HEAR Who Said What: Unlocking Speaker-Attributed Reasoning via Counterfactual Voice Grounding

Dongwook Lee[1] **Sangkwon Park**[2] **Eunwoo Song**[3] **Che Hyun Lee**[2]
Youngho Cho[1] **Junho Kim**[1] **June Young Yi**[4] **Heeseung Kim**[5†] **Sungroh Yoon**[1,2,6†]

[1]IPAI, Seoul National University (SNU) [2]Department of ECE, SNU
[3]Department of EE, Yonsei University [4]Department of CSE, SNU
[5]Department of AI, University of Seoul [6]AIIS, ASRI, INMC, and ISRC, SNU

Abstract

Speech Language Models (SLMs) are increasingly deployed in multi-speaker environments, yet their ability to attribute speech to the correct speaker and reason over speaker identities remains unclear. Hence, we introduce **HEAR**, a conceptually hierarchical benchmark diagnosing the foundational capabilities of speaker-attributed reasoning, comprising 2.4K human-verified samples from 887 diverse multi-party audio clips. Evaluating 20 leading SLMs on HEAR reveals they struggle with these foundational tasks, often relying on semantic priors rather than actual vocal cues. To address this, we present **A2R**, a 30B model optimized on *Counterfactual Audio with Speaker-level Hard negatives* (**CASH**), a dataset designed to guide the model to prioritize acoustic vocal cues over linguistic signals. A2R achieves strong performance on HEAR and exhibits zero-shot generalization to diverse multi-speaker downstream tasks, demonstrating that learned speaker attribution unlocks the model's latent capacity for speaker-aware reasoning. All resources are available on our project page.

1 Introduction

Understanding speech in real-world environments often requires handling multi-speaker scenarios, where multiple voices are present and interact dynamically (Lee et al., 2026). Identifying who said what, a task commonly referred to as speaker attribution (Kanda et al., 2021), is a fundamental prerequisite for any higher-level understanding or reasoning in multi-party scenarios. This challenge becomes critical as speech language models (SLMs) are increasingly deployed in multi-party environments (Nguyen et al., 2025). However, prior works (Kumar et al., 2025; Sakshi et al., 2024) often evaluate speaker attribution only through its downstream tasks, without directly testing the foundational abilities required to support it. This makes it difficult to localize where models actually fail: whether they struggle to discriminate between distinct voices, attribute utterances to the correct speakers, or simply lack the reasoning competency for the downstream task.

To address this gap, we introduce a benchmark, **HEAR**, designed to evaluate the foundational abilities underlying speaker-attributed reasoning in SLMs. HEAR is grounded in Erber's Auditory Hierarchy (Erber, 1982), which organizes human auditory understanding as a progression from perceptual discrimination to identification and higher-level comprehension. It maps Erber's view into three dimensions: **(i) Discrimination**, the ability to distinguish distinct voice identities; **(ii) Attribution**, the capacity to bind linguistic content to the exact voice identity; and **(iii) Reasoning**, the capacity to perform basic reasoning over speaker-attributed utterances. By mirroring this human cognitive progression, we provide a biologically inspired taxonomy to diagnose whether SLMs' failures stem from low-level acoustic perception or high-level semantic reasoning. The benchmark consists of 887 real-world audio clips and 2.4K human-verified multiple-choice question-answer pairs, covering diverse real-world acoustic settings such as daily conversations, meetings, and documentaries.

Our evaluations on the HEAR benchmark expose vulnerabilities in most leading SLMs. These models fail at basic attribution and exhibit an over-reliance on semantic priors, ignoring acoustic vocal cues when semantic context is available. To mitigate this semantic over-reliance, we introduce *Counterfactual Audio with Speaker-level Hard negatives* **(CASH)**, a 60K-scale corpus spanning all task dimensions of HEAR, designed to decouple acoustic identity from semantic content. By using voice cloning to swap speakers for specific utterances while preserving the original transcript

†Corresponding authors.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 56905–56932
October 24-29, 2026 ©2026 Association for Computational Linguistics

(Figure 2-(A)), CASH yields hard negatives where the correct answer flips solely based on vocal cues, thereby compelling models to ground their predictions in acoustic evidence rather than textual priors.

Building on this dataset, we present **A2R** (Attribution-to-Reasoning), a speech language model optimized from Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b) via Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to strengthen speaker attribution while preserving the base model's strong inherent reasoning ability. It is trained under a transcription-first objective: it first produces a speaker-tagged transcript and then reasons over it.

Consequently, A2R not only excels on HEAR but also demonstrates robust zero-shot transfer across various downstream multi-speaker benchmarks that require speaker attribution. These results suggest that speaker attribution is not merely an auxiliary sub-task, but a key mechanism for resolving the acoustic bottleneck in multi-party auditory comprehension. By explicitly forcing the model to ground utterances in vocal cues, we show that speaker-attributed reasoning can be elicited from the model in multi-speaker environments.

Our contributions can be summarized as follows: **(1)** We present **HEAR**, a benchmark comprising 2.4K human-annotated samples, designed to evaluate the foundational capabilities for speaker-attributed reasoning. **(2)** We reveal that current leading SLMs exhibit *semantic hallucination* in speaker-attributed reasoning, exploiting language priors rather than the voice cues that actually determine who said what. **(3)** We propose *Counterfactual Audio with Speaker-level Hard negatives* (**CASH**), a 60K-scale dataset designed to guide models to prioritize vocal cues over semantic content. **(4)** We introduce **A2R**, a 30B model trained with a transcription-first objective that rewards accurate transcriptions before reasoning, leading to strong improvements on HEAR and transfer gains on benchmarks requiring speaker attribution.

2 HEAR Benchmark

In this section, we introduce **HEAR**, a benchmark for Hierarchical Evaluation of Attribution and Reasoning, designed to evaluate the foundational capabilities underlying speaker-attributed reasoning in SLMs. The benchmark is formatted as Multiple-Choice Question Answering (MCQA) and organized into a hierarchy that mirrors human auditory processing (Erber, 1982), as illustrated in Figure 1: (i) Discrimination, (ii) Attribution, and (iii) Reasoning. To analyze the impact of overlapping speech, we distinguish cases based on whether the overlapping utterance segment is needed to answer the question (Appendix B.3).

2.1 Taxonomy

2.1.1 Discrimination

This axis evaluates the capacity to discriminate different voice identities and localize their temporal boundaries.

Voice Cardinality (VC) This task evaluates the model's ability to determine the total number of unique speakers present within a given audio clip. Representing the most fundamental level of speaker-identity clustering, VC assesses whether the model can accurately quantify the underlying voice population of an acoustic scene.

Voice Localization (VL) Given a reference audio snippet of a target speaker, the task is to identify the temporal intervals in which the target speaker is present or absent. By incorporating both active speech detection and absence verification, this task evaluates target-voice activity detection.

Voice Change Detection (VCD) This task requires the model to identify speaker transition points. For non-overlapping speech samples, the focus is on detecting turn-taking boundaries, whereas for overlapping samples, the task involves detecting time intervals containing simultaneous speech.

2.1.2 Attribution

This axis evaluates the ability to bind semantic content to its corresponding voice identity in both directions: text to audio and audio to text.

Content-to-Voice Attribution (CVA) Given a transcribed utterance as a textual query, the model must identify the corresponding speaker from a set of five candidate voices.

Voice-to-Content Attribution (VCA) In this task, the model is given a brief audio sample of a reference voice and must identify which of five textual candidates was spoken by that voice.

2.1.3 Reasoning

This axis examines basic *speaker-attributed reasoning* capabilities. Every Reasoning item is a semantic-hallucination pair: each original clip is

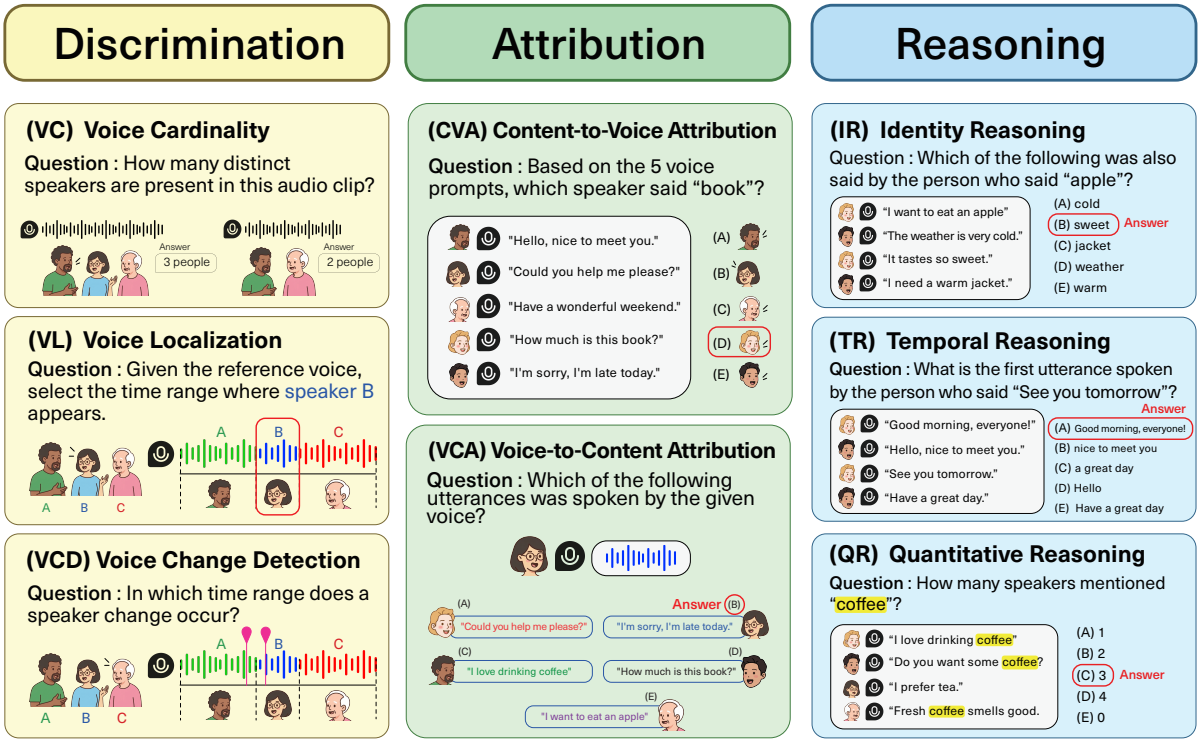


Figure 1: **Illustration of HEAR benchmark.** Inspired by Erber's auditory hierarchy (Erber, 1982), HEAR breaks speaker-attributed reasoning down into several foundational capabilities: discrimination, attribution, and reasoning.

paired with a counterfactual variant in which one answer-relevant utterance is re-spoken in another speaker's voice, leaving the transcript unchanged while flipping the correct answer, so that a voice-blind model necessarily fails the pair (Section 4).

Identity Reasoning (IR) It evaluates whether a model can determine which utterances were spoken by the same person. Given an anchor utterance (text), the model must identify another utterance (text) produced by the same speaker.

Temporal Reasoning (TR) This task assesses the model's ability to discern sequential relationships and temporal dynamics conditioned on speaker identity. We decompose this evaluation into two granular levels. (a) *Intra-speaker temporal ordering* requires the model to navigate a single speaker's history to spot the first or the last utterance of the speaker. (b) *Inter-speaker temporal alignment* evaluates the ability to align timelines across multiple participants by locating a target speaker's utterance relative to an anchor utterance from a different speaker (e.g., After someone said 'X', what was the first thing said by the person who said 'Y'?).

Quantitative Reasoning (QR) This task examines the model's capability to aggregate per-speaker utterance statistics to deduce numerical and comparative relationships. We evaluate this through two subtasks. (a) *Conditional counting* requires the model to quantify the number of distinct speakers who utter a specific word. (b) *Ordinal ranking* assesses the ability to rank speakers by cumulative speaking time and identify the speaker with the k-th longest total speaking time (e.g., "Which of the following was said by the speaker with the k-th longest total speaking time?").

2.2 Benchmark Construction

HEAR consists of approximately 2.4K MCQA queries derived from 887 audio clips, totaling roughly 23 hours of audio. The clips average ∼93 seconds in length, with a maximum duration of 150 seconds. We curate audio clips exclusively from the test splits of AMI (Carletta et al., 2005), ICSI (Janin et al., 2003), and VoxMM (Kwak et al., 2024), where speaker diarization and transcripts are already annotated by humans, ensuring reliable speaker-speech alignments across various real-world acoustic conditions such as meetings, daily conversations, broadcast media, and lectures. Each clip contains 4.05 speakers on average, and all samples are annotated by human annotators as overlap or non-overlap, based on whether answering the

question requires focusing on regions with overlapping speech (Appendix B.3). We generate queries through a rule-based pipeline and conduct an additional manual verification with eight annotators. As a result, every item in HEAR benefits from two stages of human validation: first during the construction of the source datasets, and again through our additional verification process, yielding a high-quality dataset with strong annotation reliability.

3 Training

3.1 Motivation

Unimodal collapse, where a model over-relies on a single modality, has long been a recognized issue in vision-language tasks (Koishigarina et al., 2026), and a similar phenomenon is observed in the speech domain. A recent study (Lee et al., 2026) highlights that SLMs often exhibit *semantic hallucinations*, bypassing acoustic cues in favor of their textual priors, which is a critical bottleneck for speaker-attributed reasoning in multi-party audio scenes. Determining "who said what" based on semantic plausibility may suggest who *might* speak, but fails in dynamic conversations where any participant can make any statement, thus requiring acoustic vocal cues to be prioritized over textual context to determine who *actually* spoke.

3.2 Forcing Models to Rely on Vocal Cues

To encourage models to ground speaker attribution in acoustic vocal cues rather than semantic content, we introduce *Counterfactual Audio with Speaker-level Hard negatives* (**CASH**). CASH is a 60K-scale dataset for multi-party audio, spanning all task dimensions of HEAR. Given an original multi-speaker clip, we construct counterfactual variants by replacing a target utterance with a voice-converted version of the *same transcript* by a different speaker, while preserving chronological structure and other speech content (Figure 2-(A)). We build on the VoxMM (Kwak et al., 2024) train split, extracting ∼5.2K real-world audio clips drawn from diverse audio scenes. All synthetic voices are generated with VoxCPM2 (Zhou et al., 2026), yielding ∼20K *mixed real-synthetic* audio clips and ∼60K queries in total. Details are provided in Appendix A.

3.3 Rewarding Attribution Before Reasoning

We present **A2R** (Attribution-to-Reasoning), a model developed by optimizing the baseline model via GRPO to enhance speaker attribution capabilities while preserving general performance. By bridging newly acquired vocal grounding ability with its inherent inferential capacity, we unlock the model's ability to perform speaker-attributed reasoning. During training, as illustrated in Figure 2-(C), the model first transcribes the input audio with speaker tags and then generates both the reasoning and the final answer based on its transcript.

We compute the reward from three aspects of the model response: the speaker-tagged transcription (R_{tr}), the final answer (R_{ans}), and the response format (R_{fmt}). The speaker-tagged transcription not only captures how accurately the model attributes utterances to speakers, but also serves as the basis for the reasoning needed to answer the question, whereas the final answer captures direct task-level correctness. Accordingly, for a model response Y_i, we parse its predicted speaker-tagged transcription and answer as $\hat{\tau}_i$ and $\hat{a}_i$, and compare them against the ground-truth transcription $\tau^\star$ and answer $a^\star$. The overall reward is then defined as

$$
\begin{aligned}
&R(Y_i;\tau^\star,a^\star)\\
&\quad= R_{\text{tr}}(\hat{\tau}_i;\tau^\star)+\lambda_{\text{ans}}R_{\text{ans}}(\hat{a}_i;a^\star)+\lambda_{\text{fmt}}R_{\text{fmt}}(Y_i),\\
&R_{\text{tr}}(\hat{\tau}_i;\tau^\star)\\
&\quad= \lambda_{\text{cp}}R_{\text{cp}}(\hat{\tau}_i;\tau^\star)+\lambda_{\text{cnt}}R_{\text{cnt}}(\hat{\tau}_i;\tau^\star)+\lambda_{\text{ord}}R_{\text{ord}}(\hat{\tau}_i;\tau^\star)
\end{aligned}
$$

Here, the transcription reward R_{tr} is further decomposed into three terms: a cpWER-based (von Neumann et al., 2024) attribution reward (R_{cp}), which measures whether the model correctly captures what was said by whom; a speaker-count reward (R_{cnt}), which evaluates whether the predicted transcription contains the correct number of speakers; and a speaker-order reward (R_{ord}), which assesses whether the temporal sequence of speaker turns is preserved (Appendix E).

During training, we sample G responses per input x, compute rewards, and normalize them as group-relative advantages $A_i = (R_i - \mu_R)/(\sigma_R + \delta)$, where μ_R and σ_R are the group reward mean and standard deviation. We then optimize the current policy π_θ using the objective:

$$
\begin{aligned}
\mathcal{L}_{\text{GRPO}} &= -\frac{1}{G}\sum_{i=1}^{G}\min(r_iA_i,\bar{r}_iA_i)\,,\\
\bar{r}_i &= \text{clip}(r_i,1-\epsilon_{\text{low}},1+\epsilon_{\text{high}}),
\end{aligned}
$$

where $r_i = \pi_\theta(Y_i \mid x)/\pi_{\theta_{\text{old}}}(Y_i \mid x)$, with $\pi_{\theta_{\text{old}}}$ denoting the old policy.

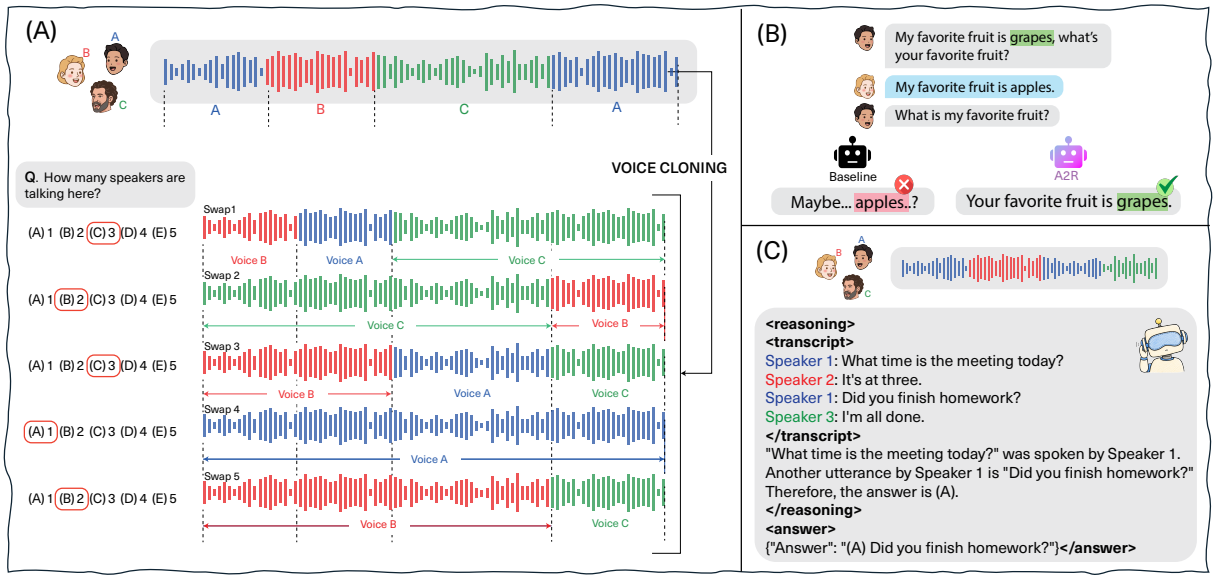


Figure 2: **Overview of our dataset and method on speaker-attributed reasoning.** **(A)** Our dataset, CASH, randomly reassigns utterances to different speakers via voice cloning while preserving the whole utterance, yielding counterfactual variants whose ground-truth answer always flips. **(B)** We show an elementary speaker-attribution task from the *What Do You Like?* benchmark (Wu et al., 2024), where the query is unsolvable without identifying which speaker said what. **(C)** Our model first transcribes the multi-party audio scene, then reasons over the transcript for more reliable speaker-aware reasoning.

4 Experiments

We design our experiments to answer the following three questions: (i) How do current leading SLMs perform on HEAR, and do they truly exploit acoustic speaker cues rather than relying on semantic priors for speaker-attributed reasoning? (ii) Do the resulting gains transfer to other multi-party audio understanding benchmarks where speaker attribution is essential, while preserving performance in dyadic conversational settings? (iii) How much does CASH-60K improve speaker-attributed reasoning, and what are the contributions of speaker-level hard negatives and training method?

Baseline Models. We evaluate 20 leading SLMs, covering both open-source and proprietary models. We assess (a) omni-modality models, (b) speech/audio language models, and (c) proprietary models as closed-source baselines. The evaluated models and their inference hyperparameters are summarized in Appendix D and Table 7. Among these models, we choose Qwen3-Omni-30B-A3B-Instruct as our base model for developing speaker-attributed reasoning capabilities, as its demonstrated zero-shot capability on HEAR indicates a promising starting point (Table 1). We train LoRA adapters (Hu et al., 2021) with $r = 64$ on NVIDIA H200 GPUs, while freezing the audio encoders and modality aligners. Details are provided in Appendix F.

Probing Semantic Hallucination. As introduced in Section 2.1.3, each "Reasoning" pair consists of (x, x'), where x' is a CASH-style voice-swapped variant of audio clip x. Pair construction enforces five constraints simultaneously: (1) the transcript and temporal structure of x' are preserved relative to x; (2) the speaker similarity (Desplanques et al., 2020) between the swapped target utterance and its reference voice exceeds 0.7; (3) the WER between the Whisper-large-v3 (Radford et al., 2022) transcription of the synthesized utterance and the reference transcript is 0, ensuring linguistic content is preserved; (4) the answer is required to be different between x and x', so that any model that ignores acoustic identity would necessarily answer identically on both, thereby failing the question; and (5) (x, x') pairs are retained only if human annotators verify that the synthesized utterance matches the intended target speaker. Target utterances are synthesized with IndexTTS 2 (Zhou et al., 2025), a different synthesizer from the one employed in CASH to avoid synthesizer-specific artifacts. Each pair is independently checked by expert annotators for acoustic naturalness, transcript fidelity, and answer-flip validity; pairs failing any criterion are discarded (Appendix B).

Model	Size	Discrimination			Attribution		Reasoning			Avg. (%)
		VC	VCD	VL	CVA	VCA	IR	QR	TR	
Proprietary Models										
Gemini-3.1-Pro-preview	-	**68.4**	**89.5**	**93.2**	**94.1**	**86.4**	**96.2 (47.6)**	**89.3 (53.6)**	**91.3 (51.3)**	**88.8 (75.8)**
Gemini-3-Flash-preview	-	52.6	66.0	81.4	73.9	60.1	91.9 (40.5)	82.5 (39.3)	80.9 (38.3)	73.1 (58.5)
GPT-4o-audio-preview	-	40.5	21.0	28.3	60.3	43.5	80.0 (15.7)	53.6 (15.7)	69.6 (17.4)	48.5 (32.8)
Open-Source Models										
Qwen3-Omni-A3B-Thinking	30B	34.2	33.0	31.2	45.0	34.6	71.4 (17.3)	58.2 (27.5)	65.2 (12.2)	45.0 (31.4)
Step-Audio-R1	32B	29.5	22.5	23.6	41.0	34.6	71.4 (5.4)	66.8 (22.1)	67.8 (19.1)	43.2 (26.5)
Qwen3-Omni-A3B-Instruct	30B	31.1	34.0	39.2	26.1	28.9	59.5 (8.6)	28.6 (5.7)	57.4 (15.7)	35.4 (24.1)
MiniCPM-o-4.5	9B	17.4	22.5	34.2	35.5	29.2	59.5 (8.6)	52.5 (15.4)	47.8 (10.4)	36.8 (23.5)
Fun-Audio-Chat	8B	55.3	20.0	25.3	32.2	21.9	53.0 (5.9)	31.1 (6.8)	45.2 (7.8)	33.4 (22.5)
Kimi-Audio-Instruct	7B	16.8	18.0	21.1	36.8	26.9	49.2 (8.1)	32.5 (2.1)	53.9 (13.9)	30.6 (19.2)
Voxtral-Small	24B	20.0	19.0	30.8	26.1	22.3	46.5 (3.2)	36.4 (4.3)	33.9 (0.0)	28.8 (17.3)
Qwen2.5-Omni	7B	12.1	20.0	31.6	25.4	24.6	48.6 (2.7)	35.4 (1.4)	34.8 (2.6)	28.6 (16.6)
Voxtral-Mini	3B	30.5	20.0	17.7	24.8	25.6	33.0 (1.1)	28.6 (1.4)	34.8 (1.7)	26.1 (16.6)
MiDashengLM	7B	34.7	19.0	26.6	19.5	20.3	24.3 (0.5)	33.2 (4.3)	26.1 (0.9)	25.1 (16.6)
Step-Audio 2 mini	8B	13.2	20.0	24.1	23.1	19.9	35.1 (3.2)	28.9 (6.4)	21.7 (2.6)	23.4 (15.4)
Audio-Flamingo-3	7B	30.5	19.5	19.8	11.7	21.6	28.6 (3.8)	24.3 (6.1)	29.6 (2.6)	22.0 (15.0)
Gemma-4-E4B-it	4B	11.6	19.5	24.1	22.5	20.3	45.4 (2.2)	24.3 (5.4)	28.7 (3.5)	23.9 (14.9)
Qwen2.5-Omni	3B	10.5	21.5	22.8	26.1	21.6	35.1 (1.6)	24.3 (1.4)	33.0 (0.9)	23.9 (14.9)
Phi-4-MM	6B	21.6	20.0	21.1	19.5	20.3	28.6 (1.1)	31.1 (2.5)	30.4 (0.9)	23.5 (14.4)
MiniCPM-o-2.6	8B	8.4	16.0	21.9	14.3	18.3	52.4 (3.8)	41.8 (9.6)	40.9 (2.6)	25.3 (13.0)
Qwen2-Audio	7B	7.9	18.5	14.8	15.6	23.9	31.4 (0.5)	46.8 (2.5)	20.0 (0.0)	23.1 (11.8)
Random Choice	-	*20.0*	*20.0*	*20.0*	*20.0*	*20.0*	*20.0 (4.0)*	*20.0 (4.0)*	*20.0 (4.0)*	*20.0 (14.9)*
A2R (Ours)	30B	**57.5**	**38.7**	**48.4**	**80.1**	**70.1**	**91.6 (70.8)**	**68.9 (47.4)**	**84.9 (66.9)**	**67.1 (60.5)**

Table 1: **Evaluation results of 20 leading Speech Language Models on the HEAR benchmark.** Values in parentheses in the Reasoning columns report scores under the paired setting, where a prediction is counted as correct only when both the original audio clip and its corresponding semantic hallucination counterpart are correct. A2R results are averaged over three seeds.

Zero-Shot Transfer and Dyadic Evaluation. To evaluate zero-shot transfer beyond HEAR, we consider three held-out benchmarks that require speaker attribution under diverse task settings.

The *What Do You Like? (WDYL)* (Wu et al., 2024) tests whether a model can associate spoken content with the correct speaker. We focus on its Identity-Critical Questions (ICQ) category, where two speakers describe their preferences and the model must determine whose preference is queried (Figure 2-(B)). We synthesize the last query sentence using a reference voice identical to one of the speakers. The model must therefore identify the queried speaker from vocal characteristics rather than relying on semantic content alone.

We additionally evaluate on *Gaokao* (Hu et al., 2024), a multi-turn dialogue QA benchmark derived from English listening examinations for the Chinese college entrance exam. To introduce an explicit speaker-attribution requirement, we synthesize the final query using the voice of a randomly selected character from the preceding dialogue, such that answering correctly requires associating the query voice with that character's earlier utterances. We further filter the examples using GPT-5.5 (OpenAI, 2026a) to retain only questions that cannot be solved without speaker attribution.

We also utilize the *Find the Spy (FTS)*, social deduction game (Xu et al., 2025c). In this task, 3-6 players introduce themselves and describe their assigned word, while one spy receives a different word. The players take turns giving clues about their word, and the group (model) must vote to identify the spy (Figure 3). Solving the task requires speaker attribution to track which player provided each description, as well as higher-level reasoning to identify the spy whose description is semantically inconsistent with the others.

Apart from the main audio already provided in the original benchmark, the text query segments are synthesized using a TTS model (Zhou et al., 2025), with speaker similarity constrained to be over 0.7 (Appendix C).

Lastly, in order to assess whether our model preserves dyadic spoken-language ability, we further evaluate on VoiceBench (Chen et al., 2024), which covers eight benchmarks in dyadic conversational settings (Appendix G).

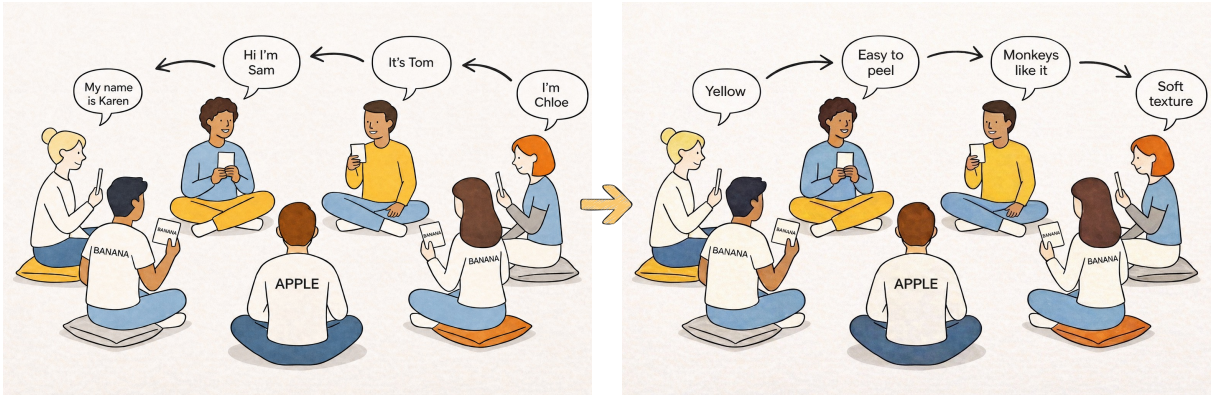


Figure 3: **Illustration of Find the Spy, a benchmark where speaker attribution is essential.** Each player receives a word (e.g., *Banana*), except for one spy who receives a different word (e.g., *Apple*), and players describe their words in a random speaking order. To identify the spy, a model must do more than recognize which clue is inconsistent with the others: it must also determine *who produced that clue*.

5 Results & Analysis

Limitations of Leading SLMs on HEAR. Across discrimination and attribution dimensions, most open-source models perform close to random chance, suggesting that current speech-language training does not reliably endow models with the basic capacity to segregate voices and localize them. As illustrated in Figure 4, this difficulty is exacerbated in scenarios with overlapping speech; performance consistently degrades across six leading baselines. While several recent open-source models achieve modest improvements over chance, the closed-source Gemini series performs substantially better, indicating that the core skills evaluated by HEAR are attainable but remain underdeveloped in current open-source SLMs. Notably, A2R (Ours) largely mitigates this limitation, bringing performance substantially similar to Gemini Flash and significantly narrowing the gap between open- and closed-source models.

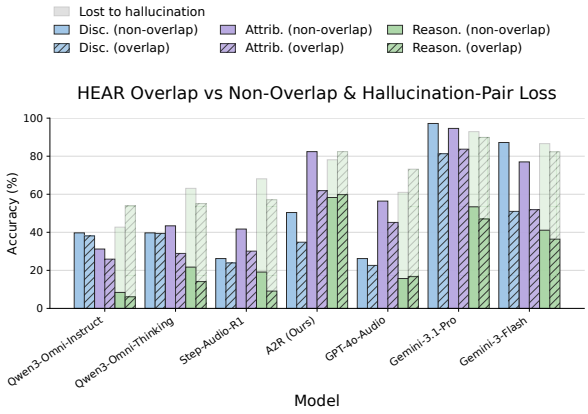


Figure 4: **Performance degradation under overlap samples and the effect of semantic hallucination.**

Overestimated Reasoning Performance Due to Linguistic Priors. Open-source models achieve strong reasoning scores, yet their performance drops sharply under the paired accuracy metric reported in parentheses in Table 1. The parenthesized score requires a model to answer both an original clip and its semantic-hallucination counterpart correctly (Appendix H). The same pattern also appears in proprietary models, including the Gemini models, suggesting that their reasoning also partly relies on semantic cues rather than voice-grounded inference.

Transfers to Unseen Attribution-Critical Tasks. As shown in Table 2, even on elementary tasks such as *What Do You Like?*, the Qwen3 series collapses to near-random performance, reflecting its lack of speaker attribution ability. In contrast, A2R shows substantial gains of +36.4 (WDYL), +22.6 (GAOKAO), and +19.0 (FTS) percentage points. These results suggest that learning speaker attribution provides a transferable inductive bias, enabling the model to generalize speaker-attributed reasoning to unseen tasks without task-specific training.

Model	WDYL	GAOKAO	FTS
Random Choice	50.0	50.0	23.8
Qwen3-Omni-30B-A3B-Instruct	61.3	65.6	29.2
Qwen3-Omni-30B-A3B-Thinking	67.1	54.8	17.2
A2R (CASH w/o HN)	95.0	78.5	40.5
A2R (CASH)	**97.7**	**88.2**	**48.2**

Table 2: **Generalization across three benchmarks where speaker attribution is essential.** "HN" denotes voice-swapped hard negative variants in CASH.

Transfers to Real-World Human Voices. To examine whether A2R's gains transfer to real-world human voices, we randomly sample 100 conversations used in the unseen-distribution evaluation above (50 WDYL, 30 GAOKAO, and 20 FTS) and collect corresponding human recordings for each instance. Specifically, five participants recorded the same conversational transcripts underlying the original synthesized inputs, with speakers randomly assigned to conversational roles. As shown in Table 3, A2R's substantial improvements over the base model persist when synthesized speech is replaced by human recordings: on WDYL, GAOKAO, and FTS, A2R improves over the base model by +48.0, +20.0, and +30.0 percentage points, respectively. These results provide evidence that the attribution capability learned from CASH transfers to real-world human voices beyond synthesized speech.

Model	Audio	WDYL	GAOKAO	FTS
Random Choice	-	50.0	50.0	23.8
Qwen3-Omni-30B-A3B-Instruct	Synthetic	50.0	76.7	50.0
A2R (CASH)	Synthetic	**96.0**	**93.3**	**65.0**
Qwen3-Omni-30B-A3B-Instruct	Human	50.0	63.3	40.0
A2R (CASH)	Human	**98.0**	**83.3**	**70.0**

Table 3: **Evaluation on human-recording attribution-critical tasks.** We compare the 100 synthetic samples with corresponding recordings produced by human speakers.

Ablation Study. To isolate the contribution of each component, we conduct ablations along three axes: (i) the inclusion of speaker-level hard negatives in CASH; (ii) reward shaping, where the model is prompted to reason but is not required to follow our structured reasoning trace, with rewards assigned only to the final answer; and (iii) the training regime, where the baseline model is trained with LoRA-based supervised fine-tuning (SFT). As shown in Figure 5, removing the hard negatives from CASH substantially reduces HEAR reasoning pair accuracy from 60.0 to 43.0 (for one seed) and also leads to a higher same-answer rate. These results suggest that the HEAR-axis data itself promotes acoustic sensitivity, while the hard-negative structure further strengthens this capability by encouraging the model to rely on speaker-specific voice cues. Importantly, the benefits of this structure extend beyond HEAR: as shown in Table 2, incorporating hard negatives also improves performance on zero-shot benchmarks. For the second axis, removing the structured reasoning-trace constraint and rewarding only the final answer similarly degrades overall HEAR performance and pair consistency, as shown in Figure 6. Finally, replacing our training objective with LoRA-based SFT yields weaker results not only on HEAR but also across most VoiceBench settings, as shown in Table 4. All ablation models are trained for one epoch and evaluated after validation performance has stabilized.

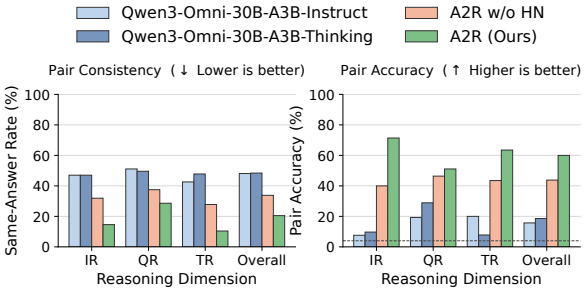


Figure 5: **Results on semantic hallucination pairs in the "Reasoning" category of HEAR.** "HN" denotes voice-swapped hard negative variants in CASH. Lower same-answer rates indicate higher acoustic sensitivity, as each pair is designed to elicit different answers.

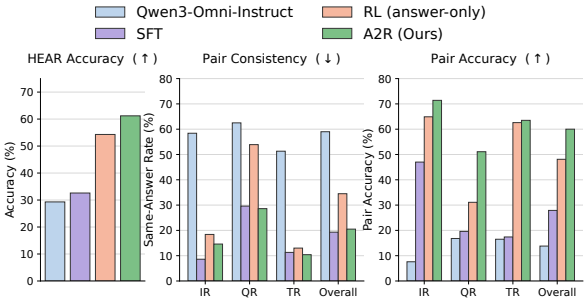


Figure 6: **Ablation of training regimes on HEAR.**

Minimal Trade-off on General Dyadic Tasks. As shown in Table 4, A2R achieves performance broadly on par with the baseline across eight general dyadic benchmarks, demonstrating that speaker-attributed reasoning can be acquired with only a marginal impact on general capability.

Datasets	Model	Performance
AlpacaEval | *BBH* | *AdvBench* | *CommonEval*	Baseline	4.38 | 94.00 | 99.42 | 4.07
	A2R	4.31 | 93.90 | 96.35 | 3.75
	Ablation (SFT)	3.19 | 68.10 | 89.62 | 1.90
WildVoice | *IFEval* | *MMSU* | *OpenBookQA*	Baseline	4.15 | 87.97 | 80.84 | 95.38
	A2R	3.93 | 86.01 | 76.87 | 92.53
	Ablation (SFT)	2.14 | 78.05 | 75.73 | 90.33

Table 4: **Evaluation on VoiceBench.** We assess performance drop under dyadic conversational settings.

Prompt Analysis. To assess A2R's learned speaker attribution capability independent of transcript-based reasoning, we remove the reasoning trace at inference time. Figure 7 compares *Reasoning w/ Transcript*, which explicitly generates a transcript before reasoning, with a less structured *Reasoning* prompt that elicits only a brief rationale. While transcript-first reasoning provides an additional benefit, A2R continues to outperform the baselines on the majority of benchmarks even without this scaffold. This suggests that A2R has internalized speaker attribution rather than merely learning to exploit a particular prompting strategy.

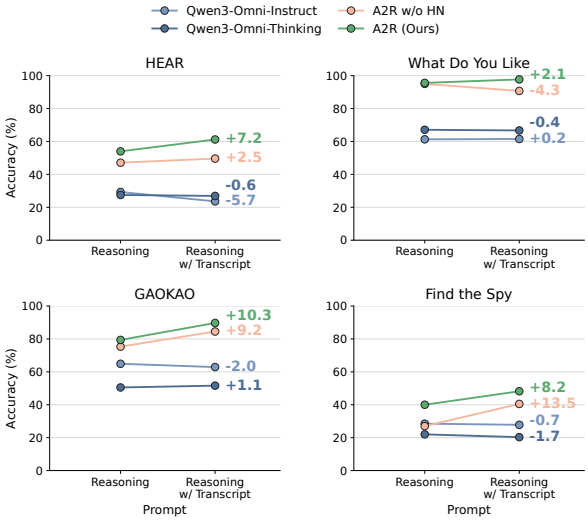


Figure 7: **Prompt Analysis on benchmarks.** "Reasoning" denotes a prompt that asks the model to provide brief reasoning before answering, while "Reasoning w/ transcript" asks for a transcript before answering.

Why A2R Works. In our answer-only ablation experiments, we observed the model relying on subjective adjectives (e.g., "high pitch", "documentary-like voice") to track speakers. This mapping from continuous acoustic signals to descriptive text can induce identity hallucination. While merely prompting baselines to use structural identifiers (e.g., "Speaker 1:") does not help (Figure 7), explicitly reinforcing a 1:1 alignment between distinct voices and these tags establishes a reliable anchor for the model's reasoning process. Replacing vague vocal descriptors with discrete tags prevents identity confusion throughout extended chain-of-thought generation.

6 Related Works

Speaker Attribution: From Speaker Recognition to Reasoning. The question of *who said what* (Kanda et al., 2021) is traditionally addressed through speaker diarization and multi-talker ASR, which identify *who spoke when* and *what was said*, respectively. Recent LLM-based recognizers improve this pipeline by incorporating speaker information or supporting instruction-guided transcription (Wang et al., 2024; Li et al., 2026; Meng et al., 2025). Yet these systems mainly produce speaker-labeled transcripts, leaving speaker-content reasoning largely unexplored. We refer to this capability as *speaker-attributed reasoning*: answering questions that require linking what was said to who said it, such as identifying the speaker of an utterance or determining who spoke more or first.

Evaluation of Speaker-Attributed Understanding. Recent SLM benchmarks have expanded from single-speaker QA to broader audio and multi-party understanding, but speaker attribution remains only partially isolated. Broad benchmarks such as MMAU (Sakshi et al., 2024; Kumar et al., 2025), AudioMarathon (He et al., 2025), and ChronosAudio (Luo et al., 2026) include multi-speaker scenarios, yet treat them as a small subset of tasks (e.g., Voice Counting) among many evaluation axes. M3-SLU (Kwon et al., 2025) is closest to our setting, but its evaluation is limited to a narrow task (e.g., similar to IR dimension of HEAR), mainly true/false or single utterance–speaker matching. MSU-Bench (Wang et al., 2025) evaluates multi-party conversational understanding, but does not separate speaker attribution as a dedicated axis. In contrast, HEAR isolates speaker attribution itself as the primary axis, organizing it into discrimination, attribution, and reasoning stages that prior benchmarks leave entangled.

7 Conclusion

We introduced HEAR, a benchmark for evaluating speaker-attributed reasoning in SLMs, and showed that current models often fail to bind utterances to distinct voices, instead relying on semantic priors over vocal evidence. To address this limitation, we proposed CASH, a 60K-scale dataset of voice-cloned hard negatives that decouples acoustic cues from semantic content, and developed A2R to explicitly learn speaker attribution. A2R substantially improves performance on HEAR and transfers to unseen multi-speaker tasks. Together, we hope our findings establish speaker attribution as a key step toward multi-party auditory comprehension.

Limitations

While explicit reasoning via transcript generation maximizes speaker attribution performance, it introduces latency that limits real-time applications. Importantly, externalizing transcripts is not essential: as shown in Figure 7, A2R significantly outperforms baselines even without this step, albeit with slightly lower performance. Bridging this latency–performance gap through implicit reasoning over internal representations, without generating explicit reasoning traces, is a promising direction for future work.

Ethics Statement

All verification in our work was conducted by eight authors of this paper. Our work also involves voice cloning samples that pose risks of impersonation and speaker re-identification. Because dataset or copyright licenses do not themselves constitute consent for voice cloning, and no specific consent was obtained from source speakers, synthetic audio is restricted to approved, non-commercial research under safeguards against misuse.

Dual-use risks. Given the potential misuse of voice cloning for impersonation, deceptive audio generation, fraud, harassment, speaker re-identification, and other harmful applications, we treat the synthetic speech component of our work as a controlled research resource rather than an unrestricted public dataset.

Voice, likeness, and consent. Although the source corpora provide licenses or research-use conditions governing the underlying recordings, these conditions do not necessarily constitute consent to generate novel utterances in an individual's voice, and we did not obtain additional task-specific consent from the source speakers. We also do not regard pseudonymous speaker identifiers as providing complete anonymity, because a recognizable voice can itself convey identity.

Release scope and access control. Code, evaluation protocols, non-identifying annotations and metadata, and aggregate statistics can be released publicly, whereas synthetic waveform data are available only through a gated research repository for approved, non-commercial research related to speaker-aware speech understanding and reasoning. Requesters must provide their identity, institutional affiliation, and intended research use and agree to the Data Use Agreement (DUA) before access is granted.

Data Use Agreement. The DUA restricts use of the synthetic audio to approved non-commercial research and prohibits redistribution, attempts to re-identify source speakers, impersonation, deceptive or misleading use, commercial voice replication, deployment of cloned voices in interactive or production systems, and other uses intended to harm or misrepresent individuals. Access is revocable when these conditions are violated.

Non-attribution and documentation. All synthetic utterances are documented as artificially generated samples. They are not statements actually made by the corresponding source speakers and must not be interpreted as representing those individuals' views, intentions, beliefs, or endorsements. Personally identifying metadata that are unnecessary for the research purpose, including available speaker names and source-speaker mappings, are removed from the released resource.

Takedown procedure. Both rights holders and individuals whose voices occur in the resource, or their authorized representatives, can request removal of corresponding entries. Requests can identify the relevant source recording or other information sufficient to locate the affected samples. Valid requests result in removal of the associated source-derived and synthetic entries from subsequent distributions of the resource. Takedown records are maintained in the dataset documentation without disclosing unnecessary personal information about the requester.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University); No.2022-0-00959, RS-2022-II220959], National Research Foundation of Korea (NRF) grant [No.2022R1A3B1077720, 2022R1A5A7083908], BK21 FOUR Program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2026, Mobile eXperience(MX) Business, Samsung Electronics Co., Ltd., NVIDIA Academic Grant Program and the Research Grant from Seoul National University(800-20250397).

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: a pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, page 28–39, Berlin, Heidelberg. Springer-Verlag.

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

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

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, and 17 others. 2026. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *Preprint*, arXiv:2604.27393.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, interspeech 2020, page 3830–3834. ISCA.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2026. Midashenglm: Efficient audio understanding with general audio captions. *Preprint*, arXiv:2508.03983.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2025. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Norman P. Erber. 1982. *Auditory training / Norman P. Erber.* Alexander Graham Bell Association for the Deaf, Washington, D.C.

Yiming Gao, Bin Wang, Chengwei Wei, Shuo Sun, and AiTi Aw. 2025. Ifeval-audio: Benchmarking instruction-following capability in audio-based large language models. *Preprint*, arXiv:2505.16774.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Google DeepMind. 2026a. Gemini 3 flash model card. Model card. Accessed: 2026-08-26.

Google DeepMind. 2026b. Gemini 3.1 pro model card. Model card. Accessed: 2026-08-26.

Peize He, Zichen Wen, Yubo Wang, Yuxuan Wang, Xiaoqian Liu, Jiajie Huang, Zehui Lei, Zhuangcheng Gu, Xiangqi Jin, Jiabing Yang, Kai Li, Zhifei Liu, Weijia Li, Cunxiang Wang, Conghui He, and Linfeng Zhang. 2025. Audiomarathon: A comprehensive benchmark for long-context audio understanding and efficiency in audio llms. *Preprint*, arXiv:2510.07293.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. Wavllm: Towards robust and adaptive speech large language model. *Preprint*, arXiv:2404.00656.

Adam L. Janin, Don Baron, Jane Edwards, Daniel P. W. Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters. 2003. The icsi meeting corpus. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, 1:I–I.

Naoyuki Kanda, Xiong Xiao, Jian Wu, Tianyan Zhou, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2021. A comparative study of modular and joint approaches for speaker-attributed asr on monaural long-form audio. *Preprint*, arXiv:2107.02852.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. 2026. Clip behaves like a bag-of-words model cross-modally but not uni-modally. *Preprint*, arXiv:2502.03566.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón,

Satvik Dixit, and 15 others. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *Preprint*, arXiv:2508.13992.

Doyeop Kwak, Jaemin Jung, Kihyun Nam, Youngjoon Jang, Jee-Weon Jung, Shinji Watanabe, and Joon Son Chung. 2024. Voxmm: Rich transcription of conversations in the wild. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12551–12555.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yejin Kwon, Taewoo Kang, Hyunsoo Yoon, and Changouk Kim. 2025. M3-slu: Evaluating speaker-attributed reasoning in multimodal large language models. *Preprint*, arXiv:2510.19358.

Dongwook Lee, Eunwoo Song, Che Hyun Lee, Heeseung Kim, and Sungroh Yoon. 2026. Still between us? evaluating and improving voice assistant robustness to third-party interruptions. *Preprint*, arXiv:2604.17358.

Li Li, Ming Cheng, Weixin Zhu, Yannan Wang, Juan Liu, and Ming Li. 2026. Dm-asr: Diarization-aware multi-speaker asr with large language models. *Preprint*, arXiv:2604.22467.

Alexander H. Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, Sanchit Gandhi, Soham Ghosh, Srijan Mishra, Thomas Foubert, Abhinav Rastogi, Adam Yang, Albert Q. Jiang, Alexandre Sablayrolles, Amélie Héliou, and 87 others. 2025. Voxtral. *Preprint*, arXiv:2507.13264.

Kaiwen Luo, Liang Lin, Yibo Zhang, Moayad Aloqaily, Dexian Wang, Zhenhong Zhou, Junwei Zhang, Kun Wang, Li Sun, and Qingsong Wen. 2026. Chronosaudio: A comprehensive long-audio benchmark for evaluating audio-large language models. *Preprint*, arXiv:2601.04876.

Lingwei Meng, Shujie Hu, Jiawen Kang, Zhaoqing Li, Yuejiao Wang, Wenxuan Wu, Xixin Wu, Xunying Liu, and Helen Meng. 2025. Large language model can transcribe speech in multi-talker scenarios with versatile instructions. *Preprint*, arXiv:2409.08596.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and Yong Jae Lee. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *Preprint*, arXiv:2512.02231.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2026a. GPT-5.5 System Card. `https://openai.com/index/gpt-5-5-system-card/`. Accessed: 2026-08-27.

OpenAI. 2026b. Introducing GPT-5.4 mini and nano. `https://openai.com/index/introducing-gpt-5-4-mini-and-nano/`. Accessed: 2026-08-27.

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026a. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. 2026b. Fun-audio-chat technical report. *Preprint*, arXiv:2512.20156.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. 2025. Step-audio-r1 technical report. *Preprint*, arXiv:2511.15848.

Thilo von Neumann, Christoph Boeddeker, Marc Delcroix, and Reinhold Haeb-Umbach. 2024. Meeteval: A toolkit for computation of word error rates for meeting transcription systems. *Preprint*, arXiv:2307.11394.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2026. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *Preprint*, arXiv:2506.04779.

Jinhan Wang, Weiqing Wang, Kunal Dhawan, Taejin Park, Myungjong Kim, Ivan Medennikov, He Huang, Nithin Koluguri, Jagadeesh Balam, and Boris Ginsburg. 2024. Meta-cat: Speaker-informed speech embeddings via meta information concatenation for multi-talker asr. *Preprint*, arXiv:2409.12352.

Shuai Wang, Zhaokai Sun, Zhennan Lin, Chengyou Wang, Zhou Pan, and Lei Xie. 2025. Msu-bench: Towards understanding the conversational multi-talker scenarios. *Preprint*, arXiv:2508.08155.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024. Just asr + llm? a study on speech large language models' ability to identify and understand speaker in spoken dialogue. *Preprint*, arXiv:2409.04927.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Zixiang Xu, Yanbo Wang, Yue Huang, Jiayi Ye, Haomin Zhuang, Zirui Song, Lang Gao, Chenxi Wang, Zhaorun Chen, Yujun Zhou, Sixian Li, Wang Pan, Yue Zhao, Jieyu Zhao, Xiangliang Zhang, and Xiuying Chen. 2025c. Socialmaze: A benchmark for evaluating social reasoning in large language models. *Preprint*, arXiv:2505.23713.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *Preprint*, arXiv:2408.01800.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2024. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 2025. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Preprint*, arXiv:2506.21619.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Jiancheng Gui, Jiaheng Wu, Ziyang Wang, Xudong Shen, Runchuan Ye, Zhisheng Zhang, Jiuyang Zhou, Bingsong Bai, Weiyue Sun, Mengyuan Deng, Qundong Shi, Zhiyong Wu, and Zhiyuan Liu. 2026. Voxcpm2 technical report. *Preprint*, arXiv:2606.06928.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Curation of CASH & Statistics

CASH-60K. We build on the VoxMM (Kwak et al., 2024) train split, extracting 5,242 multi-speaker clips drawn from diverse multi-speaker scenarios such as daily conversations, interviews and broadcast scenes. From these originals we synthesize 14,180 voice-swap variant clips, about 2.7 variants per original on average, yielding 19,422 audio clips in total. We then use both originals and variants to instantiate the task dimensions defined in Section 2, producing 59,762 training queries (roughly three queries per audio clip): 31,045 for attribution, 20,094 for reasoning, and 8,623 for discrimination. All voice cloning is performed with VoxCPM2 (Zhou et al., 2026), one of the current state-of-the-art TTS systems. To ensure the linguistic and acoustic fidelity of the cloned speech, we apply a two-stage filtering protocol to every synthesized clip. First, the clip is transcribed with Whisper-large-v3 (Radford et al., 2022) and we retain only samples whose word error rate (WER) against the target transcript falls below 0.1. Second, we compute the cosine similarity between ECAPA-TDNN (Desplanques et al., 2020) embeddings of the synthesized clip and the reference voice, and keep only samples whose speaker similarity exceeds 0.7.

B Curation Process of HEAR Benchmark

This section details the data curation methodology for the HEAR benchmark. The curation pipeline operates in two primary stages: an automated process to generate a comprehensive pool of candidate questions from three established datasets, followed by a rigorous manual review phase wherein human annotators validate and filter the questions to ensure high data quality.

B.1 Rule-Based Question Construction

Source Data and Extraction. We leverage three publicly available datasets characterized by human-verified text transcripts: AMI and ICSI (multi-party meetings) and VoxMM (CC-BY-4.0-licensed YouTube videos). To maintain ground-truth fidelity, we strictly rely on the original human annotations, explicitly excluding any AI-generated transcriptions. The source audio is segmented into 30- to 150-second clips (averaging approximately 93 seconds), with each segment curated to contain at least two distinct speakers, 4.05 on average and ranging from 2 to 12.

Voice References. All voice samples utilized as multiple-choice options are sourced externally to the test clip. These external reference samples consistently feature a single, non-overlapping speaker and range in duration from 3 to 10 seconds.

B.2 Taxonomy

The automated pipeline generated three categories of questions:

Discrimination. The Discrimination dimension is generated mechanically from the transcript and speaker timing. VC (*voice cardinality*) asks for the number of distinct speakers in a clip. VL (*voice localization*) presents a reference voice and asks the model to select a time range on the basis of that speaker's activity. Each item takes one of two polarities: a presence variant, whose gold option is an interval in which the target speaker is active while the four distractors are intervals in which the speaker is not, and an absence variant, which reverses this so that the gold option is the only interval containing no speech from the target. VCD (*voice change/overlap detection*) asks for the time range that contains either a turn change or an overlap; the generator enumerates all eligible 3–7 s windows and rejects any that violate the regime constraint.

Attribution. Attribution items are produced by jointly sampling a 5-way option set on top of the per-clip speaker pool. CVA (*content→voice*) shows a transcript span and asks which voice uttered it; VCA (*voice→content*) shows a voice and asks which transcript span matches it. Both come in non-overlap and overlap variants determined by the surrounding speech context. Distractors are drawn from the same clip to enforce voice-level discrimination; each option's voice reference is the external segment described above.

Reasoning. Reasoning items follow a hallucination-pair design. For every original question, we synthesize a counterfactual variant in which a single answer-relevant utterance has been re-spoken in a *different* speaker's voice via voice cloning, while leaving all other audio untouched. Voice swaps are produced by IndexTTS 2 (Zhou et al., 2025) conditioned on external 3–10 s voice references from the long source clip, and only swaps whose Whisper-large-v3 (Radford et al., 2022) ASR transcript matches the target text with $\mathrm{WER} = 0$ are admitted. Three sub-types are produced: QR (Quantitative Reasoning),

IR (Identity Reasoning), and TR (Temporal Reasoning). A model that ignores voice identity must answer the original and the hallucinated variant identically and therefore fails the pair; this is the mechanism by which the Reasoning split exposes voice-blind shortcutting.

B.3 Overlap Labeling Criteria

To further evaluate robustness under acoustically mixed conditions, we provide overlap-labeled subsets for all taxonomy dimensions. Overlap samples are annotated by human evaluators based on whether the overlapping speech segment is required to derive the correct answer, along with additional criteria specific to each dimension. Below, we describe the overlap criteria and task-specific objectives for each category.

Voice Localization (VL-Overlap) The overlap subset is defined as samples where the target speaker's active interval temporally overlaps with speech from at least one additional speaker. The objective is to localize the target voice despite acoustic interference and concurrent speech activity.

Voice Change Detection (VCD-Overlap) For overlap samples, the target interval contains simultaneous speech activity from multiple speakers. Unlike the standard setting, which focuses on clean turn-taking boundaries, this subset evaluates whether the model can detect overlap onset and offset boundaries within acoustically mixed regions.

Content-to-Voice Attribution (CVA-Overlap) The overlap subset consists of utterances spoken during intervals containing concurrent speech from multiple speakers. The model must correctly associate the queried textual utterance with its corresponding speaker identity despite acoustic mixtures.

Voice-to-Content Attribution (VCA-Overlap) The overlap subset is defined by cases where the ground-truth utterance occurs within overlapping speech intervals. The task evaluates whether the model can correctly retrieve the linguistic content associated with a target voice under concurrent speech conditions.

Identity Reasoning (IR-Overlap) IR-Overlap samples contain anchor or candidate utterances that temporally intersect with overlapping speech regions. The model must resolve speaker identity consistently despite acoustic interference.

Temporal Reasoning (TR-Overlap) The overlap subset includes temporal reasoning chains involving utterances occurring during concurrent speech. Models must correctly infer speaker-conditioned temporal relationships while disentangling overlapping voices.

Quantitative Reasoning (QR-Overlap) QR-Overlap samples involve speaker statistics computed from sessions containing overlapping speech intervals. The model must aggregate speaker-specific evidence robustly despite acoustically mixed observations.

B.4 Human Verification

Eight members of the authors served as annotators. All were graduate students in artificial intelligence and comfortable reading English transcripts and listening to long-form multi-party audio. Annotators worked through a custom local web tool that played the clip audio, displayed the word-level transcript, exposed every option (including each voice reference), and offered Pass/Fail buttons together with a free-text reason field as described in Figure 8.

Filtering criteria.

1. **Solvability.** The reviewers are requested to answer the question correctly using only the clip audio, transcript, and the offered voice references. Items that are ambiguous, underspecified, or require information outside the clip are rejected.

2. **Unique correct answer.** Exactly one option must be correct and the remaining four distractors must each be falsifiable from the clip. Items with multiple defensible answers, with options that are mutual paraphrases, or whose distractors collapse onto the gold answer are rejected.

3. **Voice-identity preservation (Reasoning split only).** For each TTS-synthesized swap, the reviewer plays both the original and the swapped clip and confirms that the swapped voice is perceptually *the same person* as the conditioning voice reference, and that the synthesis is free of obvious artifacts (mispronunciation, prosodic discontinuity at the splice, audible bandwidth change). Pairs that fail either criterion are rejected even if WER $= 0$.

4. **Transcript–option consistency.** Although AMI, ICSI, and VoxMM are all distributed

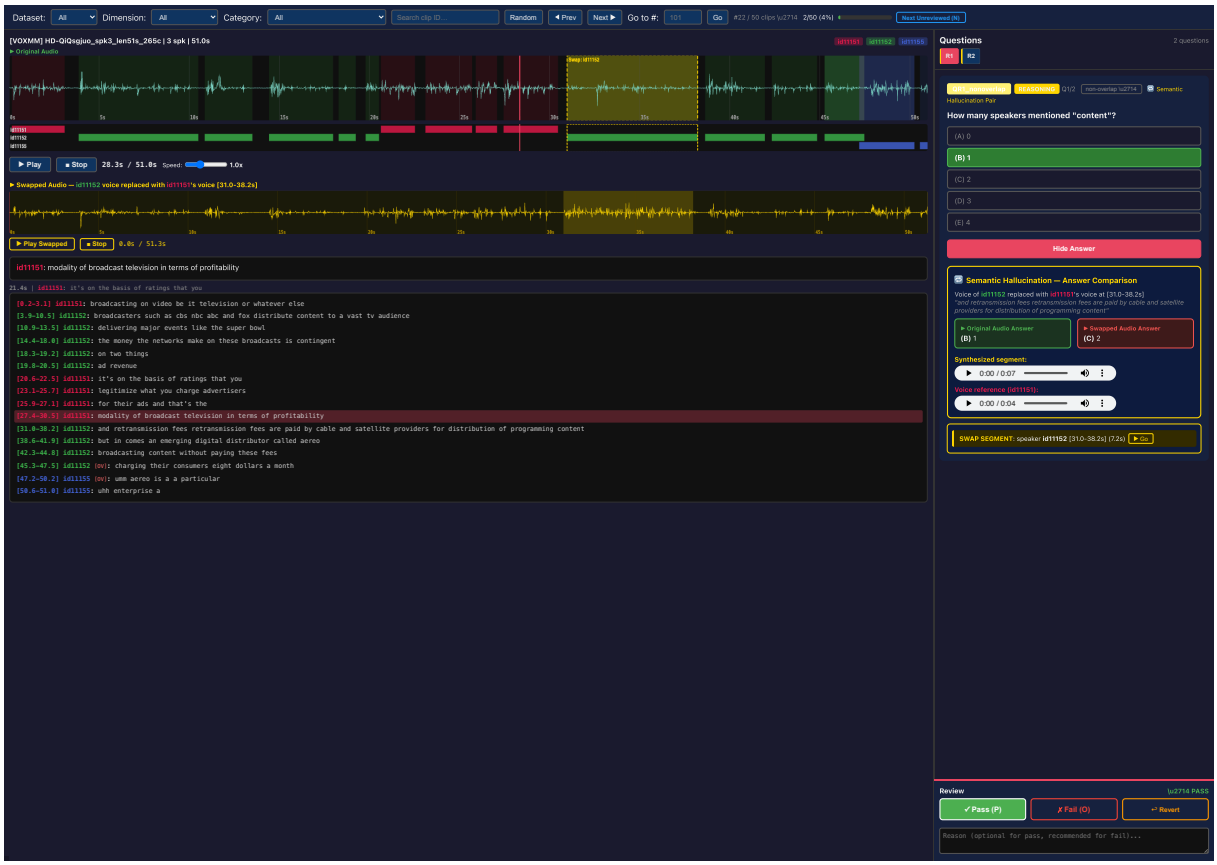


Figure 8: Screenshot of the interface shown to human annotators during verification.

with human-produced transcripts, boundary errors (off-by-a-word starts/ends, missed back-channels, mis-attributed overlapping speech) still survive into our candidate pool. Any item whose option text or query text fails to match the audio at the specified timestamps, even by a single word, is filtered out.

After the verification pass, 887 clips and 2,395 questions survive from a candidate pool of roughly three times that size. The resulting benchmark therefore reflects two compounded layers of human quality control: the original transcripts of the three source corpora, and the per-question audit described above.

B.5 Benchmark Statistics

Table 5 reports the per-domain composition of HEAR. The benchmark covers 887 clips, totals $\sim$22.8 h of audio, and exposes a model to 290 unique speakers in aggregate. VoxMM contributes the bulk of the speaker diversity (245 speakers across 34 CC-BY-4.0-licensed videos (audio) in 11 genre categories), and AMI and ICSI contribute longer, denser meeting clips (3.9–4.7 speakers/clip on average).

Domain	#Clip	#Src.	#Spk.	Dur.	Spk/clip	Hours
VoxMM	360	34	245	87.8	3.64	8.78
AMI	241	23	24	97.4	3.93	6.52
ICSI	286	6	21	94.9	4.65	7.54
TOTAL	**887**	**63**	**290**	**92.7**	**4.05**	**22.84**

Table 5: HEAR per-domain composition. **#Src.**: distinct source recordings. **#Spk.**: speakers, deduplicated within each corpus's identity namespace. **Dur.**: average clip duration, in seconds. **Hours**: total speech-clip-hours.

Domain and dimension distribution. Figure 9 (left) shows that no single domain dominates the benchmark: VoxMM accounts for 40.6% of clips, ICSI for 32.2%, and AMI for 27.2%. The skew toward meeting-style audio is intentional and follows the original corpus sizes; we counter-balance it on the question side by saturating the VoxMM category budget (Figure 11). Figure 9 (right) reports the question count per evaluation dimension: 627 Discrimination, 608 Attribution, and 1,160 Reasoning items (original + swapped), for a total of 2,395 questions.

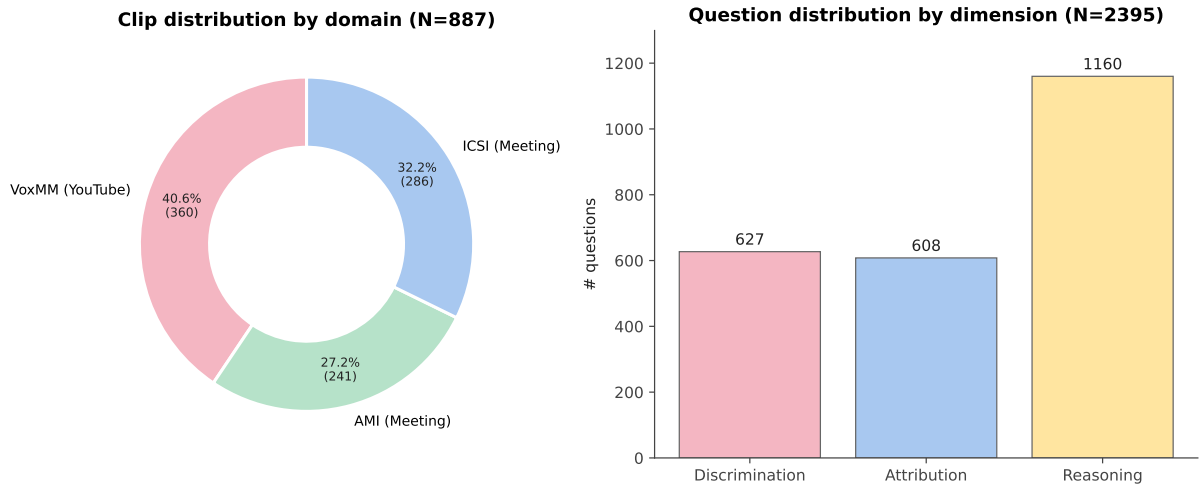


Figure 9: HEAR composition. Left: clip distribution across the three source domains ($N = 887$). Right: question distribution across the three evaluation dimensions ($N = 2{,}395$).

Sub-dimension breakdown. Figure 10 unrolls the three top-level dimensions into the eight sub-dimensions that the benchmark scores. Discrimination splits into VC (7.9%), VL (9.9%), and VCD (8.4%); Attribution splits into CVA (12.8%) and VCA (12.6%); Reasoning splits into QR (23.4%, the dominant Reasoning sub-type), IR (15.4%), and TR (9.6%).

VoxMM category coverage. Figure 11 reports the genre composition of the VoxMM slice: *Remote* (104 clips) and *Documentary* (72) are the two heads, followed by *Entertainment*, *Commercial*, *Interview*, and *News* (32–40 each), with *Politics*, *Conversation*, *Lecture*, *Presentation*, and *Sports* forming a long tail.

Per-domain summary and clip-level distributions. Figure 12 consolidates four headline statistics across the three domains: clip count, distinct source recordings, distinct speakers, and average clip duration. Clip durations are evenly distributed as described in Figure 13.

C Details of Benchmarks Where Speaker Attribution Is Essential

We provide additional details on the construction and statistics of the three voice-dependent benchmarks used in our evaluation: WDYL (*What Do You Like?*), GAOKAO, and FTS (*Find the Spy*). All three benchmarks are designed such that correctly answering the question requires not only understanding the spoken content, but also identifying and tracking *who* is speaking.

WDYL (What Do You Like?). WDYL evaluates whether a model can resolve first-person references using speaker identity. Each example contains a conversation between two speakers, followed by a first-person question such as "What is my favorite city?" spoken in the cloned voice of one of the two speakers. Because the referent of "my" depends entirely on the identity of the question speaker, the correct answer cannot be determined from the linguistic content of the question alone.

The benchmark contains 793 questions over 793 audio clips, totaling 2.36 hours of audio. The average clip duration is 10.7 s, with a range of 8.1–13.2 s. Each question has three answer choices: one corresponding to each speaker's stated preference and one distractor not mentioned by either speaker. Since the distractor can be eliminated from the transcript alone, the effective chance accuracy is 50%; accordingly, we use 50% as the random-choice baseline in Table 2. The dialogue audio is taken from (Wu et al., 2024). For each example, we append a 0.3 s pause followed by a question synthesized with IndexTTS 2, using a cloned voice from one of the speakers in the preceding dialogue.

We initially synthesized 1,000 examples and retained 793 after speaker-similarity quality control using ECAPA embeddings, requiring a speaker embedding cosine similarity (SECS) of at least 0.70. The retained examples have a mean SECS of 0.778. The cloned question voice corresponds to the *asker* in 476 examples and the *answerer* in 317 examples.

GAOKAO. GAOKAO tests speaker-conditioned interpretation in natural English listening-

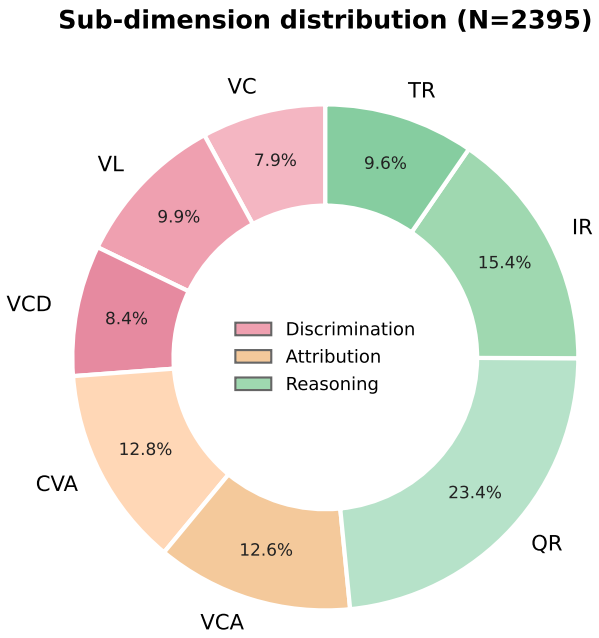


Figure 10: Sub-dimension distribution of HEAR questions ($N = 2{,}395$). VC/VL/VCD make up Discrimination (top three slices), CVA/VCA make up Attribution, and QR/IR/TR make up Reasoning.

comprehension conversations. Each example consists of a short English dialogue followed by a first-person question synthesized in the cloned voice of one of the speakers.

The benchmark contains 93 questions over 93 clips, totaling 0.52 hours (30.9 minutes) of audio. The average duration is 20.0 s, ranging from 12.0 s to 29.9 s. Each question has two answer choices, corresponding to a random-guess accuracy of 50%. The source audio consists of Gaokao English short-conversation recordings from the ICQ subset of (Hu et al., 2024). We append a 0.35 s pause and an IndexTTS 2-generated question whose voice is cloned from a speech segment of the target speaker within the same dialogue.

We synthesized 97 examples and retained 93 after quality control. Examples were required to achieve an ECAPA speaker-similarity score of at least 0.7 and a Whisper word error rate (WER) of at most 0.20. The target speaker is male in 46 examples and female in 47 examples. Gold-answer positions are distributed as A: 40 and B: 53.

FTS (Find the Spy). FTS evaluates multi-speaker voice tracking in a social-deduction setting. Each example is a 3–6-player Find-the-Spy game in which every player receives the same secret word except for one designated spy. The model observes the players' spoken clues and must identify which participant is the spy.

Crucially, players first introduce themselves by name, after which they provide their clues in a *different, randomly shuffled order* and without explicitly stating their names. Consequently, associating each clue with the corresponding player requires tracking speaker identity across temporally separated turns; textual content alone does not reveal which player produced which clue.

The benchmark contains 400 questions over 400 clips, totaling 3.67 hours of audio. The mean duration is 33.1 s, with clips ranging from 17.2 s to 56.6 s. There are 100 games for each player count from three to six, yielding chance accuracies of 33.3%, 25.0%, 20.0%, and 16.7%, respectively.

Game configurations and clues are generated using the SocialMaze (Xu et al., 2025c) framework. Each player's utterances are synthesized with IndexTTS 2, while keeping a single reference voice fixed for that player throughout the game.

D Experiment Details

As introduced in Section 4, we evaluate 20 leading SLMs categorized into three distinct families. The detailed list of models and their references is as follows:

- **Omni-modality models**: MiniCPM-o series (Yao et al., 2024; Cui et al., 2026),

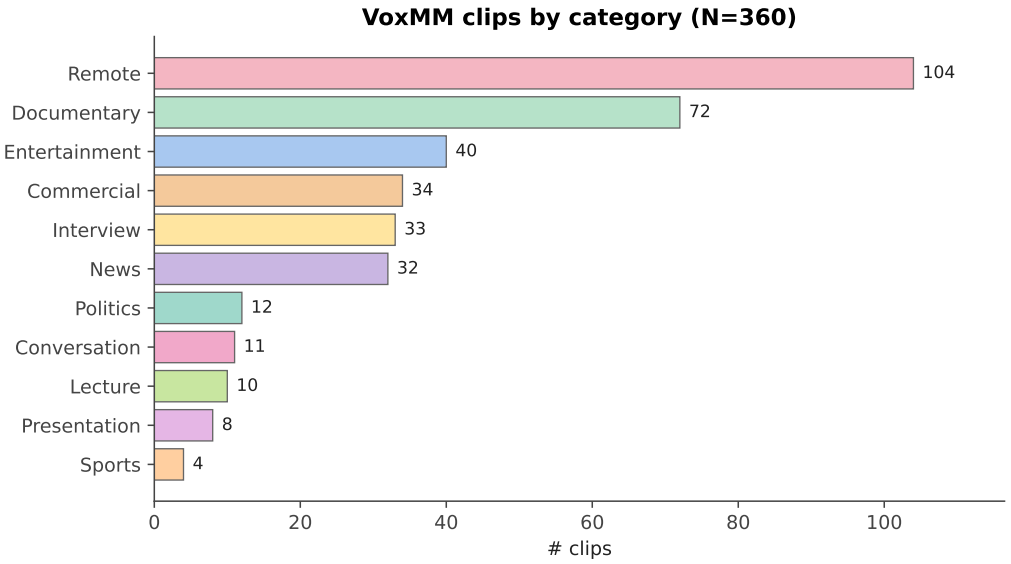


Figure 11: Distribution of VoxMM clips in HEAR by genre category ($N = 360$).

Qwen2.5-Omni series (Xu et al., 2025a), Qwen3-Omni-30B-A3B series (Xu et al., 2025b), Gemma-4-E4B-it (Team et al., 2026a), and Phi-4-Multimodal (Microsoft et al., 2025).

- **Spoken or Audio language models**: MiDashengLM (Dinkel et al., 2026), Voxtral (Liu et al., 2025), Fun-Audio-Chat (Team et al., 2026b), Kimi-Audio (KimiTeam et al., 2025), Qwen2-Audio (Chu et al., 2024), Audio-Flamingo 3 (Goel et al., 2025), and Step-Audio series (Tian et al., 2025; Wu et al., 2025).

- **Proprietary models**: Gemini-3 series (Google DeepMind, 2026b,a) and GPT-4o-audio-preview (OpenAI, 2024).

For every open-source baseline, we adopt the inference hyperparameters specified in its official repository; for proprietary systems, we use the default API settings. The concrete experimental settings are comprehensively described in Table 7. All open-weight baselines are served through vLLM or the Hugging Face inference path, while the three closed-source endpoints (*gpt-4o-audio-preview*, *gemini-3-flash-preview*, *gemini-3.1-pro-preview*) are queried with default setting. We report the value used in our configuration for reproducibility. Table 7 summarizes the per-model defaults.

E Reward Composition Details

This appendix details the coefficients of the composite reward (Section 3.3) and the dynamic-programming implementation of the speaker-order reward R_{ord}.

E.1 Per-term Coefficients

The five reward terms are linearly combined with fixed coefficients λ_x:

$$\begin{aligned} R(Y_i; \tau^\star, a^\star) = \lambda_{\text{ans}} R_{\text{ans}}(\hat{a}_i; a^\star) \\ + \lambda_{\text{fmt}} R_{\text{fmt}}(Y_i) + \lambda_{\text{cp}} R_{\text{cp}}(\hat{\tau}_i; \tau^\star) \\ + \lambda_{\text{cnt}} R_{\text{cnt}}(\hat{\tau}_i; \tau^\star) + \lambda_{\text{ord}} R_{\text{ord}}(\hat{\tau}_i; \tau^\star). \end{aligned} \tag{1}$$

Table 6 lists the coefficients used in our experiments.

Reward term	Symbol	Coefficient λ
Answer correctness	R_{ans}	1.0
Response format	R_{fmt}	0.2
Speaker-tagged transcript (cpWER)	R_{cp}	0.5
Speaker-count	R_{cnt}	0.5
Speaker-order (LCS)	R_{ord}	0.5

Table 6: Coefficients for the composite reward terms.

E.2 Speaker-Order Reward via LCS

To compute R_{ord}, we extract the sequence of speaker labels in order of appearance from both the ground-truth ($Q = (q_1, \ldots, q_m)$) and the prediction ($P = (p_1, \ldots, p_n)$).

We define R_{ord} as the length-normalized

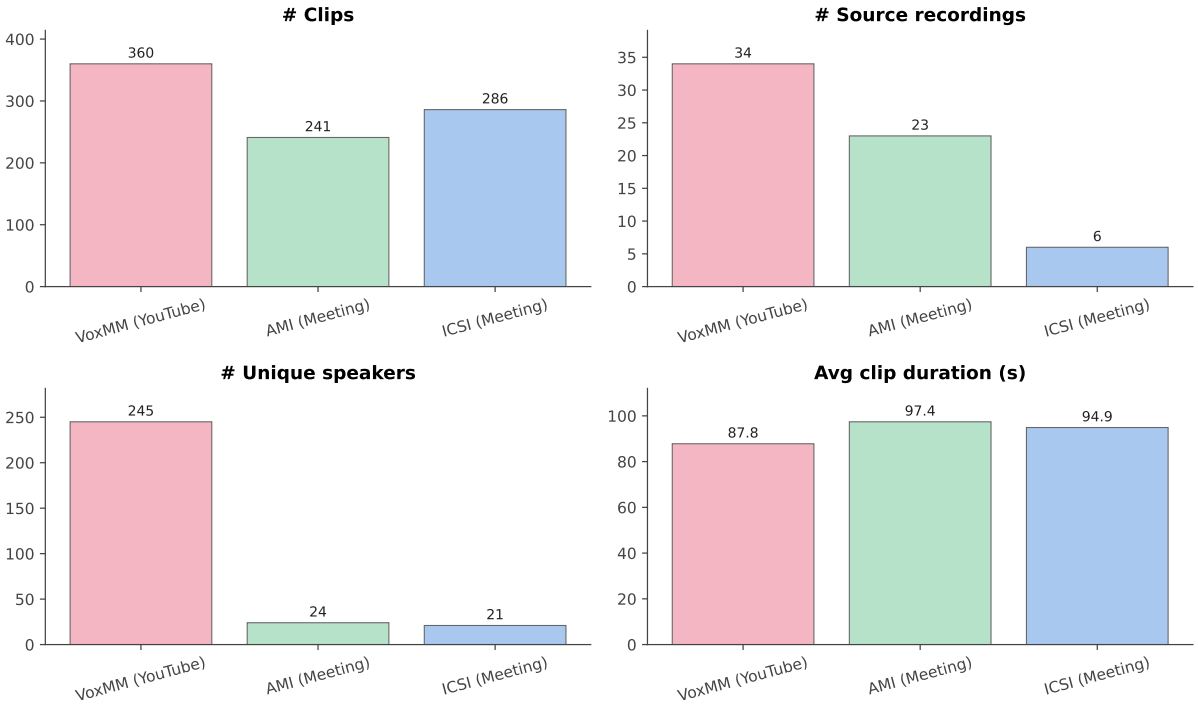


Figure 12: Per-domain summary statistics for HEAR.

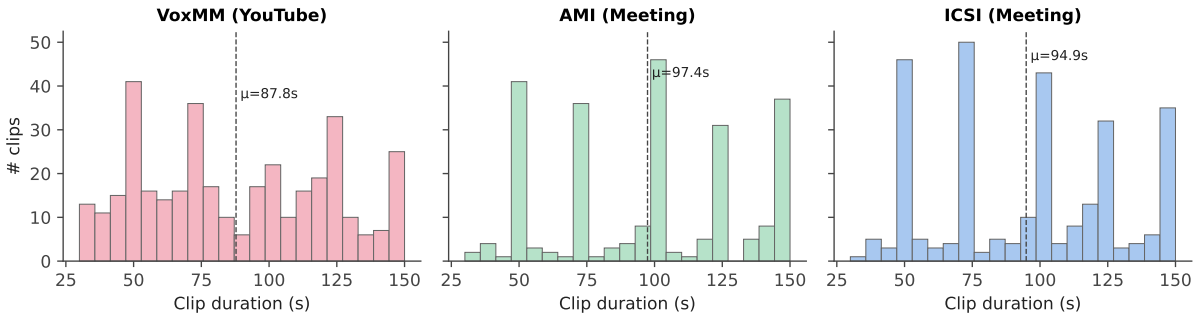


Figure 13: Statistics of clip duration for HEAR.

Longest Common Subsequence (LCS):

$$R_{\text{ord}}(\hat{\tau}_i; \tau^\star) = \frac{\text{LCS}(P, Q)}{\max(|P|, |Q|, 1)} \in [0, 1]. \quad (2)$$

$\text{LCS}(P, Q)$ represents the longest sequence appearing in both P and Q in the same relative order. Algorithm 1 outlines the $O(nm)$ dynamic-programming procedure.

From DP to Speaker Turns. The LCS algorithm directly aligns predicted (P) and ground-truth (Q) speaker sequences. When $p_i = q_j$, the DP matches a correctly attributed turn in temporal order ($+1$). When $p_i \neq q_j$, the $\max$ operation resolves misalignments by skipping either a hallucinated prediction or a missing GT turn. This ensures that isolated errors do not disrupt the global turn alignment.

Example. Suppose the maximum reward R_{ord} is normalized to 1 in this example. Given a 5-turn dialogue $Q = (s_1, s_2, s_1, s_3, s_2)$, our metric gracefully penalizes conversational errors. *Case 1 (misattribution):* Predicting $P_1 = (s_1, s_2, \mathbf{s_2}, s_3, s_2)$ yields $\text{LCS} = 4$ and $R_{\text{ord}} = 4/5 = 0.8$. *Case 2 (insertion):* Emitting an extra turn ($|P_2| = 6$) keeps $\text{LCS} = 5$ but increases the denominator, yielding $R_{\text{ord}} = 5/\max(6, 5) \approx 0.83$. *Case 3 (deletion):* Missing a turn ($|P_3| = 4$) yields $R_{\text{ord}} = 4/5 = 0.8$. Thus, misattributions, insertions, and deletions all incur proportional penalties, providing a robust turn-alignment signal.

F Training Details

We fine-tune Qwen3-Omni-30B-A3B-Instruct with GRPO on 4×NVIDIA H200 GPUs. LoRA

Algorithm 1 Speaker-Order Reward R_{ord} (Length-normalized LCS)

Require: Predicted speaker labels $P = (p_1, \ldots, p_n)$; GT speaker labels $Q = (q_1, \ldots, q_m)$
Ensure: Reward $r \in [0, 1]$
$D[i, 0] \leftarrow 0,\ D[0, j] \leftarrow 0 \quad \forall i, j$
for $i = 1$ **to** n **do**
  for $j = 1$ **to** m **do**
    if $p_i = q_j$ **then**
      $D[i, j] \leftarrow D[i-1, j-1] + 1$
    else
      $D[i, j] \leftarrow \max\big(D[i-1, j],\ D[i, j-1]\big)$
    end if
  end for
end for
return $D[n, m]\ /\ \max(n, m, 1)$

adapters are applied to all language-model linear projections with rank r=64, α=128, and dropout 0, while the audio/vision encoders and modality aligners remain frozen. Expert parallelism (EP=4) is used to fit the MoE backbone across GPUs.

Training runs for 1 epoch with an effective batch size of 64, formed by 16 unique prompts and 4 GRPO rollouts per prompt. We use bfloat16 precision and a learning rate of 5×10^{-6} with cosine scheduling, 239 warmup iterations, and cosine decay to 5×10^{-7}. Rollouts are generated with vLLM (Kwon et al., 2023) at temperature 1.0 using sequence-level importance sampling, asymmetric PPO clipping (ε_{low}=3×10^{-4}, $\varepsilon_{\text{high}}$=$4\times10^{-4}$), and β=0 KL regularization. Training is implemented on top of the MEGATRON-SWIFT (Zhao et al., 2024) backend, which provides the expert-parallel MoE training support required for the Qwen3-Omni-30B-A3B backbone.

G Descriptions of VoiceBench

To evaluate the dyadic conversational capabilities of our fine-tuned A2R models, we utilize VoiceBench (Chen et al., 2024), a comprehensive benchmark featuring both real and synthetic spoken instructions. Following standard protocols, AlpacaEval, CommonEval, and WildVoice samples are assessed using gpt-5.4-mini (OpenAI, 2026b). We evaluate our models across five key dimensions:

Open-Ended QA (AlpacaEval (Dubois et al., 2025), CommonEval (Ardila et al., 2020), WildVoice) assesses general knowledge and conversational capability. While **AlpacaEval** uses synthetic instructions to measure response quality, **CommonEval** and **WildVoice** (1,000 human-recorded samples) evaluate information-seeking capabilities under diverse, real-world human speech conditions.

Reference-Based QA (BBH) (Suzgun et al., 2022) measures reasoning accuracy against reference answers. We use a human-recorded spoken version of **BBH** (BIG-Bench Hard) to evaluate multi-step reasoning (e.g., logical, arithmetic, and temporal) on long and complex real spoken instructions.

Multiple-Choice QA (MMSU (Wang et al., 2026), OpenBookQA) (Mihaylov et al., 2018) evaluates general knowledge, elementary science, and common-sense reasoning by requiring models to select the correct option from a predefined set.

Instruction Following (IFEval) (Gao et al., 2025) tests the model's strict adherence to specific structural and formatting constraints using synthetic spoken instructions.

Safety (AdvBench) (Zou et al., 2023) measures the model's refusal rate against malicious prompts using synthetic speech, operating on the premise that a safe assistant must firmly reject harmful instructions.

H Paired Reasoning Accuracy

For the reasoning axis, we evaluate each original clip jointly with its semantic-hallucination counterpart. Let c_i and $\tilde{c}_i$ denote binary correctness indicators for the original and hallucinated clips in pair i, respectively. We define paired reasoning accuracy as

$$R_{\text{pair}} = \frac{1}{N_R} \sum_{i=1}^{N_R} c_i \tilde{c}_i, \tag{3}$$

where N_R is the number of reasoning clip pairs. Thus, each pair constitutes a single evaluation unit and contributes one correct prediction only when both clips are answered correctly.

Discrimination and attribution are evaluated with standard accuracy, whereas reasoning uses R_{pair}. To account for the different numbers of evaluation units across the three axes, we compute the overall average by weighting each axis by its corresponding number of evaluation units:

$$\text{Avg.} = \frac{N_D D + N_A A + N_R R_{\text{pair}}}{N_D + N_A + N_R}, \tag{4}$$

where N_D and N_A denote the numbers of discrimination and attribution examples, respectively, and N_R denotes the number of reasoning clip pairs. In particular, each reasoning pair is counted as one evaluation unit rather than as two individual clips.

I Licenses and Intended Use

We use all assets in accordance with their stated terms. AMI (Carletta et al., 2005), ICSI (Janin et al., 2003), and VoxMM (Kwak et al., 2024) are released under CC BY 4.0. Our base model, Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), is released under Apache 2.0.

J AI Tool Use

We used AI-based writing assistance tools only for language editing, grammar correction, and improving the clarity of the manuscript.

VOICE COUNTING

VC

▶ **Main Audio**

Listen to the main audio and answer the following multiple-choice question.
How many distinct speakers are present in this audio clip?

Options: (A) 2 (B) 3 (C) 4 (D) 5 (E) 6

VOICE LOCALISATION

VL

▶ **Main Audio + Reference Voice**

Listen to the main audio, followed by a reference voice of the target speaker...
Given the reference voice, select the time range where this speaker appears.

Options A–E are time ranges, e.g., [3.0s – 5.5s].

VOICE LOCALISATION

VL

▶ **Main Audio + Reference Voice**

Listen to the main audio, followed by a reference voice...
Given the reference voice, select the time range where this speaker does NOT appear.

Options A–E are time ranges.

VOICE-CHANGE DETECTION

VCD

▶ **Main Audio**

Listen to the main audio and answer the following multiple-choice question.
In which time range does a speaker change (voice transition) occur?

Options A–E are time ranges.

VOICE-CHANGE DETECTION

VCD

▶ **Main Audio**

Listen to the main audio and answer the following multiple-choice question.
In which time range does speaker overlap (multiple speakers talking simultaneously) occur?

Options A–E are time ranges.

Figure 14: Sample queries of *Discrimination* dimension in HEAR benchmark.

CONTENT-TO-VOICE ATTRIBUTION

CVA

▶ **Main Audio + Voice-Options (A–E)**

Listen to the main audio, followed by the voice options audio...

Which speaker said: $\langle X \rangle$?

Options (A)–(E) correspond to the voice options.

VOICE-TO-CONTENT ATTRIBUTION

VCA

▶ **Main Audio + Reference Voice**

Listen to the main audio, followed by a reference voice of the target speaker...

Which of the following utterances was spoken by the given voice?

(A) $\langle X_1 \rangle$ (B) $\langle X_2 \rangle$ (C) $\langle X_3 \rangle$ (D) $\langle X_4 \rangle$ (E) $\langle X_5 \rangle$

Figure 15: Sample queries of *Attribution* dimension in HEAR benchmark.

QUANTITATIVE REASONING

QR1

▶ **Main Audio**

Listen to the main audio and answer the following multiple-choice question.

How many speakers mentioned $\langle X \rangle$?

Options: (A) 1 (B) 2 (C) 3 (D) 4 (E) 5

QUANTITATIVE REASONING

QR2

▶ **Main Audio**

Listen to the main audio and answer the following multiple-choice question.

Which of the following was said by the speaker with the k-th longest total speaking time?

(A) $\langle X_1 \rangle$ (B) $\langle X_2 \rangle$ (C) $\langle X_3 \rangle$ (D) $\langle X_4 \rangle$ (E) $\langle X_5 \rangle$

Figure 16: Sample queries of *Reasoning* dimension in HEAR benchmark.

TEMPORAL REASONING TR1

▶ Main Audio

Listen to the main audio and answer the following multiple-choice question.

What was the very first utterance spoken by the person who said $\langle Y \rangle$?

Options A–E are candidate utterances.

TEMPORAL REASONING TR1

▶ Main Audio

Listen to the main audio and answer the following multiple-choice question.

What was the very last utterance spoken by the person who said $\langle Y \rangle$?

Options A–E are candidate utterances.

TEMPORAL REASONING TR2

▶ Main Audio

Listen to the main audio and answer the following multiple-choice question.

After someone said $\langle X \rangle$, what was the first thing said by the person who said $\langle Y \rangle$?

Options A–E are candidate utterances.

IDENTITY REASONING IR

▶ Main Audio

Listen to the main audio and answer the following multiple-choice question.

Which of the following was also said by the person who said $\langle Y \rangle$?

Options A–E are candidate utterances.

Figure 17: Sample queries of *Reasoning* dimension in HEAR benchmark.

Model	Checkpoint (HF / source)	Temp.	Top-p	Max tokens
Audio-Flamingo 3	`nvidia/audio-flamingo-3-hf`	0.7	0.9	512
Fun-Audio-Chat-8B	`FunAudioLLM/Fun-Audio-Chat-8B`	0.6	0.95	512
Gemma 4 E4B-it	`google/gemma-4-E4B-it`	1.0	0.95	512
Kimi-Audio-7B-Instruct	`moonshotai/Kimi-Audio-7B-Instruct`	0.0	–	2,048
MiDashengLM-7B	`mispeech/midashenglm-7b`	0.0	–	512
MiniCPM-o-2.6	`openbmb/MiniCPM-o-2_6`	0.3	–	512
MiniCPM-o-4.5	`openbmb/MiniCPM-o-4_5`	0.3	0.7	512
Phi-4-Multimodal	`microsoft/Phi-4-multimodal-instruct`	0.0	–	512
Qwen2-Audio-7B-Instruct	`Qwen/Qwen2-Audio-7B-Instruct`	0.7	0.5	512
Qwen2.5-Omni-3B	`Qwen/Qwen2.5-Omni-3B`	0.0	–	512
Qwen2.5-Omni-7B	`Qwen/Qwen2.5-Omni-7B`	0.0	–	512
Qwen3-Omni-30B-A3B-Instruct	`Qwen/Qwen3-Omni-30B-A3B-Instruct`	0.6	0.95	8192
Qwen3-Omni-30B-A3B-Thinking	`Qwen/Qwen3-Omni-30B-A3B-Thinking`	0.6	0.95	16,384
Step-Audio-2-mini	Step-Audio-2-mini (StepFun)	0.1	–	2,048
Step-Audio-R1	Step-Audio-R1 (StepFun, thinking mode)	0.7	0.9	16,384
Voxtral-Mini-3B	`mistralai/Voxtral-Mini-3B-2507`	0.2	0.95	512
Voxtral-Small-24B	`mistralai/Voxtral-Small-24B-2507`	0.2	0.95	1,024

Table 7: **Decoding hyperparameters used for the open-weight baselines in our HEAR evaluation.** "–" denotes parameters left unset, using the vLLM engine's default value (1.0) (Kwon et al., 2023). For A2R, we use the same parameters as the Qwen3-Omni-30B-A3B-Instruct model.

A2R (Ours)

Reason in two stages inside `<reasoning>...</reasoning>`. First, transcribe the multi-speaker conversation in the main audio inside `<transcript>...</transcript>` placed at the very start of `<reasoning>`. Use the format `'Speaker 1: ...\nSpeaker 2: ...'`, labeling speakers in the order they first appear and joining each speaker's utterances on a single line in chronological order. Then continue reasoning grounded in that transcript, and finally provide the answer in JSON inside `<answer>...</answer>`.

Example:

```
<reasoning><transcript>Speaker 1: ...\nSpeaker 2: ...</transcript>
...your detailed reasoning grounded in the transcript...</reasoning>

<answer>{"Answer": "(A)"}</answer>
```

Ablation

Reason briefly inside `<reasoning>...</reasoning>`, then output your answer in JSON inside `<answer>...</answer>`.

Example:

```
<reasoning>Your brief reasoning here.</reasoning>

<answer>{"Answer": "(A)"}</answer>
```

Figure 18: Instruction prompts used for GRPO experiments.

Inference Prompt (HEAR)

VC / VCD / IR / QR / TR	Listen to the main audio and answer the following multiple-choice question.
VL / VCA	Listen to the main audio, followed by a reference voice of the target speaker, and answer the following multiple-choice question.
CVA	Listen to the main audio, followed by the voice options audio where each option (A through E) is announced before its voice sample, and answer the following multiple-choice question.

Reason in two stages inside <reasoning>...</reasoning>. First, transcribe the multi-speaker conversation in the main audio inside <transcript>...</transcript> placed at the very start of <reasoning>. Use the format 'Speaker 1: ...\nSpeaker 2: ...', labeling speakers in the order they first appear and joining each speaker's utterances on a single line in chronological order. Then continue reasoning grounded in that transcript, and finally provide the answer in JSON inside <answer>...</answer>. Example:

```
<reasoning><transcript>Speaker 1: ...\nSpeaker 2: ...</transcript>
...reasoning...</reasoning>

<answer>{"Answer": "(A) example text"}</answer>
```

Figure 19: Instruction prompt template used for HEAR benchmark, consisting of a task-specific listening prefix and a common reasoning suffix.

Inference Prompt (What Do You Like, Gaokao)

Reason in two stages inside <reasoning>...</reasoning>. First, transcribe the multi-speaker conversation in the main audio inside <transcript>...</transcript> placed at the very start of <reasoning>. Use the format 'Speaker 1: ...\nSpeaker 2: ...', labeling speakers in the order they first appear and joining each speaker's utterances on a single line in chronological order. Then continue reasoning grounded in that transcript, and finally provide the answer in JSON inside <answer>...</answer>. Example:

```
<reasoning><transcript>Speaker 1: ...\nSpeaker 2: ...</transcript>
...reasoning...</reasoning>

<answer>{"Answer": "(A) example text"}</answer>
```

Figure 20: Instruction prompt template used for What Do You Like and Gaokao benchmarks, consisting of a task-specific listening prefix and a common reasoning suffix.

Instruction Prompt (Find the SPY)

You are playing the game "Find the Spy." There are 3 players: {Player Name 1}, {Player Name 2} .... All players except one received the same word; one player, the spy, received a different word. Each player gives a short description of their word without saying it directly. Your task is to identify which player is describing a different word.

Based on a recording of the game, decide which player is the spy.

Options:
(A) Rachel
(B) Benjamin
(C) Grace
(D) ...

Reason in two stages inside `<reasoning>...</reasoning>`. First, transcribe the multi-speaker conversation in the main audio inside `<transcript>...</transcript>` placed at the very start of `<reasoning>`. Use the format `'Speaker 1: ...\nSpeaker 2: ...'`, labeling speakers in the order they first appear and joining each speaker's utterances on a single line in chronological order. Then continue reasoning grounded in that transcript, and finally provide the answer in JSON inside `<answer>...</answer>`. Example:

```
<reasoning><transcript>Speaker 1: ...\nSpeaker 2: ...</transcript>
...reasoning...</reasoning>

<answer>{"Answer": "(A) example text"}</answer>
```

Figure 21: Instruction prompt used for the Find the Spy benchmark.