

SWEAR: SECTION-WISE EFFICIENT AUDIO REASONING VIA SELF-RATED DIFFICULTY

Yixuan Li^{1,2} Minghao Lv¹ Xuenan Xu³ Zihao Zheng¹ Mengyue Wu^{1*}

¹ MoE Key Lab of Artificial Intelligence, X-LANCE Lab, Shanghai Jiao Tong University, China

² Shanghai AI Lab, China

³ KlingAI Research, China

ABSTRACT

Recent work on large audio-language models (LALMs) explores adaptive reasoning to balance accuracy and efficiency. However, existing approaches often rely on coarse-grained difficulty estimation, overlooking the diverse factors that determine reasoning effort. We address this limitation by deriving five-dimensional difficulty scores and compressed reasoning traces from complementary offline assessments. Based on this characterization, we propose SWEAR, a difficulty-aware reasoning framework that learns difficulty signals during supervised fine-tuning and uses reinforcement learning to adapt reasoning allocation across reasoning stages. Experiments on MMAU-test-mini and MMAR show that SWEAR improves accuracy while substantially reducing generated tokens, shortening reasoning for easy instances without sacrificing performance on difficult ones.

Index Terms— Large Audio-Language Models, Structured Difficulty Modeling, Adaptive Reasoning

1. INTRODUCTION

LALMs increasingly use explicit reasoning to connect acoustic observations with answers [1, 2, 3]. As reasoning incurs additional computational cost, characterizing reasoning difficulty is essential for adaptive reasoning allocation, allowing more budget for challenging cases while reducing unnecessary computation for easier ones. However, accurately estimating input difficulty and determining the corresponding reasoning effort remain challenging, motivating difficulty-aware reasoning strategies that dynamically adjust reasoning allocation.

Addressing this question, a parallel line of work has explored input-dependent reasoning effort rather than fixed CoT behaviors. In text reasoning, AdaptThink [4], AutoThink [5], and related methods [6] adapt reasoning length or computational budgets based on instance difficulty. Similar ideas have been investigated in visual reasoning, where FAST [7] and CAR [8] adjust reasoning effort according to input cues. These studies demonstrate the potential of adaptive reasoning for balancing accuracy and efficiency across modalities.

Recent studies have explored difficulty-aware reasoning adaptation in LALMs. Audio-Thinker [9] and Omni-

AutoThink [10] use reinforcement learning to adapt reasoning activation, while Think Smart, Not Hard [11] and CESAR [12] adjust reasoning length based on difficulty-related signals. However, these approaches typically characterize difficulty with a single overall measure, such as rollout performance, attention statistics, or trace-level rewards. Such global estimates indicate whether more reasoning is needed but provide limited guidance on where to allocate reasoning effort. This limitation is important for audio reasoning, where different stages, including question interpretation, acoustic evidence extraction, and answer inference, may introduce distinct and uncorrelated difficulties. Therefore, fine-grained and multi-dimensional difficulty characterization across reasoning sections is needed for adaptive allocation of reasoning effort within structured reasoning traces.

To address this gap, we propose SWEAR, a framework for Section-Wise Efficient Audio Reasoning. In addition, We built DiCoTA from CoTA [13], which supplies section-wise supervision and compressed traces from five offline assessments. SWEAR learns the ratings during Supervised Fine-Tuning (SFT) and uses them during Reinforcement Learning (RL) to control section-wise reasoning length and reference-policy regularization, together with a feedback-driven curriculum. This encourages concise reasoning on easy inputs while preserving sufficient flexibility for challenging ones.

Overall, our contributions are summarized as follows:

- We construct DiCoTA, a 575K-example audio reasoning dataset with five-dimensional difficulty scores conditioned on reasoning sections and compressed reasoning traces.
- We develop SWEAR, a difficulty-aware curriculum learning framework for multi-dimensional reasoning adaptation. SWEAR learns difficulty self-ratings during SFT and leverages them in RL for adaptive reasoning length control.
- Experiments on MMAU and MMAR show that SWEAR improves accuracy by 2.6 and 6.3 percentage points while reducing generated tokens by 60.1% and 62.3%, respectively, compared with SFT without difficulty supervision.

2. METHOD

SWEAR learns to allocate reasoning effort across trace sections through a three-stage pipeline, as illustrated in Fig. 1. First, we construct DiCoTA, an audio reasoning dataset with multi-dimensional difficulty annotations, by assigning dif-

*Corresponding author.

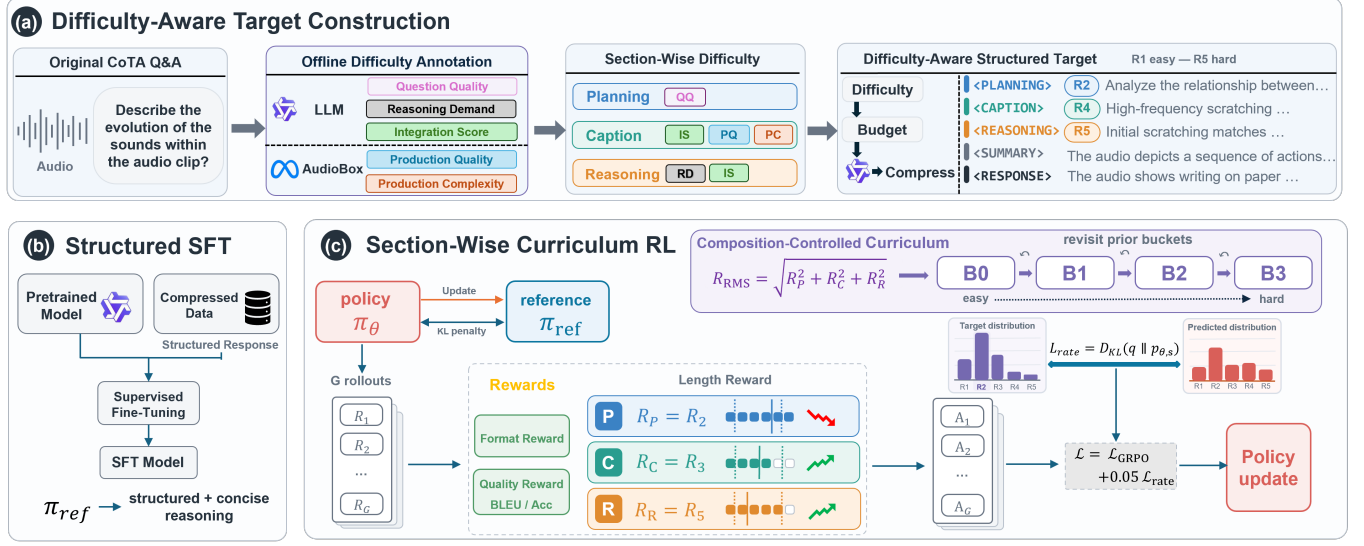


Fig. 1. Overview of our multi-dimensional difficulty-aware audio reasoning framework. (a) Offline linguistic and acoustic annotations estimate Planning, Caption, and Reasoning difficulty, which determine section ratings and compression budgets for structured targets. (b) Structured SFT teaches the model to produce concise, rated reasoning traces. (c) Curriculum RL progresses from easy to hard while jointly optimizing task quality, format, rating-conditioned section-length rewards, and teacher-anchored rating calibration under reference-policy KL regularization. Offline annotators are used only during training.

difficulty scores to individual sections and compressing their content under difficulty-dependent budgets. and compressing each section under an adaptive budget. Second, SFT teaches the policy to predict each section’s difficulty rating before generating its content. Finally, difficulty-conditioned curriculum RL progressively aligns section-wise reasoning effort with the policy’s own difficulty predictions.

2.1. DiCoTA: Difficulty-Aware Trace Construction

We construct DiCoTA, a difficulty-annotated extension of CoTA [13]. Each trace comprises Planning (P), Caption (C), Reasoning (R), and Summary (S), followed by a Response (A). We assign difficulty targets to P , C , and R , while S uses a fixed consolidation budget.

We characterize audio reasoning difficulty through five dimensions spanning acoustic quality, audio complexity, and reasoning demand: production quality (PQ), production complexity (PC), integration score (IS), reasoning demand (RD), and question quality (QQ). PQ and PC are obtained from Audiobox Aesthetics [14], measuring technical quality and the complexity of audio components, respectively. IS, RD, and QQ are provided by Qwen3-Omni-30B-A3B [15], capturing reliance on fine-grained acoustic evidence, inference complexity, and question formulation quality. Together, these dimensions characterize complementary sources of difficulty, ranging from acoustic perception and evidence extraction to question understanding and answer reasoning.

With $g_m(x) = (x - 1)/(m - 1)$ for an m -point score, we construct difficulty proxies by reversing QQ and PQ: lower

question clarity and audio quality receive higher proxy scores.

$$\begin{aligned} q &= 1 - g_5(QQ), & i &= g_5(IS), & r &= g_5(RD), \\ p &= 1 - g_{10}(PQ), & c &= g_{10}(PC). \end{aligned} \quad (1)$$

We define the integrated difficulty of each trace section by aggregating the relevant normalized proxies according to its functional role:

$$D_P = q, \quad D_C = \frac{1}{3}(i + p + c), \quad D_R = \frac{1}{2}(r + i). \quad (2)$$

Specifically, D_P captures question interpretation difficulty. D_C reflects acoustic description difficulty by considering task-relevant evidence, audio quality, and scene complexity. D_R characterizes reasoning difficulty through inference demand and acoustic evidence requirements.

The continuous difficulties are quantized into five equal-width levels, where b_s denotes the corresponding discrete difficulty rating index for section $s \in \mathcal{G} = \{P, C, R\}$.

$$b_s = \min \{5, \lfloor 5D_s \rfloor + 1\} \quad (3)$$

represented by rating tokens $R_1, \dots, R_5$. The resulting target response is

$$y = \langle P \rangle \langle R_{b_P} \rangle y_P; \langle C \rangle \langle R_{b_C} \rangle y_C; \langle R \rangle \langle R_{b_R} \rangle y_R; \langle S \rangle y_S; \langle A \rangle y_A. \quad (4)$$

Unlike uniform trace compression, we allocate a fixed trace budget according to the difficulty of different sections. Let B denote the total budget for P , C , R , and S (excluding the final answer), and let B_S be the fixed Summary budget. For $s \in \mathcal{G}$, the target budget is

$$B_s = (B - B_S) \frac{D_s + \epsilon}{\sum_{j \in \mathcal{G}} (D_j + \epsilon)}, \quad \epsilon > 0, \quad (5)$$

where the difficulty floor prevents degenerate zero allocation.

tions. We use continuous difficulty scores D_s rather than quantized ratings b_s for finer-grained budget allocation. Conditioned on (D_s, B_s) , the teacher removes redundant steps while preserving each section’s function and consistency with the answer. This allocation assigns more tokens to difficult sections while producing more concise traces for easier ones.

Given an audio–question input x and its structured target y , we perform SFT to learn the target response generation. Since each rating is predicted before its corresponding section, it serves as a causal conditioning token rather than a post-hoc annotation. SFT thus establishes the section-wise difficulty interface later optimized by curriculum RL.

2.2. Difficulty-Adaptive Curriculum RL

To encourage difficulty-aware reasoning adaptation and improve the efficiency of reasoning generation, we optimize the SFT policy with the clipped Group Relative Policy Optimization (GRPO) objective [16], using group-normalized advantages and KL regularization to the frozen SFT reference.

$$\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\text{clip}} - \beta_i D_{\text{KL}}(\pi_\theta \| \pi_{\text{SFT}}), \quad (6)$$

where $\mathcal{L}_{\text{clip}}$ denotes the clipped policy objective with group-normalized advantages.

Reward Function. For a sampled response y_i , we define

$$r_i = r_{\text{fmt}}(y_i) + r_{\text{qual}}(y_i) + r_{\text{len}}(y_i), \quad (7)$$

where r_{fmt} checks the required section and rating-token format, and r_{qual} uses normalized BLEU for speech translation or answer accuracy for closed-choice tasks.

For each section $s \in \mathcal{G}$, let $L_{i,s}$ and $C_{i,s}$ denote the generated and compressed DiCoTA reference lengths, respectively. Given the predicted rating $b_{i,s} \in \{1, \dots, 5\}$, we define a rating-dependent target length ratio $\tau_{i,s}$, which increases with $b_{i,s}$, and compute the length reward as

$$r_{\text{len}}(y_i) = - \sum_{s \in \mathcal{G}} \left(\frac{L_{i,s}}{C_{i,s}} - \tau_{i,s} \right)^2. \quad (8)$$

Here $\tau_{i,s}$ is the target ratio between the generated and compressed reference lengths for section s ; higher ratings favor longer generated sections.

Feedback-driven curriculum. We use task-quality feedback to gradually introduce harder samples. The samples are ranked by $D_{\text{rms}} = \sqrt{(D_P^2 + D_C^2 + D_R^2)/3}$ and divided into K_c equal-frequency buckets from easy to hard. The competence score $c \in [0, 1]$ is a scalar estimate of the model’s mastery of the current bucket, updated from recent reward feedback. Following AdaCuRL [17], training starts with $\mathcal{B}_1$ and a competence score $c^{(0)} = 0$, updated as

$$c^{(t+1)} = \text{clip}_{[0,1]} \left(c^{(t)} + (\bar{r}_{\text{qual}}^{(t)} - 0.5) \max\{1 - c^{(t)}, \gamma\} \right). \quad (9)$$

Here, $\bar{r}_{\text{qual}}^{(t)}$ averages the latest M prompt-level quality scores from the hardest active bucket, each computed by averaging the quality rewards of that prompt’s rollouts, and γ sets a minimum update scale. When $c \geq (k - 1)/K_c$, we activate the

next bucket $\mathcal{B}_k$ and reset c to that threshold. Earlier buckets remain active with exponentially downweighted sampling.

Adaptive KL regularization. We adapt the KL coefficient as $\beta_i = \beta_0(1 - 0.5R_{\text{RMS}})$, where R_{RMS} aggregates the normalized *predicted* section ratings. Higher predicted difficulty allows greater deviation from the frozen SFT reference.

Rating calibration. To improve the accuracy of predicted difficulty, we calibrate the control ratings against the unquantized offline difficulty. Let $\mu_s = 5D_s \in [0, 5]$ (without rounding) be the unquantized offline difficulty, and define $q_s(j) \propto \exp[-(j - \mu_s)^2/(2\sigma^2)]$ over the integer rating levels $j \in \{1, \dots, 5\}$, normalized across these levels. We obtain $p_{\theta,s}$ by applying softmax only to the logits of the rating tokens 1–5 immediately before the rating digit, conditioned on the audio, question, and preceding generated content. The calibration loss for section s is $\mathcal{L}_{\text{cal},s} = D_{\text{KL}}(q_s \| p_{\theta,s})$.

The final loss is $\mathcal{L} = \mathcal{L}_{\text{GRPO}} + 0.05 \sum_{s \in \mathcal{G}} \mathcal{L}_{\text{cal},s}$. At inference, the model generates ratings, sections, and the answer autoregressively, without external difficulty scoring or reference traces.

3. EXPERIMENT AND RESULTS

3.1. Experimental Setup

Training Data. From DiCoTA, we use 402,496 examples for SFT and a separate set of 113,410 reward-compatible examples for RL. The RL set covers speech-to-text transformation and multiple-choice audio reasoning tasks, split by D_{rms} into four equal-sized difficulty buckets.

Evaluation. We evaluate on the 1,000-example test sets of MMAU-test-mini and MMAR, using answer accuracy as the effectiveness metric and average generated tokens as the efficiency metric. For difficulty, we score each benchmark example using the same procedure as DiCoTA, partition each benchmark independently into 3 equal-sized groups from easy to hard. Answers are extracted by jointly matching the predicted option label and option content. We compare with representative open-source audio reasoning baselines and SFT variants based on CoTA and DiCoTA, denoted as CoTA-SFT and DiCoTA-SFT, respectively. All models follow the same generation and scoring protocol.

Training Settings. We use Qwen2.5-Omni-7B as the backbone. SFT is conducted for one epoch with LoRA rank 64, scaling factor 128, a global batch size of 64, and a learning rate of 1×10^{-5} . RL continues LoRA optimization from the SFT checkpoint for 2,000 steps with a learning rate of 1×10^{-6} . Each prompt samples $G = 6$ responses, and the curriculum uses $K_c = 4$ difficulty buckets. For the length reward in Eq. (8), we set $\tau_{i,s} = (b_{i,s} + 27)/30$, which increases linearly from 0.93 to 1.07 as $b_{i,s}$ goes from 1 to 5.

3.2. Results and Analysis

As shown in Table 1, our full model achieves 73.5% accuracy on MMAU-test-mini, tying the best result, and leads on MMAR with 62.0%. These gains are achieved with only

Model	MMAU-test-mini								MMAR							
	Easy	Medium	Hard	Sound	Music	Speech	Overall	Token	Easy	Medium	Hard	Sound	Music	Speech	Overall	Token
Qwen2.5-Omni-7B [18]	73.87	70.96	61.26	71.47	66.47	68.17	68.7	14.34	60.18	60.36	56.16	58.18	50.00	62.93	58.9	13.18
<i>Public Baselines</i>																
Audio-Reasoner-7B [13]	69.37	69.16	61.26	66.67	64.67	68.77	66.7	504.36	49.40	43.84	43.84	45.45	39.81	48.98	45.6	537.18
Ke-Omni-R-7B [19]	79.88	75.75	64.86	78.08	70.36	72.07	73.5	76.64	58.38	56.46	49.85	55.15	42.23	58.84	55.0	70.34
Audio-Flamingo 3 [20]	<u>78.68</u>	76.65	<u>65.17</u>	77.48	69.46	73.27	<u>73.4</u>	199.90	61.68	55.26	45.95	49.70	48.06	58.16	54.5	<u>189.10</u>
Omni-CLST [21]	77.78	74.85	63.66	77.48	<u>71.56</u>	67.27	72.1	204.03	60.78	60.06	<u>56.46</u>	63.49	52.45	56.17	59.1	368.01
Omni-R1 [22]	77.18	73.65	<u>68.47</u>	80.48	67.96	70.87	73.2	7.11	66.17	<u>63.66</u>	54.37	58.81	<u>53.46</u>	65.14	<u>61.4</u>	6.82
<i>Our Models</i>																
CoTA-SFT	75.68	74.25	62.46	71.17	67.37	<u>74.17</u>	70.9	494.58	59.28	57.06	50.76	56.36	45.63	59.86	55.7	519.82
DiCoTA-SFT	74.17	<u>76.35</u>	65.47	70.27	69.76	75.38	71.8	216.18	58.68	56.46	53.45	56.36	48.06	57.82	56.0	222.74
SWEAR	<u>78.68</u>	72.46	69.37	<u>78.38</u>	73.65	68.47	73.5	<u>197.13</u>	<u>62.28</u>	66.07	57.66	<u>62.42</u>	54.85	<u>64.29</u>	62.0	195.85

Table 1. Accuracy (%) across predicted difficulty levels and audio categories on MMAU-test-mini and MMAR. Overall denotes accuracy on the full test set, and Token denotes the average number of generated tokens. The best and second-best results are highlighted in **bold** and underlined, respectively. Token rankings include only models with explicit CoT.

197.13 and 195.85 generated tokens, reducing token usage by 60.1% and 62.3% relative to CoTA-SFT. These results demonstrate that SWEAR achieves a stronger accuracy-efficiency trade-off than prior explicit-CoT audio reasoning models, including Audio-Reasoner, Audio-Flamingo 3, Omni-CLST, and CoTA-SFT, consistently reducing reasoning tokens while maintaining or improving task performance.

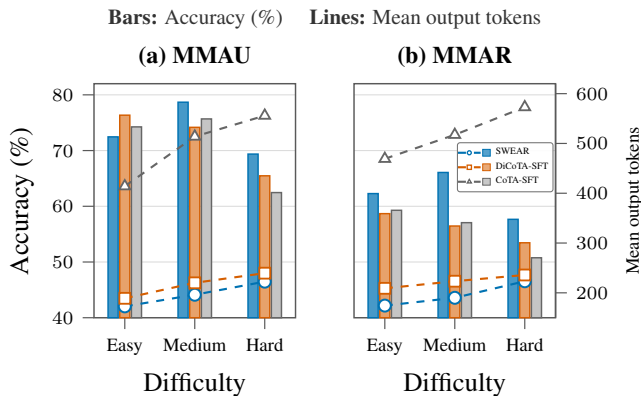


Fig. 2. Accuracy (bars) and mean output tokens (lines) across difficulty under DiCoTA standard on MMAU and MMAR.

Figure 2 exhibits the intended difficulty-aware allocation behavior. As difficulty increases, our full model raises its average output length from 172.8 to 222.5 tokens on MMAU and from 174.4 to 223.2 on MMAR, while remaining 58.3–63.3% shorter than CoTA-SFT and 5.4–16.6% shorter than DiCoTA-SFT across all splits. With this smaller budget, it maintains competitive accuracy on easy MMAU examples and improves hard accuracy by 5.6 points over CoTA-SFT. On MMAR, it consistently outperforms both SFT variants across all difficulty levels. This trend verifies that SWEAR does not apply a uniform reasoning budget, but dynamically adjusts generation length according to estimated difficulty.

3.3. Ablation Studies

We further investigate the impact of different designs and curriculum buckets. As shown in Table 2, removing rating cali-

Setting	K	MMAU		MMAR	
		Acc.↑	Token↓	Acc.↑	Token↓
SWEAR	4	<u>73.5</u>	197.1	62.0	195.9
<i>Component ablations</i>					
w/o rating calibration	4	70.3	188.5	59.4	193.6
w/o length shaping	4	74.2	216.6	60.4	225.5
w/o adaptive KL	4	71.6	198.4	<u>61.5</u>	197.7
<i>Number of curriculum buckets</i>					
uniform training	1	69.3	199.0	58.2	<u>188.2</u>
3 buckets	3	70.1	198.7	60.5	191.4
5 buckets	5	71.2	196.0	60.8	195.8
10 buckets	10	70.4	180.2	59.3	185.7

Table 2. Curriculum RL component ablations and sensitivity to the number of curriculum buckets.

bration reduces accuracy by 3.2 and 2.6 points on MMAU and MMAR. Removing length shaping increases output length by 19.5 and 29.6 tokens, improving MMAU accuracy by 0.7 points but reducing MMAR accuracy by 1.6 points. Static KL also lowers accuracy on both benchmarks. Among curriculum settings, four buckets achieve the highest accuracy. These ablations confirm that rating prediction, adaptive length control, and curriculum scheduling each contribute to the final performance, validating the design of the proposed training strategy.

4. CONCLUSION

We proposed SWEAR, a structured difficulty modeling framework for adaptive audio reasoning. Building on CoTA, we augment each trace with multi-dimensional difficulty annotations and distill them into the LALM’s self-ratings during SFT. These ratings subsequently guide curriculum RL, adapting both reasoning length and sectional demands. Experiments on MMAU and MMAR show that SWEAR produces concise traces for easy problems while preserving sufficient deliberation for hard ones. Extending this difficulty-aware principle to more diverse datasets and multimodal reasoning domains is our future direction.

5. COMPLIANCE WITH ETHICAL STANDARDS

This is a numerical simulation study for which no ethical approval was required.

6. REFERENCES

- [1] Z. Ma, Z. Chen, Y. Wang, E.-S. Chng, and X. Chen, “Audio-CoT: Exploring chain-of-thought reasoning in large audio language model,” in *Proc. ASRU*, 2025, pp. 1–6.
- [2] C. Wen, T. Guo, S. Zhao, W. Zou, and X. Li, “SARI: Structured audio reasoning via curriculum-guided reinforcement learning,” *arXiv:2504.15900*, 2025.
- [3] F. Tian, X. T. Zhang, Y. Zhang, H. Zhang, Y. Li, D. Liu, et al., “Step-audio-R1 technical report,” *arXiv:2511.15848*, 2025.
- [4] J. Zhang, N. Lin, L. Hou, L. Feng, and J. Li, “Adapt-Think: Reasoning models can learn when to think,” in *Proc. EMNLP*, 2025, pp. 3716–3730.
- [5] S. Tu, J. Lin, Q. Zhang, X. Tian, L. Li, X. Lan, et al., “Learning when to think: Shaping adaptive reasoning in R1-style models via multi-stage RL,” in *Proc. NIPS*, 2025, vol. 38, pp. 16989–17015.
- [6] Y. Shen, J. Zhang, J. Huang, S. Shi, W. Zhang, J. Yan, et al., “DAST: Difficulty-adaptive slow-thinking for large reasoning models,” in *Proc. EMNLP (Industry Track)*, 2025, pp. 2322–2331.
- [7] W. Xiao and L. Gan, “Fast-slow thinking GRPO for large vision-language model reasoning,” in *Proc. NIPS*, 2025, vol. 38, pp. 190068–190098.
- [8] J. Lu, H. Yu, S. Xu, S. Ran, G. Tang, S. Wang, et al., “Prolonged reasoning is not all you need: Certainty-based adaptive routing for efficient LLM/MLLM reasoning,” *arXiv:2505.15154*, 2025.
- [9] S. Wu, C. Li, W. Wang, H. Zhang, H. Wang, M. Yu, et al., “Audio-thinker: Guiding large audio language model when and how to think via reinforcement learning,” in *Proc. AACL*, 2026, vol. 40, pp. 33962–33970.
- [10] D. Yang, S. Liu, D. Wang, Y. Wang, G. Wan, and H. Meng, “Omni-AutoThink: Adaptive multimodal reasoning via reinforcement learning,” *arXiv:2512.03783*, 2025.
- [11] Z. Sheng, S. Zhou, C. Gong, and Z. Li, “Think smart, not hard: Difficulty adaptive reasoning for large audio language models,” in *Findings of ACL*, 2026, pp. 32771–32790.
- [12] J. Fan, R. Ren, J. Li, R. Pandey, P. G. Shivakumar, I. Bulko, et al., “Incentivizing consistent, effective and scalable reasoning capability in audio LLMs via reasoning process rewards,” in *Proc. ICLR*, 2026.
- [13] Z. Xie, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, “Audio-reasoner: Improving reasoning capability in large audio language models,” in *Proc. EMNLP*, 2025, pp. 23829–23851.
- [14] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, et al., “Meta audiobox aesthetics: Unified automatic assessment for speech, music and sound,” in *Proc. ASRU*, 2025, pp. 1–8.
- [15] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, et al., “Qwen3-Omni technical report,” *arXiv:2509.17765*, 2025.
- [16] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, et al., “DeepSeekMath: Pushing the limits of mathematical reasoning in open language models,” *arXiv:2402.03300*, 2024.
- [17] R. Li, H. Huang, F. Wei, F. Xiong, Y. Wang, and X. Chu, “AdaCuRL: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting,” *arXiv:2511.09478*, 2025.
- [18] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, et al., “Qwen2.5-Omni technical report,” *arXiv:2503.20215*, 2025.
- [19] S. Zhao, T. Guo, C. Wen, B. Xiang, W. Zou, and X. Li, “Ke-Omni-R: Achieving advanced audio reasoning with a concise 50-words think process,” GitHub repository: <https://github.com/shuaijiang/Ke-Omni-R>, 2025, Accessed: 2026-09-04.
- [20] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, et al., “Audio flamingo 3: Advancing audio intelligence with fully open large audio language models,” in *Proc. NIPS*, 2025, vol. 38.
- [21] J. Zhao, H. Su, L. Fan, Z. Luo, H. Wang, H. Sun, et al., “Omni-CLST: Error-aware curriculum learning with guided selective chain-of-thought for audio question answering,” *arXiv:2509.12275*, 2025.
- [22] A. Rouditchenko, S. Bhati, E. Araujo, S. Thomas, H. Kuehne, R. Feris, et al., “Omni-R1: Do you really need audio to fine-tune your audio LLM?,” in *Proc. ASRU*, 2025, pp. 1–7.