens exhibit very small divergence, while only a small number of critical tokens have substantially larger KL values. Specifically, we visualize the reverse KL distribution on the MMSU benchmark to investigate this phenomenon. As illustrated in Figure 2 (top-left), the distribution follows a heavy-tailed pattern, where the 80th percentile corresponds to a remarkably low divergence of only 0.23. This indicates that the vast majority of tokens are already well-aligned across modalities. Furthermore, the scatter plot (bottom-left) shows that these high-discrepancy states are concentrated in early decoding stages ($r = -0.139$). Misalignments at these pivotal early states often trigger a cascade effect, leading to cumulative reasoning failures.

In contrast, the word clouds in Figure 2(a) reveal that high-KL tokens are predominantly concentrated on semantically critical reasoning words and multiple-choice options (e.g., A, B). The misalignment at these pivotal states is a primary driver of incorrect responses; moreover, errors occurring at these early high-discrepancy tokens tend to trigger a cascade effect, leading to cumulative failures in subsequent decoding steps.

As a result, uniform averaging leads to a low overall loss magnitude, causing gradients from high-KL tokens to be diluted by numerous low-KL tokens. This weakens corrective updates precisely at positions that dominate semantic misalignment. To address this issue, we introduce an importance-aware token-level weighting scheme.

3.4 Importance-aware Token-level Alignment

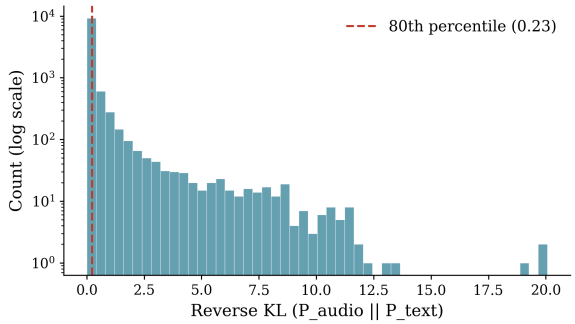
To address the gradient dilution issue, CORD introduces a multi-dimensional weighting scheme that prioritizes semantically critical and early-stage tokens.

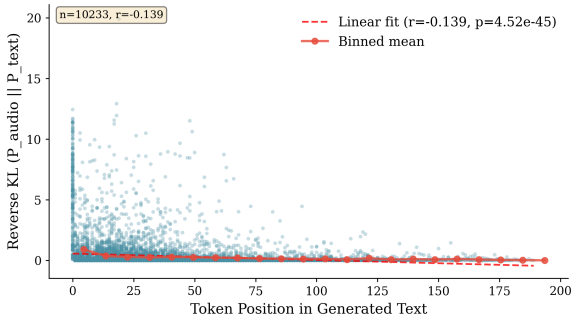
Top-K KL-based Importance Weighting. We first select the K tokens with the largest divergence values, where $K = 20$ in all experiments. Let $\mathcal{I}_K^*$ denote the index set of the top-K tokens ranked by D_t. We define a KL-based importance weight

$$w_t^{\mathrm{KL}} = \begin{cases} \alpha, & t \in \mathcal{I}_K^*, \\ 1, & \text{otherwise}, \end{cases} \tag{4}$$

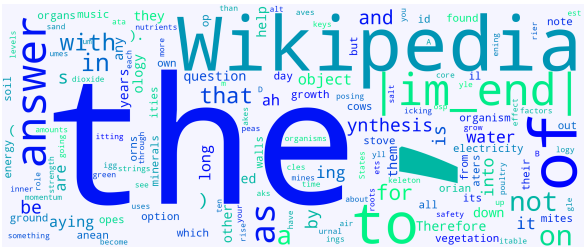
where $\alpha > 1$ is a hyperparameter that controls the strength of emphasis on high-divergence tokens. This hard selection prevents low-divergence tokens from dominating the optimization and ensures that gradient updates focus on tokens that exhibit significant cross-modal semantic mismatch.

Sequential Decay Weighting. We further observe that early decoding errors are more detrimental than later ones: once an incorrect semantic





(a) Tokens from high-KL regions.



(b) Tokens from low-KL (bottom) regions.

Figure 2: Statistical and Semantic Analysis of Token-Level Reverse KL Divergence. Top-left: Histogram of $\mathrm{KL}(P_{\text{audio}} \| P_{\text{text}})$ for all generated tokens (log scale), with the 80th percentile threshold marked by the red dashed line. Higher KL values highlight pivotal states where audio-conditioned reasoning significantly deviates from the text-conditioned teacher. Bottom-left: Scatter plot of KL divergence versus token position, revealing that major cross-modal discrepancies are concentrated in earlier reasoning stages. (a)-(b): Qualitative word cloud visualizations. High-KL regions (a) are enriched with semantically critical reasoning tokens (e.g., "Therefore", "answer") and choice options (e.g., A, B), whereas low-KL regions (b) primarily consist of common functional words and background information.

decision is made at an early step, subsequent tokens are unlikely to fully correct the reasoning trajectory. To emphasize early alignment, we introduce a position-dependent decay weight. Specifically, we assign a higher weight to earlier tokens and linearly decay it over time:

$$w_t^{\text{pos}} = \beta - (\beta - 1)\frac{t-1}{T-1}, \qquad t = 1, \ldots, T, \quad (5)$$

where $\beta > 1$ is a hyperparameter controlling the relative importance of early decoding steps. This formulation ensures that $w_1^{\text{pos}} = \beta$ and $w_T^{\text{pos}} = 1$.

Final Token-level Alignment Objective. The final token weight is defined as the product of the KL-based importance weight and the positional decay weight:

$$w_t = w_t^{\text{KL}} \cdot w_t^{\text{pos}} \quad (6)$$

The token-level alignment loss is then given by

$$\mathcal{L}_{\text{tok}} = \mathbb{E}_{y \sim p_\theta(\cdot|x^a)} \left[\sum_{t=1}^{T} w_t, D_t \right] \quad (7)$$

This objective amplifies supervision on a small set of semantically critical tokens while prioritizing early decoding stages, effectively mitigating gradient dilution and improving cross-modal reasoning alignment.

3.5 Sequence-level Alignment via Reward-guided Optimization

While token-level alignment corrects local semantic deviations, it does not explicitly constrain global reasoning behavior. In practice, locally aligned token distributions may still lead to globally inconsistent or incorrect final answers. To address this limitation, CORD introduces a sequence-level alignment objective that provides global supervision over complete audio-conditioned reasoning trajectories.

Judge-based Global Alignment Reward. Given an audio input x^a and its semantically equivalent text input x^t, we sample an audio-conditioned output sequence $y \sim p_\theta(\cdot \mid x^a)$ and generate a text-conditioned reference $\hat{y} \sim p_\theta(\cdot \mid x^t)$. We employ a judge model $J(\cdot, \cdot)$ to directly evaluate whether

the two sequences are semantically aligned at the answer level. The judge produces a binary reward:

$$r_{\text{seq}}(y; x^a, x^t) = J(y, \hat{y}) \in \{0, 1\}, \tag{8}$$

where $r_{\text{seq}} = 1$ indicates that the audio-conditioned answer is judged to be semantically consistent with the text-conditioned answer, and $r_{\text{seq}} = 0$ otherwise. This reward captures global reasoning alignment beyond local token-wise similarity.

GRPO Optimization. To optimize the model under this sequence-level reward, we adopt Group Relative Policy Optimization (GRPO). For each audio input x^a, we sample a group of N on-policy trajectories $\{y^{(i)}\}_{i=1}^{N}$ from the current policy $p_\theta(\cdot \mid x^a)$. Each trajectory is evaluated by the judge model, yielding rewards $\{r_{\text{seq}}^{(i)}\}_{i=1}^{N}$.

GRPO computes a relative advantage by comparing each trajectory's reward to the group average:

$$A^{(i)} = r_{\text{seq}}^{(i)} - \frac{1}{N} \sum_{j=1}^{N} r_{\text{seq}}^{(j)}. \tag{9}$$

Following the approach in DAPO (Yu et al., 2025), we omit the explicit KL divergence penalty. Consequently, the sequence-level optimization objective is defined as:

$$\mathcal{L}_{\text{seq}} = -\mathbb{E}_{\{y^{(i)}\}} \left[\frac{1}{N} \sum_{i=1}^{N} A^{(i)} \log p_\theta\left(y^{(i)} \mid x^a\right) \right]. \tag{10}$$

By optimizing this objective, the model increases the likelihood of audio-conditioned trajectories that achieve higher global alignment rewards relative to other on-policy rollouts. Crucially, GRPO operates entirely on-policy, ensuring that global supervision is applied to the exact inference states encountered during audio-conditioned reasoning.

Together with token-level alignment, this judge-guided GRPO objective explicitly constrains global reasoning outcomes and mitigates failure cases where locally aligned tokens still result in semantically inconsistent answers.

3.6 Overall Objective

CORD jointly optimizes local and global alignment objectives:

$$\mathcal{L}_{\text{CORD}} = \mathcal{L}_{\text{tok}} + \mathcal{L}_{\text{seq}}, \tag{11}$$

where $\mathcal{L}_{\text{seq}}$ denotes the GRPO-based sequence-level loss.

4 Experiments

4.1 Experimental Setting

Baselines. In our experiments, we adopt `Qwen2-Audio-7B-Instruct` (Chu et al., 2024) and `Step-Audio-2-Mini` (Wu et al., 2025) as base models. Following the previous distillation method (Cuervo et al., 2025), We implement two distillation objectives using teacher rollouts:

- **Supervised Fine-Tuning (SFT):** Given a teacher rollout $y \sim p_\theta(\cdot \mid x^t)$, the student is optimized to maximize the likelihood of the audio-conditioned trajectory:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{y \sim p_\theta^t} \left[\sum_{t=1}^{|y|} \log p_{\theta,t}^a \right] \tag{12}$$

- **Forward KL Divergence:** We minimize the distribution discrepancy between modalities over teacher rollouts:

$$\mathcal{L}_{\text{FKL}} = \text{KL}\big(p_\theta(\cdot \mid y_{<t}, x^t) \,\big\|\, p_\theta(\cdot \mid y_{<t}, x^a)\big) \tag{13}$$

Dataset. We curate 80,000 instances from **NuminaMath** (LI et al., 2024) to construct our training data. By employing the Kokoro (hexgrad, 2024) model to synthesize audio for each text instruction, we generate semantically equivalent pairs across the *text* and *audio* modalities.

Judge Model. Our judge model for GRPO was developed by distilling the evaluation outputs of proprietary frontier models on millions of text-based instruction-following samples. This judge model exhibits exceptional performance, with self-evaluation accuracy consistently exceeding 99%.

Evaluation Benchmarks. To evaluate the model's knowledge-based question answering capabilities, we utilize MMSU and OpenBookQA (OBQA) from VoiceBench (Chen et al., 2024). GSM8K (Cobbe et al., 2021) is employed to assess the model's mathematical reasoning performance. Furthermore, we adopt MMAU (Sakshi et al., 2024) to evaluate acoustic and paralinguistic reasoning; this benchmark is further categorized into three distinct subsets: *Music*, *Speech*, and *Sound*.

Experimental Details. Please refer to Appendix A for more details.

Method	MMSU		OBQA		GSM8K		AVG.
	Acc.	Δ_{base}	Acc.	Δ_{base}	Acc.	Δ_{base}	
Qwen2-Audio-7B-Instruct	36.04	8.42	51.20	17.59	20.73	19.74	15.25
+ SFT	36.47	7.99 $\downarrow$0.43	49.49	19.30 $\uparrow$1.71	33.68	6.79 $\downarrow$12.95	11.36
+ Forward KL	36.77	7.69 $\downarrow$0.73	48.21	20.58 $\uparrow$2.99	36.05	4.42 $\downarrow$15.32	10.90
+ CORD *(ours)*	**38.06**	**6.40** $\downarrow$**2.02**	**52.77**	**16.02** $\downarrow$**1.57**	**36.20**	**4.27** $\downarrow$**15.47**	**8.90**
Step-Audio2 mini	52.31	8.16	72.30	7.70	43.75	16.72	10.86
+ SFT	54.68	5.79 $\downarrow$2.37	74.72	5.28 $\downarrow$2.42	27.89	32.58 $\uparrow$15.86	14.55
+ Forward KL	53.12	7.35 $\downarrow$0.81	72.08	7.92 $\uparrow$0.22	46.57	13.90 $\downarrow$2.82	9.72
+ CORD *(ours)*	**57.63**	**2.84** $\downarrow$**5.32**	**77.74**	**2.26** $\downarrow$**5.44**	**47.56**	**12.91** $\downarrow$**3.81**	**6.00**

Table 1: Comparison of audio-conditioned reasoning performance across different backbones. **Base Models** (first row of each block) refer to the original instruction-tuned Qwen2-Audio-7B and Step-Audio2-mini. Δ_{base} represents the modality gap for each method, defined as $\text{Acc}_{\text{text}}^{\text{Base}} - \text{Acc}_{\text{audio}}^{\text{Method}}$, where $\text{Acc}_{\text{text}}^{\text{Base}}$ is the fixed text-conditioned accuracy of the Base Model. For Qwen2-Audio, these text baselines are 44.46, 68.79, and 40.47; for Step-Audio2, they are 60.47, 80.00, and 60.47. **AVG.** is the mean of the modality gaps. Arrows indicate the reduction ($\downarrow$) or increase ($\uparrow$) in the gap compared to the Base Model's initial Δ_{base}.

4.2 Main Results

Tables 1 and 2 report the performance of CORD across multiple benchmarks using two representative LALMs, Qwen2-Audio-7B-Instruct and Step-Audio2-mini. The results show that CORD substantially **reduces the performance gap** between audio and text modalities, while effectively **preserving auxiliary audio capabilities** beyond speech.

Cross-modal gap reduction. As shown in Table 1, CORD achieves consistent and substantial reductions in the audio–text performance gap across both backbone models. Across all benchmarks, CORD consistently outperforms SFT and Forward-KL, demonstrating the effectiveness of on-policy, trajectory-level alignment. Specifically, on Qwen2-Audio-7B-Instruct, CORD reduces the average audio–text gap by **41.6%** relative to the base model, whereas Forward-KL yields only a 28.5% reduction. On Step-Audio2-mini, CORD achieves a **44.8%** gap reduction, while Forward-KL leads to only a 10.5% reduction.

In addition, by comparing results between different models, we observe a clear trend: Step-Audio2-mini benefits more from CORD (44.8% $\rightarrow$ 41.6%), exhibiting larger gap reductions and stronger alignment. This suggests that a stronger text-conditioned teacher leads to more effective cross-modal alignment. This observation indicates that CORD naturally scales with the quality of the base LALM and can more effectively exploit improvements in text reasoning to enhance audio-conditioned performance.

Method	Music	Sound	Speech	Avg.
Base Model	58.98	64.74	58.73	60.81
+ SFT	56.29	64.44	51.51	57.39
+ Forward KL	55.99	61.70	53.01	56.90
+ CORD *(ours)*	60.18	64.44	55.42	60.01

Table 2: Fine-grained performance on the MMAU benchmark across music, sound, and speech categories based on Qwen2-Audio-7B-Instruct.

Emergent Cross-domain Generalization. Although CORD is trained exclusively on a math-focused dataset, both models exhibit more pronounced improvements on general-domain benchmarks such as MMSU and OBQA than on the math-intensive GSM8K benchmark. This observation indicates that CORD does not merely acquire domain-specific knowledge, but instead learns a transferable meta-capability of cross-modal alignment.

Preserving Auxiliary Audio Capabilities. Table 2 shows that CORD effectively preserves general audio understanding capabilities while aligning for complex reasoning, whereas baseline methods suffer from significant performance degradation. Specifically, the Forward-KL baseline exhibits a noticeable performance tax, with scores dropping by 2.99 points in *music* and 3.04 points in *sound*, suggesting that conventional distillation may inadvertently lead to the catastrophic forgetting of non-speech acoustic patterns. In contrast, CORD demonstrates remarkable robustness: it not

only maintains near-parity with the base model in the *sound* category but even yields a slight improvement in *music*. These results indicate that trajectory-level on-policy alignment effectively mitigates collateral damage to auxiliary audio modalities, ensuring a more stable and balanced cross-modal alignment that retains pre-trained general audio intelligence.

4.3 Ablation Studies

We conduct ablation studies to analyze the contribution of each component in CORD and to investigate the sensitivity of key hyperparameters. All ablations are performed on Qwen2-Audio-7B-Instruct under the same training setup as the main experiments.

4.3.1 Component-wise Ablation

In Table 3, we conduct an incremental ablation study to evaluate the efficacy of each module within the CORD framework.

We observe that while initial reinforcement learning via GRPO yields performance gains at 500 steps, it suffers from severe model collapse as training extends to 1000 steps. This degradation is most pronounced on GSM8K (35.59 $\rightarrow$ 19.89), where performance falls even below the base model baseline. Notably, our OPD module acts as a powerful regularizer that eliminates this instability, ensuring sustained optimization for up to 3000 steps without quality loss (35.59 $\rightarrow$ 36.12 in GSM8K). Finally, by augmenting OPD with cross-modal token-level weighting, the complete CORD framework achieves the best overall performance, with an average improvement of 6.35 over the base model (i). This validates the synergy between stable sequence-level reinforcement learning, on-policy anchoring, and fine-grained alignment.

4.3.2 Sensitivity to Weighting Intensity

To efficiently navigate the hyperparameter space while balancing the contributions of token-level importance (α in Equation. 4) and positional weighting (β in Equation. 5), we couple these parameters by setting $\alpha = \beta$. As illustrated in Figure 3, the performance across all benchmarks exhibits a characteristic bell-shaped trend, achieving an optimal trade-off at $\alpha = \beta = 2.0$. This configuration sufficiently accentuates pivotal reasoning states without over-concentrating gradients on a sparse subset of tokens, which could otherwise compromise optimization stability. In contrast, smaller values (e.g., 1.0) revert toward uniform KL optimization, while excessively large values (e.g., 2.5) lead to marginal performance drops due to the over-suppression of long-tail semantic information. Consequently, we adopt $\alpha = \beta = 2.0$ as the default setting for all primary experiments.

Method	Step	MMSU	OBQA	GSM8K
Base Model (i)	–	36.04	51.20	20.73
Stability Analysis (GRPO only)				
+ GRPO (ii)	500	36.92	50.54	35.59
+ GRPO (ii)	1000	27.87	36.48	19.89
Incremental Ablation (Cumulative)				
+ GRPO + OPD (iii)	3000	37.41	51.20	36.12
+ Full (iv)	3000	**38.06**	**52.77**	**36.20**

Table 3: Ablation study of individual components and training stability. We evaluate the incremental contributions of: (i) the base model, (ii) sequence-level **GRPO**, (iii) On-Policy Distillation (**OPD**), and (iv) the **Full** CORD framework, which integrates all previous modules with **token-level importance weighting** for fine-grained alignment. Red values denote model collapse where performance falls below the base baseline.

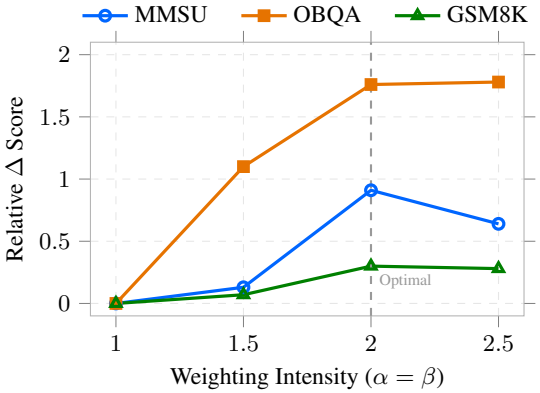


Figure 3: Sensitivity analysis of the weighting intensity α and β. To reduce hyperparameter complexity, we set $\alpha = \beta$. The scores represent relative improvements over the baseline ($\alpha = \beta = 1.0$). A value of 2.0 yields the most consistent gains across all tasks.

5 Conclusion

We introduce **CORD**, a weighted on-policy cross-modal self-distillation framework designed to bridge the audio–text gap in LALMs. By aligning audio-conditioned reasoning with text baselines at both token and sequence levels, CORD employs an importance-aware KL objective and a judge-guided GRPO objective to ensure local semantic accuracy and global trajectory consistency. Extensive benchmarks demonstrate that CORD sig-

nificantly narrows the modality gap, highlighting on-policy trajectory alignment as a robust paradigm for cross-modal alignment.

6 Limitations

Despite the significant performance gains achieved by CORD, our study has several limitations. First, due to computational resource constraints, our framework was primarily evaluated on a dataset of 80,000 instances focused on mathematical reasoning. While this scale is sufficient to demonstrate the efficacy of our method, exploring the behavior of CORD under larger and more diverse multimodal data remains an important direction for future work. Second, the current experiments are largely centered on audio understanding benchmarks. The generalizability of our importance-aware weighting and on-policy distillation to broader, more diverse domains, such as general audio-visual scene understanding.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. *Preprint*, arXiv:2306.13649.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025. Retaining by doing: The role of on-policy data in mitigating forgetting. *Preprint*, arXiv:2510.18874.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Santiago Cuervo, Skyler Seto, Maureen de Seyssel, Richard He Bai, Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly, and Zakaria Aldeneh. 2025. Closing the gap between text and speech understanding in llms. *arXiv preprint arXiv:2510.13632*.

Wenqian Cui, Xiaoqi Jiao, Ziqiao Meng, and Irwin King. 2025. Voxeval: Benchmarking the knowledge understanding capabilities of end-to-end spoken language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 16735–16753, Vienna, Austria. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2025. Minillm: Knowledge distillation of large language models. *Preprint*, arXiv:2306.08543.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. *Preprint*, arXiv:2407.05361.

William Held, Yanzhe Zhang, Minzhi Li, Weiyan Shi, Michael J. Ryan, and Diyi Yang. 2025. Distilling an end-to-end voice assistant without instruction training data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7876–7891, Vienna, Austria. Association for Computational Linguistics.

hexgrad. 2024. Kokoro: Open-weight tts model with 82 million parameters. https://github.com/hexgrad/kokoro.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu,

Jiamang Wang, Wenbo Su, Bo Zheng, and Junchi Yan. 2025. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *Preprint*, arXiv:2510.13554.

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

Gallil Maimon, Amit Roth, and Yossi Adi. 2025. Salmon: A suite for acoustic language model evaluation. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Christoph Minixhofer, Ondrej Klejch, and Peter Bell. 2025. Scaling laws for synthetic speech for model training. In *Proceedings of Interspeech 2025*. ISCA.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

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*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*.

Liang-Hsuan Tseng, Yi-Chang Chen, Kuan-Yi Lee, Da-Shan Shiu, and Hung-Yi Lee. 2025. Taste: Text-aligned speech tokenization and embedding for spoken language modeling. *Preprint*, arXiv:2504.07053.

Chen Wang, Minpeng Liao, Zhongqiang Huang, and Jiajun Zhang. 2024. Blsp-kd: Bootstrapping language-speech pre-training via knowledge distillation. *Preprint*, arXiv:2405.19041.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Preprint*, arXiv:2506.01939.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tianzhu Ye, Li Dong, Zewen Chi, Xun Wu, Shaohan Huang, and Furu Wei. 2025. Black-box on-policy distillation of large language models. *Preprint*, arXiv:2511.10643.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. 2025. Scaling speech-text pre-training with synthetic interleaved data. In *International Conference on Learning Representations (ICLR)*.

A Experiment Details

For CORD and Forward-KL distillation baselines, we generate a single on-policy rollout per prompt at each update step. In contrast, for GRPO, we sample a group of four rollouts per prompt (group size

N=4). The maximum generated sequence length is set to 200 tokens. All models are optimized using the AdamW optimizer with a learning rate of 3×10^{-5} and share the same training schedule and decoding configurations unless otherwise specified. For rollout sampling, we use a temperature of 1.0 for CORD and Forward-KL, while a higher temperature of 1.5 is adopted for GRPO to encourage trajectory diversity. For the token-level alignment objective, we fix the KL-based importance scaling factor and the positional decay factor to $\alpha = 2$ and $\beta = 2$, respectively, in all experiments.