Closing the Modality Reasoning Gap for Speech Large Language Models

Anonymous ACL submission

Abstract

Although speech large language models have achieved notable progress, a substantial modality reasoning gap remains: their reasoning performance on speech inputs is markedly weaker than on text. This gap could be associated with representational drift across Transformer layers and behavior deviations in long-chain reasoning. To address this issue, we introduce a reinforcement-learning framework that aligns text-conditioned and speech-conditioned trajectories through an asymmetric reward design. The framework employs two dense and complementary signals: representation alignment, which measures layer-wise hidden-state similarity between speech- and text-conditioned trajectories, and behavior alignment, which evaluates semantic consistency between generated outputs and reference text completions. Experiments on challenging reasoning benchmarks, including MMSU and OBQA, show that our approach significantly narrows the modality reasoning gap and achieves state-of-the-art performance among 7B-scale Speech LLMs.

1 Introduction

Recent advances in Speech Large Language Models (Speech LLMs) enable a unified framework for spoken language processing tasks such as automatic speech recognition (ASR), speech translation, and speech QA. Most Speech LLMs follow a three-stage architecture, consisting of a pretrained speech encoder, lightweight adapters, and a decoder-only text LLM (Peng et al., 2025; Cui et al., 2025). The encoder transforms raw speech into high-resolution acoustic representations, which are then projected into the text-embedding space through learned adapters, enabling the downstream LLM to process speech inputs using its text-native reasoning stack. Through shared representations across modalities, this architecture allows speech inputs to leverage the generation and reasoning capabilities of text-based LLMs.

However, Speech LLMs exhibit a persistent and critical challenge: the modality reasoning gap, denoting a substantial decline in reasoning performance on speech inputs compared to text, as evidenced by empirical analyses (Xiang et al., 2025) on VoiceBench and SpeechMMLU benchmarks (Chen et al., 2024; Xiaomi, 2025).

To close modality reasoning gap, prior works have primarily focused on input-side fusion and output-side supervision. Input-level modality fusion methods aim to reduce the discrepancy between speech representations and text embeddings at input stage, by freezing the LLM backbone and training lightweight adapters (Anonymous, 2025; Lu et al., 2025a; Xu et al., 2025c). As speech naturally contains paralinguistic cues absent in text, strict input equivalence may not be desirable. However, for complex reasoning tasks, the underlying logical progression should remain invariant regardless of the input modality (Mousavi et al., 2025). Relying solely on inputs can cause subtle discrepancies to propagate and amplify through Transformer layers, leading to significant representational drift. Another line of work provides output-level supervision. They focus on knowledge distillation or prompt-switching training to encourage speech-conditioned generations to mimic text-conditioned behaviors (Wang et al., 2025; Ding et al., 2025). However, these methods enforce strict token-level supervision in a off-policy manner. Since the speech-conditioned distribution differs from the text one, forcing the model to generate exact text-conditioned tokens targets an unreachable objective. Furthermore, this rigid supervision suffers from exposure bias: a single token error shifts the model into an unsupervised state, causing it to fall into behavioral divergence.

To address these limitations, we introduce our trajectory alignment framework, which combines representation alignment to mitigate drift and behavior alignment with more flexible objective. By

leveraging on-policy exploration, our method enables the model to mitigate exposure bias and maintain reasoning stability, preventing the model from drifting into out-of-distribution states. Specifically, representation alignment is computed from the cosine similarity of layer-wise hidden states, providing coarse-grained representation-level feedback. Complementarily, behavior alignment is derived from the semantic consistency of the final outputs, providing token-level but comparatively sparser feedback. These signals effectively steer the speech modality toward the text reasoning trajectory.

Under an asymmetric reward design with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), we jointly optimize task accuracy and these two alignment rewards. Even when all samples in the generated group exhibit zero task accuracy, a common outcome for speech-conditioned reasoning given its greater difficulty than text-based inference, the alignment signals remain informative, enabling direct trajectory alignment between speech and text. Empirical results on complex reasoning benchmarks, such as MMSU and OBQA, demonstrate that our method outperforms existing baselines, achieving state-of-the-art performance among 7B models. The contributions of our work are summarized as follows:

- We propose an on-policy reinforcement-learning framework for trajectory alignment that aligns speech-conditioned reasoning trajectory with its text-conditioned counterpart, closing the modality reasoning gap without architectural modifications.
- We introduce an asymmetric dense reward with two complementary alignment signals: representation alignment that reduces layer-wise hidden-state drift, and behavior alignment that enforces semantic consistency.
- Our method achieves the state-of-the-art performance on reasoning benchmarks (MMSU, OBQA) among 7B-scale models. We release datasets and code to facilitate reproducibility.

2 Related Works

Speech LLMs. Speech LLMs have progressed from cascaded pipelines (ASR + text LLM + optional TTS) to end-to-end architectures that couple speech perception with LLM-style generation, enabling spoken dialogue and spoken QA while better leveraging paralinguistic cues beyond transcripts (Peng et al., 2025; Cui et al., 2025). A dominant design follows a three-stage paradigm: a pretrained speech encoder extracts acoustic features, which are mapped into the text embedding space via lightweight projectors to condition a decoder-only LLM, preserving text-pretrained reasoning priors while extending to speech tasks. Recent open and proprietary omni systems further target low-latency and multi-modal interaction. Open-weight examples such as Qwen2.5-Omni and Qwen3-Omni integrate unified perception and generation, and introduce modality-specialized capacity (e.g., MoE routing) to improve scalability (Xu et al., 2025a,b). Audio-centric models like Kimi-Audio and MiniCPM-o emphasize practical voice interaction and general audio understanding (Ding et al., 2025; Yao et al., 2024).

Modality Alignment and Reasoning Gap. Despite the unified architecture, a performance disparity between speech and text modalities persists, termed the modality reasoning gap. Empirical studies (Chen et al., 2024; Xiaomi, 2025; Mousavi et al., 2025) and representational analyses (Xiang et al., 2025) reveal that speech-conditioned hidden states often drift from their text counterparts, leading to degraded reasoning capabilities. Existing efforts to bridge this gap generally fall into two categories: (1) Architectural Adaptations. To prevent the degradation of text-based capabilities, a prominent line of work adopts a frozen-backbone strategy. Methods such as AlignChat (Anonymous, 2025), DeSTA (Lu et al., 2025a), OTReg (Xu et al., 2025c), and MTBI (Xie et al., 2025) keep the LLM parameters fixed, focusing exclusively on refining the input-side projector. For instance, DeSTA utilizes speech captioning tasks to force acoustic features into the text embedding space. However, this approach yields only a surface-level alignment. By freezing the backbone, the model cannot adapt to speech-specific dynamics, causing reasoning trajectories to diverge even when inputs are projected closely. (2) Supervised Alignment Strategies. Beyond architectural changes, other works employ data-driven objectives to align model behaviors at the output level. Approaches such as Kimi-Audio (Ding et al., 2025) utilize prompt-switching strategies to bridge generation divergence, while others apply cross-modal knowledge distillation (Wang et al., 2025) or data selection as in SALAD (Cuervo et al., 2025) to en-

courage speech-conditioned probabilities to match text-based teachers. Works like SSR (Tan et al., 2025) combine theses two, try to align representation and output behavior. However, these methods rely on cross-entropy or KL divergence on static targets. This off-policy supervision forces the model to mimic the final answer but does not teach it how to correct its own reasoning trajectory, leading to compounding errors in complex tasks. In contrast, our RL-based framework aligns the reasoning trajectory itself via on-policy exploration.

Reinforcement Learning for Reasoning. RL has proven essential for enhancing the reasoning capabilities of LLMs beyond standard supervised fine-tuning (Liu et al., 2025a). Techniques like GRPO enable models to learn from sparse rewards and self-exploration, significantly improving performance on math and logic tasks (Shao et al., 2024; Yu et al., 2025b). In the speech domain, however, RL application remains nascent, primarily limited to aligning paralinguistic attributes or general helpfulness rather than reasoning logic (Li et al., 2025; Liu et al., 2025b). Most importantly, standard binary rewards are sparse and insufficient for modality alignment. Our work bridges this gap by adapting RL with dense alignment signals–leveraging the text modality as a stable reference to guide the speech reasoning trajectory.

3 Method

3.1 Problem Formulation

We define a Speech LLM π_θ as a composite architecture consisting of an audio encoder, a modality projector, and a decoder-only LLM initialized from a text-pretrained LLM π_{base}. For a given query $q \in \mathcal{D}$, the model accepts either its speech representation q_{speech} or text representation q_{text} as input to generate a text response y. Despite extensive alignment training on large-scale speech-text pairs, a significant modality reasoning gap persists, where the model's performance on speech inputs lags behind its text capabilities.

We quantify this gap using Modality Recovery Rate (MRR). Let $y_{\text{speech}} = \pi_\theta(q_{\text{speech}})$ be the completion generated by the current model, and $y_{\text{text}}^{\text{base}} = \pi_{\text{base}}(q_{\text{text}})$ be the reference completion from the base model. Given a reasoning metric $\mathcal{S}$ (e.g., QA accuracy), MRR measures the extent to which the Speech LLM retains the original reasoning capability:

$$\text{MRR}(\pi_\theta) = \frac{\mathbb{E}_{q\in\mathcal{D}}[\mathcal{S}(y_{\text{speech}})]}{\mathbb{E}_{q\in\mathcal{D}}[\mathcal{S}(y_{\text{text}}^{\text{base}})]} \times 100\%. \quad (1)$$

Our objective is to optimize parameters θ such that MRR $\geq 100\%$, effectively closing the gap.

3.2 Reward Modeling

We propose an asymmetric reward design to align reasoning trajectories across modalities. During training, we use text-conditioned completions y_{text} generated by the current policy π_θ as a moving reference. We optimize the policy on both text-conditioned and speech-conditioned completions, allowing the text branch to continue improving under base reward while providing an increasingly strong reference for aligning speech trajectories. As a result, the speech modality co-evolves with the model's improving text reasoning capability.

For a speech-conditioned completion y_{speech}, the total reward is defined as:

$$R_{\text{total}} = R_{\text{base}} + \alpha \cdot R_{\text{rep}} + \beta \cdot R_{\text{beh}}, \quad (2)$$

where R_{rep} and R_{beh} correspond to representation alignment and behavior alignment signals, respectively. We apply R_{total} to speech-conditioned completions, while text-conditioned completions are optimized using R_{base}. We set $\alpha = 1.0$ and $\beta = 1.0$ in our experiments to simultaneously align internal representations and external behaviors.

Base Reward. Following the formulation in DeepSeek-R1 (Guo et al., 2025), we design the base reward to optimize task accuracy and output format. This configuration serves as the Standard GRPO baseline in our ablation studies. It is computed as:

$$R_{\text{base}} = R_{\text{acc}} + \lambda R_{\text{fmt}}, \quad (3)$$

where $R_{\text{acc}} \in \{0, 1\}$ indicates whether the answer extracted by xFinder (Yu et al., 2025a) matches the ground truth, and $R_{\text{fmt}} \in \{0, 1\}$ rewards format compliance (see Appendix A.1). We set $\lambda = 0.5$.

Representation Alignment Reward. This component focuses on aligning the internal latent representations between speech and text. We compute the geometric similarity between the layer-wise hidden states of the generated speech completion and text reference. Let $\mathbf{H}^{(l)} \in \mathbb{R}^{T\times d}$ denote the hidden states at layer l, where T is the total sequence length and d is the hidden dimension. We

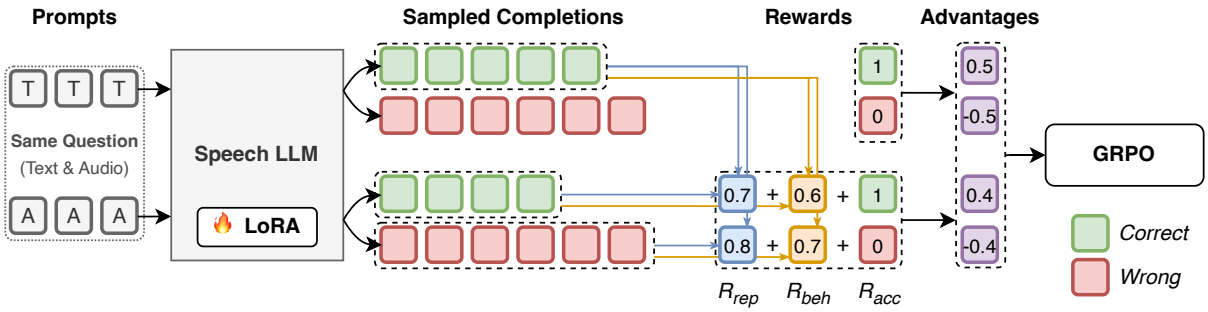

Figure 1: Overview of our framework. We introduce a reinforcement learning approach for trajectory alignment by optimizing an asymmetric reward function composed of representation alignment and behavior alignment.

exclude the first n prompt tokens to focus solely on the generated reasoning chain, computing the mean vector of the remaining tokens:

$$\bar{\mathbf{h}}^{(l)} = \frac{1}{T-n} \sum_{t=n+1}^{T} \mathbf{H}^{(l)}_{t,:}. \tag{4}$$

For each speech completion, we randomly sample a correct (i.e., $R_{\text{acc}} = 1$) text completion y^*_{text} as reference from the current group, and compute the average cosine similarity across all L layers:

$$R_{\text{rep}} = \frac{1}{L} \sum_{l=1}^{L} \text{CosSim}(\bar{\mathbf{h}}^{(l)}_{\text{speech}}, \bar{\mathbf{h}}^{(l)}_{\text{text}}). \tag{5}$$

If no correct text completion is available, we set $R_{\text{rep}} = 0$ and fall back to base reward only. This dense signal encourages the speech modality to emulate the internal thinking representation of the text modality.

Behavior Alignment Reward. To ensure behavior consistency at the output level, we employ an external embedding model $\mathcal{E}$ (e.g., `Qwen3-Embedding-0.6B` (Zhang et al., 2025)) to measures the semantic equivalence between the final generated speech completion y_{speech} and text reference y^*_{text}:

$$R_{\text{beh}} = \text{CosSim}(\mathcal{E}(y_{\text{speech}}), \mathcal{E}(y^*_{\text{text}})). \tag{6}$$

Similarly, if no correct text completion is available, we set $R_{\text{beh}} = 0$. This objective allows the model to learn from diverse valid reasoning trajectories, provided the final semantic behavior remains consistent with the teacher.

3.3 Reinforcement Learning Framework

As illustrated in Figure 1, we employ GRPO to optimize our proposed reward. For a given prompt q, we generate a group of G completions, composed of equal numbers of speech-conditioned and text-conditioned completions. Following Dr. GRPO (Liu et al., 2025c), we define the advantage $\hat{A}_i$ for the i-th completion as normalizing its reward against the group's mean. The model's parameters θ are updated using the DAPO loss (Yu et al., 2025b). Compared to using only the base reward, our alignment reward provides richer guidance by supplying a continuous similarity-based signal that remains effective even when task accuracy rewards are uniformly zero for speech-conditioned reasoning trajectories.

Modality Specific Normalization. A naive implementation of GRPO normalizes reward across the entire group. However, text-conditioned completions inherently achieve higher base rewards than speech-conditioned, which would cause speech-conditioned completions to consistently receive negative advantages, suppressing learning. To address this, we introduce modality-specific normalization, calculating advantages for text and speech completions in separate groups:

$$\hat{A}_{i,m} = r_{i,m} - \mu_m, \quad m \in \{\text{speech}, \text{text}\}, \tag{7}$$

where μ_m is the mean of rewards within modality m. This ensures that each modality is optimized relative to its own baseline, allowing continuous improvement in modality alignment.

4 Experiments

Our experiments aim to verify whether our approach can reduce the modality reasoning gap between speech and text inputs. We report performance under both Audio (A) and Text (T) input settings, and use MRR as defined in Equation 1 to quantify how effectively reasoning ability is recovered in the speech modality relative to

its text counterpart. The core of our method is an alignment-aware reinforcement learning framework, combining with an asymmetric reward design with modality-specific normalization to optimize reasoning behaviors.

4.1 Experimental Setup

Foundation Models. We select two Speech LLMs for our experiments: Qwen2.5-Omni-7B and Phi-4-Multimodal-Instruct-7B (Phi-4-MM-7B). Qwen2.5-Omni-7B is initialized from an official released checkpoint, while Phi-4-MM-7B is a internal model. Both models follow the composite Speech LLM architecture as defined in Section 3.1. For all post-training experiments, we start from the same corresponding checkpoints to ensure fair comparision across methods.

Training Data. We utilize the UnifiedQA training set (Khashabi et al., 2020)[1] as our primary dataset. To construct paired speech-text input for alignment, we synthesized speech using two high-quality TTS systems: CosyVoice2 (Du et al., 2024) and `openaudio-s1-mini`[2]. To ensure speaker diversity, we randomly sample reference speakers from the EN subset of Emilia-YODAS (He et al., 2025). For quality control, all synthesized audio is transcribed by `whisper-medium` (Radford et al., 2023) and filtered using a WER threshold of 10%. After filtering, our training set contains 9,953 samples spanning 203 hours.

Evaluation Benchmarks. We evaluate on two spoken multiple-choice QA benchmarks from VoiceBench (Chen et al., 2024): MMSU and OBQA. MMSU contains 3,074 examples, derived from MMLU-Pro (Wang et al., 2024) covering 12 diverse domains, primarily measuring multi-domain general knowledge and reasoning. OBQA consists of 455 examples, focusing on elementary science facts and commonsense reasoning. We additionally report WER on LibriSpeech (Panayotov et al., 2015) (test-clean & test-other) as an auxiliary diagnostic of ASR capability, obtained by prompting the model to transcribe the audio and computing standard WER with greedy decoding.

Baselines. To ensure a fair comparison, all post-training baselines utilize the same foundation models, training data subsets, freezing strategy and LoRA configuration. We compare our approach against two categories of methods: (1) cross-modal alignment methods that explicitly target speech-text alignment, and (2) general post-training methods that optimize task performance (e.g. accuracy) without explicitly modeling the modality gap.

For cross-modal alignment methods, we include SALAD (Cuervo et al., 2025), DeSTA2.5-Audio (Lu et al., 2025b), AlignChat (Anonymous, 2025), and Knowledge Distillation (KD) (Wang et al., 2025). SALAD and DeSTA2.5-Audio represent cross-modal alignment, where SALAD focuses on sample-efficient distillation or targeted data selection, while DeSTA2.5-Audio utilizing self-generated text completion as alignment target. AlignChat represents frozen-backbone method, focusing on speech adapters' alignment, without altering backbone LLM. KD serves as a distillation-based transfer baseline from a text teacher to speech student.

For general-purpose post-training, we compare against strategies optimized solely for task accuracy, including SFT, DPO, and Standard GRPO. For SFT and DPO, we construct preference data via reject sampling, where SFT trains only on the chosen completions while DPO leverages chosen-rejected pairs. Standard GRPO corresponds to the base-reward-only RL baseline in Section 3.3.

Additionally, we report results from cascaded systems as another pipeline baselines. We use `whisper-large-v3` as the ASR front-end and pair it with Llama3.1-8B, Qwen2.5-7B, and Phi-4-7B (the text backbone π_{base} of Phi-4-MM-7B). For audio evaluation, we first transcribe the speech using the ASR model and then feed the resulting text into the corresponding text LLM, reflecting the impact of ASR errors on reasoning. For text evaluation, we directly feed the clean text to the same LLM.

RL Training Protocol We follow the GRPO training protocol, using a group size of $G = 8$ completions per prompt. Each group contains an equal number of speech-conditioned and text-conditioned completions. During RL, as described in Section 3.2, the reference for alignment is text completions generated by the current policy, serving as a teacher signal; correspondingly, text completions also participate in gradient updates. Finally, we compute advantages using modality-specific normalization, avoiding consistently negative advantages for speech completions as discussed in Section 3.3.

[1] https://huggingface.co/datasets/cais/mmlu

[2] https://huggingface.co/fishaudio/openaudio-s1-mini

Implementation Details. We conduct training with ms-swift[3] for Qwen2.5-Omni experiments and HuggingFace TRL[4] for Phi-4-MM. We adpot parameter-efficient fine-tuning using LoRA on all linear layers, while freezing the audio encoder and projector. For reinforcement learning, we use DAPO loss estimator (Yu et al., 2025b) and follow the settings in Section 3.3. We utilize sampling decoding during RL training to encourage exploration, and greedy decoding for evaluation to ensure deterministic outputs. To reduce formatting noise, we apply xFinder (Yu et al., 2025a), an LLM-based answer extractor, to extract the predicted options from model's completion for robust accuracy computation. Training takes approximately 55 hours for Qwen2.5-Omni and 35 hours for Phi-4-MM on 4×A100 or 8×H200 GPUs. Detailed hyperparameters and prompt templates are provided in Appendix A.2 and Appendix A.3, respectively.

4.2 Main Results

Table 1 presents the performance of our framework against a suit of baselines. As shown by the base Speech LLMs, performance under speech inputs consistently lags behind text, revealing a clear modality reasoning gap. Existing cross-modal alignment baselines can narrow this gap, yet most still fall short of full recovery with MRR < 100%. Our proposed framework demonstrates state-of-the-art performance among 7B models. For the Qwen2.5-Omni with Qwen2.5-7B as backbone, our approach achieves an average audio accuracy of **76.84%**, substantially outperforming other end-to-end alignment methods such as SALAD (66.30%) and MiniCPM-o (66.40%). It also reaches an MRR of **98.89%**. Our RL-based method proves more effective than supervised mimicry, surpassing the Knowledge Distillation (KD) baseline (72.87%) by a large margin. For the Phi-4-MM with Phi-4-7B as backbone, our method achieves the best performance with an accuracy of **79.80%**, even surpassing the original text accuracy of 78.39% and achieving the MRR = **100.45%**.

These results show that our method not only narrows the modality reasoning gap, but also improves text performance, from 76.17% to 78.56% for Qwen2.5-Omni and from 78.39% to 83.82% for Phi-4-MM, indicating that gains in speech are not obtained at the expense of text reasoning, instead, the knowledge learned from speech can further strengthen text-based reasoning. Notably, the improved text accuracy remains higher than the corresponding audio accuracy (76.84% and 79.80% respectively), suggesting that residual differences are likely due to imperfect speech representations and cross-modal projection noise, making text inputs a natural upper bound.

Cascaded systems are often considered strong baselines and can outperform end-to-end models. However, our end-to-end models on Qwen2.5 (76.84%) and Phi-4-MM (79.80%) exceed the performance of the ASR + Qwen2.5-7B pipeline (75.55%) and ASR + Phi-4-7B (73.40%), respectively. This suggests that directly processing speech signals can avoid certain ASR-induced errors, leading to a more robust reasoning process.

4.3 Effectiveness of Training Strategies

Table 2 compares different training strategies on the same backbone (Phi-4-MM), including inference-time prompting, SFT, DPO, Standard GRPO, and our method. Chain-of-Thought (CoT) prompting yields a clear gain on speech inputs, improving the average audio accuracy from 63.16% to 70.06% and increasing MRR from 79.60% to 88.29%. However, it is unable to fully eliminate the gap, suggesting that prompting alone is insufficient to resolve the cross-modal misalignment.

Post-training with supervised or preference-based objectives further improves performance, yet still falls short of full recovery. SFT and DPO raise the average audio accuracy to 72.52% and 75.37%, respectively. This indicates that while supervision and preference optimization help, they do not explicitly align cross-modal reasoning behavior. Standard GRPO, trained with the base reward R_{base}, provides additional improvements (MRR = 92.21%) but still underperforms DPO, highlighting the limitation of sparse, outcome-centric rewards. In contrast, our approach achieves the highest performance, reaching **79.57%** average audio accuracy and MRR = **100.28%**, demonstrating the effectiveness of proposed asymmetric dense alignment reward. Finally, we monitor ASR-related capability via WER and observe it remains unchanged (≈4.16%–4.24%), supporting the conclusion that gains primarily stem from reasoning alignment rather than improved speech recognition.

4.4 Reward Components

Table 2 presents an ablation study on Phi-4-MM backbone, starting from Standard GRPO trained

[3] https://github.com/modelscope/ms-swift
[4] https://github.com/huggingface/trl

Model	Backbone	MMSU		OBQA		Average		MRR (%)
		A	T	A	T	A	T	
Proprietary & Cascaded Systems								
GPT-4o-mini-Audio	GPT-4o-mini	72.90	81.23	84.84	90.11	78.87	85.67	92.06
ASR† + Llama3.1-8B	Llama3.1-8B	58.78	65.65	72.53	80.88	65.66	73.27	89.61
ASR† + Qwen2.5-7B	Qwen2.5-7B	67.1	71.65	84.0	83.74	75.55	77.70	97.23
ASR† + Phi-4-7B	Phi-4-7B	69.00	74.92	77.80	83.96	73.40	79.44	92.40
Existing Baselines								
DeSTA2.5-Audio	Llama3.1-8B*	60.87	65.65	74.06	80.88	67.47	73.27	92.08
SALAD-7B	Qwen2.5-7B	57.5	71.6	75.1	90.1	66.30	80.85	85.33
MiniCPM-o 2.6	Qwen2.5-7B	54.78	59.42	78.02	82.86	66.40	71.14	85.46
Knowledge Distillation	Qwen2.5-7B	63.09	69.15	82.64	84.62	72.87	76.89	93.78
AlignChat	Qwen2.5-7B*	69.65	71.65	85.49	83.74	77.57	77.70	99.83
Base & Aligned Models								
Qwen2.5-Omni	Qwen2.5-7B	61.51	67.94	81.09	84.40	71.30	76.17	91.76
Phi-4-MM	Phi-4-7B	54.81	72.15	71.65	84.62	63.23	78.39	79.59
Ours (Qwen2.5-Omni)	Qwen2.5-7B	67.96	68.54	85.71	88.57	76.84	78.56	98.89
Ours (Phi-4-MM)	Phi-4-7B	**70.14**	**75.76**	**89.45**	**91.87**	**79.80**	**83.82**	**100.45**

Table 1: Reasoning Benchmarks Results. Accuracy (%) on MMSU and OBQA are reported using VoiceBench evaluator for Audio (**A**) and Text (**T**) input. Underlined scores denote the values used as the denominator when computing MRR. † cascaded systems; * frozen LLM backbone.

with the base reward R_{base} only, and then adding R_{rep}, R_{beh}, or their combination. Incorporating representation alignment reward consistently improves performance, increasing MRR from 92.21% to 95.56%, suggesting that aligning layer-wise hidden-state representations provides a denser signal for RL. Alternatively, adding the behavior alignment reward pushes the model close to full recovery (MRR = 99.22%), indicating that semantic-consistency supervision constrains speech outputs toward correct text-conditioned behaviors. Combining both rewards achieves the best result (MRR = 100.28%), showing that representation and behavior signals are complementary: representation alignment mitigates representation drift, while behavior alignment enforces semantic target consistency. This trend aligns with our trajectory alignment objective–jointly aligning internal representation and external semantic behaviors.

4.5 Layer Sensitivity

Figure 2 investigates the sensitivity of the representation-alignment reward to different depths within the 32-layer Phi-4-MM backbone. We partition the model into Shallow (layers 1–10), Middle (11–20), Deep (21–30), and Last (31–32) groups, comparing these against an All-layer baseline. Results indicate that the Middle layers are the most critical localized region, achieving **75.48%** accuracy. Conversely, applying the reward exclusively to Shallow or Last layers is less effective. This suggests that representation drift is primarily in the mid-to-late reasoning stages, rather than during early perceptual processing or final logit alignment. Selecting All layers yields the highest accuracy (**75.83%**), this confirms that global cross-layer similarity offers advantages over localized alignment.

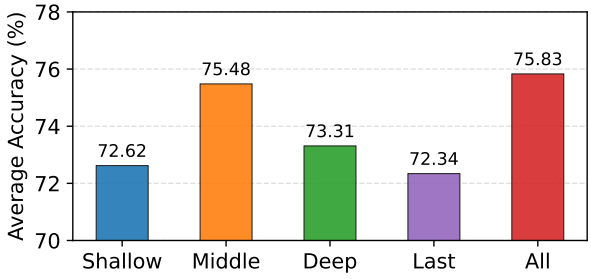

Figure 2: Sensitivity Analysis of Representation Reward Layers. Average audio accuracy on MMSU and OBQA across different layer groups.

4.6 Representation Alignment Analysis

Figure 3 presents a layer-wise representation alignment study. We perform a teacher-forcing analysis: we feed the exact same text-conditioned generated CoT response tokens to both the text-conditioned

Method	MMSU		OBQA		Average		MRR (%)	WER (%)
	A	T	A	T	A	T		
Inference Baseline								
Phi-4-MM	54.00	71.15	72.31	85.05	63.16	78.10	79.60	**4.16**
+ CoT Prompting	60.77	70.85	79.34	86.15	70.06	78.50	88.29	-
Post-training Alignment								
SFT	63.50	70.85	81.54	87.47	72.52	79.16	91.37	4.18
DPO	66.33	74.72	84.40	91.43	75.37	83.08	94.98	4.23
Reinforcement Learning								
Standard GRPO (R_{base})	63.04	72.54	83.30	89.45	73.17	81.00	92.21	4.24
+ Representation (R_{rep})	66.82	76.09	84.84	88.35	75.83	82.22	95.56	4.18
+ Behavior (R_{beh})	69.55	**76.19**	87.91	90.99	78.73	**83.59**	99.22	4.20
Ours	**69.90**	75.47	**89.23**	**91.65**	**79.57**	83.56	**100.28**	4.20

Table 2: Analysis of Training Strategies and Reward Components. Comparisons on the Phi-4-MM backbone using xFinder evaluator. WER reports the average Word Error Rate (↓) on Librispeech.

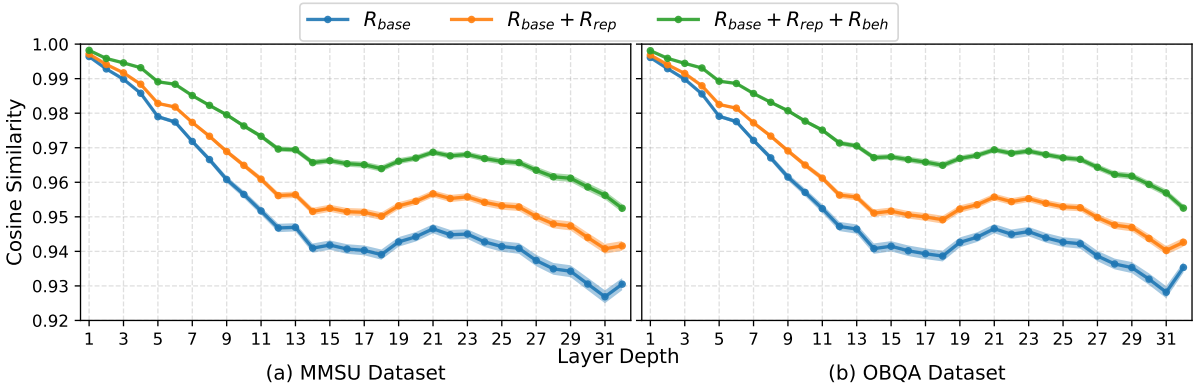

Figure 3: Layer-wise Representation Alignment Analysis. Shaded areas indicate 95% confidence intervals.

and audio-input branches, align the response span starting from the last <|assistant|> token, and compute the cosine similarity of hidden states at each token position for every layer. The reported curves are obtained by averaging similarities over tokens and samples, with 95% confidence intervals, on both MMSU and OBQA.

Under this setting, similarity naturally decreases with depth due to compounding transformations, where minor modality-specific differences accumulate as they propagate through the network. The key metric is not the downward trend, but the relative separation between methods across depth. Adding R_{rep} consistently lifts the similarity trajectory across layers, indicating that representation alignment reduces the representational drift. Furthermore, our joint strategy achieves the highest similarity, suggesting that behavior alignment acts as a complementary constraint that guides the speech branch toward semantically consistent reasoning paths. These internal observations align with the external improvements in MRR, supporting that our method performs genuine reasoning behavior transfer. The consistency of this effect across MMSU and OBQA further validates its robustness.

5 Conclusion

We introduced an on-policy trajectory alignment framework that mitigates representational drift and improves semantic consistency in speech reasoning. By combining dense representation and behavior alignment rewards under an asymmetric RL objective, our method substantially narrows the modality reasoning gap and achieves state-of-the-art performance on MMSU and OBQA among 7B-scale Speech LLMs.

Limitations

Despite its effectiveness, our trajectory alignment framework has several limitations. First, we evaluate alignment only at the 7B scale, and it remains unclear how the proposed reward design behaves for larger or more complex architectures. Second, our method focuses on single-turn reasoning, whereas multi-turn, interactive, or dialogue-driven speech reasoning may introduce additional dynamics not captured by our current formulation. Finally, while our alignment rewards mitigate modality drift, they still rely on text-only reference completions and may not fully account for paralinguistic cues, such as emotion, prosody, or intent, that do not have explicit textual counterparts.

References

Anonymous. 2025. Alignchat: Endowing LLMs with end-to-end speech-to-text chat capability through token-level representation alignment. In *Submitted to ICLR*. Under review.

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*.

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, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y Guo, and Irwin King. 2025. Recent advances in speech language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13943–13970.

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*.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

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

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. 2025. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4044–4054.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Pengcheng Li, Botao Zhao, Zuheng Kang, Junqing Peng, Xiaoyang Qu, Yayun He, and Jianzong Wang. 2025. Emo-rl: Emotion-rule-based reinforcement learning enhanced audio-language model for generalized speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18744–18754.

Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. 2025a. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*.

Yansong Liu, Jiateng Li, and Yuan Liu. 2025b. Enhancing speech large language models through reinforced behavior alignment. *arXiv preprint arXiv:2509.03526*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2025a. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Sung-Feng Huang, Chih-Kai Yang, Chee-En Yu, Chun-Wei Chen, Wei-Chih Chen, Chien-yu Huang, and 1 others. 2025b. Desta2. 5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *arXiv preprint arXiv:2507.02768*.

Pooneh Mousavi, Yingzhi Wang, Mirco Ravanelli, and Cem Subakan. 2025. Alas: Measuring latent speech-text alignment for spoken language understanding in multimodal llms. *arXiv preprint arXiv:2505.19937*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech*

and signal processing (ICASSP), pages 5206–5210. IEEE.

Jing Peng, Yucheng Wang, Bohan Li, Yiwei Guo, Hankun Wang, YanGui Fang, Yu Xi, Haoyu Li, Xu Li, Ke Zhang, Shuai Wang, and Kai Yu. 2025. A survey on speech large language models for understanding. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–32.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

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*.

Weiting Tan, Hirofumi Inaguma, Ning Dong, Paden D. Tomasello, and Xutai Ma. 2025. SSR: Alignment-aware modality connector for speech language models. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 56–75, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Enzhi Wang, Qicheng Li, Zhiyuan Tang, and Yuhang Jia. 2025. Cross-modal knowledge distillation for speech large language models. *arXiv preprint arXiv:2509.14930*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202, Suzhou, China. Association for Computational Linguistics.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jingran Xie, Xiang Li, Hui Wang, Yue Yu, Yang Xiang, Xixin Wu, and Zhiyong Wu. 2025. Enhancing generalization of speech large language models with multi-task behavior imitation and speech-text interleaving. *arXiv preprint arXiv:2505.18644*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Wenze Xu, Chun Wang, Jiazhen Yu, Sheng Chen, Liang Gao, and Weihong Deng. 2025c. Optimal transport regularization for speech text alignment in spoken language models. In *Asian Conference on Pattern Recognition*, pages 280–294. Springer.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qingchen Yu, Zifan Zheng, Shichao Song, Zhiyu li, Feiyu Xiong, Bo Tang, and Ding Chen. 2025a. xfinder: Large language models as automated evaluators for reliable evaluation. In *International Conference on Representation Learning*, volume 2025, pages 59850–59892.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025b. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Implementation Details

A.1 Format Reward Regular Expression

The format reward, R_{format}, is a binary reward assigned based on a regular expression match. The regular expression is:

```
^<think>.*?</think>\s*<answer>.*The
↪  answer is [ABCD][\.:,].*</answer>$
```

A.2 Hyperparameters

Table 3 lists the detailed hyperparameters used for the SFT and RL training stages across all experiments. We utilize LoRA for efficient fine-tuning to reduce computational overhead.

A.3 Prompt Templates

We employ a unified prompt format across all experiments to ensure consistency. The system instruction enforces a Chain-of-Thought reasoning structure. Table 4 illustrates the specific templates used for Audio and Text inputs. The content of `[QUESTION TEXT]` (and the spoken content of `<|audio_1|>`) follows the multiple-choice format:

Hyperparameter	Value
LoRA Configuration	
Rank (r)	8
Alpha (α)	32
Dropout	0.05
Target Modules	All Linear Layers
Training Configuration	
Learning Rate	1e-5 (Qwen), 2e-5 (Phi)
LR Scheduler	Cosine
Warmup Ratio	0.01
Num Epochs	3
Batch Size (Global)	64
Gradient Accumulation	4
Optimizer	AdamW
Weight Decay	0.01
Max Grad Norm	1.0
Precision	bfloat16
GRPO / DAPO Configuration	
Generations per Prompt (G)	8
Temperature	1.0
Max Completion Length	1024
Epsilon High (ϵ_{high})	0.28
KL Coefficient (β)	0.0

Table 3: Detailed hyperparameters for training.

```
{question}

Option A: {option_a}
Option B: {option_b}
Option C: {option_c}
Option D: {option_d}
```

Component	Content
System	<|system|>A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think>[THINKING PROCESS]</think><answer>The answer is [CHOICE].</answer><|end|>
User (Audio)	<|user|><|audio_1|><|end|>
User (Text)	<|user|>[QUESTION TEXT]<|end|>
Assistant	<|assistant|>

Table 4: Prompt templates used for training and inference. For text inputs, [QUESTION TEXT] replaces the <|audio_1|> token with the semantic equivalent content.