

Hear, Invoke, and Understand: A Skill-Calling Multimodal Agent for Large Audio Language Models

Yuwen Wang^{1,2*}, Tian-Hao Zhang^{2*}, Minghao Cai¹, Yilin Ren¹, Ziyang Jiang¹,
Xin Wang², Zhichao Wang², Zhou Pan², Kun Zhan², Xinyuan Qian¹

¹ University of Science and Technology Beijing, China

² Li Auto Inc., Beijing, China

Abstract

Complex acoustic problems may require models to perform acoustic operations, interact with external tools and reason over the resulting textual or processed-audio observations rather than answer directly from a fixed audio input. We study such problems as tool-interactive audio reasoning and develop *SpeechAgent-R*, an audio agent that coordinates its intrinsic multimodal understanding with external skills and tools. To support this capability, we construct *HIU-Corpus*, comprising 65,492 interaction trajectories and 507.6 hours of audio across 24 tasks, 8 skills and 9 tools. *SpeechAgent-R* first learns structured interaction behaviors through trajectory-based supervised fine-tuning and then improves its decisions through multi-turn reinforcement learning. We further introduce *HIU-Bench* to jointly evaluate task performance, interaction quality and generalization to diverse task settings. It contains 1,395 samples across 56 tasks, including in-distribution (ID) and out-of-distribution (OOD) splits with substantial shifts in tool usage and workflow composition. *SpeechAgent-R* achieves 84.17 on ID tasks and 70.94 on OOD tasks, improving over the base model under the same agent harness by 15.40 and 14.23 points. These results demonstrate that learning skill and tool coordination improves audio agents' ability to handle diverse task settings and adaptive tool interactions.

1 Introduction

Recent advances in large audio language models (LALMs) have improved their ability to understand diverse audio content (Chu, Xu, and Yang 2024; Tian, Zhang, and Zhang 2025). Meanwhile, benchmarks such as MMAU and MMAR (Sakshi, Tyagi, and Kumar 2024; Ma, Ma, and Zhu 2025) have introduced more challenging problems to evaluate these capabilities. However, two important limitations remain. First, existing LALMs mainly rely on internal inference and parametric knowledge acquired during training. When a task requires acoustic operations or reasoning procedures beyond their learned capabilities, they lack mechanisms to acquire such capabilities at inference time. Second, unlike text-based agents that can learn new skills from task instructions and tool descriptions, current multimodal models generally lack mechanisms to extend their capabilities through such instructions at inference time. This ability to acquire and apply new skills remains underexplored in audio understanding.

These limitations are particularly consequential in complex acoustic scenes, where task-relevant evidence must be

*These authors contributed equally.

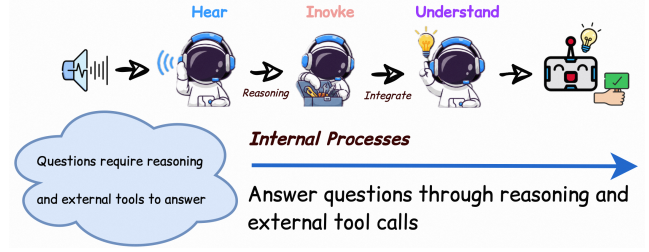


Figure 1: Overview of the *Hear-Invoke-Understand* framework. *SpeechAgent-R* hears the acoustic scene, invokes required skills through task reasoning, and understands the problem by integrating the returned evidence.

actively acquired and integrated (Sussman 2017). For example, recovering the speech of a target speaker from a spatial recording may require localization, speaker identification, speech extraction and semantic interpretation, with each operation depending on preceding results. Recent audio agents have explored external tools, but do not yet support this adaptive process: AudioToolAgent (Wijngaard, Formisano, and Dumontier 2025) uses a text-based LLM to coordinate audio models but lacks native audio reasoning; AuTAgent (Tong, Li, and Wang 2026) learns tool selection through reinforcement learning but does not compose tools based on intermediate outcomes; Echo (Wu, Zhang, and Yang 2026) enables audio re-listening but remains restricted to the original input; and EChO-Agent (Zhang, Zong, and Wang 2026) improves evidence organization within predefined interaction workflows. Consequently, existing approaches remain limited in their ability to adapt tool usage and compose workflows for previously unseen tasks and newly introduced skills. The central challenge is therefore to enable audio models to reason over multi-step interactions, coordinate diverse tools based on skill descriptions, and integrate tool-generated audio and textual observations into coherent decision making.

To address this challenge, we formulate complex audio problem solving as an iterative *Hear-Invoke-Understand* (HIU) process, as illustrated in Figure 1. Given an audio query, the model first *hears* the input audio with its native perception, *invokes* appropriate tools when needed and *understands* the resulting textual or audio observations before deciding the next action. Learning this process requires com-

plete execution trajectories, which existing audio datasets do not provide. We therefore construct *HIU-Corpus*, comprising 65,492 samples and 507.6 hours of audio across 24 tasks, 8 skills and 9 tools. Using this corpus, we develop *SpeechAgent-R*, a skill-calling multimodal agent trained with trajectory-based supervised fine-tuning (SFT) and multi-turn reinforcement learning (RL) to learn skill selection and tool coordination. We further introduce *HIU-Bench* to evaluate task performance, interaction quality and generalization across diverse task and skill settings. It contains 1,395 samples spanning 56 tasks and 26 tools, with ID and OOD subsets sharing only two tools and no annotated workflows. *SpeechAgent-R* achieves an overall score of 80.05, outperforming its base model under the same agent harness by 15.04 points and demonstrating the potential of LALMs to move beyond direct audio understanding toward interactive problem solving with external skills. Our contributions are:

- We formulate complex audio problem solving as an adaptive multi-step interaction in which audio agents combine external tools with their multimodal understanding.
- We construct *HIU-Corpus* and develop *SpeechAgent-R* with trajectory-based SFT and multi-turn RL, enabling the model to learn skill-based tool coordination.
- We introduce *HIU-Bench* to evaluate audio agents from task performance, interaction quality and generalization to unseen configurations.

2 Related Work

2.1 LALMs and Audio Reasoning Benchmarks

LALMs have evolved toward unified understanding of speech, environmental sounds and music. Models such as Qwen2-Audio, Kimi-Audio and Qwen3-Omni support general audio understanding and instruction following (Chu, Xu, and Yang 2024; KimiTeam, Ding, and Ju 2025; Team 2025), while Audio-Reasoner, Audio-Cogito and Step-Audio-R1 further improve complex reasoning through audio-grounded reasoning data and post-training (Xie, Lin, and Liu 2025; Li et al. 2026; Tian, Zhang, and Zhang 2025). These advances improve reasoning over fixed audio inputs, yet the models still operate on the original input alone, without acquiring additional evidence through acoustic operations or incorporating tool-generated observations into subsequent reasoning.

Audio benchmarks have accordingly expanded from general understanding to complex reasoning. AIR-Bench and AudioBench evaluate the understanding of speech, sounds and music (Yang, Xu, and Liu 2024; Wang, Zou, and Lin 2025), while MMAU, MMAR and MMAU-Pro introduce professional knowledge, multi-step reasoning, multiple audio inputs and spatial audio problems (Sakshi, Tyagi, and Kumar 2024; Ma, Ma, and Zhu 2025; Kumar, Sedláček, and Lokegaonkar 2026). MSU-Bench further focuses on speaker-centric understanding in multi-speaker conversations (Wang, Sun, and Lin 2025). Despite increasingly challenging tasks and more detailed reasoning evaluation, these benchmarks primarily evaluate models’ ability to reason over provided audio inputs. They do not evaluate whether models can select appropriate skills and tools, reason over intermediate observations and solve audio tasks through interaction.

2.2 Audio and Voice Agents

Recent studies have explored tool-based audio understanding from two directions. One line of work enhances audio reasoning through specialized tools, including model orchestration, tool routing, audio re-listening and evidence organization (Wijngaard, Formisano, and Dumontier 2025; Tong, Li, and Wang 2026; Chen, Chen, and Cai 2026; Wu, Zhang, and Yang 2026; Wang, Peng, and Li 2026; Rong, Li, and Yu 2025; Zhang, Zong, and Wang 2026). Although these methods improve audio reasoning, they treat capabilities in isolation rather than learning to dynamically coordinate acoustic skills based on intermediate observations. Another line of work develops general-purpose voice assistants that invoke external tools for digital tasks such as information retrieval, travel booking, and device control (Maben, Lakshmy, and Radhakrishnan 2025), accompanied by benchmarks that evaluate tool selection, dependent API calls, multi-turn interaction and robustness under realistic speech conditions (Jain, Shukla, and Rajeew 2025; Pahwa, Beedu, and Priye 2026; Lin, Chen, and Chen 2026). However, existing audio agents remain limited to isolated acoustic capabilities, while voice agents mainly treat speech as a command interface rather than audio evidence for reasoning. In contrast, *SpeechAgent-R* learns multi-step coordination of acoustic skills and tools for complex audio problem solving, while *HIU-Bench* evaluates generalization to unseen task settings.

3 Problem Formulation

An audio agent should go beyond answering from fixed acoustic inputs: it should determine when additional skills are needed, retrieve relevant skills, invoke tools and use returned information to guide subsequent actions. Given audio inputs $\mathcal{A} = \{A_1, A_2, \dots, A_M\}$ and a user query Q , the agent has access to a user-defined skill library $\mathcal{S} = \{s_1, s_2, \dots, s_N\}$. Each skill $s_i = (d_i, \mathcal{T}_i)$ consists of a natural-language description d_i for skill selection and a set of associated tools $\mathcal{T}_i = \{t_{i1}, t_{i2}, \dots, t_{iL_i}\}$. A tool processes input x with parameters ϕ and returns observation $o = t(x, \phi)$, which can contain textual information or processed audio. Rather than following a predefined workflow, the agent retrieves the corresponding skill document before invoking associated tools. At step k , its context is defined as $h_k = (Q, \mathcal{A}, \mathcal{D}, a_{<k}, o_{<k})$, where $\mathcal{D}$ denotes available skill information during interaction. Based on h_k , the agent either retrieves the selected skill document, invokes a tool via action a_k , or terminates the interaction by generating response y . The resulting interaction forms a decision trajectory $\tau = \{(h_k, a_k, o_k)\}_{k=1}^K$. Since tool observations are produced by external tools, the agent learns to optimize its actions and final response as:

$$(\hat{\tau}, \hat{y}) = \arg \max_{\tau, y} \left[\prod_{k=1}^K P_{\theta}(a_k | h_k) \right] P_{\theta}(y | h_{K+1}). \quad (1)$$

This formulation captures an adaptive *Hear-Invoke-Understand* process, where agents select skills and tools based on intermediate evidence rather than fixed workflows.

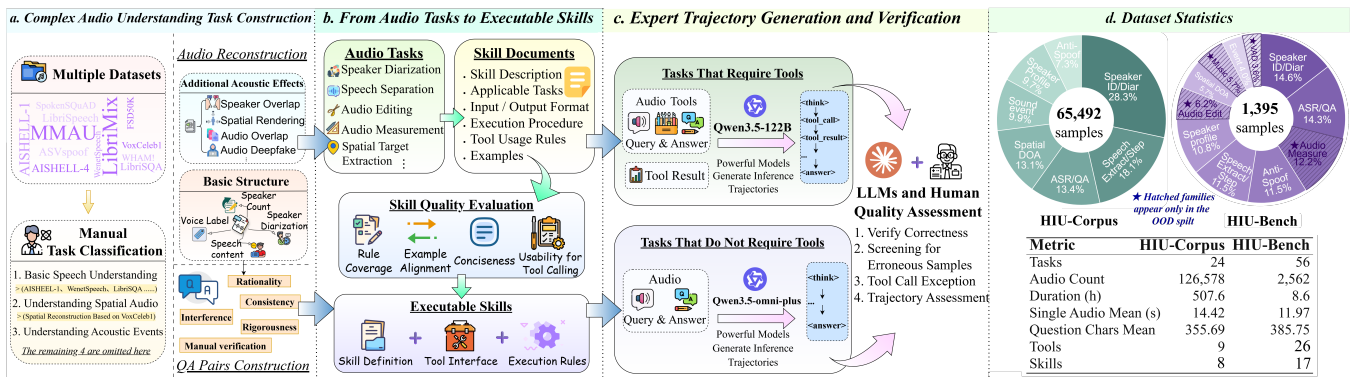


Figure 2: Construction pipeline of HIU-Corpus and HIU-Bench, covering task collection, skill construction, trajectory generation and verification, and dataset statistics.

4 Method

4.1 HIU-Corpus Construction

To provide diverse, high-quality trajectories for training generalizable audio agents, we follow three principles: tasks should i) require multi-step reasoning beyond direct audio understanding to capture meaningful decision processes; ii) combine multimodal understanding with external skills when needed and reason over their outputs rather than rely on tool execution alone; and iii) span diverse skills and workflows to encourage generalizable interaction beyond fixed workflows. Guided by these principles, we construct HIU-Corpus through a three-stage pipeline (Fig. 2).

Complex Audio Understanding Task Construction

Guided by these principles, we collect public datasets spanning speech recognition and audio QA (Zhang, Lv, and Guo 2021; Bu, Du, and Na 2017; Zhao, Jiang, and Liu 2023; Sakshi, Tyagi, and Kumar 2024; Li, Wu, and Liu 2018), multi-speaker and mixed speech (Fu, Cheng, and Lv 2021; Cosentino et al. 2020), speaker recognition and audio authenticity (Nagrani, Chung, and Zisserman 2017; Liu, Wang, and Sahidullah 2022) and acoustic event understanding (Fonseca, Favory, and Pons 2020). We select 24 tasks that involve complex acoustic understanding, support reliable QA construction from source annotations, and cover scenarios where external skills may be required. For each task, we manually design a template that maps source annotations to QA pairs and specifies the expected solution process, including whether external skills are needed. To increase acoustic complexity, we create additional acoustic scenes through audio mixing, temporal overlap and spatial rendering, while preserving task-relevant metadata such as speaker identities, timestamps and spatial positions. We instantiate the templates using the original or constructed audio and use the original annotations as reference answers. Finally, Claude-Opus-4.7 checks each QA pair against its template and annotations, with flagged samples manually corrected or removed.

From Audio Tasks to Executable Skills For tasks requiring external skills, related scenarios with similar problem-solving requirements are grouped into skills. Each skill repre-

sents a task-oriented capability abstraction that organizes related tools, usage conditions and interaction patterns. Within each skill, different scenarios may require different workflows, where one or more tools return text or processed audio for subsequent reasoning. We therefore create a skill document for each skill, describing its tools, interfaces and calling rules, while providing guidance on when and how to select or combine tools across scenarios. We then use Claude-Opus-4.7 to evaluate each document against real usage samples covering its tasks and major workflows, scoring interface accuracy, rule coverage, example alignment, conciseness and usability for tool calling. Based on these evaluations, we manually refine the documents over multiple rounds.

Trajectory Generation and Verification We construct

direct-answer and skill-assisted trajectories following the interaction format used for agent execution. For direct-answer tasks, Qwen3.5-omni-plus generates reasoning and responses from the audio, question and task prompt. For skill-assisted tasks, we execute tools specified by the assigned skill, record their calls and outputs, and provide the QA instance, skill document and interactions to Qwen3.5-122B for trajectory generation. Tool outputs may contain text or processed audio serve as observations for subsequent reasoning. To prevent incorrect tool feedback from affecting SFT, we correct inaccurate observations using verified annotations while preserving the original tool-response format. Claude-Opus-4.7 evaluates each trajectory for reasoning quality, decision quality and reasoning-answer consistency, with flagged samples manually corrected or removed. Across skill and trajectory evaluations, scores range from 4.32 to 4.80 out of 5. Within HIU-Corpus, 87.6% of trajectories contain at least one tool call, and 10.7% contain multiple calls. Among tool-assisted trajectories, sequences average 3.24 turns (median 3, maximum 10), covering diverse tool-use patterns beyond direct audio answering. HIU-Corpus is split into training and validation sets with a 95/5 ratio for training and model selection.

4.2 SpeechAgent-R Training

Trajectory-based SFT We initialize SpeechAgent-R from Qwen3-Omni-Thinking and perform SFT on verified in-

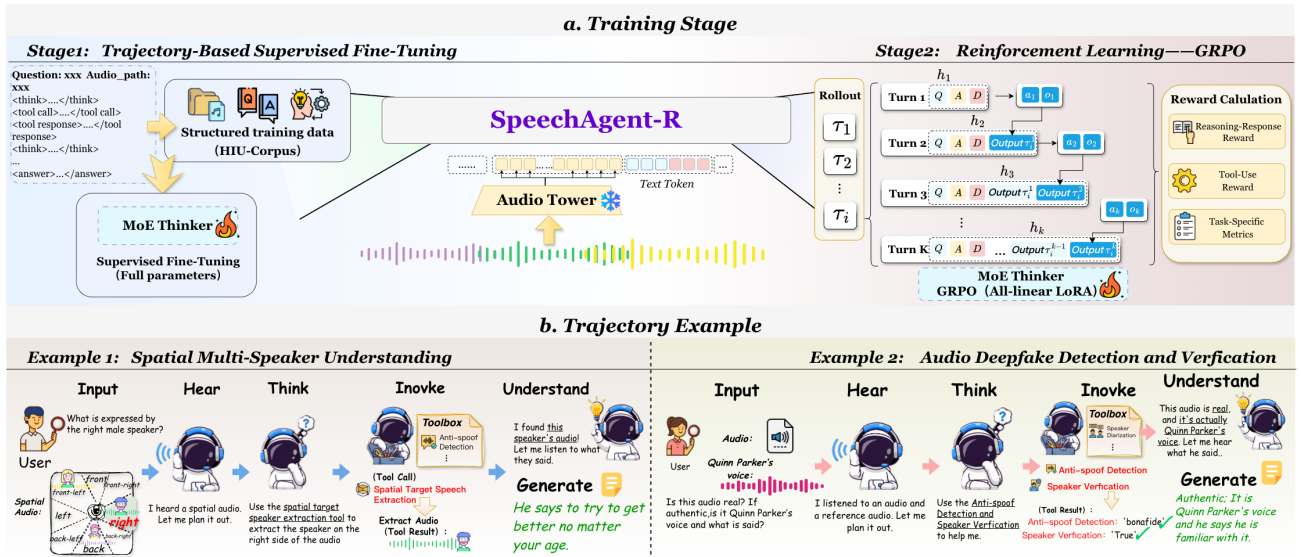


Figure 3: Training and interaction workflow of SpeechAgent-R. (a) Two-stage training with trajectory-based SFT and multi-turn GRPO. (b) Example trajectories showing skill invocation and tool-output integration for two different tasks.

teraction trajectories from HIU-Corpus to learn structured decision-making and tool-use behaviors. The sequence-modeling objective supervises only agent-generated tokens while masking user inputs and tool observations:

$$\mathcal{L}_{\text{SFT}} = - \sum_{t=1}^T m_t w_t \log P_{\theta}(y_t | x, y_{<t}) \quad (2)$$

where x denotes the input, $y = \{y_t\}_{t=1}^T$ the interaction sequence and P_{θ} the model probability. The mask m_t excludes user and tool-observation tokens, while w_t emphasizes agent actions such as skill selection and tool invocation.

GRPO-based RL We further optimize SpeechAgent-R with GRPO through multi-turn rollouts involving tool interaction. For each audio-query pair, the model samples a group of trajectories by selecting skills, invoking tools and reasoning over their outputs. We maximize the GRPO objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\ell_{i,t}^{\text{clip}} - \beta D_{\text{KL}}^{i,t} \right) \right], \quad (3)$$

$$\ell_{i,t}^{\text{clip}} = \min \left(\rho_{i,t} \hat{A}_i, \bar{\rho}_{i,t} \hat{A}_i \right).$$

where G is the group size, T_i is the number of model-generated tokens, $\rho_{i,t}$ is the current-to-old policy ratio, $\bar{\rho}_{i,t}$ is its ϵ -clipped value, $\hat{A}_i$ is the group-relative advantage and β is the KL weight. We define the composite reward:

$$R(\tau_i) = \lambda_f R_f(\tau_i) + \lambda_t R_t(\tau_i) + \lambda_a R_a(\tau_i). \quad (4)$$

R_f rewards valid reasoning-response structures and penalizes malformed or incomplete outputs. R_t starts from 1 and penalizes missing or incorrect tools, redundant calls and execution errors. Both scores are clipped to $[0, 1]$. R_a uses normalized task-specific metrics, including EM for short answers, $1 - \text{CER}$ for ASR, BERTScore for open-ended responses, a combined transcription and turn-overlap score

for diarization and path-set F1 for processed-audio outputs, with BLEU-4 and Token-F1 as fallback metrics. For the final model, we set $(\lambda_f, \lambda_t, \lambda_a) = (0.05, 0.25, 0.70)$. The composite rewards are then normalized within each group:

$$\hat{A}_i = \frac{R(\tau_i) - \mu_R}{\sigma_R} \quad (5)$$

where μ_R and σ_R are the group reward mean and standard deviation. Because the reward is trajectory-level, $\hat{A}_i$ is shared across all model-generated tokens.

5 HIU-Bench

5.1 Benchmark Construction and Statistics

HIU-Bench comprises an ID split that follows the same task, skill and tool distributions as HIU-Corpus while using independent audio instances, and an OOD split constructed through the same pipeline with new task scenarios and skill compositions. It contains 1,395 samples across 56 tasks, including 960 samples from 24 ID tasks and 435 samples from 32 OOD tasks. We check data independence between HIU-Corpus and HIU-Bench at the audio and task levels. File-level matching and MD5 content verification find no binary-identical audio overlap in model inputs, while OOD tasks are disjoint from HIU-Corpus at the task level. Speaker overlap in ID follows the splits of the source datasets, with no additional speaker leakage introduced during benchmark construction. Furthermore, we compare tool coverage and workflow composition between ID and OOD. Fig. 4(a) shows different tool-family coverage, with OOD introducing additional tool categories, while Fig. 4(b) shows 12 and 21 unique gold tool chains with no overlap. These differences enable HIU-Bench to evaluate tool coordination across diverse task scenarios and workflow compositions.

Model	In-domain				Out-of-domain				Overall	
	Tool	Answer	Total	Time	Tool	Answer	Total	Time	Total	Time
<i>(a) Direct setting</i>										
Kimi-Audio	–	–	30.44	2.04	–	–	44.96	5.32	34.97	3.06
Step-Audio-R1	–	–	37.99	11.16	–	–	24.86	11.39	33.90	11.23
MiDashengLM	–	–	38.07	1.93	–	–	45.91	1.24	40.51	1.71
Qwen2-Audio	–	–	37.78	1.07	–	–	39.65	0.73	38.36	0.96
Qwen2.5-Omni	–	–	37.46	13.72	–	–	48.82	14.71	41.01	14.03
Qwen3-Omni-Thinking	–	–	40.67	14.51	–	–	45.25	12.31	42.10	13.83
Gemini-3.1-Pro-Preview	–	–	42.76	8.85	–	–	52.78	7.96	45.89	8.57
Gemini-3-Flash-Preview	–	–	42.63	5.73	–	–	41.31	4.90	42.22	5.47
<i>(b) Agent harness setting</i>										
Kimi-Audio	57.89	21.39	34.06	58.84	22.09	12.48	16.05	77.25	26.60	64.58
Step-Audio-R1	79.97	49.26	58.90	25.32	32.78	15.86	22.14	24.82	46.24	25.16
MiDashengLM	70.11	35.88	48.39	15.16	26.18	25.75	27.35	17.75	42.30	15.97
Qwen2-Audio	71.25	44.66	53.17	1.50	31.95	32.31	33.10	0.79	47.18	1.28
Qwen2.5-Omni	72.00	44.89	52.34	8.05	26.92	25.62	27.73	9.92	44.66	8.63
Qwen3-Omni-Thinking	78.98	63.53	68.77	48.21	65.56	51.12	56.71	51.33	65.01	49.18
Gemini-3.1-Pro-Preview	92.28	62.83	69.33	16.23	52.51	60.89	58.02	10.71	65.81	14.51
Gemini-3-Flash-Preview	91.08	71.21	75.13	16.37	40.55	56.93	52.50	10.39	68.08	14.51
SpeechAgent-SFT (ours)	97.94	<u>70.74</u>	<u>78.92</u>	19.24	<u>78.51</u>	<u>60.56</u>	<u>66.98</u>	18.97	<u>75.20</u>	19.15
SpeechAgent-R (ours)	<u>96.74</u>	78.59	84.17	20.85	79.61	65.80	70.94	25.45	80.05	22.29

Table 1: Main results on HIU-Bench, reported separately on ID and OOD splits. Overall summarizes performance across both splits and all Time values are reported in seconds. Boldface and underlining denote the best and second-best results in each column, respectively. Gray shading highlights our models.

Setting	In-domain (ID)			Out-of-domain (OOD)			Overall
	Tool	Answer	Total	Tool	Answer	Total	
Base	–	–	40.67	–	–	45.25	42.10
+ Agent harness	78.98	63.53	68.77	65.56	51.12	56.71	65.01
+ SFT (answer-only)	71.25	67.26	69.89	58.85	56.52	59.22	66.57
+ SFT	97.94	<u>70.74</u>	<u>78.92</u>	<u>78.51</u>	<u>60.56</u>	<u>66.98</u>	<u>75.20</u>
+ SFT + RL (Ours)	<u>96.74</u>	78.59	84.17	79.61	65.80	70.94	80.05

Table 2: Ablation of the agent framework and training strategies on HIU-Bench. Best and second-best results are shown in bold and underlined, respectively, gray denotes our final model.

5.2 Evaluation Protocol

We define S_{total} on HIU-Bench from three evaluation dimensions: format correctness (S_f), tool interaction (S_t) and answer quality (S_a), weighted by 0.05/0.25/0.70, with their effects analyzed in Section 6.3. For direct-answer evaluation, only S_a is used. For agent evaluation, S_f and S_a follow the same criteria as R_f and R_a , while S_t is evaluated separately for ID and OOD tasks. For ID tasks, S_t follows the training-time scoring function, whereas OOD tasks adopt a workflow-aware rubric: expected tool usage receives full scores, valid alternatives partial scores, and invalid calls or failed executions limited scores. Overall performance is computed as the sample-weighted average of ID and OOD scores. The evaluation metrics follow the same task definitions as training rewards but are independently computed on held-out annotations, preventing direct reuse of training signals.

6 Experiments

6.1 Experimental Setup

SpeechAgent-R is initialized from Qwen3-Omni-Thinking and trained on 24 NVIDIA A100 GPUs using ms-swift in BF16. SFT fully fine-tunes the Thinker with a learning rate of 1×10^{-6} and a global batch size of 32. For GRPO, we apply LoRA to all Thinker linear layers with rank 16 and alpha 32. The training uses a learning rate of 5×10^{-7} , a group size of 4, $\epsilon = 0.2$, $\beta = 0.04$, and reward weights (0.05, 0.25, 0.70) for 150 steps. The step-80 checkpoint is selected based on validation performance, and all models are evaluated under the same inference budget and agent environment. We evaluate SpeechAgent-R on HIU-Bench against Kimi-Audio (KimiTeam, Ding, and Ju 2025), Step-Audio-R1 (Tian, Zhang, and Zhang 2025), MiDashengLM (Dinkel, Li, and Liu 2025), Qwen2-Audio (Chu, Xu, and Yang 2024), Qwen2.5-Omni (Team 2026), Gemini-3.1-Pro-Preview and Gemini-3-Flash-Preview in direct and agent settings using the same scoring protocol.

Tool	Metric	Base	SFT	Ours
Speaker diarization	$100 \cdot (1 - \text{DER})$	75.7	76.6	77.4
Speaker verification	Verified acc.	84.3	89.5	88.7
Anti-spoof detection	Label acc.	52.5	71.2	71.2
Sound event detector	Top-5 acc.	100.0	86.8	87.2
Speech DOA estimator	Acc. @ 10°	100.0	100.0	100.0
Target-speaker extraction	SI-SDRi	61.9	62.2	62.2
Spatial target-speech extraction	SI-SDRi	26.9	32.0	41.6
Speech separation	SI-SDRi	74.5	77.4	75.9
Overall		72.0	74.5	75.5

Table 3: Per-tool performance on the ID split of HIU-Bench. Each tool is evaluated using the metric specified in the second column, and Overall denotes the macro-average across all tools. Higher values indicate better performance for all metrics.

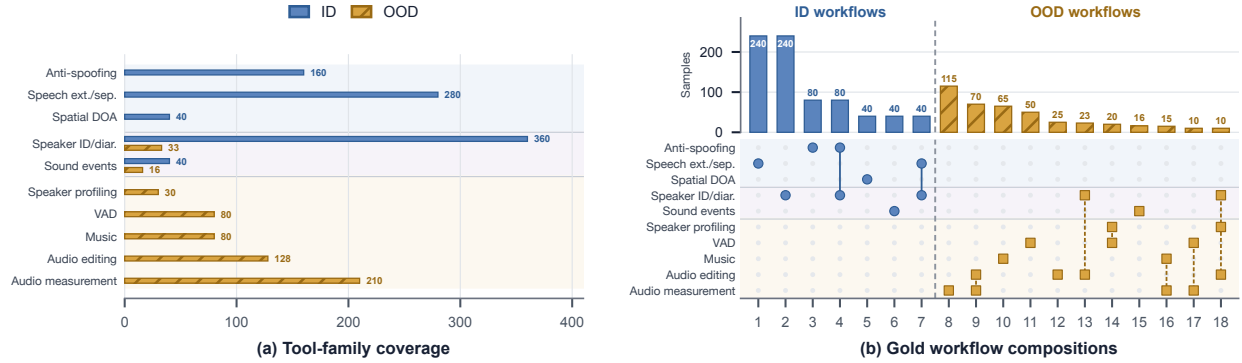


Figure 4: Comparison of tool coverage and workflow compositions between ID and OOD splits in HIU-Bench. (a) Tool-family distribution. (b) Gold workflow compositions and their overlap across splits.

6.2 Main Results on HIU-Bench

Table 1 presents the results under two settings. Under the direct setting, all models achieve limited performance, with overall scores ranging from 33.90 to 45.89. These limitations are particularly evident in complex audio scenarios, such as spatial audio understanding and overlapping-audio analysis, where models must extract and integrate task-relevant acoustic evidence. The shared agent harness improves most models, with Step-Audio-R1 rising from 33.90 to 46.24 and Gemini-3-Flash-Preview from 42.22 to 68.08. However, the improvements vary across models: access to the same skills and tools does not lead to effective tool use by itself. For example, Kimi-Audio achieves only 26.60 overall score, as unsuccessful tool interactions and format violations offset the benefits of tool access. Moreover, some models also incur additional inference latency due to longer interaction trajectories. These results highlight the need for training audio agents that can effectively utilize skills and tools and generalize such capabilities to unseen scenarios.

After SFT on verified interaction trajectories from HIU-Corpus, SpeechAgent-SFT improves the overall score from 65.01 to 75.20 under the same agent harness, with gains on ID and OOD splits. However, OOD analysis shows that SFT still faces tool-use challenges: 32.4% of failures come from incorrect tool selection or poor tool-result use, and 18.2% from ineffective interaction strategies. With GRPO optimization, SpeechAgent-R achieves the highest overall score of 80.05, outperforming the base model with agent

harness and SpeechAgent-SFT by 15.04 and 4.85 points. OOD error analysis shows that GRPO improves agent behavior beyond supervised trajectories: compared with SFT, SpeechAgent-R reduces incorrect tool selection failures by 32.8% and tool-result utilization failures by 16.3%, enabling more effective skill use and response generation.

6.3 Experiments Analysis

Ablation studies of different training strategies Table 2 analyzes the effects of training strategies. Answer-only SFT, which supervises only final responses, achieves 66.57 overall, providing only a small improvement over the agent baseline, whereas trajectory-based SFT reaches 75.20 with gains on both ID and OOD tasks. Building on SFT, GRPO further improves the score by 4.85 points to 80.05, showing that interaction trajectories establish agent behaviors and GRPO further refines skill selection and tool use. Reward weight ablations show that 0.05/0.25/0.70 weighting achieves the best overall score, while higher answer weights degrade tool-use performance. Increasing the answer weight to 0.8 achieves 79.99 but causes more tool-use failures, highlighting the need to balance task performance and interaction reliability.

Tool Interaction Analysis Table 3 reports tool-level performance on the ID test set. Across different training strategies, tool execution performance remains comparable, with average scores of 72.0, 74.5, and 75.5 for Base, SFT, and SpeechAgent-R. These results suggest that the gains mainly

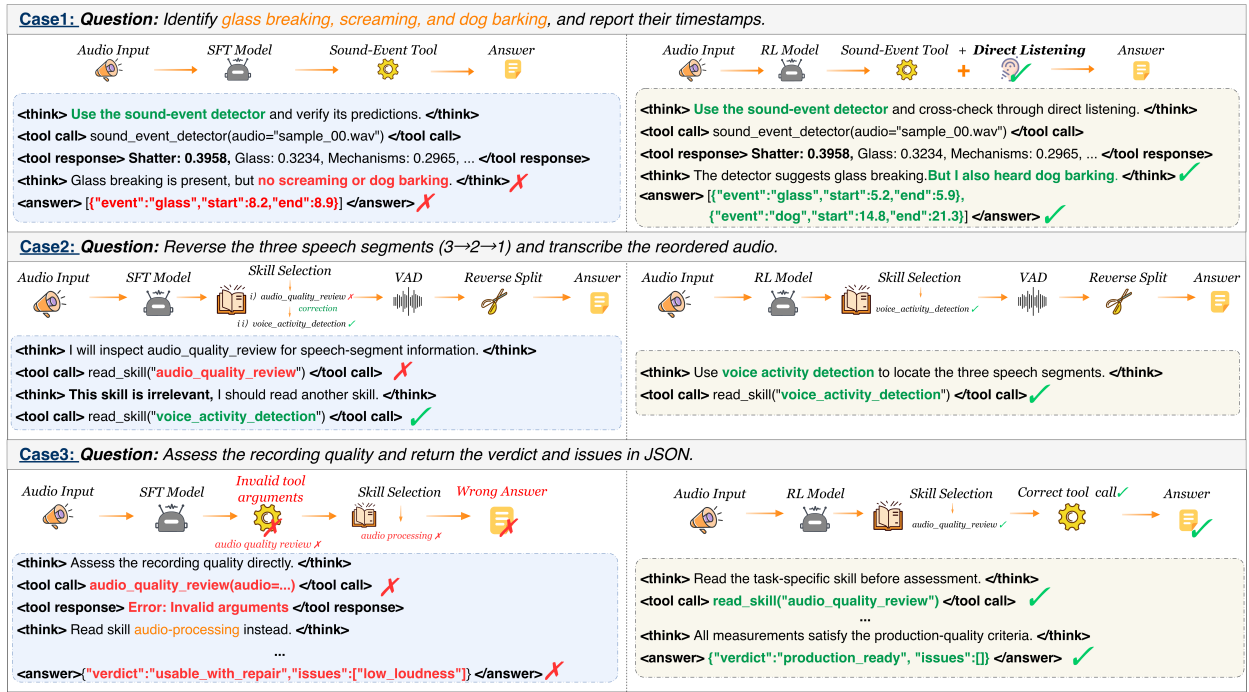


Figure 5: Representative cases comparing the trajectories generated by the SFT and RL models, illustrating differences in skill selection, tool usage and response generation.

Component	Normal	Oracle	Δ
Format	99.5	99.3	-0.2
Tool-call	96.7	97.9	+1.2
Answer	78.6	88.5	+9.9
Total	84.2	91.4	+7.2

Table 4: Oracle-tool analysis on the HIU-Bench ID split. Δ denotes Oracle minus Normal.

come from improved agent-side skill and tool utilization rather than changes in tool capabilities. To investigate the remaining performance bottleneck of SpeechAgent-R, we evaluate its upper bound under an oracle tool setting, where ground-truth tool outputs replace actual tool responses to eliminate tool execution errors. As shown in Table 4, the overall score improves from 84.2 to 91.4, mainly driven by the increase in Answer score from 78.6 to 88.5, while Tool-call remains nearly unchanged. These results show that, after reliable skill selection and tool invocation are achieved, further improvements mainly depend on tool output quality and the model’s multimodal understanding and reasoning ability.

Further Analysis Figure 5 presents three representative cases showing how SpeechAgent-R improves tool interaction beyond SFT. In Case 1, SpeechAgent-R improves tool-result utilization by integrating tool outputs with its own audio understanding, using multimodal evidence to refine predictions when tool responses are incomplete. In Case 2, SpeechAgent-R better understands task requirements and selects the appropriate skill, while SFT initially invokes an ir-

relevant skill and requires additional exploration. In Case 3, SpeechAgent-R further demonstrates skill-guided tool execution by understanding the relationship between skills and executable tools. It follows the skill requirements to invoke the necessary tools for the evaluation workflow, whereas SFT directly treats the skill name as an executable tool and produces an invalid tool call. Together, these cases show that SpeechAgent-R learns a reliable interaction process, covering skill selection, tool execution, and tool-result reasoning, which enables better generalization to complex audio tasks.

7 Conclusion

In this paper, we present SpeechAgent-R, a skill-calling multimodal agent that enables audio models to handle complex acoustic tasks through adaptive interaction. Through trajectory-based SFT on large-scale interaction data from HIU-Corpus and subsequent multi-turn RL, SpeechAgent-R learns to select skills, coordinate tools and use tool feedback for subsequent decisions beyond fixed-input audio understanding. We further introduce HIU-Bench to evaluate audio agents under both in-distribution and out-of-distribution settings. SpeechAgent-R achieves an overall score of 80.05 on HIU-Bench, improving over its base model under the same agent harness by 15.40 points and demonstrating improved skill and tool coordination across diverse task and workflow configurations. One limitation is that, although HIU-Bench covers diverse task settings, extending it to more complex and realistic audio agent scenarios remains an important direction for future work. We will release the code, models, and datasets upon publication to support future research.

References

- Bu, H.; Du, J.; and Na, X. 2017. AISHELL-1: An Open-Source Mandarin Speech Corpus and A Speech Recognition Baseline. *CoRR*, abs/1709.05522.
- Chen, L.; Chen, H.; and Cai, Y. 2026. AudioRouter: Data Efficient Audio Understanding via RL based Dual Reasoning. *CoRR*, abs/2602.10439.
- Chu, Y.; Xu, J.; and Yang, Q. 2024. Qwen2-Audio Technical Report. *CoRR*, abs/2407.10759.
- Cosentino, J.; Pariente, M.; Cornell, S.; Deleforge, A.; and Vincent, E. 2020. LibriMix: An Open-Source Dataset for Generalizable Speech Separation. *arXiv:2005.11262*.
- Dinkel, H.; Li, G.; and Liu, J. 2025. MiDashengLM: Efficient Audio Understanding with General Audio Captions. *CoRR*, abs/2508.03983.
- Fonseca, E.; Favory, X.; and Pons, J. 2020. FSD50K: an Open Dataset of Human-Labeled Sound Events. *CoRR*, abs/2010.00475.
- Fu, Y.; Cheng, L.; and Lv, S. 2021. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario. *CoRR*, abs/2104.03603.
- Jain, D.; Shukla, H.; and Rajeev, G. 2025. VoiceAgent-Bench: Are Voice Assistants ready for agentic tasks? *CoRR*, abs/2510.07978.
- KimiTeam; Ding, D.; and Ju, Z. 2025. Kimi-Audio Technical Report. *CoRR*, abs/2504.18425.
- Kumar, S.; Sedláček, S.; and Lokegaonkar, V. 2026. MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, 22688–22697. AAAI Press.
- Li, C.; Wu, S.; and Liu, C. 2018. Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension. *CoRR*, abs/1804.00320.
- Li, L.; Chen, H.; Li, Z.; Hu, Q.; Kang, J.; Li, J.; Xie, L.; and Li, Y. 2026. Audio-Cogito: Towards Deep Audio Reasoning in Large Audio Language Models. *arXiv:2604.12527*.
- Lin, G.; Chen, C.; and Chen, Z. 2026. Full-Duplex-Bench-v3: Benchmarking Tool Use for Full-Duplex Voice Agents Under Real-World Disfluency. *CoRR*, abs/2604.04847.
- Liu, X.; Wang, X.; and Sahidullah, M. 2022. ASVspoof 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *CoRR*, abs/2210.02437.
- Ma, Z.; Ma, Y.; and Zhu, Y. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. *CoRR*, abs/2505.13032.
- Maben, L. M.; Lakshmy, G. G.; and Radhakrishnan, S. 2025. AURA: Agent for Understanding, Reasoning, and Automated Tool Use in Voice-Driven Tasks. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, 1–4. IEEE.
- Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: A Large-Scale Speaker Identification Dataset. In Lacerda, F., ed., *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*, 2616–2620. ISCA.
- Pahwa, R.; Beedu, A.; and Priye, P. 2026. Audio2Tool: Speak, Call, Act - A Dataset for Benchmarking Speech Tool Use. *CoRR*, abs/2604.22821.
- Rong, Y.; Li, C.; and Yu, D. 2025. AudioGenie-Reasoner: A Training-Free Multi-Agent Framework for Coarse-to-Fine Audio Deep Reasoning. *CoRR*, abs/2509.16971.
- Sakshi, S.; Tyagi, U.; and Kumar, S. 2024. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. *CoRR*, abs/2410.19168.
- Sussman, E. S. 2017. Auditory Scene Analysis: An Attention Perspective. *Journal of Speech, Language, and Hearing Research*, 60(10): 2989–3000.
- Team, Q. 2025. Qwen3-Omni Technical Report. *CoRR*, abs/2509.17765.
- Team, Q. 2026. Qwen3.5-Omni Technical Report. *CoRR*, abs/2604.15804.
- Tian, F.; Zhang, X. T.; and Zhang, Y. 2025. Step-Audio-R1 Technical Report. *CoRR*, abs/2511.15848.
- Tong, S.; Li, X.; and Wang, Y. 2026. AuTAgent: A Reinforcement Learning Framework for Tool-Augmented Audio Reasoning. *CoRR*, abs/2602.13685.
- Wang, B.; Zou, X.; and Lin, G. 2025. AudioBench: A Universal Benchmark for Audio Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 4297–4316. Association for Computational Linguistics.
- Wang, S.; Sun, Z.; and Lin, Z. 2025. MSU-Bench: Towards Understanding the Conversational Multi-talker Scenarios. *CoRR*, abs/2508.08155.
- Wang, Y.; Peng, J.; and Li, H. 2026. Audio-Mind: An Auditable Agentic Framework for Audio Understanding. *CoRR*, abs/2605.28480.
- Wijngaard, G.; Formisano, E.; and Dumontier, M. 2025. AudioToolAgent: An Agentic Framework for Audio-Language Models. *CoRR*, abs/2510.02995.
- Wu, D.; Zhang, X.; and Yang, D. 2026. Echo: Towards Advanced Audio Comprehension via Audio-Interleaved Reasoning. *CoRR*, abs/2602.11909.
- Xie, Z.; Lin, M.; and Liu, Z. 2025. Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, 23829–23851. Association for Computational Linguistics.
- Yang, Q.; Xu, J.; and Liu, W. 2024. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), *ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 1979–1998. Association for Computational Linguistics.

Zhang, B.; Lv, H.; and Guo, P. 2021. WenetSpeech: A 10000+ Hours Multi-domain Mandarin Corpus for Speech Recognition. *CoRR*, abs/2110.03370.

Zhang, S.; Zong, J.; and Wang, J. 2026. EChO-Agent: Evidence Chain Orchestration Agent for Audio Reasoning. *CoRR*, abs/2606.15141.

Zhao, Z.; Jiang, Y.; and Liu, H. 2023. LibriSQA: Advancing Free-form and Open-ended Spoken Question Answering with a Novel Dataset and Framework. *CoRR*, abs/2308.10390.