AURAL: TOWARDS SPOKEN-CONTENT-AWARE SPATIAL AUDITORY SCENE UNDERSTANDING WITH LARGE AUDIO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large audio language models (LALMs) have advanced from recognizing *what is present* to understanding *where sources are*. Existing spatial LALMs can model both spoken content and general sound sources; however, these capabilities are largely treated as separate tasks, without explicitly connecting spoken content to spatially grounded sound events. This task separation limits joint understanding of speech and the surrounding auditory scene, where spoken content can be crucial for interpreting the situation. For example, a crash followed by speech may indicate very different situations depending on whether a speaker calls for help or comments on a dropped plate. To address this gap, we first curate **AuralSpeech**, a corpus for spatially localizing and transcribing speakers. We then introduce **AuralScene**, a multiple-choice question-answering dataset for spoken-content understanding, spatial grounding, and relational reasoning. Unlike existing datasets that primarily study these capabilities in isolation, AuralScene explicitly couples spoken content with spatially grounded sound events to support joint scene-level reasoning. To enable its systematic construction, we propose **Spatial Auditory Graph (SAG)**, a structured representation of spatial auditory scenes that enables the generation of diverse and controllable spatial audio scenes and question-answer pairs. Finally, we develop **AuralListener**, a baseline spatial LALM for the proposed tasks that combines monaural and spatial audio token streams, enabling controlled analysis of their complementary roles in spatial auditory understanding. Through comprehensive evaluations, we analyze how spoken content, monaural, and spatial audio representations contribute to understanding and identify remaining challenges in the proposed tasks, providing a step toward **spoken-content-aware spatial audio understanding**.

1 INTRODUCTION

Large audio language models (LALMs) combine pretrained audio encoders with large language models (LLMs) to enable instruction following and open-ended understanding of speech, environmental sounds, and music (Tang et al., 2024a; Kong et al., 2024; Ghosh et al., 2025; Chu et al., 2023). However, most LALMs rely on monaural inputs and remain unaware of where sounds originate.

Recent work extends LALMs to spatial audio by leveraging binaural or ambisonic inputs for sound source localization (Zheng et al., 2024; Jiang et al., 2026). Such capability is essential for embodied agents operating in 3D sound fields, including robots, assistive hearing devices, and AR systems (Gan et al., 2020). In this work, we focus on binaural audio, which closely matches human auditory perception and supports diverse real-world applications (Gierlich, 1992; Yang et al., 2022).

BAT (Zheng et al., 2024), the first spatial LALM, answers questions about sound classes and source locations in binaural mixtures. DSpAST (Wilkinghoff & Tan, 2026) enhances BAT's encoder by feature disentanglement. OWL (Biswas et al., 2026) introduces room impulse response reconstruction to pretrain the audio encoder. However, these studies construct scenes by randomly placing sources, lacking realistic semantic and causal relations among different events. Therefore, scenes are modeled as collections of independent sources rather than coherent situations, and these models treat speech only as a sound class, leaving the understanding of spoken content unexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Several spatial LALMs have explored spoken content understanding, but mainly with ambisonic or multi-channel recordings rather than binaural audio. Specifically, Tang et al. (Tang et al., 2024b) study direction-aware speech transcription, while PhaseCoder (Dementyev et al., 2026) further considers speaker localization, and TWNM (You et al., 2026) formulates speech content recognition as a multiple-choice task. However, similar to BAT and OWL, these methods construct scenes by randomly placing sound sources, without explicit causal or interactional structure. Due to this random placement, spoken content is not explicitly related to surrounding sound events, leaving joint reasoning about their interactions and the resulting auditory situation underexplored.

These gaps raise a natural question: *can a spatial LALM understand an auditory scene as a coherent situation*, *where spoken content and sound events are interpreted together rather than recognized in isolation*? This ability is important for real-life applications, where understanding an auditory situation often requires connecting what is said with what is happening around the speakers. To answer this question, we present **Aural**, a new dataset and benchmark suite for **spoken-content-aware spatial auditory scene understanding**, which requires jointly reasoning about speech and spatially grounded sound events (Figure 1).

Aural includes **AuralSpeech**, the binaural audio dataset that supports two tasks: joint speaker transcription and localization, and target speaker transcription conditioned on spatial cues. AuralSpeech systematically controls the number of speakers and conditioning attributes, enabling versatile evaluation of spatial speech understanding and serving as a foundational step before higher-level understanding (e.g., speech and event interactions).

However, understanding an auditory scene as a coherent situation requires connecting spoken content with surrounding events and reasoning about their interactions. To the best of our knowledge, no public dataset targets the interaction between spoken content and spatially grounded sound events. A further limitation lies in how a scene is represented. Most existing datasets typically represent scenes with captions or simple metadata, which are insufficient for generating questions that require relational reasoning (Zheng et al., 2024; Biswas et al., 2026). TWNM (You et al., 2026) extends this direction by converting simulator metadata into textual descriptions, but still treats sources as independent entities without realistic causal relations.

To address these limitations, we propose **Spatial Auditory Graph** (**SAG**), a structured representation that explicitly models sound sources along with their temporal, spatial, and causal relations. SAG serves as a unified specification from which realistic spatial auditory scenes are rendered and diverse question-answer pairs are automatically generated. Building upon SAG, we develop an automated pipeline to construct **AuralScene**, a multi-choice question-answer (MCQA) dataset for spoken-content-aware spatial auditory scene understanding. By leveraging the structured scene graph representation, our pipeline automatically synthesizes realistic auditory scenes and generates question-answer pairs that require jointly reasoning over spoken content, environmental sounds, along with their temporal, spatial, and causal relationships. AuralScene systematically evaluates spatial auditory scene understanding across different tasks, ranging from sound source perception to complex relation reasoning among different objects.

Our contributions are threefold: (1) we introduce **spoken-content-aware spatial auditory scene understanding**, a new research problem that jointly interprets spoken content, sound events, and their relationships within a coherent 3D auditory scene; (2) we present **AuralSpeech** and **AuralScene**, a dataset suite for this task, and propose **Spatial Auditory Graph (SAG)**, a structured scene representation that enables the automatic generation of coherent spatial audio scenes and diverse question-answer pairs; and (3) we develop **AuralListener**, a spatial LALM baseline for the proposed tasks. We release code, data, performance overview, and demonstrations of different QA tasks on our project page to support future research[1].

2 Related Work

Spatial Large Audio Language Models. Most existing LALMs operate on monaural audio inputs. Recently, researchers have begun to explore spatial auditory understanding with LALMs by using either First-Order Ambisonics (FOA) or binaural audio. Tang et al. (Tang et al., 2024b) incorporate spatial features from FOA signals for speaker transcription and sound source localization. For bin-

[1]Project Page: `https://anonymous.4open.science/w/Aural/`

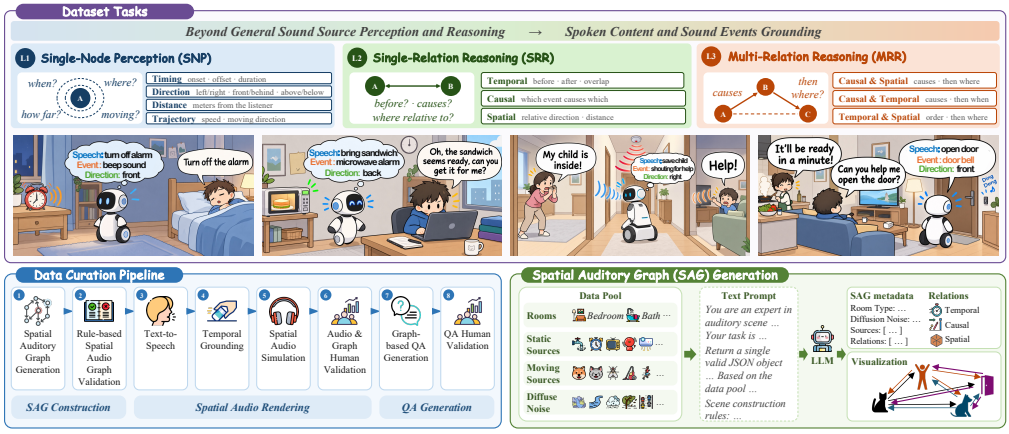


Figure 1: Overview of the AuralScene tasks and dataset curation pipeline.

aural audio, BAT (Zheng et al., 2024) supports questions about sound categories and locations, while OWL (Biswas et al., 2026) further improves spatial representation learning through room impulse response reconstruction. AudioMotionBench (Sun et al., 2025) evaluates motion and trajectory reasoning in binaural spatial audio but has not yet been open-sourced.

Spoken-Content-Aware Spatial Auditory Understanding. Spoken content is a key component of monaural LALMs, as speech often provides semantic context beyond acoustic events (Sakshi et al., 2025b; Yin et al., 2026b). Some benchmarks are proposed for jointly understanding speech and sound events, but they focus exclusively on monaural audio (Yin & Choi, 2026; Zhang et al., 2026). Recent spatial LALMs have extended speech understanding to spatial audio by localizing and transcribing speakers (Tang et al., 2024b; Dementyev et al., 2026; Zhu et al., 2026). However, these spatial models largely treat speech perception and environmental sound understanding as separate tasks, leaving their interactions unexplored. SPUR (Sakshi et al., 2025a) models the spatial and temporal alignment between speech and sound events, but does not target the semantic content of speech itself. In the binaural domain, BAT (Zheng et al., 2024) and OWL (Biswas et al., 2026) treat speech as a generic sound source without explicitly modeling its spoken content.

Spatial Auditory Graph. Scene graphs have been widely used in visual reasoning to represent entities and their relationships (Wei et al., 2024; Liu et al., 2025). In spatial audio understanding, however, there is no explicit relational representation for organizing interactions among sound sources and their associated attributes. Therefore, we introduce SAG, a structured representation that explicitly models spatially grounded sound sources, including speakers with transcribed content, together with their temporal, spatial, and causal relations. SAG provides a unified representation for generating coherent auditory scenes and constructing diverse question-answer pairs.

3 AURALSPEECH

In AuralSpeech, each sample contains one or two speakers rendered with binaural room impulse responses from SpatialSoundQA (Zheng et al., 2024) using SoundSpaces 2.0 (Chen et al., 2022). We use disjoint training and evaluation environments from the official split, resample all audio to 32 kHz, and apply loudness normalization, following (Zheng et al., 2024). Training speech comes from Common Voice 25.0 (Ardila et al., 2020) and GigaSpeech (Chen et al., 2021), while evaluation uses VCTK (Yamagishi et al., 2019), ensuring disjoint speakers and transcripts across splits. For two-speaker scenes, we require different spoken content, and for GigaSpeech, select speaker pairs with cosine distance > 0.5 using 256-dimensional WeSpeaker embeddings (Wang et al., 2023).

Speaker Transcription and Localization (STL) Task. STL requires the model to jointly predict the utterance and location of every speaker. Following the convention of BAT (Zheng et al., 2024), we discretize directions into eight octants using azimuth and elevation: $|\text{azimuth}| \leq 90°$ denotes *right* and otherwise *left*, while $\text{azimuth} \geq 0°$ denotes *front* and otherwise *behind*; $\text{elevation} \geq 0°$ denotes *above* and otherwise *below*. We use multiple STL question templates and randomly sample one per sample for robustness (Appendix A.1).



Figure 2: AuralScene jointly represents speech, sound events, and their relations for reasoning.

Target Speaker Transcription (TST) Task. TST evaluates whether a spatial LALM can selectively attend to a specified speaker. Specifically, the target is queried either by direction or by distance. Following previous work (Dementyev et al., 2026), we represent spatial conditions with natural language (e.g., *left*, *front*, and *closer*) instead of exact numerical values. As in STL, we use a pool of semantically equivalent question templates and randomly sample one for each example (Appendix A.1). The model is expected to output only the transcription of the queried speaker.

Statistics. Overall, AuralSpeech contains approximately 7.6k hours of binaural speech samples for training and 29 hours for evaluation, covering single- and two-speaker scenes with both overlapping and non-overlapping speech. The dataset provides 9.36 million training QA samples and 48k evaluation QA samples across seven controlled conditions. Details are provided in Appendix A.2.

4 AuralScene

AuralScene targets holistic auditory scene understanding, whereas AuralSpeech focuses specifically on speech. Beyond recognizing individual sounds, AuralScene requires models to jointly reason about spoken content, environmental sounds, along with their relationships, treating an auditory scene as a coherent situation rather than a collection of independent events. Building such a benchmark is challenging. Large-scale collection and annotation of real-world spatial audio are expensive, making simulation a more practical solution. However, existing simulation pipelines typically construct scenes by randomly placing sound sources, resulting in acoustically valid but semantically implausible scenes with limited relational structure. To address this limitation, we introduce SAG, a structured representation of sound sources and their interactions that enables realistic scene generation and automatic QA construction. As shown in Figure 1, we define room layouts, sound sources, and scene rules. We then use an LLM to generate SAGs for spatial audio rendering and QA generation. Figure 2 illustrates the core elements of AuralScene.

4.1 SAG Generation and Validation

Data Pool. We first assemble a data pool defining the candidate elements for scene generation, including room layouts, sound sources, and diffuse background sounds. Room layouts are adopted from Matterport3D (Chang et al., 2017), while sound sources are drawn from Hive (Li et al., 2026) and AudioSet (Gemmeke et al., 2017), with official training and evaluation splits.

Graph Generation. Conditioned on the data pool, we prompt an LLM to generate SAGs describing the room, speech transcripts, sound sources, their spatial and temporal attributes, and the relations among sources. A key design is to couple spoken content with surrounding sound events during graph generation. The LLM first constructs the sound sources and their temporal, spatial, and causal relations, and then generates speech conditioned on these relations. Each scene is required to contain at least one meaningful connection between a speaker's utterance and a non-speech event, such that the utterance can describe, refer to, explain, or respond to the event or its relation to the speaker. This construction makes spoken content informative for interpreting the surrounding auditory scene, rather than treating speech as an independent sound source. Since the source pool and relation ontology are predefined, the LLM is responsible only for generating semantically coherent scenarios while the admissible scene space remains constrained. We use different LLMs to generate SAGs for the training and evaluation splits. Details about SAG generation are in Appendix B.1.

Table 1: Comparison of AuralScene with open-sourced binaural audio understanding datasets. "SEI": Speech-Event Interactions. "Rel.": "Relation". ▲: SEI may exist in the audio, but is not explicitly defined and evaluated. ✳: the task is only supported with monaural audio.

Dataset	Train	Eval	SAG	SEI	Causal Rel.	Temporal Rel.	Spatial Rel.	Human Validation
SpatialSoundQA	✔	✔	✗	✗	✗	✗	✔	✗
BiDepth	✔	✔	✗	✗	✗	✗	✔	✗
STAR-Bench	✗	✔	✗	▲	✳	✳	✔	✔
AuralScene	✔	✔	✔	✔	✔	✔	✔	✔

Rule-based SAG Validation. Every generated graph is verified before rendering. A rule-based validator checks the schema, room and source labels, relation types, and physical consistency. Temporal relations are validated using onset/offset times; spatial relations using coordinates. Only validated graphs are used for the following audio rendering and QA generation. Details in Appendix B.2.

4.2 Spatial Audio Rendering

Speech Generation. We synthesize graph-specified speech using GPT-SoVITS (RVC-Boss, 2024), a zero-shot TTS model that generates 32 kHz audio. Speaker references are sampled from LibriSpeech (Panayotov et al., 2015), with disjoint speakers across training and evaluation splits. We transcribe the synthesized **dry** speech using Whisper-large-v3 (Radford et al., 2023), obtaining an average Word Error Rate (WER) of 1.72%, indicating high transcript fidelity before spatialization.

Temporal Grounding. Because zero-shot TTS does not provide precise duration control, synthesized speech may differ from the durations specified in the generated graphs. We therefore replace each speech duration with its measured value while preserving its onset, and iteratively shift the onsets of other sources to restore the specified temporal relations. For the few unresolved cases, we extend non-speech intervals and adjust the speeds of moving sources while preserving their trajectory lengths. We then update the affected metadata and re-validate the graphs based on rules.

Spatial Audio Simulation. We render binaural audio using SoundSpaces 2.0 (Chen et al., 2022) in Matterport3D (Chang et al., 2017). To better reflect the diverse acoustic intensities of real-world events, we assign each event class a predefined loudness range based on its typical real-world intensity. The audio simulation process is detailed in Appendix B.3.

4.3 QA Generation, Human Validation and Statistics

From each SAG, as shown in Figure 1, we generate multiple-choice QAs at three understanding levels: Single-Node Perception (SNP), Single-Relation Reasoning (SRR), and Multi-Relation Reasoning (MRR). All tasks are further divided into two categories: **with** and **without Spoken Content Understanding (SCU)**, depending on whether spoken content understanding is required.

Single-Node Perception. SNP evaluates the perception of a single attribute of an object. Specifically, we consider four groups of attributes: *timing*, *direction*, *distance*, and *trajectory*. SNP questions are generated using pre-defined templates and LLMs, with answers retrieved from the graph and distractors constrained to be physically consistent.

For *with SCU*, the target object is a speech utterance, and an LLM is used to paraphrase the spoken content into a concise description of its intent or semantic content, which is then used to construct the question and answer choices. For *without SCU*, the target object is a non-speech event and is directly identified by its event class name.

Single-Relation Reasoning. SRR evaluates reasoning over a single relation between two objects. We consider three relation types: *causal*, *temporal*, and *spatial*. For causal relations, an object may have multiple causes, and we retain all relations when constructing the corresponding questions. When either target event is speech, the question is categorized as *with SCU*, as the model must first identify the corresponding speech source from its spoken content. When both target objects are non-speech events, the target objects can be determined directly from their event classes, and the question is *without SCU*. For spatial questions, we do not query the direction of an object relative to a speaker's facing direction, since the orientation of the speaker cannot be estimated from audio.

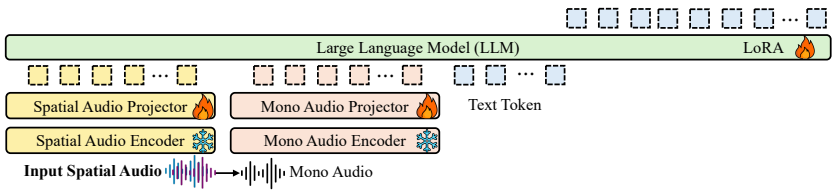


Figure 3: Overview of the proposed spatial LALM baseline, AuralListener.

Multi-Relation Reasoning. MRR is designed to jointly reason over multiple relations among objects. We consider combinations of two relation types selected from causal, temporal, and spatial relations. When at least one target event is speech, the question is categorized as *with SCU*, requiring the model to identify the relevant speech source from its spoken content before reasoning over the target relations. When none of the related objects is speech, the question is *without SCU*. As in SRR, we do not query the direction of an object relative to a speaker's facing direction.

Similar to SNP, we construct SRR and MRR questions using a fixed set of manually designed question templates for each relation type. An LLM is used to rephrase the speech utterances. Refer to Appendix B.4 for details about question generation. As summarized in Table 1, AuralScene is the first binaural dataset with SAG annotations, speech-event interactions, and relational reasoning. STAR-Bench (Liu et al., 2026) is the closest benchmark to AuralScene. However, it contains only partial real-life spatial audio without explicit definitions and evaluations of interactions between spoken content and spatially grounded sound events; it serves as an evaluation set in our study.

Human Validation. We manually verify all SAGs by building an annotation platform that visualizes each graph and provides playback of both mixture and individual sources. Annotators assess scene plausibility and audio-graph consistency. We further verify the evaluation QA set to ensure that each QA is reasonable. Details about human validation are presented in Appendix B.5.

Statistics. We initially generate 12,300 SAGs, including 12,150 for training and 150 for evaluation. Rule-based validation retains 10,082 training and 125 evaluation SAGs (83%), followed by human validation that retains 9,120 and 113 SAGs, respectively (90% of rule-valid SAGs). We further validate the evaluation QA samples against their corresponding validated SAGs, checking questions, answers, and distractors. All evaluation QA samples are retained (100%) as their answers and underlying relations are directly grounded in the validated SAGs. The final dataset contains 420k training and 2.7k evaluation QA samples across SNP, SRR, and MRR, with an evaluation set comparable in scale to STAR-Bench. Detailed statistics are presented in Appendix B.6.

5 AURALLISTENER

We introduce AuralListener, a spatial LALM designed for spoken-content-aware spatial auditory scene understanding. Existing binaural LALMs focus on independent sound sources, with limited modeling of spoken content and cross-object interactions. AuralListener provides a baseline for our benchmarks and enables controlled analysis of how different audio representations contribute.

As shown in Figure 3, AuralListener uses two branches with complementary spatial and mono audio tokens. The spatial branch uses SAGE (Biswas et al., 2026), initialized from the public checkpoint, followed by a Q-Former projector that produces 64 spatial tokens. The mono branch uses a pretrained SPEAR-XLarge encoder (Yang et al., 2025) on the channel-averaged waveform, followed by a convolutional projector that produces mono tokens at 25 Hz. The two token streams are concatenated with text tokens and fed into a LLaMA-2-7B decoder (Touvron et al., 2023).

Motivated by the effectiveness of curriculum learning in prior LALM studies (Zheng et al., 2024; Biswas et al., 2026), we adopt a three-stage curriculum that progressively increases the complexity. **Stage 1** trains the projectors using SpatialSoundQA (1-source) and AuralSpeech (1-speaker). SpatialSoundQA provides audio tagging (AT) and sound source localization, while AuralSpeech uses only the automatic speech transcription (ASR) task for speech-text alignment. **Stage 2** enables LLM fine-tuning with LoRA, introducing the full SpatialSoundQA and AuralSpeech datasets for multi-source understanding. **Stage 3** incorporates AuralScene for holistic scene reasoning over spoken content, sound events, and their relationships. Training details are provided in Appendix C.1.

Table 2: Ablation study of AuralListener (Stage 1) with different input audio tokens.

System	SpatialSoundQA (1-Source) AT (mAP↑)	DOA (ACC↑)	DP (DER↓)	AuralSpeech (1-Speaker) ASR (WER↓)
AuralListener	**28.50**	**75.68**	**26.11**	4.63
w/o Mono Token	26.27	75.27	29.59	177.83
w/o Spatial Token	7.57	23.85	35.68	**3.50**

Table 3: Spatial perception performance on SpatialSoundQA evaluation set.

System	Perception Detection (mAP↑) 1-Source	2-Source	DOA (ACC↑) 1-Source	2-Source	DP (DER↓) 1-Source	2-Source	Binary QA (ACC↑) Direction	Distance	Avg.
OWL†	26.76	12.73	78.31	**43.15**	**26.14**	43.21	71.21	86.91	**79.06**
BAT†	26.34	9.89	75.54	37.65	29.16	47.90	69.77	84.01	76.89
DSpAST‡	27.15	10.62	**78.84**	38.69	28.41	47.89	**72.52**	82.91	77.71
Kimi-Audio (M)	9.05	2.59	12.29	11.48	79.50	84.18	46.68	55.43	51.06
Qwen2-Audio (M)	12.68	3.25	11.15	12.49	77.02	81.35	49.70	47.09	48.40
MidashengLM (M)	18.05	5.25	10.86	11.78	74.92	83.56	46.99	49.00	48.00
Audio-Flamingo-3 (M)	12.45	3.21	11.42	9.74	78.55	87.32	41.83	49.42	45.63
Audio-Flamingo-Next (M)	17.08	4.06	14.12	12.42	77.91	86.74	49.22	53.52	51.37
AuralListener	**30.83**	**14.56**	77.68	39.25	26.19	**41.63**	66.15	**87.45**	76.80
w/o Mono Token	28.03	11.39	77.37	38.65	31.14	49.92	63.33	82.44	72.89
w/o Spatial Token	9.56	3.29	19.50	17.07	40.00	54.73	56.18	73.75	64.97

†: numbers from (Biswas et al., 2026); ‡: numbers from (Wilkinghoff & Tan, 2026). (M): Monaural LALMs.

6 Experiments

6.1 Implementation Details

Benchmarks. We evaluate AuralListener on in-domain and out-of-domain (OOD) benchmarks. For in-domain evaluation, we use SpatialSoundQA (Zheng et al., 2024), AuralSpeech, and AuralScene. For OOD evaluation, we use BiDepth (Biswas et al., 2026) and STAR-Bench (Liu et al., 2026).

Evaluation Metrics. Following prior work (Zheng et al., 2024), audio tagging is evaluated by mean Average Precision (mAP); predictions and labels are matched by cosine similarity of text embeddings (OpenAI, b). DOA estimation uses accuracy (ACC), and distance prediction (DP) uses Distance Error Rate (DER). For STL, we use ACC and DER for speaker localization and WER for transcription. For TST, we report Success Rate (SR) and WER on correctly identified targets (sWER), with target-speaker verification by open-sourced GPT-OSS-120B (OpenAI, a). Binary and Multi-Choice QA tasks use ACC. Metrics are in percentages; best and second-best results are marked in bold and underlined, respectively. Evaluation details are in Appendix C.2.

Other Baselines. We also evaluate representative pretrained monaural and spatial LALMs to assess their capabilities; we do not evaluate BAT, DSpAST, and OWL on speech understanding or our MCQA tasks, as these tasks are outside their original task designs and output formats. For AuralScene, we construct two pure-text baselines with an LLM (Sonnet5 (Anthropic, 2026b)). Sonnet5 (SAG) receives the SAGs together with the question and options, providing an upper bound based on the scene representation and verifying that the questions are answerable from their source SAGs. Sonnet5 (Blind) receives only the question and options, providing a lower bound based on language priors, without access to audio and scene metadata. Baseline details are in Appendix C.3.

6.2 Results and Discussions

Impact of Audio Tokens. We first conduct a Stage 1 ablation study (Table 2). The results show that the two streams are complementary: spatial tokens benefit spatial understanding, while mono tokens benefit speech recognition. SAGE's AT pretraining (*w/o Mono Token*) yields stronger AT performance than the self-supervised SPEAR-XLarge (*w/o Spatial Token*), while combining both streams achieves the best overall performance. Subsequent results use fully curriculum-trained models.

Audio Tagging and Sound Source Localization. As shown in Table 3, monaural LALMs perform substantially worse than spatial LALMs on spatial perception, despite being pretrained on large-scale audio-text corpora, suggesting that monaural pretraining alone is insufficient for spatial auditory understanding. In contrast, spatial LALMs achieve stronger DOA and DP performance, demon-

Table 4: Spatial speech understanding I: speaker transcription and localization (STL) on AuralSpeech evaluation set. Each cell reports WER↓ | ACC↑ | DER↓.

System	1-Speaker	2-Speaker Non-Overlap	2-Speaker Overlap
Kimi-Audio	**1.52** | 9.87 | 78.12	3.34 | 9.59 | 78.68	**24.22** | 8.33 | 80.89
Qwen2-Audio	42.83 | 11.98 | 91.62	54.74 | 8.97 | 92.44	61.64 | 7.74 | 94.55
MidashengLM	28.31 | 10.00 | 78.44	5.49 | 9.88 | 80.99	52.26 | 10.21 | 82.09
Audio Flamingo 3	10.82 | 0.01 | 70.53	49.23 | 0.05 | 75.53	83.89 | 0.08 | 82.33
Audio Flamingo Next	3.70 | 8.51 | 84.15	53.67 | 9.10 | 83.23	49.66 | 5.83 | 87.10
AuralListener	1.58 | **69.90** | **38.01**	**2.44** | **34.55** | **36.85**	29.60 | **31.41** | **46.36**
w/o Mono Token	117.4 | 68.90 | 51.81	117.3 | 30.53 | 62.74	109.4 | 26.09 | 65.20
w/o Spatial Token	1.54 | 14.86 | 46.98	2.12 | 15.98 | 41.34	31.11 | 14.75 | 43.87

Table 5: Spatial speech understanding II: target speaker transcription (TST) on non-overlapping and overlapping speeches of AuralSpeech evaluation set. Each cell reports SR↑ | sWER↓ | WER↓. Con-DOA and Con-DIS denote targets specified by direction and by distance queries, respectively.

System	Non-Overlap Con-DOA	Non-Overlap Con-DIS	Overlap Con-DOA	Overlap Con-DIS
Kimi-Audio	11.16 | 1.84 | 91.94	11.72 | 2.10 | 92.65	45.22 | 4.56 | 64.11	45.59 | **3.86** | 63.47
Qwen2-Audio	21.47 | 5.78 | 86.56	29.74 | 6.29 | 80.59	37.98 | 11.76 | 71.70	37.72 | 12.32 | 72.67
MidashengLM	6.13 | 6.84 | 96.78	9.02 | 3.53 | 94.94	31.84 | 14.50 | 79.45	33.99 | 11.76 | 72.98
Audio Flamingo 3	1.43 | 21.60 | 132.8	4.15 | 17.21 | 134.8	30.92 | 18.32 | 85.07	33.51 | 15.58 | 83.14
Audio Flamingo Next	4.35 | 5.72 | 98.58	14.94 | 2.55 | 90.35	34.47 | 8.65 | 71.24	36.88 | 7.93 | 71.10
AuralListener	**50.68** | **0.71** | **58.05**	**88.95** | **1.02** | **13.37**	**46.53** | **4.44** | **62.71**	**55.42** | 6.04 | **46.95**
w/o Mono Token	0.04 | 120.0 | 155.6	0.00 | - | 161.1	0.03 | 69.23 | 141.3	0.09 | 107.0 | 143.6
w/o Spatial Token	39.49 | 0.65 | 71.49	78.40 | 0.96 | 25.79	41.16 | 4.55 | 69.10	49.67 | 6.24 | 55.28

strating the benefit of spatial audio representations. Although AuralListener slightly underperforms compared to OWL and DSpAST on direction and distance perception, it achieves the best detection performance. Ablation variants show that removing spatial tokens causes a substantial drop in spatial perception, with DOA accuracy falling to 19.50% and 17.07%, while removing mono tokens has a smaller effect on DOA. Combining both streams therefore provides more balanced performance across spatial perception and Binary QA.

Spatial Speech Understanding. For STL, as shown in Table 4, monaural LALMs achieve relatively low WER in most settings but exhibit poor spatial localization with DOA ACC below 15% and DP DER above 70% in most conditions. AuralListener achieves stronger spatial localization across all settings while maintaining competitive transcription performance. These results highlight that AuralListener jointly models speech content and spatial information, while conventional monaural LALMs primarily focus on the spoken content. For TST (Table 5), AuralListener shows the highest SR. However, DOA-conditioned and overlapping TST still remain challenging for all models, with SR around 50%, consistent with findings from prior studies (Dementyev et al., 2026) and highlighting a persistent challenge for future work. The ablations further confirm that spatial tokens support target identification, while mono tokens are important for transcription.

Spatial Auditory Understanding. As shown in Table 6, AuralListener reaches 58.62% and 65.00% overall ACC without and with SCU, respectively. Since AuralListener is trained on AuralScene while the other LALMs are evaluated zero-shot, the contribution of each token stream is better assessed through the ablations rather than this direct comparison. Sonnet5 (SAG) achieves over 97% accuracy, demonstrating that SAGs are well aligned with the corresponding QAs, while Sonnet5 (Blind) remains close to the random baseline, confirming that AuralScene questions cannot be answered from language priors alone. The human reference shows that listeners readily answer timing, direction, causal, and temporal questions, while absolute distance and inter-source spatial relations are harder. AuralListener is comparable to or above humans on distance, trajectory, and speech-anchored spatial relations, but lags behind on other tasks. The ablation study shows that spatial tokens are crucial for spatial grounding, while mono tokens become particularly important when understanding spoken content is required. Scene-level 95% confidence intervals are reported in Appendix D.2.2. Details about human evaluation are in Appendix D.2.3.

Out-of-Distribution Evaluation. For a fairer comparison with other baselines, we further evaluate AuralListener in a zero-shot condition on STAR-Bench. As shown in Table 7, AuralListener achieves competitive results across both foundational perception and spatial reasoning, demonstrating that its spatial representations generalize beyond our synthetic training scenes. Additional OOD results on BiDepth are provided in Appendix C.4.

Table 6: Spatial auditory understanding performance on AuralScene evaluation set (ACC ↑).

System	SNP				SRR			MRR			Micro
	Timing	Direction	Distance	Trajectory	Causal	Temporal	Spatial	C&T	C&S	T&S	Overall
Random	25.00	25.00	25.00	25.00	25.00	33.33	25.00	33.33	25.00	25.00	26.04
without Spoken Content Understanding											
Human	70.67	85.00	61.00	78.50	90.00	93.00	65.00	92.00	54.00	64.00	72.85
Sonnet5 (SAG)	98.00	100.00	98.00	99.50	98.00	98.00	99.00	98.00	99.00	95.00	98.38
Sonnet5 (Blind)	32.33	28.00	23.00	22.50	34.00	36.00	20.00	24.00	21.00	26.00	26.54
Kimi-Audio	23.67	24.00	27.00	26.00	32.00	37.00	24.00	34.00	24.00	21.00	25.92
Qwen2-Audio	23.67	39.00	22.00	25.50	**56.00**	42.00	30.00	28.00	27.00	39.00	30.23
MidashengLM	12.00	33.00	28.00	41.50	44.00	36.00	34.00	22.00	36.00	33.00	29.69
Audio Flamingo 3	29.67	22.00	35.00	50.00	46.00	45.00	24.00	**38.00**	26.00	31.00	33.69
Audio Flamingo Next	23.67	17.00	26.00	51.00	42.00	47.00	20.50	22.00	38.00	25.00	30.69
AuralListener	**48.00**	**66.00**	**67.00**	**83.50**	50.00	**63.00**	**59.00**	**38.00**	**45.00**	**48.00**	**58.62**
w/o Mono Token	49.67	64.00	65.00	84.00	42.00	60.00	57.00	38.00	42.00	47.00	57.62
w/o Spatial Token	44.00	42.00	56.00	69.00	48.00	61.00	34.50	42.00	39.00	38.00	47.69
with Spoken Content Understanding											
Human	81.67	80.00	58.00	-	91.00	95.00	67.50	94.00	56.00	61.50	73.79
Sonnet5 (SAG)	99.33	100.00	100.00	-	95.00	99.00	98.00	97.00	95.00	93.50	97.29
Sonnet5 (Blind)	29.00	27.00	25.00	-	39.00	28.00	26.50	27.00	29.50	25.00	28.21
Kimi-Audio	24.67	27.00	20.00	-	43.00	39.00	28.50	43.00	21.50	28.00	28.71
Qwen2-Audio	26.33	27.00	28.00	-	59.00	44.00	21.00	45.00	31.50	25.00	31.21
MidashengLM	24.00	24.00	26.00	-	57.00	44.00	29.00	46.00	23.50	30.50	31.07
Audio Flamingo 3	34.67	29.00	37.00	-	65.00	42.00	28.00	51.00	32.00	23.00	35.29
Audio Flamingo Next	32.00	28.00	46.00	-	**70.00**	**58.00**	24.50	59.00	27.00	21.50	35.93
AuralListener	**59.33**	**61.00**	**72.00**	-	61.00	55.00	**70.00**	**64.00**	**63.00**	**76.50**	**65.00**
w/o Mono Token	55.67	52.00	51.00	-	54.00	48.00	61.00	52.00	49.00	63.50	55.07
w/o Spatial Token	58.00	34.00	52.00	-	51.00	57.00	42.50	63.00	45.00	49.50	50.36

Table 7: Out-of-distribution performance: zero-shot performance on STAR-Bench (ACC ↑).

System	Foundational Perception			Spatial Reasoning		
	Azimuth	Elevation	Distance	Localization	Relation	Trajectory
BAT†	26.04	41.67	23.53	0.00	0.00	0.00
Kimi-Audio†	18.12	38.19	18.13	27.56	38.94	44.03
Qwen2-Audio†	22.50	48.16	12.75	21.32	24.78	15.09
MiDashengLM†	20.63	48.26	29.90	43.11	45.43	46.23
Audio Flamingo 3†	27.92	28.82	32.84	37.22	38.35	44.03
Audio Flamingo Next	25.00	50.00	49.02	18.02	40.71	45.28
AuralListener	**39.58**	**52.78**	**50.98**	**60.78**	**53.98**	**55.66**

†: numbers from (Liu et al., 2026).

Impact of Speech Supervision in AuralSpeech. To evaluate the contribution of explicit spatial speech understanding supervision, we progressively remove the supervision signals provided by AuralSpeech. Specifically, we first simplify AuralSpeech to standard single-speaker ASR by removing the STL and TST supervision, and then eliminate AuralSpeech entirely.

As shown in Figure 4, removing AuralSpeech supervision degrades performance on AuralScene, particularly for the QAs with SCU. ASR-only supervision recovers part of this gap, while the full STL and TST supervision achieves the best overall performance. These results indicate that AuralSpeech supervision provides additional benefits for grounding spoken content with spatial and relational information.

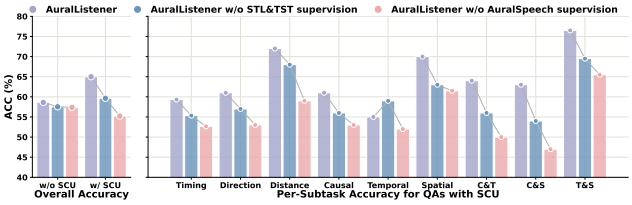


Figure 4: Impact of speech supervision on AuralScene.

Case Studies. See Appendices D and E for detailed case studies, limitations, and future work.

7 Conclusion

In this work, we introduce spoken-content-aware spatial auditory scene understanding, along with AuralSpeech and AuralScene for spatial speech and scene understanding. We also develop spatial auditory graphs for controllable scene and QA generation, and AuralListener as a spatial LALM baseline. Experiments show that spatial and monaural representations complement each other, while speech supervision improves spoken-content grounding in spatial and relational contexts.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets: Claude Opus 4.8 and Claude Fable 5 generated the Spatial Auditory Graphs for the training and evaluation splits, respectively; GPT-5.6 Luna and GPT-5.6 Terra generated training and evaluation QA pairs, speech paraphrases, and causal-answer distractors; and GPT-SoVITS synthesized the speech in AuralScene. We used GPT-OSS-120B as an LLM judge for TST evaluation. Additionally, we used generative AI tools for language revision and for creating illustrative figures.

We have not used generative AI tools to develop the research methodology, implement methods, or interpret results, and the formulation or proof of mathematical claims is not applicable to this work. We have reviewed all AI-assisted work: all SAGs and all 2,700 evaluation QA pairs were verified by human annotators, and the LLM judge was validated against human labels on 200 samples with 100% agreement. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our study does not involve the collection of personal or sensitive information. The speech data in AuralSpeech are derived from publicly available datasets and used in accordance with their respective licenses and terms of use. The speech in AuralScene is synthetically generated using text-to-speech models and does not include recordings of real participants in the constructed scenes. Human annotation is conducted by paid annotators for dataset quality validation. We release data where permitted by the licenses of the underlying resources, while complying with applicable privacy and data-sharing requirements.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed datasets, data generation and validation procedures, model architecture, training strategy, and evaluation protocols in the main paper and appendix. Implementation details, training configurations, prompt templates, and additional experimental results are also provided in the appendix. We also provide code, data, performance overview and detailed demonstrations (including audio) of various QA tasks on our project page.

REFERENCES

Anthropic. Claude fable 5. URL https://platform.claude.com/docs/en/models/fable-5/overview.

Anthropic. Introducing claude opus 4.8, May 2026a. URL https://www.anthropic.com/news/claude-opus-4-8.

Anthropic. Introducing claude sonnet 5, June 2026b. URL https://www.anthropic.com/news/claude-sonnet-5.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4218–4222, 2020.

Subrata Biswas, Mohammad Khan, and Bashima Islam. Owl: Geometry-aware spatial reasoning for audio large language models. In *International Conference on Learning Representations*, volume 2026, pp. 20685–20710, 2026.

Simon Carlile and Johahn Leung. The perception of auditory motion. *Trends in hearing*, 20: 2331216516644254, 2016.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor

environments. In *2017 International Conference on 3D Vision (3DV)*, pp. 667–676, 2017. doi: 10.1109/3DV.2017.00081.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Artem Dementyev, Wazeer Zulfikar, Sinan Hersek, Pascal Getreuer, Anurag Kumar, and Vivek Kumar. Phasecoder: Microphone geometry-agnostic spatial audio understanding for multimodal llms. *arXiv preprint arXiv:2601.21124*, 2026.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*, 2025.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9701–9707. IEEE, 2020.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19358–19405. PMLR, 13–19 Jul 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, Siddharth Gururani, Sang-gil Lee, Jaehyeon Kim, Aya Aljafari, et al. Audio flamingo next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026a.

Sreyan Ghosh, Arushi Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Advances in Neural Information Processing Systems*, 38:41819–41886, 2026b.

HW Gierlich. The application of binaural technology. *Applied acoustics*, 36(3-4):219–243, 1992.

Emanuël AP Habets, Israel Cohen, and Sharon Gannot. Generating nonstationary multisensor signals under a spatial coherence constraint. *The Journal of the Acoustical Society of America*, 124 (5):2911–2917, 2008.

Xilin Jiang, Hannes Gamper, and Sebastian Braun. Sci-phi: A large language model spatial audio descriptor. *IEEE Open Journal of Signal Processing*, 2026.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: a novel audio language model with few-shot learning and dialogue abilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kai Li, Jintao Cheng, Chang Zeng, Zijun Yan, Helin Wang, Zixiong Su, Bo Zheng, and Xiaolin Hu. A semantically consistent dataset for data-efficient query-based universal sound separation. *arXiv preprint arXiv:2601.22599*, 2026.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Ding Wang, and Botian Shi. Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 981–992. IEEE, 2025.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, et al. Star-bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 134703–134731, 2026.

OpenAI. Gpt-oss-120b, a. URL https://huggingface.co/openai/gpt-oss-120b.

OpenAI. text-embedding-ada-002, b. URL https://developers.openai.com/api/docs/models/text-embedding-ada-002.

OpenAI. Gpt-5.6 luna, c. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna.

OpenAI. Gpt-5.6 terra, d. URL https://developers.openai.com/api/docs/models/gpt-5.6-terra.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

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

RVC-Boss. GPT-SoVITS. https://github.com/RVC-Boss/GPT-SoVITS, 2024. GitHub repository.

S Sakshi, Vaibhavi Lokegaonkar, Neil Zhang, Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. Spur: A plug-and-play framework for integrating spatial audio understanding and reasoning into large audio-language models. *arXiv preprint arXiv:2511.06606*, 2025a.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025b.

Zhe Sun, Yujun Cai, Jiayu Yao, and Yiwei Wang. Audiomotionbench: Evaluating auditory motion perception in audio llms. *arXiv preprint arXiv:2511.13273*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024a.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. Can Large Language Models Understand Spatial Audio? In *Interspeech 2024*, pp. 4149–4153, 2024b. doi: 10.21437/Interspeech.2024-2419.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Hongji Wang, Chengdong Liang, Shuai Wang, Zhengyang Chen, Binbin Zhang, Xu Xiang, Yanlei Deng, and Yanmin Qian. Wespeaker: A research and production oriented speaker embedding learning toolkit. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James Kwok, and Yu Zhang. Gita: Graph to visual and textual integration for vision-language graph reasoning. *Advances in neural information processing systems*, 37:44–72, 2024.

Kevin Wilkinghoff and Zheng-Hua Tan. Dspast: Disentangled representations for spatial audio reasoning with large language models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14747–14751. IEEE, 2026.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive*, 2019.

Jing Yang, Amit Barde, and Mark Billinghurst. Audio augmented reality: A systematic review of technologies, applications, and future research directions. *journal of the audio engineering society*, 70(10):788–809, 2022.

Xiaoyu Yang, Yifan Yang, Zengrui Jin, Ziyun Cui, Wen Wu, Baoxiang Li, Chao Zhang, and Phil Woodland. Spear: A unified ssl framework for learning speech and audio representations. *arXiv preprint arXiv:2510.25955*, 2025.

Han Yin and Jung-Woo Choi. Can large audio language models understand audio well? speech, scene and events understanding benchmark for lalms. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13042–13046. IEEE, 2026.

Han Yin, Yafeng Chen, Chong Deng, Luyao Cheng, Hui Wang, Chao-Hong Tan, Qian Chen, Wen Wang, and Xiangang Li. Speakerlm: End-to-end versatile speaker diarization and recognition with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34467–34475, 2026a.

Han Yin, Yang Xiao, Younghoo Kwon, Ting Dang, and Jung-Woo Choi. Focus then listen: An empirical study of plug-and-play audio enhancer for noise-robust large audio language models. In *ICML 2026 Workshop on Machine Learning for Audio (Learning to Listen)*, 2026b. URL https://mlforaudioworkshop.github.io/.

Yuhuan You, Lai Wei, Xihong Wu, and Tianshu Qu. The world is not mono: Enabling spatial understanding in large audio-language models. *arXiv preprint arXiv:2601.02954*, 2026.

Pavel Zahorik, Douglas S Brungart, Adelbert W Bronkhorst, et al. Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica united with Acustica*, 91(3): 409–420, 2005.

Pengfei Zhang, Hoang H Nguyen, Kazi Shaharair Sharif, Yutong Song, Wenjun Huang, Henry Peng Zou, Pinxin Liu, Honghui Xu, and Amir M Rahmani. From sounds to scenes: A benchmark for evaluating context-aware auditory scene understanding in large audio language models. *arXiv preprint arXiv:2606.25391*, 2026.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. Bat: learning to reason about spatial sounds with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zhiyuan Zhu, Yixuan Chen, Yiwen Shao, Wenxiang Guo, Changhao Pan, Yu Zhang, Yuxiang Wang, Wei Liu, Houhua Zhang, Chengkuan Zeng, et al. Spatial-omni: Spatial audio understanding integration in multimodal llms via foa encoding. *arXiv preprint arXiv:2606.10738*, 2026.

APPENDIX CONTENTS

A: **Details of AuralSpeech**

A.1: Question Prompts

A.2: Statistics

B: **Details of AuralScene**

B.1: Spatial Auditory Graph (SAG) Generation

B.2: Rule-based SAG Validation

B.3: Spatial Audio Simulation

B.4: QA Generation

B.5: Human Validation

B.6: Statistics

C: **Details of Experiments**

C.1: Training Configurations

C.2: Evaluation Metrics

C.3: Baselines

C.3.1: LALM Baselines

C.3.2: Pure-Text Baselines

C.4: Evaluation on BiDepth

D: **Case Studies**

D.1: AuralSpeech

D.2: AuralScene

D.2.1: Performance Breakdown

D.2.2: Confidence Interval Analysis

D.2.3: Human Evaluation

E: **Limitations and Future Work**

A Details of AuralSpeech

A.1 Question Prompts

In AuralSpeech, both STL and TST tasks are posed to the model as natural-language questions. To prevent the model from overfitting to a single phrasing, we write a pool of paraphrased templates for each task and sample one uniformly at random for every scene. All templates are semantically equivalent within a task and differ only in surface wording.

For STL, the model is required to output one transcription and one spatial location for each speaker. Each speaker is represented by a paired *Transcription* and *Location* field, where the location specifies the azimuth direction, elevation, and distance from the listener.

For example, for a single-speaker scene, the output follows the format:

Transcription 1: This place is good. *Location 1*: left, front, above; 2.15 m.

For a two-speaker scene, the model outputs the two speaker entries in the same format, separated by a vertical bar. For non-overlapping speech, the speaker who speaks first is assigned as Speaker 1, and the other speaker as Speaker 2. For overlapping speech, where the temporal order is ambiguous, the two speaker indices are assigned randomly. An example is as follows:

Transcription 1: This place is good. *Location 1*: left, front, above; 2.15 m | *Transcription 2*: Of course. *Location 2*: right, front, below; 3.26 m.

For TST, the model is required to output only the transcription of the target speaker, without any additional information. This output format isolates the target-speaker transcription capability: given a spatial query, the model must first identify the corresponding speaker and then transcribe their speech, while avoiding additional output fields that are not part of the task.

The full template pools are listed below.

Question Prompts for Speaker Transcription and Localization (STL)

1. What is everyone saying and where are they?
2. Transcribe everything you hear and tell me where each speaker is located.
3. Give me a full transcript of all the speech along with each speaker's position.
4. Write down what each person says and identify where they are.
5. Transcribe all the speakers and report their locations.
6. What are the people saying, and from which locations?
7. Provide the transcript of every utterance together with the speaker's position.
8. Listen to the audio, transcribe each voice, and state where it comes from.
9. For every speaker in the scene, give me their words and their location.
10. Transcribe the whole conversation and specify the position of each participant.
11. Tell me what each speaker says and where each of them is situated.
12. Report everything being said and the direction each voice comes from.
13. Can you transcribe all speech in the audio and locate every speaker?
14. Write out each person's speech and annotate it with their spatial position.
15. Identify where each speaker is and provide a transcript of what they say.
16. Give the content of each utterance along with the location of its speaker.
17. Transcribe every voice you hear and describe where each one is located.
18. What is being said by each person, and where is each person positioned?
19. Produce a transcript for all speakers, including their positions in the scene.
20. Capture everyone's words and tell me the location of each speaker.

Question Prompts for Target Speaker Transcription (TST)

Closer Speaker:

1. Please transcribe the speech spoken by the person who is closer to you.
2. Transcribe what the nearer speaker is saying.
3. Please write down the words spoken by the closer speaker.
4. What is the person closer to you saying? Please transcribe it.
5. Transcribe the speech from the speaker who is at a shorter distance.

Question Prompts for Target Speaker Transcription (TST) (continued)

6. Please transcribe the voice coming from the nearer person.
7. Write down what the closer person is saying.
8. Transcribe the utterance of the speaker who is closer to you.
9. Please provide the transcript of the nearer speaker.
10. What does the person closer to you say? Transcribe it.
11. Transcribe the speech produced by the closer speaker.
12. Please write the words spoken by the speaker nearer to you.
13. Transcribe what the closest person is saying.
14. Please provide a transcript of the speech from the closer person.
15. What is being said by the nearer speaker? Transcribe it.
16. Transcribe the words spoken by the person at a closer distance.
17. Please transcribe the sentence spoken by the closer speaker.
18. Write down the speech from the person who is closer to you.
19. Transcribe the audio spoken by the nearer person.
20. Please provide the words spoken by the closer speaker.

Farther Speaker:
1. Please transcribe the speech spoken by the person who is farther away from you.
2. Transcribe what the more distant speaker is saying.
3. Please write down the words spoken by the farther speaker.
4. What is the person farther away saying? Please transcribe it.
5. Transcribe the speech from the speaker who is at a greater distance.
6. Please transcribe the voice coming from the farther person.
7. Write down what the more distant person is saying.
8. Transcribe the utterance of the speaker who is farther from you.
9. Please provide the transcript of the farther speaker.
10. What does the person farther away say? Transcribe it.
11. Transcribe the speech produced by the more distant speaker.
12. Please write the words spoken by the speaker farther away.
13. Transcribe what the farthest person is saying.
14. Please provide a transcript of the speech from the farther person.
15. What is being said by the more distant speaker? Transcribe it.
16. Transcribe the words spoken by the person at a farther distance.
17. Please transcribe the sentence spoken by the farther speaker.
18. Write down the speech from the person who is farther away.
19. Transcribe the audio spoken by the more distant person.
20. Please provide the words spoken by the farther-away speaker.

A conditioned direction:
1. Please transcribe the speaker {direction}.
2. What is the speaker {direction} saying? Please transcribe it.
3. Transcribe the speech from the speaker {direction}.
4. Please write down what the speaker {direction} is saying.
5. What does the speaker {direction} say? Transcribe it.
6. Transcribe the utterance of the speaker {direction}.
7. Please provide a transcript of the speaker {direction}.
8. Write down the words spoken by the speaker {direction}.
9. Transcribe the audio from the speaker {direction}.
10. Please transcribe what is being said by the speaker {direction}.
11. What is being said by the speaker {direction}? Please transcribe it.
12. Transcribe the sentence spoken by the speaker located {direction}.
13. Please write the words of the speaker positioned {direction}.
14. Provide a transcription of the speaker {direction}.
15. Transcribe the voice of the speaker {direction}.
16. Please transcribe the speech of the speaker {direction}.
17. Write down the speech from the speaker {direction}.
18. What words does the speaker {direction} say? Please transcribe them.
19. Please provide the transcript of the speaker {direction}.
20. Transcribe what the speaker positioned {direction} is saying.

PS: The {direction} field is instantiated with one of six spatial descriptions that can distinguish the two speakers: on your left, on your right, in front of you, behind you, above you, and below you.

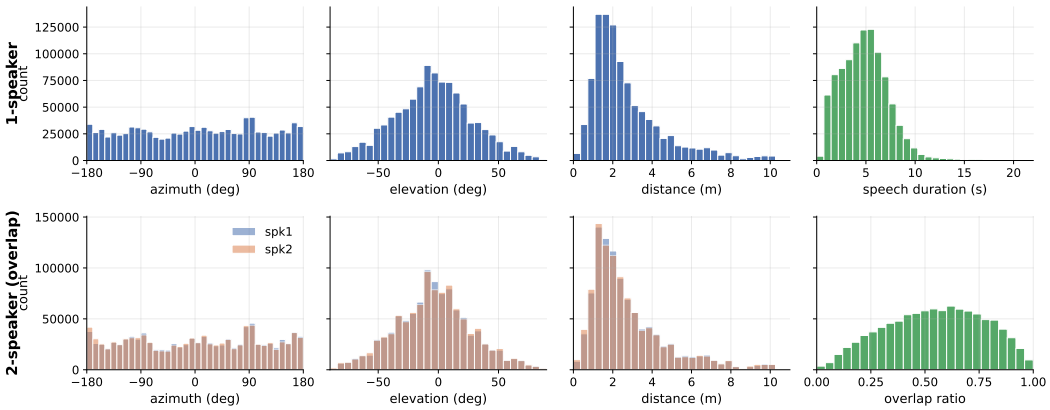


Figure 5: Data distribution of AuralSpeech.

Table 8: Statistics of AuralSpeech.

Task	# Speaker	Overlap	Condition	# Sample (train/eval)	Duration (h) (train/eval)
STL	1	–	–	1,337,481 / 6,920	1,804.00 / 7.01
	2	✗	–	1,337,122 / 6,920	3,605.15 / 14.06
	2	✔	–	1,337,122 / 6,920	2,234.01 / 8.30
TST	2	✗	Direction	1,337,122 / 6,920	3,605.15 / 14.06
	2	✗	Distance	1,337,122 / 6,920	3,605.15 / 14.06
	2	✔	Direction	1,337,122 / 6,920	2,234.01 / 8.30
	2	✔	Distance	1,337,122 / 6,920	2,234.01 / 8.30

A.2 STATISTICS

Table 8 summarizes the composition of AuralSpeech. The training set contains roughly 1.34 million samples per condition, totaling over 1,800 hours for one-speaker scenes and 2,200 to 3,600 hours for two-speaker scenes. The non-overlap condition contains longer audio samples because the two utterances are rendered sequentially rather than concurrently. Here, "concurrent" only means that both utterances are active at the same time, not that they overlap for 100%, as silence may exist within a sample. The same rendered scenes are shared across the two tasks: the two-speaker configurations for STL are reused for TST, with direction and distance conditions differing only in how the target speaker is specified. This ensures that performance differences reflect task requirements rather than variations in the underlying audio.

For evaluation, each condition contains exactly 6,920 samples, enabling balanced comparisons across conditions. The seven conditions independently vary the number of speakers, speech overlap, and target-conditioning attribute, providing controlled coverage of different scenarios.

Figure 5 visualizes the distribution using one million randomly sampled examples, illustrating the diversity of the generated spatial speech scenes. In two-speaker scenes, we define the overlap ratio as the fraction of speech-active frames during which both speakers are simultaneously active. Speech activity is estimated at a 10 ms resolution using a pretrained neural voice activity detector (Silero VAD[2]) to avoid bias from leading and trailing silence.

[2] https://github.com/snakers4/silero-vad

B Details of AuralScene

B.1 Spatial Auditory Graph (SAG) Generation

In this work, we use an LLM to generate the metadata of each SAG. We provide a fixed system instruction that specifies the coordinate system, scene construction rules, and required JSON schema. For each generation request, we randomly sample a condition from our data pools, including three candidate rooms with their room types and dimensions, three diffuse-noise classes, ten static sound-event classes, ten moving sound-event classes, a scene duration between 5 and 30 seconds, and a listener ear height between 1.5 and 1.9 meters.

Given these candidates, the LLM is instructed to construct a coherent auditory scene by selecting a room, determining the number and configuration of sound sources, placing one or two speakers and non-speaker events in a 3D space, assigning active time intervals to each source, and specifying the relations between source pairs. For scenes containing speakers, the LLM also generates realistic dialogue and associates each utterance with its corresponding speaker.

The model outputs only structured scene metadata, including Cartesian source positions, onset and offset times, speaker transcriptions, and pairwise relations. This process provides a controlled yet flexible way to generate diverse SAGs while ensuring that each scene satisfies the predefined spatial, temporal, and relational constraints. For training SAGs, we use Opus 4.8 (Anthropic, 2026a), while for evaluation SAGs, we use Fable 5 (Anthropic), introducing a different source of world knowledge and scenario priors across the two splits. The system instruction of the LLM is defined as follows:

System Instruction for LLM-Based Spatial Auditory Graph Generation

You are an expert in auditory scene understanding. Your task is to generate the metadata of a Spatial Auditory Graph (SAG) for spatial audio understanding. An SAG describes all sound sources in a scene and the relations among them. Return a single valid JSON object that follows the template at the end. No extra text, no markdown code fences, no comments.

Coordinate System
The listener is at the origin [0, 0, 0]. All distances are in meters.
- Cartesian: [x, y, z], where +x = right, +y = above, +z = behind (therefore -z = front).

Inputs
The user will provide the following variables in their message:
- ROOM LIST: candidate rooms; each entry contains a room type and a room size. [size x, size y, size z] in meters (width along x, height along y, depth along z).
- NOISE LIST: candidate diffuse noises.
- STATIC EVENT LIST: candidate static (non-moving) sound events.
- MOVE EVENT LIST: candidate moving sound events.
- DURATION: total scene duration in seconds.
- EAR HEIGHT: the listener's ear height above the floor, in meters.

Scene Construction Rules
1. Room: choose exactly one room from ROOM LIST, and copy its type and size into "Room Type" and "Room Size" in the output.
2. Room bounds: the listener stands at the center of the room, with ears at EAR HEIGHT above the floor. Since the origin is at the listener's ears, given "Room Size" = [size x, size y, size z], every source position [x, y, z] must satisfy (y = -EAR HEIGHT is the floor, y = size y - EAR HEIGHT is the ceiling):
-size x/2 $\leq$ x $\leq$ size x/2
-EAR HEIGHT $\leq$ y $\leq$ size y - EAR HEIGHT
-size z/2 $\leq$ z $\leq$ size z/2
No source may be placed outside the room.
3. Diffuse noise: set to null for quiet environment; Or choose one or multiple noises from NOISE LIST. Every chosen noise must be acoustically plausible for the chosen room (e.g., traffic noise for a room facing a street).
4. Sound events: choose one or more static events from STATIC EVENT LIST and/or moving events from MOVE EVENT LIST. Only use classes that appear in the given lists.
5. Speech: the scene must contain either (a) one person speaking, or (b) two people having a

System Instruction for Spatial Auditory Graph Generation (continued)

conversation. Choose one situation. Each speaker is a separate source with "Class": "speaker". In a conversation, alternate turns with realistic timing (turns may occasionally overlap slightly). The spoken content must be conditioned on the selected non-speech sound events and their relations, rather than generated independently.
6. Timing: "Onset" and "Offset" are in seconds relative to the start of the scene. For every source, $0 \leq$ Onset $<$ Offset $\leq$ DURATION.
7. Position: every source must have "Position" = [x, y, z] relative to the listener. For a moving source, this field describes its STARTING location.
8. Moving sources: if "Static" is false, fill in "Move Direction" and "Move Speed". The source moves from its starting position in "Move Direction" at "Move Speed" for its whole active interval (Offset - Onset seconds). Its ENTIRE trajectory must stay inside the room bounds defined in Rule 2, i.e., start position + speed · (Offset - Onset) along the moving axis must not exceed the room boundary. Verify this before finalizing the position, speed, and duration.
9. Static sources: if "Static" is true, "Move Direction" and "Move Speed" must both be null.
10. Speaker height: Vary each speaker's height naturally: some sitting (mouth below ear level, negative y), some standing (roughly at or slightly above ear level).
11. Plausibility: the combination of room, noises, events, positions, and speech content must form a coherent, realistic everyday scene, with each source at a physically sensible height (e.g., footsteps on the floor, a ceiling fan near the ceiling).

Relations
Before writing the scene, especially the spoken content, decide the relations between sources. The spoken content must meaningfully describe, refer to, explain, or respond to at least one non-speech sound event or its temporal, spatial, or causal relation to the speaker. Speech should reflect these relations naturally. All three relation types describe THE OBJECT RELATIVE TO THE ANCHOR:
1. Temporal Relation: before / after / overlap
- before: the Object's interval ends before the Anchor's interval begins.
- after: the Object's interval begins after the Anchor's interval ends.
- overlap: their intervals overlap in time (The overlap time should be no less than 50% of the shorter event duration).
- Must be consistent with the sources' Onset / Offset values.
2. Spatial Relation: a list of exactly three items: one of left / right, one of front / behind, one of above / below, describing where the Object is relative to the Anchor's position, e.g., ["left", "front", "above"].
- For moving sources, use the starting position.
- Must be consistent with the sources' coordinates.
3. Causal Relation: cause / result / independent
- cause: the Object causes the Anchor.
- result: the Object is a result of the Anchor.
- independent: no causal link.

Output Template
Return exactly this structure, as valid JSON (straight double quotes, null instead of None, lowercase true/false). Template:

```
{
    "Room Type": "string, chosen from ROOM LIST",
    "Room Size": [0.0, 0.0, 0.0],
    "Diffuse Noise": ["chosen from NOISE LIST; empty list [] if none"],
    "Duration": 0.0,
    "Sources": [
                {
                    "ID": 1,
                    "Class": "speaker, or a sound event class from the given lists",
                    "Transcription": "spoken content for speech; null for sound events",
                    "Onset": 0.0,
                    "Offset": 0.0,
                    "Position": [0.0, 0.0, 0.0],
                    "Azimuth": 0.0,
                    "Elevation": 0.0,
                    "Distance": 0.0,
                    "Static": true,
                    "Move Direction": null,
```

System Instruction for Spatial Auditory Graph Generation (continued)

```
                    "Move Speed": null
                }, ...
            ],
    "Relations": [
                {
                    "Anchor": "ID Class, e.g., 1 speaker",
                    "Relations": [
                                    {
                                      "Object": "ID Class, e.g., 2 dog bark",
                                      "Temporal Relation": "before | after | overlap",
                                      "Spatial Relation": ["left | right", "front | behind", "above |
                                                           below"],
                                      "Causal Relation": "cause | result | independent"
                                    }
                                ]
                    }, ...
                ],
    "Ear Height": 0.0,
}

## Field notes:
- "ID": integer, unique per source, starting from 1.
- "Room Size": [size x, size y, size z] in meters, copied from the chosen ROOM LIST entry.
- "Move Direction": one of "Right", "Left", "Front", "Behind", "Above", "Below", or null for static
sources.
- "Move Speed": float in m/s (realistic values, e.g., 0.5-2.0 for walking), or null for static sources.
- Every source ID referenced in "Relations" must exist in "Sources", and the "ID Class" string must
match that source's "ID" and "Class" exactly.
```

B.2 Rule-based SAG Validation

Before rendering, each generated SAG is checked by a rule-based validator to ensure that it can be reliably converted into an audio scene. First, we verify the **schema**, requiring each graph to be a valid JSON object that follows the predefined schema. Second, we check **room and source labels** against the corresponding data pools, ensuring that every room and sound source can be matched to an available resource for rendering. Third, we verify the **relation types** and required relation fields to ensure that all specified relations are complete and valid. Finally, we check **physical consistency** using the scene metadata. For example, each source must satisfy $0 \leq \text{onset} < \text{offset} \leq \text{duration}$, and the trajectory of a moving source must remain within the room boundaries. Temporal and spatial relations are further verified by computing them from the corresponding timestamps and coordinates. Only SAGs that pass all checks are retained for audio rendering and subsequent QA generation.

B.3 Spatial Audio Simulation

To simulate realistic acoustic scenes, we render binaural room impulse responses (BRIRs) with the geometric acoustic simulator, SoundSpaces 2.0 (Chen et al., 2022). Each acoustic scene is a multi-source binaural mixture: a set of point sources of speech or sound-event audio clips, each convolved with a per-position BRIR, summed and combined with spatially diffuse background noises.

We introduce three refinements to the simulation pipeline: a physically motivated amplitude protocol that prevents mixing-induced distortion, corrected floor and ceiling material assignments for more realistic room acoustics, and diffuse background-noise simulation for naturalistic audio mixtures.

Audio Source Preprocessing and Loudness Protocol. We draw source clips from Hive (Li et al., 2026) and AudioSet (Gemmeke et al., 2017). Globally normalizing all dry clips to the same loudness would remove the natural variation in the typical power of different sound sources. We therefore normalize the dry signals by semantic class rather than globally, preserving the relative loudness of everyday sounds while allowing BRIR convolution to introduce distance attenuation and reverbera-

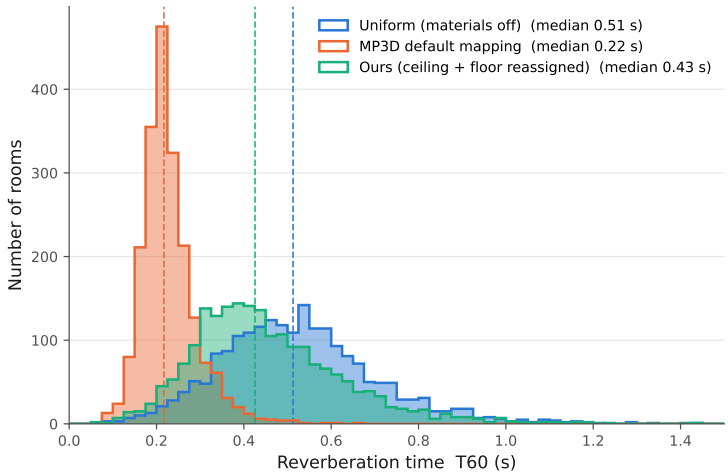


Figure 6: Distribution of broadband reverberation time (T_{60}) over all rooms under three surface-material configurations.

Table 9: RMS ranges and source distribution across loudness tiers.

Tier	RMS range	Share	Example classes
Quiet	0.05–0.10	11%	stir, humming, drip
Medium (default)	0.09–0.16	74%	guitar, bird, singing, speaker
Loud	0.14–0.24	15%	alarm clock, blender, brass instrument

tion. Specifically, we map each source class to one of three loudness tiers and uniformly sample the target RMS for each source from the corresponding tier range, as specified in Table 9.

Acoustic Material in SoundSpaces. SoundSpaces 2.0 (Chen et al., 2022) loads the mesh's per-surface semantic labels and maps each labeled surface class to an acoustic material via a label-to-material table, with 30 materials in the default MP3D configuration; walls, ceiling, floor, and furniture already carry distinct absorption coefficients. However, the "ceiling" label is bound to "Acoustic Tile" assigned to an absorption rate of 0.67 ($\alpha = 0.67$), an office ceiling material that absorbs more than usual everyday rooms; thus, when applied to every room, it over-damps the entire corpus and collapses reverberation to unrealistically low T_{60}. Moreover, the "floor" label is bound to "Carpet" for all scenes, so every room receives the same soft, absorptive floor, whereas real rooms exhibit floor variability such as ceramic tile ($\alpha \approx 0.01$), hardwood ($\alpha \approx 0.08$), and carpet ($\alpha \approx 0.25$). We therefore reassign these two surface classes, deterministically per room, so a given room's condition is reproducible across renders.

To verify that our material reassignment yields physically plausible room acoustics, we measure the reverberation time T_{60} of all simulated rooms under three configurations for comparison (Figure 6). The MP3D default mapping collapses T_{60} into an unrealistically short, narrow band that also reduces inter-room variation. Disabling materials and setting every surface to uniform absorption results in long decays and ignores how a room's actual surfaces shape its reverberation. Our reassignment restores an intermediate T_{60} with a broad, material-driven spread, recovering the inter-room acoustic diversity expected of real indoor scenes.

Diffuse Background Noise. To emulate realistic ambient noise, we add a two-channel spatially diffuse noise field to each mixture. The field is synthesized with the Arbitrary Noise Field generator (Habets et al., 2008), which imposes the theoretical spherically isotropic spatial coherence between two binaural microphones separated by the interaural spacing (0.17m). We scale the diffuse component to a per-mixture signal-to-noise ratio drawn uniformly from 5–20 dB relative to the clean spatial mixture, then sum it with the mixture.



Figure 7: AuralScene QA examples.

B.4 QA Generation

Based on the generated SAGs, we construct QA pairs at three levels of increasing reasoning complexity: Single-Node Perception (SNP), Single-Relation Reasoning (SRR), and Multi-Relation Reasoning (MRR). SNP focuses on perceiving attributes of individual objects, including their temporal, spatial, and motion-related properties. SRR requires reasoning over a relation between two objects, covering temporal, spatial, and causal relations. MRR further combines multiple relations to require joint reasoning over more complex interactions among different objects.

To examine the role of spoken content in spatial auditory reasoning, we further divide each task into two settings: with Spoken Content Understanding (SCU) and without SCU. The former requires models to identify and associate spoken content with the corresponding spatial or relational information, while the latter isolates spatial and relational reasoning from speech content. This separation allows us to distinguish general spatial auditory reasoning ability from the additional challenge of grounding spoken content within the auditory scene.

Single-Node Perception (SNP). SNP evaluates the perception of a single object and a single attribute. We consider three groups of attributes: timing (onset, offset, and duration), spatial properties (direction and distance), and trajectory (speed and moving direction). To avoid ambiguity in identifying the target event, we apply different referring strategies depending on whether the target

is speech or a non-speech event. As shown in Figure 7, for non-speech events, the target is referred to by its event class (e.g., "music" or "ball bouncing"), and SCU is not required. We exclude scenes in which multiple instances of the same class occur, ensuring that the class uniquely identifies the target event. For speech events, the target is instead described by a semantic paraphrase of its spoken content (e.g., "the speech about getting the ball"), requiring the model to understand the utterance and associate it with the corresponding spatially grounded speaker. We exclude scenes containing multiple utterances with the same semantic content to prevent trivial identification of the target. For direction, distance, and trajectory questions with SCU, we further require at least two speech sources at distinct spatial locations, ensuring that answering the question requires both spoken-content identification and spatial grounding rather than relying on a single candidate source.

For each attribute, we manually design five semantically equivalent question templates to introduce linguistic diversity while preserving the same underlying reasoning requirement. During QA generation, the target placeholder in each template is instantiated with the corresponding sound-event class name for non-speech events. For speech events, directly using the class name "speaker" would not uniquely identify the target; therefore, we use an LLM to generate a semantic paraphrase of each speech utterance in the scene. The paraphrase is then used to instantiate the target placeholder, requiring the model to identify the relevant speech content before determining the queried attribute. The resulting questions differ in surface form while maintaining consistent answer semantics across samples. The specific question templates for each attribute are provided below.

Question Prompts for Single-Node Perception (SNP)

1. Timing: Duration
   (1) How long does {ref} last? (2) What is the duration of {ref}? (3) For how long is {ref} audible?
   (4) How many seconds does {ref} last? (5) What is the total length of {ref} in the audio?
2. Timing: Onset
   (1) When does {ref} begin? (2) At what time does {ref} start?
   (3) When does {ref} first appear in the audio? (4) At what point does {ref} start?
   (5) What is the onset time of {ref}?
3. Timing: Offset
   (1) When does {ref} end? (2) At what time does {ref} stop? (3) At what point does {ref} end?
   (4) When does {ref} finish in the audio? (5) What is the offset time of {ref}?
4. Spatial: Direction
   (1) From which direction does {ref} come? (2) From what direction do you hear {ref}?
   (3) On which side does {ref} come from? (4) Where does {ref} come from?
   (5) What is the direction of {ref} relative to the listener?
5. Spatial: Distance
   (1) How far away is {ref} from you? (2) How many meters away is {ref} from you?
   (3) What is the distance between you and {ref}? (4) How far is {ref} from the listener?
   (5) At what distance is {ref}?
6. Trajectory (non-speech): Speed
   (1) At what speed is the {ref} moving? (2) How fast is the {ref} moving?
   (3) What is the speed of the moving {ref}? (4) How quickly does the {ref} move?
   (5) What is the movement speed of the {ref}?
7. Trajectory (speech): Speed
   (1) {ref}; how fast is its speaker moving? (2) {ref}; at what speed is its speaker moving?
   (3) {ref}; how quickly does its speaker move?
   (4) {ref}; what is the movement speed of its speaker?
   (5) {ref}; what is the speed at which its speaker moves?
8. Trajectory (non-speech): Direction
   (1) In which direction is the {ref} moving? (2) Which way is the {ref} moving?
   (3) What is the direction of motion of the {ref}? (4) Toward which side is the {ref} moving?
   (5) In what direction does the {ref} travel?
9. Trajectory (speech): Direction
   (1) {ref}; in which direction is its speaker moving? (2) {ref}; which way is its speaker moving?
   (3) {ref}; toward which side is its speaker moving?
   (4) {ref}; what is the direction of motion of its speaker?
   (5) {ref}; in what direction does its speaker travel?

Single-Relation Reasoning (SRR). SRR evaluates reasoning over a single relation between two objects. As shown in Figure 7, we consider three relation types: causal, temporal, and spatial. The SCU setting depends on the identities of the two target objects. When both target objects are non-

speech sounds, the relation can be determined directly from their event classes, and the question does not require SCU. When either target event is speech, the question is categorized as with SCU, as the model must first identify the corresponding speech source from its spoken content. This includes cases where one or both target events are speech. For spatial questions, we do not query the direction of an object relative to a speaker's facing direction, since the orientation of the speaker can not be available by audio. For causal relations, an object may have multiple causes, and we retain all valid causal relations when constructing the corresponding questions.

Similar to SNP, we construct SRR questions using a fixed set of manually designed question templates for each relation type. Each template contains placeholders for the target objects and the queried relation. During QA generation, these placeholders are instantiated with the corresponding source classes or speech-content paraphrases, depending on whether the targets are non-speech or speech events. This procedure keeps the underlying reasoning requirement consistent while introducing linguistic variation across QA samples. The specific question templates for causal, temporal and spatial relations are provided below.

Question Prompts for Single-Relation Reasoning (SRR)

1. Temporal Relation
   (1) What is the temporal relation between {ref 1} and {ref 2}?
   (2) In time, how does {ref 1} relate to {ref 2}?
   (3) Does {ref 1} occur before, after, or overlapping with {ref 2}?
   (4) Temporally, when does {ref 1} happen compared with {ref 2}?
   (5) How is {ref 1} ordered in time relative to {ref 2}?
2. Spatial Relation: Distance
   (1) How far apart are {ref 1} and {ref 2}?
   (2) Approximately how far is {ref 1} from {ref 2}?
   (3) What is the distance between {ref 1} and {ref 2}?
   (4) How many meters separate {ref 1} and {ref 2}?
   (5) What is the spatial distance between {ref 1} and {ref 2}?
3. Spatial Relation: Direction
   (1) In which direction is {ref 1} relative to {ref 2}?
   (2) Where is {ref 1} located relative to {ref 2}?
   (3) From the listener's view, on which side of {ref 2} is {ref 1}?
   (4) What is the direction of {ref 1} with respect to {ref 2}?
   (5) Relative to {ref 2}, in which direction is {ref 1}?
4. Causal Relation
   (1) Why does {ref} happen? (2) What is the reason for {ref}?
   (3) What causes {ref}? (4) What leads to {ref}? (5) What brings about {ref}?

Multi-Relation Reasoning (MRR). MRR evaluates whether a model can jointly reason over multiple relations among objects. As shown in Figure 7, we consider combinations of two relation types selected from causal, temporal, and spatial relations. When none of the related events is speech, the question does not require SCU. When at least one target event is speech, the question is categorized as with SCU, requiring the model to identify the relevant speech source from its spoken content before reasoning over the target relations. As in SRR, we do not query the direction of an object relative to a speaker's facing direction because speaker orientation is not perceptible from audio.

To construct MRR questions, we use manually designed templates for each combination of two relation types. Each template contains placeholders for the target objects and relations, which are instantiated with the corresponding source classes or speech-content paraphrases during QA generation. To introduce linguistic diversity, we design five semantically equivalent templates for each MRR subtask while keeping the underlying reasoning requirement unchanged. For multi-hop references, we further use multiple equivalent phrases to identify an event through an intermediate relation, such as the "*sound event that causes* {*result*}" or "*the sound that triggers* {*result*}." For example, a causal–spatial question can first identify the cause of an event and then ask for its distance from another sound source. When a causal relation is used to identify the target object, we exclude cases where multiple causes are included, since the causal description would not uniquely identify a single source. The complete MRR templates are provided below.

Question Prompts for Multi-Relation Reasoning (MRR)

1. Cause (templates to describe a target object based on a causal relation):
   (1) the sound event that causes {ref} (2) the sound event that prompts {ref}
   (3) the sound that triggers {ref} (4) the sound event responsible for {ref}
   (5) the sound event that leads to {ref}
2. Causal & Spatial (Distance)
   (1) How far away is {cause} from {ref}?
   (2) What is the distance between {cause} and {ref}?
   (3) How many meters separate {cause} and {ref}?
   (4) Approximately how far is {cause} from {ref}?
   (5) What is the spatial distance between {cause} and {ref}?
3. Causal & Spatial (Direction)
   (1) What is the direction of {cause} relative to {ref}?
   (2) Relative to {ref}, in which direction is {cause}?
   (3) On which side of {ref} is {cause}? (4) Where is {cause} located relative to {ref}?
   (5) From the listener's view, in which direction is {cause} with respect to {ref}?
4. Causal & Temporal
   (1) Does {cause} happen before, after, or overlapping with {ref}?
   (2) Temporally, when does {cause} occur compared with {ref}?
   (3) In time, how does {cause} relate to {ref}?
   (4) Is {cause} before, after, or overlapping with {ref}?
   (5) In terms of timing, does {cause} come before, after, or overlap with {ref}?
5. Temporal & Spatial (Distance)
   (1) The sound event/speech happens {time}; how far away is it/its speaker from {ref}?
   (2) The sound event/speech happens {time}; what is the distance between it/its speaker and {ref}?
   (3) The sound event/speech happens {time}; approximately how far is it/its speaker from {ref}?
   (4) The sound event/speech happens {time}; how many meters separate it/its speaker and {ref}?
   (5) The sound event/speech happens {time}; what is its/its speaker's distance from {ref}?
6. Temporal & Spatial (Direction)
   (1) The sound event/speech happens {time}; what's the direction of it/its speaker relative to {ref}?
   (2) The sound event/speech happens {time}; on which side of {ref} is it/its speaker?
   (3) The sound event/speech happens {time}; where is it/its speaker located relative to {ref}?
   (4) The sound event/speech happens {time}; in which direction is it/its speaker with respect to {ref}?
   (5) The sound event/speech happens {time}; relative to {ref}, in which direction is it/its speaker?

PS: In 5 and 6, {time} describes a target object based on a temporal relation. For example, "right after the ding sound" and "during the speech about room comfort."

Answer Generation. For all QA tasks, we derive the ground-truth answer directly from the SAG metadata and formulate each question as a multiple-choice QA problem. Temporal relation questions use three answer choices (*before*, *after*, and *overlap*), while all other questions use four choices. For causal relation questions, the answer needs to describe an event as a complete statement (e.g., "*a person is coughing*") rather than only its event class (e.g., "*cough*"). We therefore use an LLM to convert each event in the SAG into a natural-language statement and use these statements to construct the candidate answer pool for causal questions. When a scene contains fewer than four suitable events and therefore cannot provide enough answer choices, we use the LLM to generate additional distractor events based on the scene context. The generated distractors are required to be contextually plausible while being clearly distinguishable from the ground-truth event, preventing ambiguity in the answer choices.

Throughout the QA construction process, we use GPT-5.6 Luna (OpenAI, c) to assist with training QA generation and GPT-5.6 Terra (OpenAI, d) for evaluation QA generation, further reducing shared language and world-knowledge priors between the two splits.

MCQA Format. We formulate the questions in a fixed multiple-choice question answering (MCQA) format with three fields: *question*, *choice*, and *answer*. The *question* field contains the natural-language question and specifies that the model should select one option from the provided choices. The *choice* field contains four candidate answers, including one correct answer and distractors. The *answer* field specifies the correct option. LALMs are instructed to output only the selected option without additional explanation. An example is shown below:

MCQA Example

Instruction:
Where is the doorbell coming from? Choose the correct option from the following choices:
A: Left, behind, and above.
B: Left, front, and below.
C: Right, behind, and below.
D: Right, front, and above.
Please output the correct option directly without any explanation.
Answer:
D: Right, front, and above.

B.5 HUMAN VALIDATION

Our AuralScene generation pipeline involves two stages of human validation to ensure the quality and consistency of both the generated scenes and the resulting question-answer pairs. First, after constructing the Spatial Auditory Graphs (SAGs), we manually verify whether each scene is structurally plausible and acoustically consistent. We then use the validated SAGs to generate a large collection of QA samples, leading to a second round of human validation to assess the correctness and quality of the generated questions and answers.

For both stages, we develop dedicated web-based annotation platforms that provide an efficient and user-friendly interface for annotators. The platforms visualize the corresponding scene structures and spatial configurations, while also allowing annotators to directly listen to the rendered audio, enabling them to jointly inspect the semantic, spatial, and auditory consistency of each sample.

SAGs Human Validation Platform. To validate the quality of the generated SAGs, we developed a dedicated web-based annotation platform that allows annotators to jointly inspect the semantic, spatial, temporal, and causal aspects of each scene. As shown in Figure 8, the platform consists of four main components: Scene Graph, Spatial Map, Isolated Sources and Background, and Transcript, together with a Verdict panel for recording annotations. The Scene Graph visualizes the generated causal structure, while the Spatial Map presents the locations of all sound sources relative to the listener. The Isolated Sources and Background panel provides individual audio tracks for each source and the background, and the Transcript displays the corresponding speech content when applicable. During audio playback, the Spatial Map is synchronized with the audio and dynamically updates the positions of the sound sources. Moving sources are further represented using animations, allowing annotators to directly verify whether the spatial trajectories are consistent with the audio.

An important feature of the platform is that annotators can directly edit the causal structure during validation. Specifically, they can add or remove causal relations between any two objects through the Scene Graph interface. This is necessary because the causal relations initially generated by the LLM may contain hallucinated relations or omit important causal chains. Rather than restricting annotators to a binary judgment, we allow them to correct the graph and obtain a valid causal structure that better reflects the underlying scene. In addition, during validation, annotators are required to ensure that every retained causal relation is supported by evidence in the rendered audio and can be reasonably inferred from the auditory scene. Thus, the final causal structure reflects not only plausible causal relationships between events, but also relationships that are grounded in the available audio evidence. After inspecting the scene, annotators provide two types of verdicts. **Scene Plausibility** evaluates whether the overall scene is reasonable, while **Audio Source Plausibility** evaluates whether each audio source matches its assigned semantic label and location. Annotators can mark individual sources as unreasonable or accept all sources. Overall, the interactive platform enables efficient validation while providing annotators with sufficient information to assess both the generated scene representation and its corresponding audio.

SAG Human Validation Implementation. We recruit human annotators to validate the SAGs and provide each candidate annotator with detailed instructions and examples before annotation. However, even with this initial training, annotators may still make inconsistent judgments, particularly for complex scenes. To assess annotation quality, we manually select 40 relatively complex SAGs as a cross-check set. Each candidate annotator is required to annotate all 40 samples before proceeding to the main annotation task. We then collect the annotations from all candidates and measure their agreement with the other annotators along three dimensions.

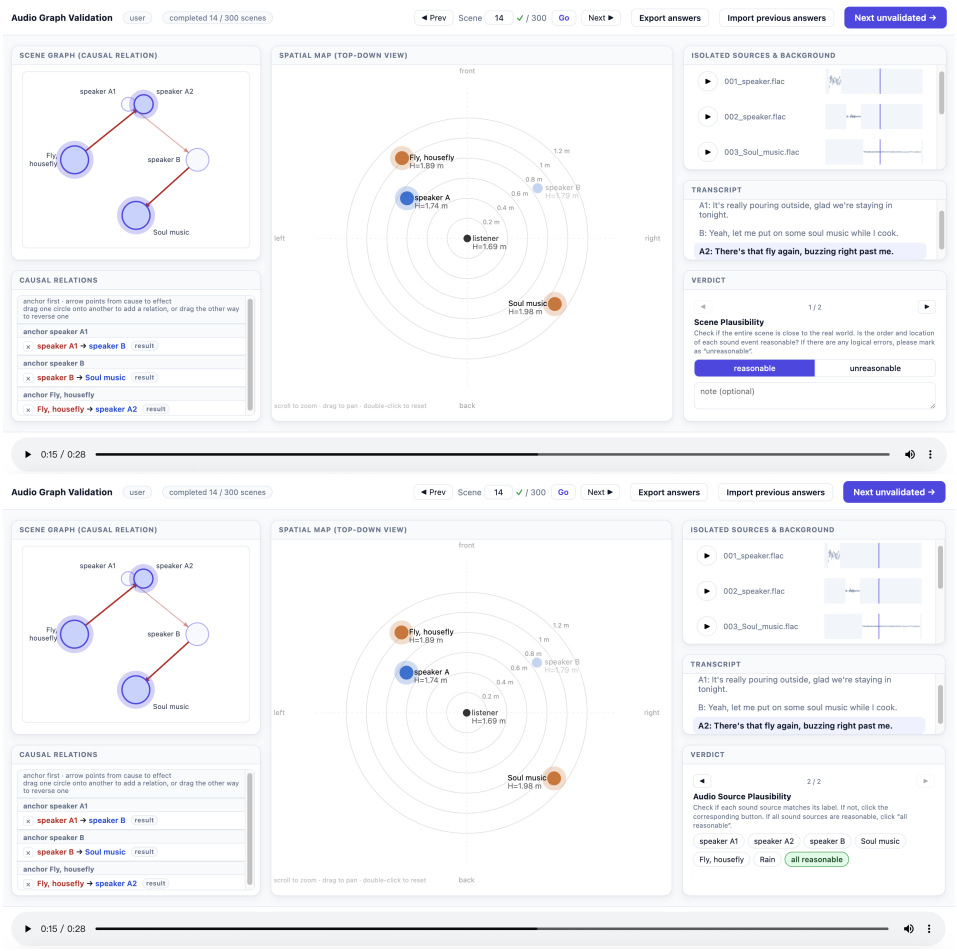


Figure 8: Spatial audio graph annotation platform (same graph, different verdicts).

First, *scene plausibility* measures whether the overall scene is reasonable, considering the sound events and their interactions. Second, *audio plausibility* measures whether individual audio sources are consistent with their semantic labels and spatial locations. Third, *causal plausibility* measures the consistency of the annotated causal structure and whether the annotated causal relations can be reasonably inferred from the audio evidence. Specifically, we compute the overlap between the causal relations retained or added by each annotator, based on their edits to the initial causal graph. For each dimension, we compute each annotator's agreement with the other annotators and use the average agreement as the annotator-level score. We consider an annotator qualified only if their average agreement exceeds 80% on all three dimensions.

After this screening process, 14 annotators are retained for the full annotation process, including 11 graduate students majoring in audio engineering and 3 undergraduate students. Figure 9 presents the agreement matrix among the qualified annotators across the three validation dimensions.

QA Human Validation Platform. After generating QA samples from the validated SAGs, we conduct a second round of human validation to assess the correctness and quality of each question-answer pair. We also develop a dedicated web-based platform that presents each QA sample together with the context needed for evaluation, allowing annotators to verify the answer against the scene, spatial configuration, and rendered audio.

As shown in Figure 10, the platform consists of five components: a Scene Graph, a Spatial Map, a Spatial Relations table, a QA panel, and a Transcript. The Scene Graph visualizes the causal structure among sound events, with each event annotated by its temporal span. The Spatial Map provides a top-down view of all sources relative to the listener, including distance rings and animated trajec-



Figure 9: SAG human validation agreement.

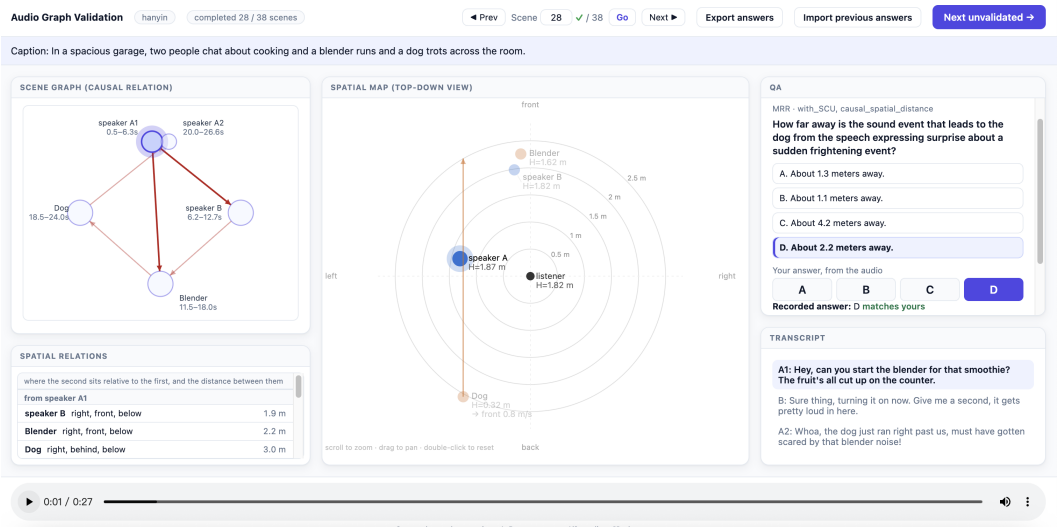


Figure 10: QA human validation platform.

tories for moving sources that remain synchronized with audio playback. The Spatial Relations table complements this view by listing the relative direction and distance between each pair of sources, while the Transcript displays the speech content of each speaker turn. The QA panel presents the current question, its task-type annotations, and the candidate options.

Importantly, the correct answer is not revealed before the annotator makes a judgment. Annotators must first independently select an answer from the candidate options based on the provided scene and audio, after which the system reveals the reference answer for comparison.

This interactive design encourages annotators to actively reason about the scene rather than simply verifying a pre-provided answer, thereby reducing confirmation bias and providing a more reliable assessment of whether the question is answerable and whether its reference answer is correct. Annotators then mark the question as "reasonable" or "unreasonable" and may provide an optional note. By exposing the semantic, spatial, temporal, and causal aspects of each sample within a single interface, the platform enables efficient and consistent validation of the generated QA pairs.

QA Human Validation Implementation. From the annotators qualified during SAG validation, we select the three with the highest mutual agreement to further validate the generated QA samples. Due to the large scale of the training set (∼420k QA pairs), we restrict human validation to the 2,700 QA samples in the evaluation set, while the large-scale training set is not manually verified.

Each test QA is independently annotated by all three annotators, and we adopt a majority vote to decide its verdict. A QA pair is marked "unreasonable" if the question is ambiguous, the reference answer is inconsistent with the scene, or more than one option is defensible. No evaluation

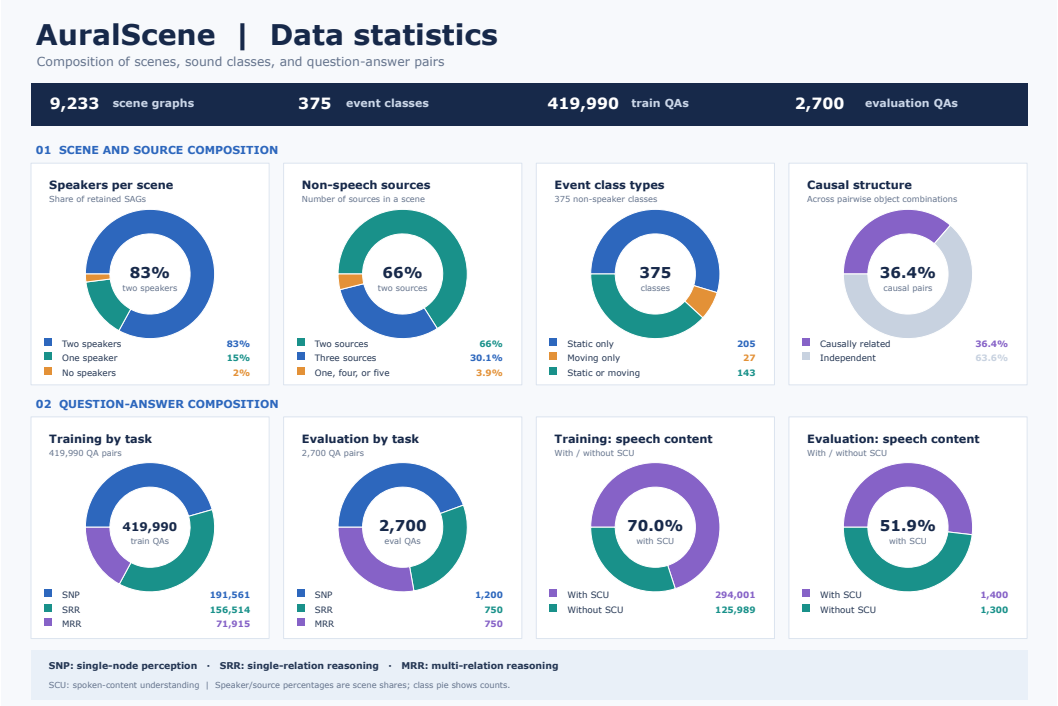


Figure 11: Overall statistics of AuralScene dataset.

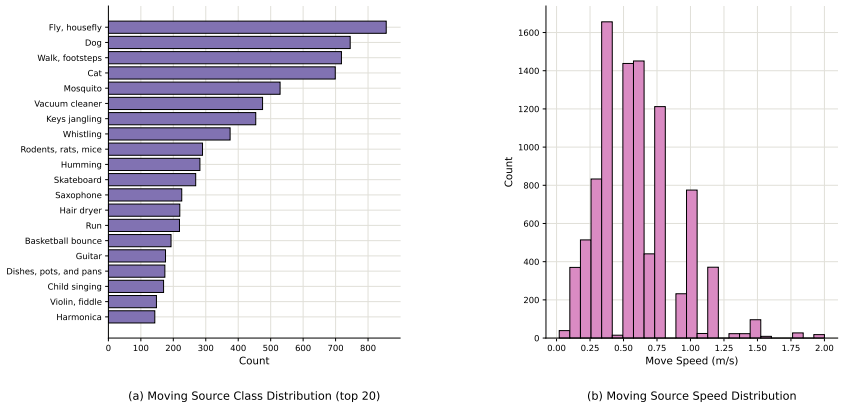


Figure 12: Audio class distribution of SAGs.

QA pair receives a majority "unreasonable" verdict, and all 2,700 pairs are therefore retained. We attribute this to two factors: problematic scenes, such as those with indistinguishable sources or causal relations lacking audio evidence, are already removed during SAG validation; while answers and distractors are constructed deterministically from the validated graphs using fixed templates.

B.6 Statistics

SAG Distribution. To characterize the diversity and complexity of the resulting dataset, we analyze all retained SAGs in terms of audio classes, source motion, spatial configurations, durations, and inter-event relations (Figure 11). Figure 12 shows the distribution of audio classes. The dataset covers a diverse range of everyday sound events, including 375 non-speaker event classes. Among them, 205 classes correspond to static sources, 27 classes correspond to moving sources (e.g., footsteps), and 143 classes can occur in either static or moving forms (e.g., basketball bouncing). Figure 13 further presents the distribution of moving-source classes and their speeds. Since the moving-source



(a) Moving Source Class Distribution (top 20)

(b) Moving Source Speed Distribution

Figure 13: Moving source distribution of SAGs.

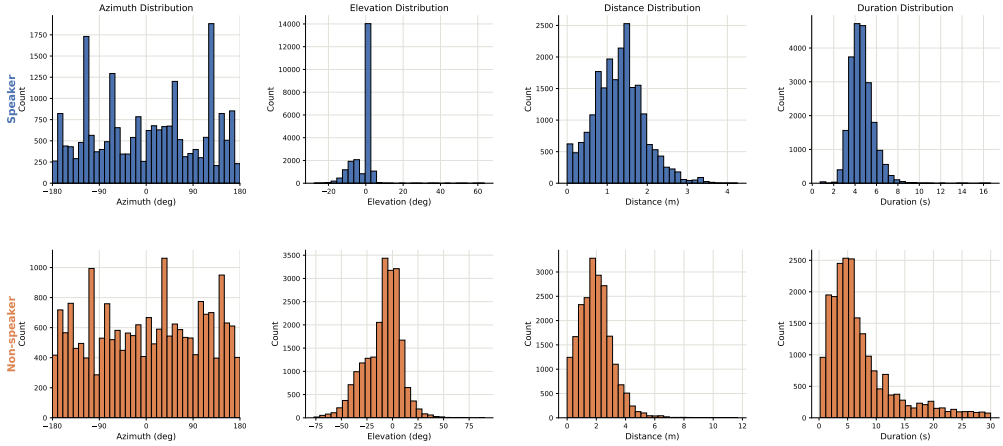


Figure 14: Spatial and duration distribution of SAGs.

taxonomy contains more than 150 classes, we visualize only the 20 most frequent classes for clarity, while the speed distribution is computed over all moving-source instances.

Although the generation prompt specifies one or two speakers per scene, a small number of generated SAGs place a speaker outside the room boundaries. In such cases, we remove all speakers from the scene, together with every relation involving them, and retain the remaining non-speech sources if the resulting scene is still plausible; otherwise, the SAG is discarded. This accounts for the scenes without speakers in the final dataset. As a result, AuralScene contains approximately 15% single-speaker scenes, 83% two-speaker scenes, and 2% scenes without speakers. Each scene contains between one and five non-speech sound sources. Most scenes contain either two sources (66%) or three sources (30.1%), while the remaining 3.9% contain one, four, or five sources. When two speakers interact, each conversation contains at most five turns, with most conversations consisting of two turns (51.9%) or three turns (29.4%). Among all pairwise object combinations, 36.4% exhibit a causal relation, while the remaining 63.6% are independent. This composition provides a range of scene complexities, with a predominance of moderately populated scenes that reflect common configurations in everyday auditory environments.

Figure 14 shows the spatial and duration distributions of speaker and non-speaker events. The elevation distribution of speakers is concentrated around 0°, as speakers in typical scenes are generally located at similar heights relative to the listener. Larger negative elevations occur in cases where a speaker is seated or otherwise positioned substantially below the listener. In contrast, non-speaker sources exhibit a broader spatial distribution, reflecting the greater diversity of source positions in

Table 10: Statistics of the QA samples in AuralScene dataset.

Task	Knowledge	# Train Sample	# Test Sample
SNP	without SCU, Timing, Onset	20,012	100
	without SCU, Timing, Offset	20,012	100
	without SCU, Timing, Duration	20,012	100
	without SCU, Direction	10,963	100
	without SCU, Distance	10,963	100
	without SCU, Trajectory, Speed	9,049	100
	without SCU, Trajectory, Direction	9,049	100
	with SCU, Timing, Onset	18,279	100
	with SCU, Timing, Offset	18,279	100
	with SCU, Timing, Duration	18,279	100
	with SCU, Direction	18,274	100
	with SCU, Distance	18,274	100
	with SCU, Trajectory, Speed	58	-
	with SCU, Trajectory, Direction	58	-
SRR	without SCU, Causal Relation	1,380	50
	without SCU, Temporal Relation	13,725	100
	without SCU, Spatial Relation, Direction	2,844	100
	without SCU, Spatial Relation, Distance	2,886	100
	with SCU, Causal Relation	18,250	100
	with SCU, Temporal Relation	61,798	100
	with SCU, Spatial Relation, Direction	14,545	100
	with SCU, Spatial Relation, Distance	41,086	100
MRR	without SCU, Causal&Spatial, Direction	320	50
	without SCU, Causal&Spatial, Distance	320	50
	without SCU, Causal&Temporal	864	50
	without SCU, Temporal&Spatial, Direction	1,785	50
	without SCU, Temporal&Spatial, Distance	1,805	50
	with SCU, Causal&Spatial, Direction	1,970	100
	with SCU, Causal&Spatial, Distance	9,132	100
	with SCU, Causal&Temporal	21,280	100
	with SCU, Temporal&Spatial, Direction	8,589	100
	with SCU, Temporal&Spatial, Distance	25,850	100
Total	-	419,990	2,700

the environment. The duration distributions also span a wide range of temporal scales, covering both short transient events and longer-lasting sound events.

Overall, the retained SAGs provide diverse coverage of sound categories, source motion, spatial configurations, temporal durations, and causal structures. This diversity enables AuralScene to capture a broad range of spatial auditory scene compositions and reasoning requirements.

QA Distribution. Table 10 summarizes the QA distribution in AuralScene, with 419,990 training samples and 2,700 evaluation samples across three levels of reasoning: Single-Node Perception (SNP), Single-Relation Reasoning (SRR), and Multi-Relation Reasoning (MRR). SNP covers event-level attributes, including timing, direction, distance, trajectory, and speed, while SRR and MRR progressively introduce single- and multi-relation reasoning over causal, temporal, and spatial relations. SCU is further incorporated into these tasks to require models to associate spatial or relational information with the spoken content of sound events.

Some task categories contain substantially fewer samples than others. In particular, SNP questions that combine SCU with trajectory-related knowledge are relatively rare. This is primarily because we do not explicitly constrain whether speakers should be static or moving, and the LLM naturally generates most speakers as stationary in everyday scenes. This composition is also consistent with typical auditory environments, where speech sources are more often stationary than moving. We therefore do not treat SCU-aware trajectory understanding as a separate evaluation category. Similarly, causal chains that do not involve speech are less frequently generated, resulting in fewer samples for several non-SCU causal reasoning categories, such as causal-only SRR and multi-relation questions involving causal relations. These variations reflect the natural composition of generated auditory scenes rather than an artificial attempt to balance each QA category. Our evaluation consequently focuses on the major task categories with sufficient coverage, while the full QA collection retains these less frequent cases for spatial LALM training.

Table 11: Training data usage in different stages.

Dataset	Tasks	Stage 1	Stage 2	Stage 3
SpatialSoundQA (1-Source)	AT & DOA & DP	100%	100%	20%
AuralSpeech (1-Speaker)	ASR	100%	100%	20%
SpatialSoundQA (2-Source)	AT & DOA & DP & Binary QA	0%	100%	20%
AuralSpeech (2-Speaker)	STL & TST	0%	100%	20%
AuralScene	Multi-Choice QA	0%	0%	100%

AT: Audio Tagging. DOA: Direction-of-Arrival estimation. DP: Distance Estimation. ASR: Automatic Speech Transcription. STL: Speaker Transcription and Localization. TST: Target Speaker Transcription.

C Details of Experiments

C.1 Training Configurations

We adopt a three-stage training strategy to progressively introduce spatial auditory perception, speech understanding, and spatially grounded reasoning.

The data usage in each stage is summarized in Table 11. In Stage 1, we train the model on single-source SpatialSoundQA and single-speaker AuralSpeech data, focusing on basic spatial auditory perception and speech transcription. In Stage 2, we additionally introduce two-source and two-speaker samples, together with the Binary QA and STL/TST tasks, to enable the model to handle multiple spatially overlapping sound sources and jointly model their spatial and speech information. In Stage 3, we further train the model on AuralScene using the full training set, directly optimizing the model for spatial auditory reasoning and spoken-content-aware spatial understanding. To maintain a balance between the general perception/speech capabilities and the reasoning objective, 20% of the Stage 1 and Stage 2 training data is retained in Stage 3.

All experiments are conducted on 8 NVIDIA A100 GPUs with an effective training batch size of 32. We use a linear warm-up schedule for the first 2,000 optimization steps, followed by linear learning-rate decay. The peak learning rate is set to 1×10^{-4} for Stages 1 and 2, and 2×10^{-5} for Stage 3. Stages 1 and 2 are trained for 2 epochs each, while Stage 3 is trained for 1 epoch. The lower learning rate and shorter training schedule in Stage 3 are used to reduce excessive adaptation to the QA-specific training distribution while preserving the spatial and speech representations acquired in the preceding training stages.

The audio encoders are frozen throughout all three stages. In Stage 1, only the audio-language projectors are optimized, allowing the model to first learn to map spatial auditory and speech representations into the language model's embedding space without altering the pretrained language backbone. In Stages 2 and 3, we additionally enable LoRA-based adaptation of the LLM backbone, using a rank of 8 and an α value of 16. This progressively increases the model's adaptation capacity as the training objectives become more complex, while keeping the number of trainable parameters relatively small and reducing the risk of disrupting the pretrained linguistic knowledge.

Overall, this training scheme follows a gradual transition from perception to multi-source understanding and finally to spatial reasoning. By first aligning the audio representations, then introducing parameter-efficient adaptation for multi-source speech and spatial understanding, and finally specializing the model on AuralScene, we aim to provide a stable training process while preserving the general capabilities of the pretrained audio and language components.

C.2 Evaluation Metrics

Here we define every metric used in our experiments. Throughout, N denotes the number of evaluation samples for the corresponding task, and $\mathbb{1}[\cdot]$ is the indicator function. All metrics are reported as percentages. Higher is better for mAP, ACC, and SR; lower is better for DER, WER, and sWER.

We treat outputs with an incorrect answer or an invalid format as incorrect when computing ACC; in particular, any output from which the answer cannot be parsed is counted as an error. This prevents models from avoiding penalties by producing malformed outputs that do not contain a valid answer. Higher ACC therefore indicates better performance.

For all transcription-based metrics, both hypotheses and references are processed using Whisper's EnglishTextNormalizer (Radford et al., 2023) to ensure consistent text normalization before scoring.

Mean Average Precision (mAP). For each class i, the average precision AP_i is the area under its precision–recall curve, and mAP averages AP_i over the C classes:

$$\mathrm{mAP} = \frac{1}{C} \sum_{i=1}^{C} \mathrm{AP}_i. \tag{1}$$

Higher values indicate more consistent retrieval and classification.

Specifically, we follow the evaluation protocol of prior work exactly (Zheng et al., 2024; Biswas et al., 2026). The ground-truth labels are represented as a multi-hot vector. Since the models generate free-form textual responses rather than class IDs, each predicted response is first encoded into a text embedding using a pretrained model (OpenAI, b). We then compute its cosine similarity with the pre-computed embedding of each class:

$$s_i = \frac{\mathbf{e}_{\mathrm{pred}}^{\top} \mathbf{e}_i}{\|\mathbf{e}_{\mathrm{pred}}\|_2 \|\mathbf{e}_i\|_2}, \tag{2}$$

where $\mathbf{e}_{\mathrm{pred}}$ denotes the embedding of the model response and $\mathbf{e}_i$ denotes the embedding of class i. These cosine similarities serve as the prediction scores for all classes. We then compute AP_i independently for each class using the corresponding ground-truth binary labels and prediction scores, and average them across all classes to obtain mAP.

Accuracy (ACC) for DOA. Directions are represented using natural-language descriptions. For each sample, the prediction is considered correct if the predicted direction matches the ground-truth direction description. ACC is computed as the proportion of samples with a match between the predicted and ground-truth directions.

Accuracy (ACC) for Binary and Multi-Choice QA. For Binary QA in SpatialSoundQA, the model is required to answer with either "yes" or "no", whereas our proposed AuralScene uses multiple-choice questions, where the model is required to select one correct answer from the provided options. We use the same ACC metric for both tasks, defined as the proportion of questions for which the model's predicted answer matches the ground-truth answer.

Distance Error Rate (DER). A distance prediction is counted as an error when it deviates from the ground truth by more than $0.5\,\mathrm{m}$, or when a valid distance value cannot be parsed from the model output. DER is the fraction of such samples:

$$\mathrm{DER} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\Big[\, |\hat{d}_i - d_i| > 0.5 \;\text{ or }\; \hat{d}_i \text{ is unparseable} \,\Big], \tag{3}$$

where $\hat{d}_i$ and d_i are the predicted and ground-truth distances. We set the threshold to 0.5, to be consistent with previous studies (Zheng et al., 2024; Biswas et al., 2026). To extract the predicted distance, we first isolate the spatial component according to the output format of each task. For STL, the output contains separate *Transcription* and *Location* fields, and we parse the first numerical value only from the *Location* field. Thus, numerical values in the transcription are not used for DER computation. For SpatialSoundQA, the spatial output is formatted as "direction; distance", and we parse the numerical value from the distance field. In both cases, numerical values appearing in the transcription or other non-spatial content are excluded from distance parsing.

Word Error Rate (WER) for Transcription. We evaluate transcription using the standard Word Error Rate (WER), micro-averaged over all utterances. Before scoring, both hypotheses and references are normalized using the Whisper English text normalizer.

For the Speaker Transcription and Localization (STL) task with two speakers, the WER computation depends on whether the two utterances overlap in time. For non-overlapping utterances, the model is instructed to output speakers in chronological order, allowing the i-th predicted speaker to be directly matched to the i-th reference speaker. For overlapping speech utterances, chronological ordering is ambiguous, and therefore no output ordering is imposed. Instead, we use a permutation-invariant protocol that finds a one-to-one assignment between predicted and reference speakers

to minimize the total word errors across the two speakers, following a concatenated minimum-permutation (cpWER)-style criterion (Yin et al., 2026a). Specifically, we enumerate all possible assignments between the predicted and reference speakers and select the assignment that minimizes the total number of word errors across all speaker pairs.

The resulting speaker assignment is also reused for localization evaluation, ensuring a consistent correspondence between predicted and reference speakers across metrics. To avoid rewarding under- or over-generation, we pad the predicted and reference speaker sets to equal size before matching. An unmatched reference speaker contributes its words as deletions, while a spurious predicted speaker contributes its words as insertions. Thus, the WER denominator remains fixed to the reference transcription, while both missing and hallucinated speakers are penalized.

LLM Judge for Target Speaker Transcription (TST). Target speaker transcription requires the model to identify the target speaker and transcribe the corresponding utterance. To distinguish speaker selection errors from transcription errors, following previous work (Tang et al., 2024b), we employ an open-sourced LLM, i.e., GPT-OSS-120B (OpenAI, a), as a judger to determine which speaker each hypothesis corresponds to. For each sample, the judge is provided with three inputs: (1) the reference utterance of the target speaker, (2) the reference utterance of the other speaker in the same scene as a distractor, and (3) the model hypothesis.

The LLM judge assigns exactly one of four labels:

- **Target**: The hypothesis is an attempt to transcribe the target speaker, allowing for typical ASR errors such as substitutions, deletions, insertions and partial truncation.
- **Other**: The hypothesis corresponds to the distractor speaker, indicating that the model selected the wrong speaker.
- **Mixed**: The hypothesis contains substantial content from both speakers, such as a concatenation or mixture of the two utterances.
- **Garbage**: The hypothesis is empty, unintelligible, or unrelated to either reference utterance.

Judgments are based on semantic correspondence rather than exact lexical matching. The judge is instructed to return a valid JSON object without additional explanations. The complete system prompt used for the LLM judge is shown below.

Prompt of LLM Judge for Target Speaker Transcription (TST)

You are a strict evaluator for a target-speaker ASR system.
Each scene contains two speakers at different spatial locations. The model was asked to transcribe ONLY the target speaker. For each item you get three fields:

- Target : the target speaker's utterance (the reference)
- Other : the other speaker's utterance in the same scene (a distractor)
- Hyp : the model output

Assign exactly one label to each item:
"Target" : hyp is a transcription attempt of the TARGET utterance. Tolerate normal ASR errors: substitutions, deletions, insertions, casing, punctuation, minor paraphrase, partial truncation. Content must correspond to the target.
"Other" : hyp is a transcription attempt of the OTHER speaker's utterance, not the target. The model picked the wrong speaker.
"Mixed" : hyp contains substantial content from BOTH utterances (concatenation or interleaving of the two speakers).
"Garbage" : hyp is empty, unintelligible, repetitive filler, or unrelated to both utterances.

Decide by meaning, not exact wording. If hyp matches the target with only ASR-level errors, label "target" even if a few words coincide with the other speaker. Only use "mixed" when both utterances are clearly present. Output strict JSON only, no explanation, no reasoning.

To evaluate the reliability of the LLM judge, we collect TST outputs from all evaluated models and randomly sample 50 examples from each of the four judgment categories, resulting in 200 samples in

total. An independent human annotator verifies each assigned label based on the target and distractor reference utterances and the model hypothesis. The human annotations agree with the LLM judge on all 200 samples (100% agreement). This provides empirical evidence that the judge reliably distinguishes the four categories across outputs from different models.

Success Rate (SR) for Target Speaker Transcription (TST). SR is the fraction of hypotheses that the judge attributes to the target speaker, capturing whether the model selects the correct speaker:

$$\mathrm{SR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[\ell_i = \textit{target}]\,. \tag{4}$$

where ℓ_i is the output from the LLM judge for target speaker transcription.

Successful WER (sWER) for Target Speaker Transcription (TST). sWER measures transcription quality conditioned on correct speaker selection. It is the WER computed only over samples labeled *target* by the TST LLM judge, aggregated at the corpus level:

$$\mathrm{sWER} = \frac{\sum_{i\in\mathcal{T}} \big(S_i + D_i + I_i\big)}{\sum_{i\in\mathcal{T}} M_{\mathrm{ref},i}}, \qquad \mathcal{T} = \{\, i : \ell_i = \textit{target} \,\}. \tag{5}$$

Here, S_i, D_i, and I_i denote the numbers of substitutions, deletions, and insertions in the transcription of sample i, respectively, and $M_{\mathrm{ref},i}$ denotes the number of words in its reference transcription. By evaluating only samples for which the target speaker is correctly selected, sWER measures how accurately the model transcribes the selected speaker's speech without mixing speaker-selection errors into the transcription metric. Thus, selecting the wrong speaker affects SR but does not affect sWER, while fluent transcription of the wrong speaker cannot improve sWER because that sample is excluded from its computation.

C.3 Baselines

C.3.1 LALM Baselines

We compare AuralListener against eight LALMs, grouped into five monaural LALMs, i.e., Kimi-Audio (Ding et al., 2025), Qwen2-Audio (Chu et al., 2024), MiDashengLM (Dinkel et al., 2025), Audio Flamingo 3 (Ghosh et al., 2026b), and Audio Flamingo Next (Ghosh et al., 2026a); and three binaural LALMs, including BAT (Zheng et al., 2024), DSpAST (Wilkinghoff & Tan, 2026), and OWL (Biswas et al., 2026).

We deliberately include monaural LALMs as baselines for three reasons. First, binaural LALMs remain scarce: beyond BAT, DSpAST and OWL, few publicly available models accept binaural audio. Restricting the comparison to spatial models would therefore provide a limited view of current LALM capabilities. Second, monaural LALMs are pretrained on much larger and more diverse audio corpora than existing spatial models and thus provide strong general audio understanding from monaural inputs. Third, their monaural inputs make them useful for diagnosing the role of spatial information. When these models fail on spatial tasks despite having strong audio understanding, the performance gap provides evidence for the contribution of spatial modeling.

For all monaural baselines, we down-mix the binaural input to a single channel before feeding it to the model. We use task-specific instructions to ensure that each model produces the required output format, while keeping the underlying task and evaluation criteria consistent across models. We use the publicly released instruction-tuned checkpoint of each baseline without additional fine-tuning.

It should be noted that existing spatial LALMs such as BAT and OWL are primarily designed for spatial perception and do not provide a speech-understanding pathway comparable to our AuralListener. Moreover, their original task designs do not support the multiple-choice question answering (MCQA) format used in AuralScene and STAR-Bench. Therefore, these models cannot directly perform the speech-understanding and MCQA tasks, rather than merely exhibiting degraded performance. This highlights the need for a model that jointly supports spatial audio understanding, spoken-content understanding, and structured reasoning.

Kimi-Audio. Kimi-Audio is an open audio foundation model that unifies audio understanding, generation, and conversation. It couples a hybrid audio tokenizer (producing discrete semantic tokens

together with continuous acoustic features) with a 7B-scale LLM initialized from a text backbone, and is pretrained on more than 13 million hours of speech, sound, and music. Its broad pretraining makes it a strong general-purpose model for speech recognition and audio question answering.

Qwen2-Audio. Qwen2-Audio is an audio-language model that pairs a Whisper-style audio encoder with a Qwen language backbone. It is trained in three stages (pretraining, supervised instruction tuning, and preference optimization) and supports both a voice-chat and an audio-analysis mode within a single set of weights. Qwen2-Audio is a widely used and competitive baseline across speech, sound, and music understanding tasks.

MiDashengLM. MiDashengLM is an efficient audio-language model that integrates the open Dasheng audio encoder with a Qwen2.5-Omni-based decoder. Instead of relying primarily on speech-to-text alignment, it is trained with a general audio captioning objective that treats speech, music, and environmental sound as parts of a single auditory scene, yielding holistic audio understanding at high inference efficiency.

Audio Flamingo 3. Audio Flamingo 3 is a fully open LALM that advances reasoning across speech, sound, and music. It combines the unified AF-Whisper encoder with a lightweight audio adaptor and a Qwen2.5-7B backbone in a LLaVA-style configuration, and adds on-demand chain-of-thought reasoning, multi-turn multi-audio chat, and long-context audio understanding of up to ten minutes. It reports state-of-the-art results on over twenty audio understanding and reasoning benchmarks, and thus serves as one of our strongest monaural reference points.

Audio Flamingo Next. Audio Flamingo Next is a recent published model in the Audio Flamingo series. Relative to Audio Flamingo 3, it scales training data beyond academic benchmarks, extends native long-audio support up to thirty minutes, and strengthens multilingual ASR, multi-talker speech understanding, and long-form captioning through timestamp-aware modeling. We include it as the most capable monaural baseline in our comparison.

BAT. BAT is a spatial LALM that operates on binaural audio. It derives spatial features with a dedicated spatial encoder (SpatialAST) and fuses them with an LLM decoder to answer questions about sound events, their directions, and distances. BAT establishes binaural spatial question answering, but it is restricted to non-speech sound events and coarse categorical spatial labels.

DSpAST. DSpAST retains BAT's overall pipeline and only replaces its SpatialAST audio encoder with a disentangled version, achieving better QA performance. However, only the encoder is open-sourced, while the LLM LoRA weights are not publicly available.

OWL. OWL is also a binaural LALM. It builds on a Spatial-Acoustic Geometry Encoder (SAGE) that aligns binaural acoustic features with 3D spatial structure using panoramic depth and room impulse responses at training time, while requiring only audio at inference. OWL only releases the weights of its SAGE encoder and the LLM LoRA adapters for Stage 1 and Stage 2, while the final-stage LORA checkpoint remains unavailable.

C.3.2 PURE-TEXT BASELINES

For AuralScene, to assess how much information is contained in the questions and the structured scene representation itself, we evaluate two pure-text baselines using **Sonnet5**. Both baselines receive the same question and answer choices as the audio-based models, but have no access to the input audio. They differ only in whether the ground-truth Spatial Auditory Graph (SAG) is provided.

Sonnet5 (SAG). This baseline receives the ground-truth SAG together with the question and answer choices. The SAG provides the structured scene information, including sound sources, spoken content, spatial attributes, temporal information, and inter-event relations. This setting measures the performance attainable when the underlying scene representation is directly available, serving as an upper bound for the corresponding QA tasks.

Sonnet5 (Blind). This baseline receives only the question and answer choices, without the audio or SAG. It therefore cannot access the scene information underlying the question and must rely only on the question wording, answer choices, and general language knowledge. This setting provides a lower bound that reflects performance obtainable from language priors and surface-level cues alone.

Detailed instruction prompts used for pure-text baselines are as follows:

Instruction Prompts Used for Pure-Text Baselines

Sonnet5 (SAG):
You answer multiple-choice questions about an audio scene using ONLY the structured scene metadata provided as JSON below, you are given no audio and no other description of the scene.

The metadata is ground truth: each source's class, speech transcription (if any), onset/offset time, position, azimuth/elevation/distance/direction relative to the listener (who sits at the origin, [0, 0, 0]), and pairwise temporal / spatial / causal relations between sources. In the source list, all directions and distances are relative to the listener at [0, 0, 0]. While in the relation list, all spatial AND temporal relationships refer to 'object' relative to 'anchor', i.e., anchors are used as the reference point. If, in some problems, the 'object' in the relation list is used as a reference point, then the spatial relation should be flipped from the metadata. Each 'Causal Relation' entry has already been resolved into an explicit, unambiguous statement, either 'source causes source' or 'no direct causal link'. For any causal question, answer strictly by looking up these edges; do NOT substitute your own judgment about which source seems narratively more likely to be the cause, the given edges are ground truth even when another source would intuitively seem like a more plausible cause. For multi-hop causal questions, trace the chain edge by edge using these explicit statements.

IMPORTANT: direction of 'Temporal Relation' edges: exactly like Spatial Relation, each 'Temporal Relation' value in the Relations list is the OBJECT's timing relative to the ANCHOR, not the other way around. 'before' means the Object occurs before the Anchor (the Object's interval ends before the Anchor's begins); 'after' means the Object occurs after the Anchor (the Object's interval begins after the Anchor's ends); 'overlap' means they overlap (see the exact overlap rule below). If a question's phrasing effectively makes the 'object' source the reference point instead (i.e. it asks about the Anchor's timing relative to the Object), flip 'before'/'after' from what the metadata says, the same way you would flip a Spatial Relation. For any 'Temporal Relation' question (before / after / overlap) between two sources, first check whether that exact pair already appears as an edge in the Relations list and use that value directly (applying the direction rule above), it is ground truth.

Only if the specific pair is not covered by any Relations edge should you derive it yourself from the two sources' Onset/Offset, and if you do, apply this EXACT rule (do not use your own intuition about what counts as overlapping): let inter = the length of the time intersection of the two [Onset, Offset] intervals (0 if they do not intersect at all), and shorter = the duration (Offset - Onset) of whichever of the two sources is shorter. The relation is 'overlap' if and only if inter $\geq 0.5 \times$ shorter, i.e. the two sources share at least 50% of the shorter source's own duration, any smaller intersection (including a small or brief overlap that does not reach that 50% threshold) does NOT count as 'overlap'. In every case that is not 'overlap', decide 'before' vs 'after' by which source starts earlier (ties broken by which ends earlier): the source with the earlier (Onset, Offset) is 'before' the other, which is 'after' it. Use the metadata to answer exactly as instructed in the question.

Sonnet5 (Blind):
You answer multiple-choice questions about an audio scene. You are given no audio, no scene metadata, and no other description of the scene, only the question text and its answer choices.

Use your best judgment and general knowledge to pick the most plausible answer. Answer exactly as instructed in the question.

C.4 Evaluation on BiDepth

We further evaluate AuralListener on BiDepth, which adopts a different azimuth convention based on clock positions, whereas SpatialSoundQA and our datasets use four directional categories for azimuth, i.e., left, right, front, and behind. Since the mapping between the two conventions is not specified in the reference and the BiDepth evaluation script is not publicly available, we do not report direction-related metrics and only evaluate distance prediction. BiDepth also differs from SpatialSoundQA, AuralSpeech, and AuralScene in its multi-source configuration: multiple sources may be located in different rooms. This setting introduces additional difficulty for distance estimation, since sources in different rooms can have different acoustic and spatial relationships to the listener.

While AuralListener is not trained on the BiDepth benchmark or its downstream tasks, its spatial branch adopts the SAGE encoder from OWL, which was pretrained on BiDepth for impulse response reconstruction and depth estimation. Therefore, this evaluation should be viewed as a benchmark-level zero-shot and cross-task transfer evaluation rather than a strictly unseen-data evaluation.

Table 12: Zero-shot performance of different systems on BiDepth evaluation set.

System	Detection (mAP ↑)		DP (DER ↓)	
	1-Source	2-Source	1-Source	2-Source
BAT	23.61	8.12	51.86	61.69
Kimi-Audio	8.21	2.38	83.88	84.81
Qwen2-Audio	11.07	2.49	82.49	82.32
MidashengLM	14.93	4.26	80.81	79.19
Audio Flamingo 3	12.05	2.99	84.63	85.01
Audio Flamingo Next	14.71	3.47	84.14	84.41
AuralListener	**26.11**	**12.28**	**43.66**	**59.62**

Nevertheless, Table 12 tests whether the proposed spatial LALM, i.e., AuralListener, can transfer its learned spatial reasoning capabilities to a new downstream task without task-specific training on BiDepth. AuralListener achieves the best results on both detection and distance prediction. For 1-source scenes, AuralListener achieves 26.11% mAP and a 43.66% DER, improving over the strongest baseline, BAT, by 2.50 and 8.20 percentage points, respectively. The advantage becomes more pronounced for 2-source scenes, where AuralListener achieves 12.28% mAP and 59.62% DER, compared with 8.12% and 61.69% for BAT.

Question:
What is everyone saying and where are they?

Ground Truth:
Transcription 1: So we talked about that for a bit. Location 1: left, behind, above; 2.09 m

Kimi-Audio:
Transcription 1: So we talked about that for a bit. Location 1: left, front, above; 2.76 m
Qwen2-Audio:
Transcription 1: He talked about that for a bit. Location 1: right, behind, above; 5.08 m

AuralListener:
Transcription 1: So we talked about that for a bit. Location 1: left, behind, above; 1.95 m

(a) STL for Single-Speaker

Question:
Tell me what each speaker says and where each of them is situated.

Ground Truth:
Transcription 1: I've got to hold my own hands up for this one. Location 1: right, front, below; 2.23 m | Transcription 2: It was all over now. Location 2: left, front, below; 1.81 m

Qwen2-Audio:
Transcription 1: I've got to hold my own hands up for this one. It was all over now. Location 1: right, behind, above; 4.16 m
MidashengLM:
Transcription 1: I got to hold my own hands up for this one. It was all over now. Location 1: left, front, above; 2.76 m | Transcription 2: We should have gone solo. Location 2: left, behind, below; 3.42 m

AuralListener:
Transcription 1: I've got to hold my own hands up for this one. Location 1: right, front, below; 2.64 m | Transcription 2: It was all over now. Location 2: left, front, below; 1.78 m

(b) STL for Two-Speaker

Figure 15: Examples of STL.

Question: What does the speaker behind you say? Transcribe it.
Target Speaker: There are so many people to thank.
Another Speaker: That's the truth.
AuralListener: There are so many people to thank.

(a) Successfully transcribe the target speaker (Target)

Question: Please provide a transcript of the speaker on your right.
Target Speaker: Sir Jeremy, from Glasgow, said they had been delighted.
Another Speaker: Scotland won by six wickets.
AuralListener: Scotland won by six wickets.

(b) Fail: transcribe the other speaker (Other)

Question: Transcribe the sentence spoken by the speaker located in front of you.
Target Speaker: We must look forward.
Another Speaker: I'm not afraid of that.
AuralListener: We must look forward, I'm not afraid of that.

(c) Fail: contains content from both speakers (Mixed)

Question: Transcribe the voice of the speaker on your left.
Target Speaker: There is great relief at this point.
Another Speaker: And we might go back.
AuralListener: Chewing, mastication; Gargling; Speech

(d) Fail: empty, unintelligible, or unrelated (Garbage)

Figure 16: Examples of TST.

D Case Studies

D.1 AuralSpeech

In AuralSpeech, we consider two tasks for spatial speech understanding: Speaker Transcription and Localization (STL) and Target Speaker Transcription (TST). Here we present successful and failure cases of AuralListener on these tasks to illustrate its spatial speech understanding capabilities and identify representative failure modes.

In Figure 15, we present examples for STL with different number of speakers. In both settings, AuralListener accurately transcribes the speakers and identifies their spatial locations. In contrast, monaural LALMs can often produce reasonable transcriptions but fail to correctly localize the speakers. In two-speaker scenes, they may also miss one speaker entirely, leading to incomplete transcription and localization. These examples highlight the importance of spatial information for jointly transcribing and localizing speakers.

Table 13: Distribution of TST output types in the AuralSpeech evaluation set. Only "Target" indicates successful target-speaker selection, while the other three modes represent failure cases.

Spatial Query	Overlap	Target	Other	Mixed	Garbage
Direction	✗	**50.68%**	49.06%	0.25%	0.01%
Distance	✗	**88.95%**	10.55%	0.27%	0.23%
Direction	✔	**46.53%**	46.43%	1.79%	5.25%
Distance	✔	**55.42%**	25.69%	1.14%	17.75%

Table 14: Performance breakdown of AuralListener in individual knowledge categories in AuralScene (ACC ↑). "Dur.", "Ons", "Off.", "Dir.", "Dis." represent "Duration", "Onset", "Offset", "Direction", and "Distance", respectively.

Timing			SNP		Trajectory		SRR		Spatial		MRR				
			Dir.	Dis.							C&T	C&S		T&S	
Dur.	Ons.	Off.	-	-	Dir.	Speed	Causal -	Temporal -	Dir.	Dis.	-	Dir.	Dis.	Dir.	Dis.
without Spoken Content Understanding															
44.00	54.00	46.00	66.00	67.00	82.00	85.00	50.00	63.00	67.00	51.00	38.00	22.00	68.00	34.00	62.00
with Spoken Content Understanding															
70.00	56.00	52.00	61.00	72.00	-	-	61.00	55.00	71.00	69.00	64.00	82.00	44.00	86.00	67.00

TST requires the model to first identify the target speaker from a spatial query and then transcribe only the target speech. Figure 16 presents four representative outcomes: *Target*, *Other*, *Mixed*, and *Garbage*. The *Target* case corresponds to correct target-speaker selection and transcription, while the other three cases reflect different failure modes, including selecting the wrong speaker, mixing speech from multiple speakers, and producing irrelevant or invalid outputs.

Table 13 further reports the distribution of these outcomes across different conditions. The results indicate that incorrect target selection is a major source of TST errors, with failure patterns becoming more pronounced as the number of speakers and speech overlap increase.

D.2 AURALSCENE

D.2.1 PERFORMANCE BREAKDOWN

AuralScene covers a diverse set of multiple-choice questions spanning spatial perception and temporal, spatial, and causal reasoning. Representative demonstrations are provided in Appendix B.4 and on our project page; therefore, we do not enumerate individual examples here. Instead, in Table 14, we provide a fine-grained breakdown of AuralListener's performance across individual knowledge categories, highlighting its strengths and remaining challenges.

The results show that individual spatial attributes are generally easier to recognize than their combinations with other relations, while distance remains challenging across settings. Notably, SCU substantially improves multi-relation reasoning, with accuracy increasing from 38% to 64% for causal-temporal reasoning, from 22% to 82% for causal-spatial reasoning in direction, and from 34% to 86% for temporal-spatial reasoning in direction.

In contrast, temporal reasoning alone decreases from 63% to 55% with SCU, indicating that spoken content is particularly beneficial when multiple relations must be jointly interpreted. These results suggest that current spatial LALMs still struggle with complex relational reasoning, despite stronger performance on individual spatial attributes.

D.2.2 CONFIDENCE INTERVAL ANALYSIS

Here we supplement the point estimates in Table 6 with interval estimates (Table 15), so that differences between systems can be interpreted against the sampling variability of the evaluation set.

Scene-level resampling. The 2,700 evaluation QA samples are derived from 113 validated SAGs, approximately 24 questions per scene, and questions from the same SAG share the same room, sound sources, spoken content, and relation graph. We therefore resample at the level of scenes rather than individual questions: in each of 1,000 iterations we draw 113 scene identifiers with replacement,

Table 15: AuralScene accuracy with scene-level 95% confidence intervals (ACC ↑), computed with 1,000 bootstrap iterations over the 113 evaluation scenes.

System	Setting	SNP	SRR	MRR	Overall
AuralListener	w/o SCU	63.43 [59.85, 66.91]	58.86 [53.78, 63.96]	44.80 [38.69, 50.58]	58.62 [55.97, 61.53]
	w/ SCU	62.20 [57.85, 66.12]	64.00 [59.30, 69.41]	68.60 [64.57, 72.75]	65.00 [62.41, 67.54]
w/o Mono Token	w/o SCU	63.71 [60.26, 67.61]	55.71 [50.43, 61.04]	43.20 [37.66, 49.07]	57.62 [54.98, 60.39]
	w/ SCU	54.00 [49.57, 58.15]	56.00 [50.84, 60.87]	55.40 [51.23, 59.36]	55.07 [52.44, 57.48]
w/o Spatial Token	w/o SCU	52.57 [49.13, 55.91]	44.00 [38.35, 50.14]	39.20 [33.02, 46.22]	47.69 [45.05, 50.57]
	w/ SCU	52.00 [47.88, 56.00]	48.25 [43.62, 53.46]	50.40 [46.44, 54.67]	50.36 [48.19, 52.70]

take all QA samples belonging to the selected scenes, and recompute the metric, and we report the 2.5th and 97.5th percentiles as the 95% interval. Inference is performed once and the resampling operates on the fixed per-question outcomes, so no model is re-evaluated.

We additionally measure the intra-class correlation of per-question correctness within a scene, and find it to be small for all three variants (ρ = 0.010, 0.006, and effectively zero for AuralListener, w/o Mono Token, and w/o Spatial Token respectively; the corresponding design effect is at most 1.2). Outcomes across questions from the same SAG are thus close to independent in practice, which we attribute to the fact that questions generated from one SAG target different attributes and relations and therefore do not share a single dominant source of difficulty. Scene-level intervals are consequently only about 10–15% wider than question-level ones.

Interpretation of the Ablations. The contribution of spatial tokens is well separated from sampling variability: removing them lowers overall accuracy from 58.62 to 47.69 without SCU and from 65.00 to 50.36 with SCU, and in both settings the two intervals are disjoint. The contribution of mono tokens is setting-dependent. With SCU, removing them lowers accuracy from 65.00 to 55.07 with disjoint intervals. Without SCU, the difference is 1.0 point and the intervals overlap almost entirely (58.62 [55.97, 61.53] vs. 57.62 [54.98, 60.39]), so no effect is measurable at this evaluation scale. This asymmetry is consistent with the construction of the two settings: questions without SCU refer to their target events by class name and can be answered without a speech-oriented representation, whereas questions with SCU require recovering the utterance semantics before any spatial or relational reasoning can be performed. The size of the evaluation set is determined by the cost of exhaustive human verification: every evaluation SAG is inspected on our annotation platform and each of the 2,700 evaluation QA samples is independently annotated by three qualified annotators. The training split, which is not manually verified at the QA level, contains 9,120 SAGs and 419,990 QA samples, indicating that the SAG construction pipeline itself scales; we plan to expand the evaluation split in future releases under the same verification standard.

D.2.3 Human Evaluation

The human reference on AuralScene is obtained from three independent normal-hearing listeners. During evaluation, the listeners are provided only with the rendered audio and question, without access to any graph structure or other metadata, ensuring that the reference reflects auditory understanding from the audio itself. The final answer for each question is determined by majority voting among the three listeners. Listeners use headphones in a quiet room, and may replay each clip without time limit. All audio is rendered with the non-individualized HRTF of SoundSpaces 2.0.

As a three-listener reference, these results should be interpreted as a reference point rather than a population-level upper bound. Nevertheless, the results show a clear distinction across question types (Table 6). Causal and temporal relations, including their combination, are the most readily answerable (90–95%), followed by direction (80–85%), timing (70.67–81.67%), and trajectory (78.50%). In contrast, absolute distance (58–61%) and spatial relations between two sources (54–67.5% across SRR spatial, C&S, and T&S) remain the most challenging.

This pattern is consistent with known limits of human spatial hearing. Absolute distance perception is poor in humans, who systematically compress perceived distance and rely on cues such as level and direct-to-reverberant ratio that are ambiguous without familiarity with the source or room (Zahorik et al., 2005). Spatial relations between two sources are harder still, since answering them

requires localizing both sources and then inferring their relative direction or distance, so that localization errors on each source compound. Human listeners are comparatively sensitive to the direction of auditory motion (Carlile & Leung, 2016), which is consistent with the higher trajectory accuracy. However, non-individualized HRTFs increase front–back and elevation confusions, which also affect judgments of motion direction, and estimating the absolute speed of a moving source in m/s is not a task human listeners perform reliably (Carlile & Leung, 2016); both factors may account for the remaining trajectory errors.

We further observe that human accuracy is similar with and without SCU (73.79% vs. 72.85%), suggesting that questions anchored by spoken content are not inherently easier to perceive. Compared with the human reference, AuralListener is comparable to or slightly above humans on distance and trajectory, and exceeds them on spatial relations anchored by speech, but still lags behind on timing, direction, causal, and temporal reasoning.

E LIMITATIONS AND FUTURE WORK

World-Knowledge Priors in AuralScene. In this work, we use different LLMs to generate the training and evaluation SAGs and QAs, reducing shared generation and language priors between the two splits. However, some world-knowledge priors are intentionally preserved to keep the generated scenes physically and semantically plausible. For example, footsteps are naturally associated with positions near the floor, making their elevation partly predictable from the event class. Such priors are difficult to avoid without sacrificing the correspondence between sound events and their typical real-world contexts. In this sense, these priors also allow models to learn useful world knowledge for spatial auditory understanding. Our STAR-Bench results further show that the learned spatial representations can transfer to real-world recordings. In future work, we plan to construct evaluation scenarios with more diverse configurations, while retaining realistic physical constraints, to better disentangle learned world knowledge from dataset-specific spatial priors.

Model Scope. AuralListener is intended as a baseline that enables controlled analysis rather than as an architectural contribution. It combines two frozen public encoders with a LLaMA-2-7B decoder, and we do not explore stronger backbones, trainable encoders, or alternative fusion mechanisms. Whether the remaining gap to the oracle reference is limited by the audio representation or the language decoder is left to future work.

Broader Task Opportunities. Beyond the tasks studied in this work, AuralSpeech and AuralScene can support a broader range of spatial auditory learning tasks. For example, they can be used to train spatial audio encoders, predict structured spatial auditory graphs, or study alternative forms of scene-level reasoning. We hope these resources will facilitate further research on spoken-content-aware spatial auditory understanding and related studies.