Audio-Thinker: Guiding Large Audio Language Model When and How to Think via Reinforcement Learning

Shu Wu[1], Chenxing Li[1*], Wenfu Wang[1], Hao Zhang[2], Hualei Wang[1], Meng Yu[2], Dong Yu[2*]

[1]Tencent AI Lab, Beijing,China
[2]Tencent AI Lab, Seattle,USA
sethwu@foxmail.com, chenxingli@tencent.com, dyu@global.tencent.com

Abstract

Recent advancements in large language models, multimodal large language models, and large audio language models (LALMs) have significantly improved their reasoning capabilities through reinforcement learning utilizing rule-based rewards. However, the explicit reasoning process has yet to yield substantial benefits for audio question answering, and effectively leveraging deep reasoning remains an open challenge, with LALMs still falling short of human-level auditory-language reasoning. To address these limitations, we propose Audio-Thinker, a reinforcement learning framework designed to enhance the reasoning capabilities of LALMs through improved adaptability, consistency, and effectiveness. Our approach introduces an adaptive think accuracy reward, enabling the model to adjust its reasoning strategies based on task complexity. Furthermore, we incorporate an external reward model to evaluate the overall consistency and quality of the reasoning process, complemented by think-based rewards that assist the model in distinguishing between valid and flawed reasoning paths during training. Experimental results demonstrate that Audio-Thinker models outperform existing reasoning-oriented LALMs across various benchmark tasks, exhibiting superior reasoning and generalization capabilities.

Extended version — https://arxiv.org/abs/2508.08039

Introduction

Recent advances in LLMs show that reasoning can be enhanced through chain-of-thought (CoT) prompting, cognitive frameworks, and reinforcement learning (RL) (Guo et al. 2025). RL-tuned models excel in complex tasks, such as math and coding, with methods like Group Relative Policy Optimization (GRPO) (Shao et al. 2024) outperforming traditional supervised learning. Smaller models benefit from structured reasoning, while larger models prefer unstructured approaches.

Recent studies (Huang et al. 2025b; Liu et al. 2025b; Pan et al. 2025; Zhou et al. 2025) extend RL to Multimodal Large Language Models (MLLMs) across domains like object recognition (Liu et al. 2025b), semantic segmentation (Liu et al. 2025a), and video analysis (Sun et al. 2025). These methods enhance MLLM capabilities in data-scarce scenarios, matching supervised fine-tuning (SFT) performance on in-domain tasks while outperforming supervised models in out-of-distribution evaluations.

The domain of audio-language reasoning and RL-based fine-tuning (RLF) remains relatively unexplored. Prominent Large Audio Language Models (LALMs) such as Audio Flamingo (Kong et al. 2024), SALMONN (Tang et al. 2023), and Qwen2-Audio (Yang et al. 2024a) significantly advance audio comprehension across various benchmarks. However, these models primarily concentrate on perception and basic question-answering tasks without incorporating explicit reasoning processes. Subsequently, Audio-Reasoner (Xie et al. 2025b) employs a structured reasoning methodology on Qwen2-Audio, while R1-AQA (Li et al. 2025a) applies the GRPO algorithm, finding that the simple addition of a reasoning chain is insufficient to yield substantial improvements. SARI (Wen et al. 2025) fine-tunes Qwen2.5-Omni using a combination of RL and both structured and unstructured reasoning. However, its performance does not match that of Omni-R1 (Rouditchenko et al. 2025), which is trained exclusively with RL. These findings underscore the ongoing challenge of effectively leveraging RL to enhance reasoning capabilities in audio question-answering tasks.

In this study, we address these challenges by introducing Audio-Thinker, a RL-based framework designed to enhance the adaptive, consistent, and effective reasoning capabilities of LALMs. As illustrated in Figure 1, Audio-Thinker employs an adaptive thinking mode policy that determines when the model should engage in "thinking", based on the query complexity. Moreover, it integrates an external LLM-based expert to provide thought-based supervision, thereby guiding the model in generating coherent and effective reasoning processes. The main contributions of this work are as follows.

- **Audio-Thinker Framework :** We present Audio-Thinker, a universal RL-based framework that empowers LALMs to explore effective reasoning policies while simultaneously enhancing reasoning quality.
- **When to Think:** We introduce an adaptive thinking accuracy reward that trains LALMs to modulate their reasoning strategies according to task complexity, directing the model to find optimal reasoning approaches.

*Corresponding author.

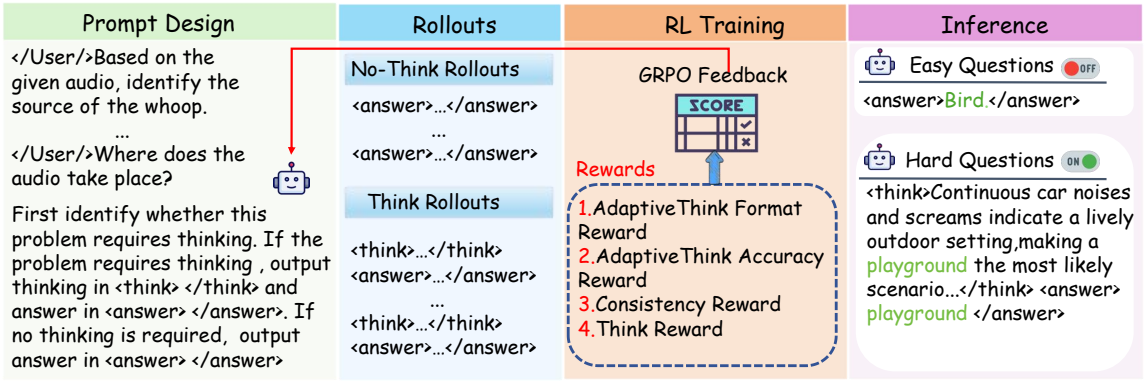

Figure 1: Overview of the Audio-Thinker framework. As illustrated in the block Inference, the LALMs trained using the Audio-Thinker framework are capable of achieving adaptive reasoning capabilities that scale according to the complexity and difficulty of the given task.

- **How to Think:** We integrate think-based rewards that evaluate the consistency and quality of reasoning, allowing the model to distinguish between sound and flawed reasoning processes during training.
- **State-of-the-Art Performance:** In the experiments, with only 40k post-training samples, Audio-Thinker models consistently outperform existing LALMs on diverse benchmarks, including MMAU (Sakshi et al. 2024), MMAR (Ma et al. 2025b), and AIR (Yang et al. 2024b), demonstrating strong reasoning and generalization capabilities.

Relate Works

Large Audio Language Models

The rapid advancement of LLMs catalyzes the evolution of MLLMs, which possess the capacity to comprehend and reason across a diverse array of data modalities, including auditory information. Exemplary instances of LALMs, such as GAMA (Ghosh et al. 2024), SALMONN (Tang et al. 2023), Qwen2-Audio (Yang et al. 2024a), Audio Flamingo (Kong et al. 2024), Audio Flamingo 2 (Ghosh et al. 2025), and Audio Flamingo 3 (Goel et al. 2025), exhibit remarkable capabilities in audio understanding.

Language and Multimodal Reasoning

Recently, models such as OpenAI-o1 (Jaech et al. 2024), Kimi K1.5 (Team et al. 2025), and DeepSeekR1 (Guo et al. 2025) draw attention for enhancing reasoning performance through RL (Jin et al. 2025; Peng et al. 2025; Face 2025). This progress spurs follow-up research, including successful method replications (Xie et al. 2025a) and efforts to improve algorithmic efficiency (Yu et al. 2025). RL is increasingly applied to vision-language models (Yang et al. 2025b; Feng et al. 2025; Huang et al. 2025a).

Audio Models with Reasoning

Recent efforts concentrate on enhancing reasoning capabilities in audio-language models. A notable example is Mellow (Deshmukh et al. 2025), a lightweight audio-language model that demonstrates exceptional reasoning abilities. Despite having only 167 million parameters and being trained on 1.24 million examples, Mellow outperforms larger State-of-the-Art (SOTA) models across various domains. Audio-CoT (Ma et al. 2025a) is the first model to explore CoT reasoning in audio-language models; however, it does not incorporate model updates and offers limited advancements for tackling complex issues. Additionally, Audio-Reasoner (Xie et al. 2025b) introduces a structured reasoning process that utilizes a large-scale dataset (CoTA) and employs a multi-phase "thinking" architecture comprising planning, captioning, reasoning, and summarization before generating its final response. Furthermore, R1-AQA (Li et al. 2025a) utilizes the GRPO algorithm to fine-tune the Qwen2-Audio model for audio question-answering tasks, enhancing reasoning accuracy with less data through reward-driven optimization. Concurrently, SARI (Wen et al. 2025) fine-tunes Qwen2.5-Omni (Xu et al. 2025) using RL, presenting a study focused on improving the reasoning capabilities of audio multimodal models by leveraging explicit CoT training and curriculum-guided RL. Finally, Omni-R1 (Rouditchenko et al. 2025) fine-tunes Qwen2.5-Omni with GRPO, employing a straightforward yet effective prompt that streamlines training and testing, ultimately achieving a new SOTA performance.

Observations and Motivations

O1: Explicit Thinking Does Not Always Yield Effective Results

Prior studies on large language models (LLMs) and multimodal large language models (MLLMs) commonly suggest that explicit reasoning mechanisms can enhance reasoning performance. However, recent works such as R1-AQA and Omni-R1 demonstrate that explicitly modeling the reasoning process does not lead to substantial improvements in automated question answering (AQA) tasks. These results suggest that the effectiveness of explicit reasoning is task-dependent and may not generalize across settings. Consequently, how to effectively leverage what we refer to as **deep**

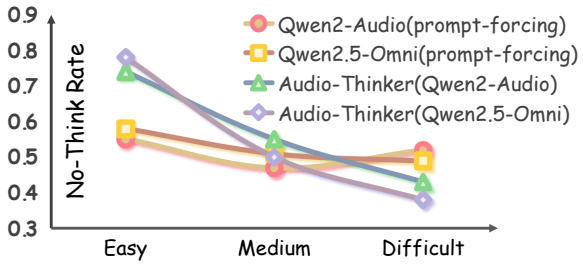

Figure 2: No-Think Rate by Difficulty on MMAU-test-mini. Prompt-forcing models show a flat distribution, indicating no sensitivity to problem complexity, while Audio-Thinker models exhibit a clear trend, demonstrating difficulty-aware reasoning.

thinking remains an open research question.

O2: Prompting Alone Does Not Enable Adaptive Thinking

One possible solution to the issue identified in O1 is the implementation of ***adaptive thinking*** (Zhang et al. 2025a; Li et al. 2025b), whereby the model dynamically determines the necessity of reasoning based on input characteristics. This can be achieved through a prompting strategy that enables context-aware adaptation to question complexity.

To evaluate performance, we use a prompt strategy (see Appendix A.1 in the extended version) and evaluate the results on the MMAU-test-mini dataset across three complexity levels. As shown in Figure 2, we analyze the no-think rate across these levels. Notably, prompt-forced models show no clear trend, indicating their reasoning activation is largely insensitive to problem difficulty. This observation underscores the limited adaptability of prompt-forced models in determining when deep thinking is necessary.

Guiding LALMs When and How to Think

Based on current observations, existing LALMs lack adaptive thinking and sufficient supervision over their reasoning processes during training, which may hinder generalization. To address this, we propose Audio-Thinker, an audio-language RL-based framework that promotes difficulty-aware, consistent, and effective reasoning. As shown in Figure 2, the models trained with Audio-Thinker demonstrate clear difficulty-aware reasoning.

Audio Thinker

As depicted in Figure 1, Audio-Thinker consists of two primary components:

- Adaptive Thinking Prompt Design: A prompting strategy that facilitates stochastic transitions between thinking and non-thinking modes in LALMs.
- RL-based Training Framework: As shown in Figure 3, our approach employs a progressively refined reward function, enabling LALMs to discern the necessity of reasoning and to follow the most effective reasoning trajectory toward the solution.

Below, we provide a detailed explanation of the implementation of each module.

Prompt Design

We prompt the model to first assess whether a query requires reasoning, and then either generate a reasoning process if needed or provide a direct answer otherwise. Details of the prompt are provided in Appendix A.1.

Progressively Refined Reward Designs

Reward 1: Adaptive Think Format Reward We prompt LALMs to decide whether reasoning is needed and then generate either a reasoned response or a direct answer accordingly. Both formats receive a format reward of 1.

Reward 2: Adaptive Think Accuracy Reward As shown in Figure 2, the prompt-only control approach presents a key limitation: without feedback, the model cannot determine when reflective thinking is necessary. Inspired by AutoThink (Tu et al. 2025), we introduce the Adaptive Think Accuracy Reward (ATAR) to guide the model in deciding whether to engage in deep reasoning based on problem complexity, as illustrated in Figure 3, Block "Adaptive Think Accuracy Reward". Specifically, we assign higher rewards for correct answers that do not require reflection and impose stricter penalties for incorrect responses. We define four cases: **Case 1**: think and correct, **Case 2**: think and incorrect, **Case 3**: no-think and correct, **Case 4**: no-think and incorrect. Each sample i receives an initial reward $R_{\mathrm{a},i} \in \{+1, 0, +2, -1\}$ for Case 1, 2, 3, and 4, respectively.

This reward structure encourages difficulty-aware behavior; however, it may cause instability in the early stages of training. The model might converge on a degenerate policy, consistently choosing either to think or to skip, depending on which option appears to yield a higher short-term expected reward. This tendency limits exploration and hampers further optimization. To mitigate this issue, we integrate the implementation of batch-level reward balancing.

Let $\lambda \in [0, 1]$ denote the proportion of think trajectories in a training batch, with $1 - \lambda$ representing the proportion of no-think samples. For both think and no-think samples, we calculate soft penalty factors as follows:

$$\gamma_{\text{think}} = e^{-\lambda}, \tag{1}$$

$$\gamma_{\text{no-think}} = e^{-(1-\lambda)}. \tag{2}$$

The introduction of soft penalty factors assists the model in maintaining behavioral stability between thinking and non-thinking modes during the initial phases of training. However, this mechanism also constrains the model's ability to evolve freely within each mode. To address this limitation, we propose a strategy that gradually reduces the impact of soft penalty factors as training progresses. This approach encourages the reasoning model to increasingly rely on the original accurate rule-based rewards in later stages, with the soft penalty factor gradually converging towards a value of 1. The final soft penalty factors are defined as follows:

$$\gamma_{\text{think}} = e^{-\lambda \cdot (1 - \frac{steps}{T})}, \tag{3}$$

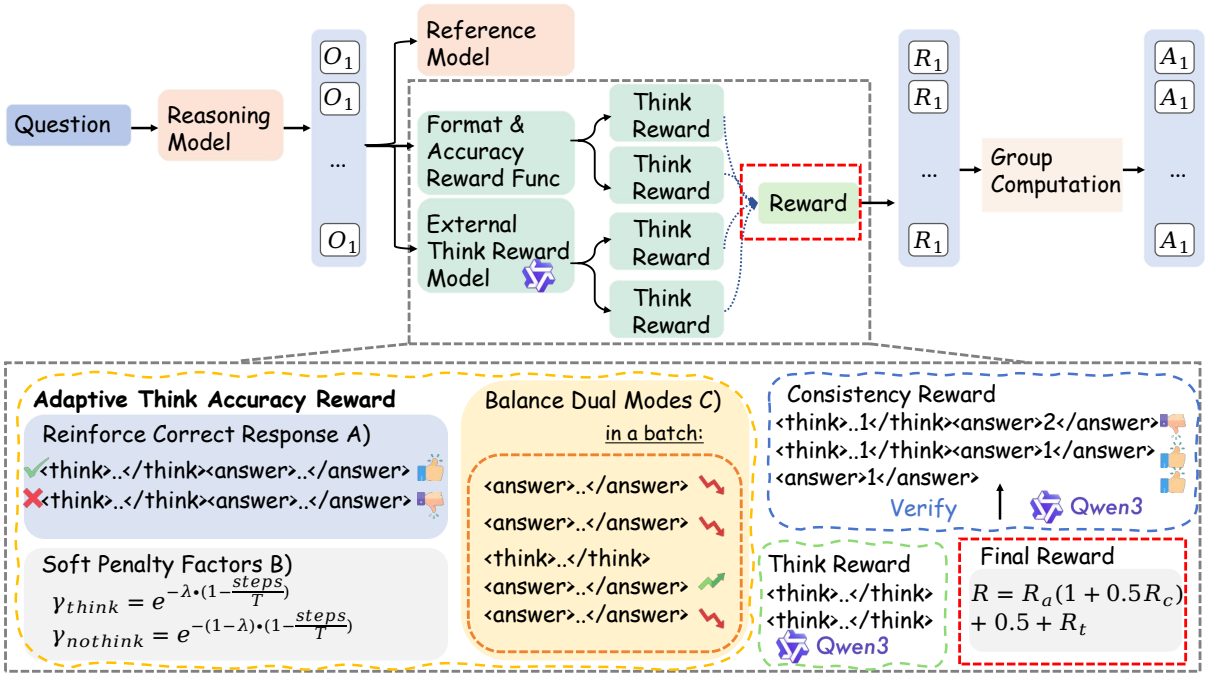

Figure 3: An illustration of Audio-Thinker RL-based training pipeline. The upper portion of the figure depicts the overall RL-based training framework, while the lower section presents a detailed breakdown of the progressively refined reward design components.

$$\gamma_{\text{no-think}} = e^{-(1-\lambda)\cdot(1-\frac{steps}{T})}, \tag{4}$$

where steps denotes the current global training step, and T represents the total training steps, thereby allowing for adjustment of the soft penalty factor's influence. Accordingly, the final reward can be defined as follows:

$$R_{\text{a},i} = \begin{cases} \gamma_{\text{think}} & \text{Case 1,} \\ -\big(1-\gamma_{\text{think}}\big) & \text{Case 2,} \\ 2\gamma_{\text{no-think}} & \text{Case 3,} \\ -\gamma_{\text{no-think}} - 2\big(1-\gamma_{\text{no-think}}\big) & \text{Case 4.} \end{cases} \tag{5}$$

When thinking processes dominate, the rewards for cognitive responses, especially incorrect ones, are subtly diminished. Similarly, when no-think responses are overly represented, their rewards also decline. In both cases, the model is encouraged to restore balance by favoring the less frequent behavior.

Reward 3: Consistency Reward Ideally, a model's reasoning should directly support its final answer. However, with accuracy-based training methods such as GRPO, inconsistencies may emerge. Specifically, while the model often produces correct answers, its CoT reasoning frequently lacks coherence. This indicates that the model has learned to generate correct outputs without developing strong reasoning skills. As demonstrated in Figure 3, the model might classify response 1 as preferable yet produce output 2 (e.g., `<think>...the final answer is 1</think><answer>2</answer>`).

This gap stems from supervising only the final answer while ignoring the reasoning trajectory. Consequently, the model fortifies faulty yet fortuitously correct chains, relegating reasoning to noise and yielding repetitive or random outputs. Although accuracy may persist, transparency and interpretability are sacrificed. Inspired by R1-Reward (Zhang et al. 2025b), we employ Qwen3-8B-Base (Yang et al. 2025a) as a supervisory model to evaluate reasoning–output alignment. This ensures alignment between the reasoning process and output, leading to the following reward function:

$$R_{\text{c},i} = \begin{cases} 1, & \text{Think is consistent with the answer,} \\ 0, & \text{Think is inconsistent with the answer.} \end{cases} \tag{6}$$

For responses in the no-think mode, the consistency reward function is set to 1.

Reward 4: Think Reward Consistency rewards have the potential to enhance the alignment between a model's reasoning process and its final answer. However, a significant challenge remains: models may produce correct answers through flawed reasoning rather than systematic deduction. Our observations indicate that when this reward is applied in isolation, GRPO training can yield scenarios in which the reasoning conclusion, although aligned with the final answer, emerges from erroneous logic or inaccurate information. SophiaVL-R1 (Fan et al. 2025) is among the first to apply a think reward in MLLMs reasoning, achieving promising results. This naturally raises the hypothesis: *Can a **think reward** that emphasizes the **thinking process** guide LALMs to improve their reasoning?*

To investigate this concept, we propose a model-generated think reward. This approach enables us to evaluate the nuanced reasoning quality of LALMs and examine their effects on final inference outcomes. We incorporate the Qwen3-8B-Base model as the think reward model, which assigns a score ranging from 0 to 1 increments solely based on the quality of intermediate reasoning, independent of the correctness of the final answer. In cases where responses are generated in a no-think mode, the think reward is computed as the average of the think rewards within the batch.

Overall Reward Integrating the consistency reward separately with previous reward functions may lead to high overall rewards even for incorrect answers due to the consistency component. Therefore, we apply this reward only when the final answer is correct to avoid over-emphasizing consistency. The think reward, in contrast, targets improvements in reasoning quality by evaluating intermediate steps, irrespective of the final answer's correctness. The final reward structure is defined as follows.

$$R = R_a \times (1 + 0.5 \times R_c) + 0.5 \times R_f + R_t. \quad (7)$$

Reinforcement Learning

Following DeepSeek-R1 (Shao et al. 2024), given an input question q, GRPO samples a group of responses $\{o_1, o_2, \cdots, o_G\}$, and their corresponding rewards corresponding rewards $\{R_1, R_2, \cdots, R_G\}$ are computed using the reward model. The advantage is subsequently computed as:

$$\hat{A}_{i,t} = \widetilde{R}_i = \frac{R_i - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})} \quad (8)$$

The policy model is subsequently optimized by maximizing the Kullback-Leibler objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\mathcal{D}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\rho_{i,t} \hat{A}_{i,t}, \right. \right. \right.$$
$$\left. \left. \left. \text{clip}\left(\rho_{i,t}, 1-\epsilon, 1+\epsilon\right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL}\left[\pi_\theta || \pi_{ref}\right] \right\} \right] \quad (9)$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$ is the probability ratio between the current policy π_θ and the policy $\pi_{\theta_{old}}$, and ϵ and β are hyper-parameters introduced in Proximal Policy Optimization (PPO) (Schulman et al. 2017).

Experiment

Experiment Setup

Dataset The training data is drawn from the AVQA dataset (Yang et al. 2022), designed for audio-visual question answering. Following R1-AQA, we extract audio from videos and construct audio-text pairs by replacing "video" with "audio" in questions, resulting in 40,176 training samples. For SFT with CoT, we first generate audio captions using Qwen2-Audio-7B-Instruct on AVQA, then employ Qwen2.5-72B-Instruct (Yang et al. 2024a) to generate CoT rationales from the caption, question, and answer. The prompt for CoT generation is provided in Appendix A.2.

Implementation Details We use Qwen2-Audio-7B-Instruct and Qwen2.5-Omni as the foundational models for our experiments. Training is conducted on the SWIFT (Zhao et al. 2025) framework. To train our models, we use a node with 8 H20 GPUs (96GB). The batch size per GPU is 1 with gradient accumulation steps of 2 for a total effective batch size of 16. The models are trained for 1000 steps on AVQA with a learning rate of 1e-6, a temperature of 1.0, 8 responses per GRPO step, and a Kullback–Leibler β of 0.04.

Evaluation Metrics

We evaluate model performance primarily by accuracy on multiple-choice questions using three main evaluation sets:

- **MMAU Benchmark** (Sakshi et al. 2024): We evaluate the model using the MMAU benchmark's test-mini and test sets, featuring complex audio-based question-answer pairs requiring expert-level reasoning. Accuracy is measured as the percentage of correctly answered multiple-choice questions. Recently, an updated version named MMAU-v05.15.25 was proposed. To maintain research continuity, we use the previous MMAU version for ablation studies and comparisons with prior work. Results for MMAU-v05.15.25 are detailed in Appendix B.1.
- **MMAR Benchmark** (Ma et al. 2025b): This benchmark assesses deep reasoning across a range of real-world audio scenarios, incorporating mixed sounds, music, and speech, with questions specifically designed to challenge reasoning abilities.
- **AIR Benchmark** (Yang et al. 2024b): We evaluate the model's audio comprehension using the foundational sections of AIR-Bench, which encompasses a variety of audio modalities, including sound, speech, and music.

Results

Ablation Study

To systematically analyze the impact of different reasoning strategies and training methodologies, we conduct ablation studies using Qwen2-Audio-7B-Instruct and Qwen2.5-Omni as baselines. Detailed experimental results are presented in Table 1. The best-performing models in each category are highlighted in **bold**, and the second-best scores are underlined.

Impact of GRPO Explicit reasoning alone provides insufficient guidance for model improvement without effective supervision. We apply SFT and GRPO to Qwen2-Audio-7B-Instruct and Qwen2.5-Omni, resulting in model variants: SFT (sft-a, sft-b, sft-c, sft-d) and GRPO (grpo-a, grpo-b, grpo-c, grpo-d). GRPO models achieve significant improvements on MMAU-test-mini, AIR Foundation, and MMAR benchmarks. However, explicit reasoning variants (grpo-b, grpo-d) do not outperform implicit counterparts (grpo-a, grpo-c), suggesting explicit reasoning alone is insufficient without effective supervision. These findings highlight the importance of proper supervision mechanisms for leveraging explicit reasoning in model enhancement.

Model	Method	MMAU Test-mini				MMAR				AIR
		Sound↑	Music↑	Speech↑	Average↑	Sound↑	Music↑	Speech↑	Average↑	Average↑
Qwen2-Audio-7B-Instruct (reproduce)	-	62.16	53.59	48.59	54.90	33.33	24.27	32.31	30.00	61.3
sft-a	SFT	63.66	56.59	54.35	58.20	52.73	37.86	49.32	48.90	63.8
sft-b	SFT+CoT	63.36	56.29	54.41	57.80	56.36	41.75	48.30	49.80	62.6
grpo-a	GRPO	68.47	62.87	60.06	63.80	56.36	39.81	48.98	50.20	64.5
grpo-b	GRPO+CoT	70.27	63.17	61.56	65.00	**58.18**	35.44	52.04	50.00	64.1
model-a	GRPO+ATAR	74.47	63.47	62.76	66.90	57.58	**54.55**	54.17	50.70	66.4
model-b	GRPO +ATAR+ CR	74.77	66.17	62.16	67.70	**58.18**	45.45	**62.50**	50.90	66.5
model-c	GRPO +ATAR+ CR + TR	**76.88**	**62.87**	**64.26**	**68.00**	56.97	45.45	57.50	**52.00**	**66.8**
Qwen2.5-Omni (reproduce)	-	69.67	67.37	61.86	66.30	61.21	49.51	57.14	58.20	64.9
sft-c	SFT	77.18	62.57	63.96	67.90	63.03	50.00	57.82	60.90	65.8
sft-d	SFT+CoT	75.98	63.47	63.06	67.50	61.21	48.06	54.08	59.80	65.2
grpo-c	GRPO	75.38	70.06	66.67	69.70	66.06	51.94	62.24	62.50	66.2
grpo-d	GRPO+CoT	76.28	69.76	66.37	69.80	64.24	53.40	59.52	61.80	65.9
model-d	GRPO+ATAR	75.08	67.66	71.77	71.50	63.64	54.85	62.93	64.20	66.8
model-e	GRPO +ATAR+ CR	76.58	68.87	72.07	72.50	66.67	**55.83**	61.22	64.40	67.0
model-f	GRPO+ATAR + CR + TR	**77.48**	**70.36**	**73.37**	**73.70**	**67.27**	53.88	**64.29**	**65.30**	**67.1**

Table 1: Ablation Study Employing Qwen2-Audio-7B-Instruct and Qwen2.5-Omini as the Base Model. ATAR stands for Adaptive Think Accuracy Reward, CR stands for Consistency Reward, and TR stands for Think Reward.

Effectiveness of Adaptive Think Accuracy Reward The adaptive thinking accuracy reward mechanism improves model reasoning performance. We compare models employing the adaptive thinking accuracy reward (model-a/d) with those trained using standard GRPO (grpo-a/c and grpo-b/d). The Qwen2-Audio-based model outperforms grpo-a by 3.10, 0.50, and 1.9 points on the MMAU-test-mini Avg, AIR Foundation Avg, and MMAR Avg metrics, respectively, and exceeds grpo-b by 1.90, 0.70, and 2.3 points on the same metrics. Similarly, the Qwen2.5-Omni-based model-d surpasses grpo-c by 1.80, 1.70, and 0.6 points, and outperforms grpo-d by 1.70, 2.40, and 0.9 points across the corresponding metrics. These results confirm that the adaptive thinking accuracy reward enhances reasoning performance.

Necessity of Consistency Reward Introducing a consistency reward further improves model performance. Models that incorporate the consistency reward (model-b/e) outperform those without it (model-a/d). Specifically, model-b achieves gains of 0.80, 0.20, and 0.10 over model-a on the MMAU-test-mini Avg, AIR Foundation Avg, and MMAR Avg metrics, respectively. Likewise, model-e demonstrates improvements of 1.00, 0.20, and 0.20 over model-d across the same metrics. This mechanism effectively mitigates inconsistencies in the reasoning process.

Impact of Think Reward The integration of think reward during RL yields further performance improvements. Models incorporating thinking rewards (model-c/f) consistently outperform their counterparts lacking the expert LLM judging mechanism (model-b/e). Specifically, model-c outperforms model-b by 0.30, 1.10, and 0.3 points on the MMAU-test-mini Avg, MMAR Avg, and AIR Foundation Avg metrics, respectively. Similarly, model-f exceeds model-e by 1.20, 0.90, and 0.1 points, respectively. These results demonstrate the effectiveness of incorporating think reward to guide model learning.

Comparison with SOTA

MMAU Audio-Thinker achieves substantial improvements over existing baselines on the MMAU benchmark. Table 2 summarizes the key results from the MMAU benchmark. Compared to the Qwen2.5-Omni baseline, Audio-Thinker (Qwen2.5-Omni) improves Test-mini performance from 66.30 to 73.70 and Test-full from 68.03 to 72.83. Compared to the Qwen2-Audio baseline, Audio-Thinker (Qwen2-Audio) also shows substantial improvements, with Test-mini performance increasing from 54.90 to 68.00, and Test-full performance rising from 52.50 to 67.90. Notably, Audio-Thinker (Qwen2.5-Omni) outperforms all previously reported models by achieving absolute improvements of 2.40 and 1.63 over Omni-R1 on Test-mini and Test-full averages, respectively, thereby establishing Audio-Thinker as the new SOTA on the MMAU benchmark.

AIR Audio-Thinker demonstrates state-of-the-art performance on the AIR-Bench foundation benchmark, particularly in sound and speech reasoning. As shown in Table 3, the benchmark evaluates audio understanding across three main categories: sound, music, and speech, where the latter includes Speech Emotion Recognition (SER), Vocal Sound Classification (VSC), and Speech Number Variation (SNV). Audio-Thinker (Qwen2.5-Omni) achieves an overall AIR Avg score of 67.1, surpassing all existing open-source models and several closed-source systems, including GPT-4o Audio (Jaech et al. 2024). It sets a new benchmark in the sound category with 75.8, and excels in speech subtasks, scoring 56.2 in SER, 94.5 in VSC (the highest reported), and 67.5 in SNV (second-best). These results highlight its robust

Model	Sound		Music		Speech		Avg	
	Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Random Guess	26.72	25.73	24.55	26.53	26.72	25.50	26.00	25.92
Most Frequent Choice	27.02	25.73	20.35	23.73	29.12	30.33	25.50	26.50
Human (Test-Mini)	86.31	-	78.22	-	82.17	-	82.23	-
GPT-4o Audio (Jaech et al. 2024)	61.56	56.27	56.29	55.27	66.37	67.20	61.40	59.58
Gemini 2.0 Flash (Narzary et al. 2025)	66.37	63.87	59.58	62.73	60.66	63.13	62.20	63.24
Gemini 2.5 Flash (Comanici et al. 2025)	67.96	65.43	62.28	65.30	62.76	63.30	64.30	64.68
GAMA 7B (Ghosh et al. 2024)	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
Qwen2-Audio-Instruct (Chu et al. 2023)	55.25	56.73	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2 Audio (Yang et al. 2024a)	62.16	45.90	53.59	53.26	48.59	45.90	54.90	52.50
Mellow (Deshmukh et al. 2025)	61.26	64.90	54.19	52.67	29.73	38.77	48.40	52.11
Audio Flamingo 2 (Ghosh et al. 2025)	61.56	65.10	73.95	**72.90**	30.93	40.26	55.48	59.42
Kimi-Audio (Team et al. 2025)	61.68	-	73.27	-	60.66	-	65.00	-
SARI (Qwen2.5-Omni) (Wen et al. 2025)	72.75	-	67.22	-	61.26	-	67.08	-
Audio-Reasoner (Xie et al. 2025b)	60.06	-	64.30	-	60.70	-	61.71	-
Audio-CoT (Ma et al. 2025a)	61.86	-	56.29	-	55.26	-	57.80	-
R1-AQA (Li et al. 2025a)	68.77	69.76	64.37	61.40	63.66	62.70	65.60	64.36
Qwen2.5-Omni-7B (Xu et al. 2025)	69.67	70.63	67.37	66.93	61.86	66.57	66.30	68.03
Omni-R1 (VGGS-GPT) (Rouditchenko et al. 2025)	73.6	74.1	**74.3**	70.8	66.1	68.7	71.3	71.2
AUDIO-THINKER QWEN2-AUDIO *(ours)*	76.88	75.13	62.87	61.83	64.26	67.03	68.00	67.90
AUDIO-THINKER QWEN2.5-OMNI *(ours)*	**77.48**	**76.30**	70.36	66.63	**73.37**	**73.27**	**73.70**	**72.83**

Table 2: Accuracy (%) comparison on MMAU. For baselines, we evaluate GPT-4o Audio, Gemini 2.0 Flash, and Gemini 2.5 Flash. The results of other previous work are sourced from the original papers or the MMAU Leaderboard (old version).

Model	AIR-Sound	AIR-Music	AIR-Speech			AIR-Avg	MMAR			
	SoundAQA	MusicAQA	SER	VSC	SNV		Sound	Music	Speech	Avg
GPT-4o Audio (Jaech et al. 2024)	68.3	67.7	51.2	90.0	61.6	62.3	53.94	50.97	70.41	63.50
Gemini 2.0 Flash (Narzary et al. 2025)	69.9	68.2	56.2	93.5	64.8	66.1	61.21	50.97	72.11	65.60
Gemini 2.5 Flash (Comanici et al. 2025)	74.8	**73.7**	56.4	94.1	**68.5**	**67.4**	55.28	53.40	**77.21**	**66.80**
SALMONN (Yang et al. 2024a)	28.4	54.6	29.9	45.3	34.3	36.8	30.91	29.61	24.35	32.80
Minmo (Chen et al. 2025)	50.3	-	**64.5**	93.0	-	-	-	-	-	-
Qwen2-Audio-Instruct (Yang et al. 2024a)	67.2	64.6	50.5	87.9	60.5	61.3	33.33	24.27	32.31	30.00
Qwen2.5-Omni-7B (Xu et al. 2025)	75.3	70.6	56.4	92.9	63.9	64.9	58.79	40.78	59.86	56.70
Audio-Reasoner (Xie et al. 2025b)	65.7	55.2	60.5	-	56.3	65.2	43.64	33.50	32.99	36.80
Omni-R1 (VGGS-GPT) (Rouditchenko et al. 2025)	-	-	-	-	-	-	67.3	51.5	64.3	63.4
AUDIO-THINKER QWEN2-AUDIO *(ours)*	75.5	68.7	55.7	94.4	64.5	66.8	56.97	45.63	57.50	52.00
AUDIO-THINKER QWEN2.5-OMNI *(ours)*	**75.8**	69.5	56.2	**94.5**	67.5	67.1	**68.48**	**53.88**	64.29	65.30

Table 3: Accuracy (%) comparison on AIR foundation and MMAR. For baselines, we evaluate GPT-4o Audio, Gemini 2.0 Flash, and Gemini 2.5 Flash on the AIR-Bench foundation, and evaluate Gemini 2.5 Flash on MMAR.

audio understanding capabilities across diverse domains.

MMAR Audio-Thinker substantially enhances deep audio reasoning capabilities on the MMAR evaluation set. Table 3 summarizes the results from the MMAR evaluation set. Audio-Thinker (Qwen2.5-Omni) outperforms all existing open-source models, including Omni-R1, which is based on Qwen2.5-Omni but utilizes a larger training dataset. The performance of Audio-Thinker models is comparable to current SOTA closed-source systems such as Gemini 2.5 Flash and GPT-4o Audio. These findings provide compelling evidence that Audio-Thinker establishes its superiority in tackling complex audio reasoning tasks.

Conclusion

We propose Audio-Thinker, an audio-language RL framework that integrates model-generated think-based rewards with adaptive outcome rewards to enable difficulty-aware, consistent, and effective reasoning. An adaptive think–accuracy reward is introduced to dynamically adjust reasoning strategies to task complexity, while think-based supervision mitigates reward hacking by assessing reasoning quality. Experiments on multiple benchmarks show that Audio-Thinker consistently surpasses existing LALMs, highlighting the importance of adaptive reasoning and supervision of the thinking process beyond final correctness.

References

Chen, Q.; Chen, Y.; Chen, Y.; Chen, M.; Chen, Y.; Deng, C.; Du, Z.; Gao, R.; Gao, C.; Gao, Z.; Li, Y.; Lv, X.; Liu, J.; Luo, H.; Ma, B.; Ni, C.; Shi, X.; Tang, J.; Wang, H.; Wang, H.; Wang, W.; Wang, Y.; Xu, Y.; Yu, F.; Yan, Z.; Yang, Y.; Yang, B.; Yang, X.; Yang, G.; Zhao, T.; Zhang, Q.; Zhang, S.; Zhao, N.; Zhang, P.; Zhang, C.; and Zhou, J. 2025. MinMo: A Multimodal Large Language Model for Seamless Voice Interaction. arXiv:2501.06282.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. arXiv:2311.07919.

Comanici, G.; Bieber, E.; Schaekermann, M.; and Pasupat, I. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Deshmukh, S.; Dixit, S.; Singh, R.; and Raj, B. 2025. Mellow: a small audio language model for reasoning. arXiv:2503.08540.

Face, H. 2025. Open R1: A fully open reproduction of DeepSeek-R1.

Fan, K.; Feng, K.; Lyu, H.; Zhou, D.; and Yue, X. 2025. SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward. arXiv:2505.17018.

Feng, K.; Gong, K.; Li, B.; Guo, Z.; Wang, Y.; Peng, T.; Wang, B.; and Yue, X. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. arXiv:2503.03983.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. arXiv:2406.11768.

Goel, A.; Ghosh, S.; Kim, J.; Kumar, S.; Kong, Z.; gil Lee, S.; Yang, C.-H. H.; Duraiswami, R.; Manocha, D.; Valle, R.; and Catanzaro, B. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. arXiv:2507.08128.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Hu, Y.; and Lin, S. 2025a. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Huang, W.; Jia, B.; Zhai, Z.; Cao, S.; Ye, Z.; Zhao, F.; Xu, Z.; Hu, Y.; and Lin, S. 2025b. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. arXiv:2503.06749.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, B.; Zeng, H.; Yue, Z.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Kong, Z.; Goel, A.; Badlani, R.; Ping, W.; Valle, R.; and Catanzaro, B. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*.

Li, G.; Liu, J.; Dinkel, H.; Niu, Y.; Zhang, J.; and Luan, J. 2025a. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. *arXiv preprint arXiv:2503.11197*.

Li, M.; Zhong, J.; Zhao, S.; Lai, Y.; Zhang, H.; Zhu, W. B.; and Zhang, K. 2025b. Think or Not Think: A Study of Explicit Thinking in Rule-Based Visual Reinforcement Fine-Tuning. arXiv:2503.16188.

Liu, Y.; Peng, B.; Zhong, Z.; Yue, Z.; Lu, F.; Yu, B.; and Jia, J. 2025a. Seg-Zero: Reasoning-Chain Guided Segmentation via Cognitive Reinforcement. arXiv:2503.06520.

Liu, Z.; Sun, Z.; Zang, Y.; Dong, X.; Cao, Y.; Duan, H.; Lin, D.; and Wang, J. 2025b. Visual-RFT: Visual Reinforcement Fine-Tuning. arXiv:2503.01785.

Ma, Z.; Chen, Z.; Wang, Y.; Chng, E. S.; and Chen, X. 2025a. Audio-CoT: Exploring Chain-of-Thought Reasoning in Large Audio Language Model. *arXiv preprint arXiv:2501.07246*.

Ma, Z.; Ma, Y.; Zhu, Y.; Yang, C.; Chao, Y.-W.; Xu, R.; Chen, W.; Chen, Y.; Chen, Z.; Cong, J.; Li, K.; Li, K.; Li, S.; Li, X.; Li, X.; Lian, Z.; Liang, Y.; Liu, M.; Niu, Z.; Wang, T.; Wang, Y.; Wang, Y.; Wu, Y.; Yang, G.; Yu, J.; Yuan, R.; Zheng, Z.; Zhou, Z.; Zhu, H.; Xue, W.; Benetos, E.; Yu, K.; Chng, E.-S.; and Chen, X. 2025b. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. arXiv:2505.13032.

Narzary, S.; Brahma, B.; Mahilary, H.; Brahma, M.; Som, B.; and Nandi, S. 2025. Comparative Study of Zero-Shot Cross-Lingual Transfer for Bodo POS and NER Tagging Using Gemini 2.0 Flash Thinking Experimental Model. arXiv:2503.04405.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning. arXiv:2502.19634.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Rouditchenko, A.; Bhati, S.; Araujo, E.; Thomas, S.; Kuehne, H.; Feris, R.; and Glass, J. 2025. Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM? arXiv:2505.09439.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. arXiv:2410.19168.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv: Learning,arXiv: Learning*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sun, G.; Yang, Y.; Zhuang, J.; Tang, C.; Li, Y.; Li, W.; MA, Z.; and Zhang, C. 2025. video-SALMONN-o1: Reasoning-enhanced Audio-visual Large Language Model. arXiv:2502.11775.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Tu, S.; Lin, J.; Zhang, Q.; Tian, X.; Li, L.; Lan, X.; and Zhao, D. 2025. Learning When to Think: Shaping Adaptive Reasoning in R1-Style Models via Multi-Stage RL. arXiv:2505.10832.

Wen, C.; Guo, T.; Zhao, S.; Zou, W.; and Li, X. 2025. SARI: Structured Audio Reasoning via Curriculum-Guided Reinforcement Learning. arXiv:2504.15900.

Xie, T.; Gao, Z.; Ren, Q.; Luo, H.; Hong, Y.; Dai, B.; Zhou, J.; Qiu, K.; Wu, Z.; and Luo, C. 2025a. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025b. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024a. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, P.; Wang, X.; Duan, X.; Chen, H.; Hou, R.; Jin, C.; and Zhu, W. 2022. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 3480–3491. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; and Zhou, J. 2024b. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. arXiv:2402.07729.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; et al. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Fan, T.; Liu, G.; Liu, L.; Liu, X.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhang, J.; Lin, N.; Hou, L.; Feng, L.; and Li, J. 2025a. AdaptThink: Reasoning Models Can Learn When to Think. arXiv:2505.13417.

Zhang, Y.-F.; Lu, X.; Hu, X.; Fu, C.; Wen, B.; Zhang, T.; Liu, C.; Jiang, K.; Chen, K.; Tang, K.; Ding, H.; Chen, J.; Yang, F.; Zhang, Z.; Gao, T.; and Wang, L. 2025b. R1-Reward: Training Multimodal Reward Model Through Stable Reinforcement Learning. arXiv:2505.02835.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Zhang, H.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2025. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zhou, H.; Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2025. R1-Zero's "Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. arXiv:2503.05132.