EMPOWERING LARGE AUDIO-LANGUAGE MODELS VIA SPATIAL QUESTION AND ANSWERINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial audio understanding requires models to connect sound identities with locations, motion, and relations over time. Most large audio-language models (ALMs), however, process monaural audio and cannot directly access the inter-channel cues available in spatial recordings. We introduce **SonicSpatial**, a parameter-efficient framework that enhances the spatial understanding capabilities of pretrained ALMs. A question-conditioned spatial selection module uses learnable queries to compress dense features from a frozen spatial audio encoder into compact spatial tokens. These spatial tokens are then combined with the backbone's semantic audio tokens and question tokens as input to the LLM for answer generation. We also introduce **SonicSpatial-Bench**, comprising 3,045 multiple-choice questions over 576 recorded and simulated FOA audio instances across eight capability categories, including source motion, temporal relations, listener-centered transformations, and room-scale inference. On this benchmark, SonicSpatial with MOSS-Audio-8B achieves 60.5% overall accuracy, improving on its backbone by 22.0 percentage points; the Qwen2.5-Omni variant achieves 63.1%, improving on its backbone by 25.4 points. The two variants reach 64.3% and 65.4%, respectively, on our multiple-choice adaptation of SO-Bench. The spatial selection module can be integrated into diverse ALM backbones without modifying their audio encoders, yielding consistent gains across five evaluated models. These results demonstrate the broad applicability of SonicSpatial as a parameter-efficient spatial adaptation framework.

1 INTRODUCTION

Large Audio-Language Models (ALMs) (Chu et al., 2023; 2024; Goel et al., 2025) have demonstrated strong capabilities in speech recognition, audio captioning, and acoustic scene reasoning, with recent work further extending these models to fine-grained temporal grounding (Wang et al., 2026; Sun et al., 2026). Nevertheless, most large ALMs are designed for monaural input and therefore lack the inter-channel spatial cues required to infer source directions, trajectories, and spatial relations, hindering their deployment in embodied agents, immersive environments, and spatial scene understanding.

First-Order Ambisonics (FOA) represents the full-sphere sound field in a listener-centric coordinate system, encoding spatial cues related to source direction, motion, and multi-source configurations through its omnidirectional and orthogonal directional components. However, adapting large pretrained ALMs to FOA remains challenging, as existing approaches often train dedicated spatial encoders from scratch or substantially modify the original architecture, which can require considerable spatial training data and computational resources (Zheng et al., 2024; Sudarsanam & Politis, 2025; Zhu et al., 2026). Meanwhile, although prior benchmarks have evaluated source localization, static spatial relations, and, more recently, source motion and spatio-temporal localization (Zheng et al., 2024; Sudarsanam & Politis, 2025; Sridhar et al., 2026), they provide limited systematic coverage of long-horizon trajectories, temporally separated events, evolving multi-source relations, ego-rotation, and environmental acoustics such as room properties and spatial scale.

We present **SonicSpatial**, a parameter-efficient framework that equips pretrained ALMs with spatial reasoning capabilities. Given four-channel FOA audio, the framework combines complementary semantic and spatial representations: the backbone's original audio encoder processes the

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.

omnidirectional W channel, while a pretrained spatial audio encoder processes all four channels to extract dense spatial features. A question-conditioned spatial selection module connects these features to the language model. It first conditions a set of learnable queries on the question, then uses them to summarize the relevant spatial information into a compact set of tokens. These spatial tokens are combined with the semantic audio and question tokens and passed to the backbone LLM for answer generation. Training updates only the spatial selection module and the LLM's LoRA parameters, keeping both pretrained audio encoders frozen and avoiding the need to train a spatial encoder from scratch.

To enable comprehensive evaluation, we introduce **SonicSpatial-Bench**, a benchmark for spatial perception and reasoning in long-form acoustic scenes. It contains 3,045 questions over 576 FOA clips, combining real-world clips from STARSS23 (Shimada et al., 2023) with spatial audio rendered under controlled room configurations. Clips have an average duration of 36.1 seconds, enabling the evaluation of long source trajectories, temporally separated events, and evolving spatial relations. The benchmark comprises 31 tasks across 8 dimensions: Absolute Position, Relative Position, Hallucination, Ego-rotation, Multi-source, Spatio-temporal Localization, Motion, and Room Acoustics. Each question is generated from the underlying scene metadata and subjected to a two-stage verification pipeline: candidate questions are first validated against source activities, trajectories, timestamps, coordinate ranges, and task-specific ambiguity margins, and are then reverified after linguistic paraphrasing to ensure that their physical meaning and answers remain unchanged. Using the same generation and verification pipeline, we further construct a training set of 22,472 spatial question–answer pairs over 2,944 FOA clips.

In summary, we make the following contributions: (i) we introduce **SonicSpatial-Bench**, a long-form FOA benchmark containing 3,045 questions over 576 audio clips and covering 31 tasks across 8 diverse and complementary dimensions of spatial audio perception and reasoning; using the same data generation and verification pipeline, we further construct 22,472 question–answer pairs over 2,944 FOA clips for spatial instruction tuning; (ii) we propose **SonicSpatial**, a parameter-efficient and plug-and-play framework that combines a pretrained spatial audio encoder with a question-conditioned spatial selection module, enabling existing large ALMs to acquire spatial reasoning capabilities without training a spatial encoder from scratch; (iii) extensive experiments reveal substantial limitations in the spatial understanding of current large ALMs and demonstrate that our approach consistently outperforms existing large ALMs and competitive baselines.

2 SonicSpatial-Bench and Dataset

We first analyze existing spatial audio benchmarks in Section 2.1. We then describe the construction of **SonicSpatial-Bench** in Section 2.2, followed by its task taxonomy, benchmark composition, and accompanying training data.

2.1 Analysis of Existing Benchmarks

As summarized in Table 1, existing spatial audio benchmarks vary in audio format, clip duration, and evaluated capabilities. Most adopt binaural audio, while FOA remains comparatively less explored in spatial audio–language benchmarks. Existing benchmarks are also generally constructed from short clips and primarily evaluate absolute/relative position, static spatial relations, or short-term motion. Consequently, they provide limited systematic coverage of long-horizon trajectories, spatio-temporal localization, evolving multi-source relations, ego-rotation, and room-acoustic properties. **SonicSpatial-Bench** addresses these limitations through long-form FOA scenes and a broad taxonomy of 31 fine-grained tasks grouped into eight capability categories.

2.2 Benchmark Construction

SonicSpatial-Bench combines two complementary subsets. The recorded subset uses long-form, multi-source FOA recordings and frame-level trajectories from STARSS23 (Shimada et al., 2023) to evaluate rich spatial understanding questions, especially those involving dynamic, source-level spatial reasoning. The simulated subset uses dry sounds from FOA-MEIR (Yasuda et al., 2022) and controlled rendering with SoundSpaces 2.0 (Chen et al., 2022) to evaluate room-acoustic and room-scale perception.

Table 1: **Comparison with existing spatial audio benchmarks.** We categorize evaluated capabilities into eight key spatial dimensions: **Abs.** (Absolute Position), **Rel.** (Relative Position), **Multi** (Multi-source), **Mot.** (Motion), **S-T** (Spatio-temporal Localization), **Hall.** (Hallucination), **Ego.** (Ego-rotation), and **Room** (Room Acoustics).

Benchmark	Format	#Tasks	Evaluated Spatial Tasks							
			Abs.	Rel.	Multi	Mot.	S-T	Hall.	Ego.	Room
SpatialSoundQA (Zheng et al., 2024)	Binaural	5	✓	✓	✓	✗	✗	✗	✗	✗
SPUR-Set (Sakshi et al., 2025)	FOA	6	✓	✓	✓	✓	✓	✓	✓	✗
Hear You Are QA (Ryu et al., 2026)	Binaural	4/9	✓	✓	✓	✗	✗	✓	✗	✗
SAVVY-Bench (Chen et al., 2026)	7ch	4	✓	✓	✗	✓	✓	✗	✓	✗
STAR-Bench (Liu et al., 2026)	Binaural	3	✓	✓	✓	✓	✓	✗	✗	✗
The World is Not Mono (You et al., 2026)	FoA	3	✓	✓	✓	✗	✓	✓	✗	✗
Sci-Phi (Jiang et al., 2026)	FOA	15	✓	✓	✓	✗	✓	✗	✗	✓
SO-Bench (Zhu et al., 2026)	FOA	3/16	✓	✓	✓	✓	✓	✗	✗	✗
SonicSpatial-Bench (Ours)	**FOA**	**8/31**	✓	✓	✓	✓	✓	✓	✓	✓

2.2.1 Long-Form Spatial Audio Construction

The original STARSS23 recordings are divided into clips of approximately 45s while preserving complete sound events whenever possible. Based on source activity and 3D trajectories, we generate questions concerning Absolute Position, Relative Position, Hallucination, Ego-rotation, Multi-source, Spatio-temporal Localization, and Motion. This design enables each clip to support questions about both spatial relationships at specific moments and changes in source position over time. All answers are computed directly from the scene metadata, with task-specific constraints applied to source activity, timestamps, trajectory changes, spatial margins, and distractor separation. Detailed coordinate conventions, task definitions, and generation criteria are provided in the Appendix.

2.2.2 Room-Acoustic Simulation

To evaluate environmental perception, we construct a simulated subset using FOA-MEIR dry sounds and SoundSpaces 2.0. As shown in Figure 1(b), sounds are placed in parameterized shoebox rooms and rendered as four-channel FOA audio at 24 kHz. Room geometries are sampled from four non-overlapping volume ranges: 15–30 m^3, 45–85 m^3, 130–250 m^3, and 380–700 m^3. The gaps between adjacent ranges reduce ambiguity near category boundaries. The subset evaluates Volume Range, Volume Estimation, and Size Comparison, covering range/categorical, numerical, and comparative room-scale perception. Detailed simulation parameters and variable controls are provided in the Appendix.

2.2.3 QA Generation and Verification

As illustrated in Figure 1(a), candidate questions are generated from structured scene metadata. A separately implemented verification program independently recomputes their answers and checks source activity, timestamp validity, comparison margins, distractor uniqueness, and answer indices. DeepSeek-V4.1-Flash (Xu et al., 2026) then paraphrases the question stems without modifying answer choices, timestamps, numerical values, or spatial relations. The paraphrased questions are verified again, and samples are removed if their structured meaning or correct answers change. Language models are used only for linguistic rewriting, while all labels are derived from scene metadata. Detailed generation and verification rules are provided in the Appendix.

2.2.4 Task Categories

As illustrated in Figure 1(c) and in Figure 1(d), **SonicSpatial-Bench** contains 31 fine-grained tasks grouped into eight capability categories: Absolute Position, Relative Position, Hallucination, Ego-rotation, Multi-source, Spatio-temporal Localization, Motion, and Room Acoustics. Together, they evaluate absolute distance, azimuth, and elevation estimation; relative spatial relations; source, directional, and temporal existence verification; rotated azimuth estimation; multi-source counting and ordering; spatio-temporal localization; motion trajectory tracking; and room-scale inference. The complete task definitions and representative examples are provided in the Appendix.

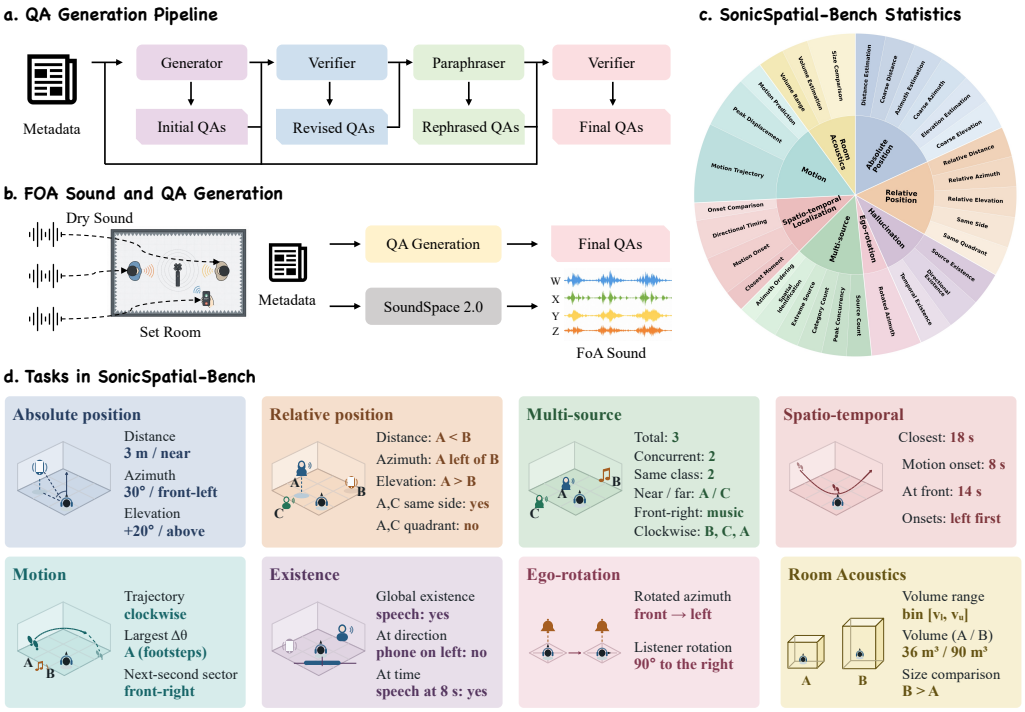


Figure 1: **Construction and composition of `SonicSpatial-Bench`.** (a) QA pairs are generated from scene metadata, physically verified, linguistically paraphrased, and verified again. (b) For the room-acoustic subset, shared scene metadata are used for both FOA rendering and QA generation. (c) Distribution of questions across the eight spatial audio capability categories. (d) Detailed task examples across the eight capability categories in **`SonicSpatial-Bench`**.

2.2.5 Benchmark and Training Data Statistics

The questions contain two to five answer choices: 1,082 have two choices, 245 have three, 1,475 have four, and 243 have five. We use the STARSS23 test partition to construct the benchmark and its training partition to construct the training data. Following the same generation procedure, we additionally construct 22,472 training QA pairs over 2,944 FOA clips. Our data feature event-rich acoustic scenes, with an average of 3.75 sound events, providing a basis for evaluating spatial reasoning in the presence of multiple events. We further balance the generated data across task categories and answer distributions by removing overrepresented samples, reducing the distributional biases.

Questions have between two and five answer choices: 1,082 questions have two choices, 245 have three, 1,475 have four, and 243 have five. To limit the influence of answer-frequency and task-frequency biases, we balance answers within each task type and downsample disproportionately large task pools when assembling the benchmark. Detailed task-level counts and answer distributions are provided in the Appendix. Using the same generation and verification procedure, we construct a spatial instruction-tuning set of 22,472 question–answer pairs over 2,944 FOA clips. These comprise 350 STARSS23 segments and 2,594 simulated scenes. The training clips have a mean duration of 31.5 s and contain an average of 3.56 sound events. Training and benchmark data follow the original dataset partitions: STARSS23 recordings are separated by recording, while the simulated subsets use disjoint anechoic source recordings.

3 Methods

In this section, we present **`SonicSpatial`**, a parameter-efficient and plug-and-play framework that equips large Audio-Language Models (ALMs) with spatial perception and reasoning capabilities over First-Order Ambisonics (FOA) recordings. We first formulate the task in Section 3.1, then

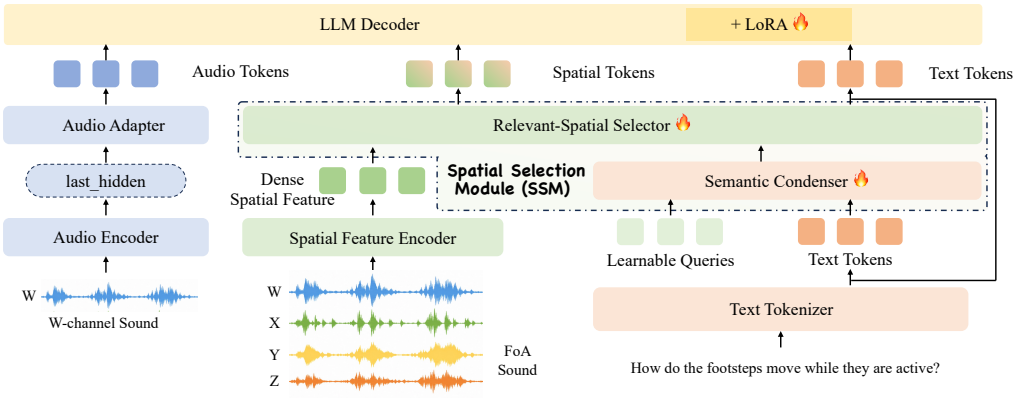


Figure 2: Overview of **SonicSpatial**. The semantic branch processes the omnidirectional W channel, while the spatial branch encodes the complete FOA signal and selects question-relevant spatial evidence. The resulting semantic audio, spatial, and textual tokens are jointly processed by the backbone LLM.

introduce the proposed architecture and spatial selection module and describe the parameter-efficient training strategy in Section 3.2.

3.1 Problem Formulation

Given a four-channel FOA recording $\mathcal{A} = [\mathcal{A}^W, \mathcal{A}^X, \mathcal{A}^Y, \mathcal{A}^Z] \in \mathbb{R}^{4\times T}$ and a textual question $\mathcal{Q}$, the goal is to generate an answer $\mathcal{Y}$ that is consistent with both the acoustic content and the spatio-temporal configuration of the scene:

$$\hat{\mathcal{Y}} = \Phi_{\texttt{SonicSpatial}}(\mathcal{A}, \mathcal{Q}), \tag{1}$$

where T is the number of time-domain samples per channel. We separately process the omnidirectional W channel for semantic audio understanding and the complete FOA signal for spatial analysis. This decomposition retains compatibility with the pretrained ALM while preserving the inter-channel cues required for spatial reasoning.

3.2 Spatially-Aware Audio-Language Framework

As illustrated in Figure 2, **SonicSpatial** augments a pretrained ALM with a spatial branch and a question-conditioned Spatial Selection Module (SSM). The original audio pathway processes the omnidirectional W channel to preserve general audio understanding, while a pretrained Spatial Feature Encoder extracts dense representations from the complete FOA signal. The SSM selects question-relevant spatial evidence, which is combined with the semantic audio and textual tokens for answer generation. By separating dense spatial encoding from question-conditioned selection, the framework preserves detailed sound-field information without passing the entire spatial sequence to the LLM.

3.2.1 Dual-Branch Audio Encoding

The semantic branch processes the omnidirectional W channel using the original audio encoder and adapter of the pretrained ALM:

$$\mathbf{H}^{\text{aud}} = \phi_{\text{adapter}}\left(\phi_{\text{audio}}(\mathcal{A}^W)\right) \in \mathbb{R}^{N_a \times d_l}, \tag{2}$$

where N_a is the number of semantic audio tokens and d_l is the LLM embedding dimension. In parallel, the pretrained Spatial Feature Encoder and a spatial adapter process the complete FOA signal:

$$\mathbf{H}^{\text{dense}} = \phi_{\text{spatial}}(\mathcal{A}) \in \mathbb{R}^{N_s \times d_s}, \tag{3}$$

where N_s and d_s denote the sequence length and feature dimension, respectively. Before being fed into the SSM, $\mathbf{H}^{\text{dense}}$ is first projected into the LLM embedding space by a spatial adapter.

The two branches provide complementary representations: the semantic branch captures general acoustic content, whereas the spatial branch retains directional and temporal sound-field cues. We retain the temporally ordered spatial representations rather than applying global pooling, thereby preserving the evolution of the sound field. Because directly passing this dense sequence to the LLM is inefficient and may introduce irrelevant information, we use the SSM to select question-relevant spatial evidence.

3.2.2 Question-Conditioned Spatial Selection

Let $\mathbf{H}^{\text{text}} \in \mathbb{R}^{N_t \times d_l}$ denote the embeddings of the textual question, where N_t is the number of question tokens. These embeddings are used both to condition the SSM and as textual input to the LLM. The SSM contains two consecutive Q-Formers. The Semantic Condenser uses the question to condition N_q learnable queries, which are then used by the Relevant-Spatial Selector to retrieve evidence from $\mathbf{H}^{\text{dense}}$:

$$\mathbf{H}^{\text{spa}} = \Phi_{\text{SSM}}\left(\mathbf{H}^{\text{dense}}, \mathbf{H}^{\text{text}}\right) \in \mathbb{R}^{N_q \times d_l}. \tag{4}$$

This two-stage design separates the semantic interpretation of the question from the retrieval of corresponding spatial evidence. Since the output contains a fixed number of N_q spatial tokens, its contribution to the LLM context length is independent of the input duration.

3.2.3 Multimodal Language Decoding

The tokens that encode semantic audio, spatial, and textual information are concatenated as $\mathbf{S} = [\mathbf{H}^{\text{aud}}; \mathbf{H}^{\text{spa}}; \mathbf{H}^{\text{text}}]$, and provided to the backbone LLM for autoregressive answer generation:

$$P\left(\mathcal{Y} \mid \mathcal{A}, \mathcal{Q}\right) = \prod_{i=1}^{N_y} P\left(y_i \mid \mathbf{S}, y_{<i}\right). \tag{5}$$

Through self-attention over the multimodal sequence, the LLM can jointly reason about acoustic semantics, selected spatial evidence, and the textual question during generation. This token-level interface introduces spatial information without modifying the original audio encoder or the internal architecture of the backbone LLM.

We freeze the original audio encoder and adapter, the Spatial Feature Encoder, and the base LLM parameters. Only the SSM and the LoRA parameters inserted into the LLM are optimised, θ_{train}. This strategy concentrates task-specific adaptation on spatial evidence selection and multimodal reasoning while retaining the pretrained representations of the backbone components. For a target answer $\mathcal{Y} = \{y_1, \ldots, y_{N_y}\}$, we minimize the autoregressive language modeling loss:

$$\mathcal{L} = -\sum_{i=1}^{N_y} \log P\left(y_i \mid \mathbf{S}, y_{<i}; \theta_{\text{train}}\right). \tag{6}$$

The loss is computed only over the target answer tokens, while the semantic audio, spatial, and question tokens serve exclusively as conditioning context.

4 Experiments

4.1 Implementation Details

We instantiate **`SonicSpatial`** with two backbone variants: **`SonicSpatial-A`**, built on the audio-language model MOSS-Audio-8B, and **`SonicSpatial-O`**, built on the omni-modal model Qwen2.5-Omni. In both variants, the backbone's semantic audio encoder processes the FOA W channel, while a pretrained spatial audio encoder, GRAM (Yuksel et al., 2026), extracts eight spatial feature streams at 12.5 Hz. These streams are flattened into a sequence of spatial tokens. The semantic audio encoder, GRAM spatial encoder, and backbone language model remain frozen; only the Spatial Selection Module and LoRA parameters are trained. Our training set comprises 22,472 synthetic multiple-choice QA pairs and 10,000 examples randomly sampled from SO-Dataset (Zhu et al., 2026). We train for three epochs using 32 learnable queries, with the LoRA rank and scaling factor set to 16 and 32, respectively. Further implementation details are provided in the Appendix.

4.2 SPATIAL AUDIO UNDERSTANDING

We evaluate both variants of **SonicSpatial** on two benchmarks, **SonicSpatial-Bench** and SO-Bench. **SonicSpatial-A** uses MOSS-Audio-8B as its audio-language backbone, whereas **SonicSpatial-O** uses Qwen2.5-Omni as its omni-modal backbone. Since SO-Bench originally consists of free-form questions, we employ DeepSeek-V4.1-Flash to generate multiple-choice options without modifying the original questions. We report accuracy for each task, capability category, and the overall benchmark. All models are evaluated using the same prompting and answer-parsing protocol.

As shown in Table 2, we draw three main observations: (i) Existing large audio-language and omni-modal models exhibit limited spatial audio understanding on **SonicSpatial-Bench**. The strongest baseline, Gemini-2.5-Pro, achieves an overall accuracy of 47.6%, while the strongest open-source audio-language and omni-modal baselines reach 41.3% and 42.7%, respectively, with limited performance on motion understanding and Ego-rotation tasks. More recent models also do not consistently outperform their predecessors, suggesting that progress in general audio understanding does not necessarily yield better spatial reasoning. (ii) Both variants of **SonicSpatial** substantially improve upon their respective backbones. **SonicSpatial-A** achieves 60.5% overall accuracy, outperforming Moss-Audio by 22.0 points, the strongest open-source audio-language baseline by 19.2 points, and the strongest baseline overall by 12.9 points. It also achieves the highest accuracy among the evaluated open-source audio-language models in all eight capability categories, with especially large gains over its backbone in Absolute Position, Spatio-temporal Localization, and Room Acoustics. **SonicSpatial-O** reaches 63.1% overall accuracy, improving upon Qwen2.5-Omni by 25.4 points and surpassing the strongest open-source omni-modal baseline by 20.4 points. These gains across two distinct backbone families support the effectiveness of question-conditioned spatial selection. (iii) Both variants improve upon their backbones on SO-Bench and surpass the strongest baseline on this benchmark. **SonicSpatial-A** improves from 43.7% to 64.3% overall accuracy, while **SonicSpatial-O** improves from 46.6% to 65.4%. The largest gains over their respective backbones occur in the Basic category: 25.0 points for **SonicSpatial-A** and 23.3 points for **SonicSpatial-O**. Both surpass the strongest SO-Bench baseline, Gemini-3.1-Pro, which achieves 58.3% overall accuracy. In particular, **SonicSpatial-O** scores 13.9 points higher than its backbone on the Relational category. These results indicate that the benefits of explicit spatial features extend to SO-Bench, including relational tasks.

Table 2: Spatial audio understanding results on **SonicSpatial-Bench** and SO-Bench. Bold fonts indicate the best performance within each open-source model group. We evaluate capabilities into eight key spatial dimensions: **Abs.** (Absolute Position), **Rel.** (Relative Position), **Multi** (Multi-source), **Mot.** (Motion), **S-T** (Spatio-temporal Localization), **Hall.** (Hallucination), **Ego.** (Ego-rotation), and **Room** (Room Acoustics) in **SonicSpatial-Bench**, and three capability groups: **Basic.** (Detection and Localization), **Rela.** (Spatial Relation Understanding), and **Reas.** (Complex Reasoning with Semantics) in SO-Bench.

Models	SonicSpatial-Bench									SO-Bench			
	Abs.	Rel.	Multi	Mot.	S-T	Hall.	Ego.	Room	Avg.	Basic.	Rela.	Reas.	Avg.
Proprietary Models													
Gemini-2.5-Pro (Comanici et al., 2025)	40.6	57.2	62.1	28.9	44.5	66.4	25.3	50.2	47.6	46.9	65.6	57.1	54.8
Gemini-3.1-Pro (Google DeepMind, 2026)	39.3	61.9	51.4	27.5	36.3	71.2	25.0	44.0	45.1	52.3	67.6	57.6	58.3
Open-Source Large Audio-Language Models													
Qwen-Audio (Chu et al., 2023)	9.1	16.5	22.3	16.5	3.1	50.5	11.0	23.8	18.3	10.4	32.2	33.5	20.9
Qwen2-Audio (Chu et al., 2024)	28.1	49.3	35.3	28.2	19.6	55.7	29.2	33.2	34.7	31.2	46.3	39.4	37.5
Kimi-Audio (Ding et al., 2025)	31.3	47.4	43.7	26.8	37.8	60.3	24.0	40.4	39.0	32.4	54.4	51.1	42.5
Step-Audio (Huang et al., 2025)	36.1	53.5	53.0	27.6	21.3	66.8	29.9	41.4	41.3	44.0	67.8	56.7	54.0
Step-Audio 2 (Wu et al., 2025)	33.0	41.3	49.8	29.7	16.0	60.9	22.7	42.0	37.3	35.3	60.7	50.4	46.1
Audio Flamingo 3 (Goel et al., 2025)	27.5	46.1	37.9	26.8	30.8	58.0	18.8	43.3	36.3	31.9	50.9	45.9	40.3
Audio Flamingo Next (Goel et al., 2025)	35.5	47.6	41.6	18.6	32.2	40.7	6.5	33.6	34.0	35.3	53.8	45.3	43.1
Moss-Audio (Yang et al., 2026a)	31.0	47.0	46.0	30.8	26.3	61.6	21.4	40.7	38.5	35.0	54.1	50.6	43.7
SonicSpatial-A	**67.2**	**58.9**	**61.9**	**38.3**	**69.2**	**73.0**	**41.6**	**70.4**	**60.5**	**60.0**	**71.5**	**61.2**	**64.3**
Open-Source Large Omni-modal Models													
Qwen2.5-Omni (Xu et al., 2025a)	30.8	50.7	49.8	22.4	25.2	61.2	27.3	33.9	37.7	37.9	58.2	50.6	46.6
Qwen3-Omni (Xu et al., 2025b)	37.3	47.0	49.3	22.4	45.7	66.1	23.4	51.5	42.7	43.6	66.0	54.3	52.9
Spatial-Omni (Zhu et al., 2026)	35.1	52.6	45.3	18.0	27.7	57.3	23.4	36.2	37.4	50.8	69.4	56.5	58.1
SonicSpatial-O	**67.9**	**63.9**	**65.1**	**42.9**	**67.5**	**73.9**	**39.6**	**77.9**	**63.1**	**61.2**	**72.1**	**64.4**	**65.4**

4.3 Ablation

To assess the contribution of each component, we compare **SonicSpatial** with three alternatives: (i) fine-tuning the backbone using only the FOA W channel, (ii) appending all spatial tokens directly to the audio tokens, and (iii) reducing the two Q-former in the Spatial Selection Module to a single Q-former that directly interacts with dense spatial features. As shown in Figure 3, W-channel fine-tuning improves both backbones, while adding spatial features yields further gains. Our full model achieves performance close to that of directly appending all spatial tokens on both benchmarks, while being substantially more efficient: it uses only 32 learned queries instead of 4k+ tokens (a count that grows with audio length), reducing per-step training from 15.7s to 4.7s and inference latency from 0.86s to 0.41s. Reducing the two-layer Q-former to a single layer lowers overall accuracy on both backbones and both benchmarks, supporting the use of the full Spatial Selection Module.

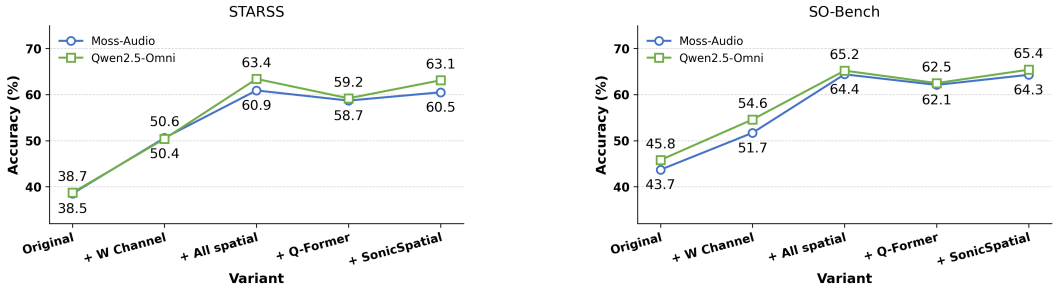


Figure 3: **A Systematic Ablation Study of SonicSpatial-A and SonicSpatial-O on SonicSpatial-Bench and SO-Bench.**

4.4 Implementation on Other Large Audio-Language Models

To examine the generality of our framework, we apply **SonicSpatial** to diverse open-source large audio-language models, including Kimi-Audio, Step-Audio 2, Qwen2-Audio, Audio Flamingo Next, and Moss-Audio. These models differ substantially in their audio encoders, language models, and multimodal alignment strategies. For each backbone, we retain the original architecture and introduce only the proposed Spatial Selection Module together with LoRA parameters, demonstrating that our method can be integrated without full-model fine-tuning.

As shown in Table 3, incorporating **SonicSpatial** generally improves the spatial understanding performance of the evaluated backbones. The gains are observed across models with different architectures and initial spatial capabilities, indicating that our method is not tied to a specific audio-language model. These results demonstrate the broad compatibility and effectiveness of **SonicSpatial** as a parameter-efficient approach for equipping existing large audio-language models with spatial reasoning abilities. Moreover, stronger backbone models generally lead to better performance after spatial adaptation, suggesting that our framework can continue to benefit from future advances in large audio-language models.

Table 3: Implementation on Other Large Audio-Language Models.

Models	SonicSpatial									SO-Bench			
	Abs.	Rel.	Multi	Mot.	S-T	Hall.	Ego.	Room	Avg.	Basic.	Rela.	Reas.	Avg.
Qwen2-Audio (Chu et al., 2024)	28.1	49.3	35.3	28.2	19.6	55.7	29.2	33.2	34.7	31.2	46.3	39.4	37.5
(+) **SonicSpatial**	64.9	55.9	48.1	35.4	61.3	60.3	31.8	48.9	52.3 (↑17.6)	57.4	70.2	57.3	61.9 (↑24.4)
Kimi-Audio (Ding et al., 2025)	31.3	47.4	43.7	26.8	37.8	60.3	24.0	40.4	39.0	32.4	54.4	51.1	42.5
(+) **SonicSpatial**	51.8	51.7	46.3	21.5	58.3	59.3	26.6	43.0	45.6 (↑6.6)	58.3	68.2	61.1	62.1 (↑19.6)
Step-Audio 2 (Wu et al., 2025)	33.0	41.3	49.8	29.7	16.0	60.9	22.7	42.0	37.3	35.3	60.7	50.4	46.1
(+) **SonicSpatial**	67.6	54.6	54.9	41.4	65.8	65.1	37.7	58.0	56.8(↑19.5)	60.0	65.2	64.5	62.4 (↑16.3)
Audio Flamingo Next (Goel et al., 2025)	35.5	47.6	41.6	18.6	32.2	40.7	6.5	33.6	34.0	35.3	53.8	45.3	43.1
(+) **SonicSpatial**	68.1	50.2	47.0	37.9	63.3	68.7	40.9	50.8	54.1 (↑20.1)	59.9	68.0	61.8	63.1 (↑20.0)
Moss-Audio (Yang et al., 2026a)	31.0	47.0	46.0	30.8	26.3	61.6	21.4	40.7	38.5	35.0	54.1	50.6	43.7
(+) **SonicSpatial**	67.2	58.9	61.9	38.3	69.2	73.0	41.6	70.4	60.5 (↑22.0)	60.0	71.5	61.2	64.3 (↑20.6)

5 Related Works

5.1 Large Audio-Language Models

Large audio-language models (ALMs), including LTU, SALMONN, Qwen2-Audio, and MOSS-Audio, combine acoustic representations with language models to support audio understanding and instruction following (Gong et al., 2024; Tang et al., 2024a; Chu et al., 2024; Yang et al., 2026a). The Audio Flamingo series (Kong et al., 2024; Ghosh et al., 2025; Goel et al., 2025) develops dialogue, long-audio understanding, and reasoning, while the Step-Audio family (Huang et al., 2025; Wu et al., 2025; Lin et al., 2026; Tian et al., 2025) advances speech interaction and audio-grounded reasoning. Beyond clip-level understanding, temporal question answering (Sridhar et al., 2025a), TimeAudio (Wang et al., 2026), and SpotSound (Sun et al., 2026) investigate temporal relationships, timestamp-aware modeling, and event grounding. These developments reflect growing attention to temporal structure, but do not by themselves provide the directional evidence required for three-dimensional spatial reasoning.

5.2 Spatial Audio Understanding

BAT and OWL combine binaural spatial representations with language reasoning (Zheng et al., 2024; Biswas et al., 2026), with OWL additionally introducing geometric supervision and spatial chain-of-thought reasoning. Complementary approaches investigate first-order Ambisonics (FOA) for spatial speech processing (Tang et al., 2024b) and audio–text alignment (Devnani et al., 2024), while SPUR and The World is Not Mono adapt ALMs using dedicated spatial representations (Sakshi et al., 2025; You et al., 2026). Their data resources include simulated scenes, real-recording question answering, and hybrid corpora (Zheng et al., 2024; Biswas et al., 2026; Sudarsanam & Politis, 2025; Sakshi et al., 2025). Motion-focused studies further introduce frame-level spatial attributes and movement reasoning (Sridhar et al., 2025b; 2026). Building on these advances, SonicSpatial combines question-conditioned selection of pretrained FOA features with parameter-efficient adaptation to extend spatial reasoning. SonicSpatial-Bench evaluates complementary spatial capabilities over long-form recorded and simulated audio, covering source localization and relations, motion and spatio-temporal reasoning, listener-centric transformations, and room-acoustic inference.

5.3 Spatial Visual Understanding

Visual spatial understanding spans static geometry, reasoning across views, and temporal dynamics. SpatialVLM, SpatialRGPT, and SpatialBot (Chen et al., 2024; Cheng et al., 2024; Cai et al., 2025) develop spatial supervision and depth-aware modeling. LLaVA-3D, Video-3D LLM, VLM-3R, and VG-LLM (Zhu et al., 2024; Zheng et al., 2025; Fan et al., 2026; Zheng et al., 2026) incorporate spatial coordinates or reconstruction priors into multimodal representations. SPAR, VSI-Bench, and MMSI-Bench (Zhang et al., 2026; Yang et al., 2025; 2026c) address spatial perception and reasoning across views, while STI-Bench, VLM4D, OmniSpatial, and 4D-RGPT (Li et al., 2025; Zhou et al., 2025; Jia et al., 2026; Yang et al., 2026b) extend evaluation and modeling toward temporal dynamics, perspective-taking, and spatial interaction. Inspired by Reason in 4D (Zhou et al., 2026), SonicSpatial selects task-relevant evidence from spatial audio representations to support reasoning about acoustic scenes.

6 Conclusion

In this paper, we introduce **SonicSpatial**, addressing the limited spatial reasoning capabilities of large audio-language models. Our approach combines a pretrained spatial audio encoder with a question-conditioned Spatial Selection Module, enabling efficient extraction of task-relevant spatial evidence from FOA audio. To evaluate spatial reasoning in complex, long-form acoustic scenes, we introduce **SonicSpatial-Bench**, a comprehensive benchmark covering eight capabilities, from source localization and motion understanding to listener-centric transformation and room-acoustic inference. Across recorded and simulated scenes, **SonicSpatial** consistently outperforms strong ALM baselines. These advances narrow the gap between holistic audio understanding and fine-grained spatial reasoning, moving ALMs toward reliable deployment in embodied, robotic, and immersive applications.

AI USE STATEMENT

We used generative AI to rephrase selected questions in the dataset and benchmark; these questions are identified in the paper. Generative AI was also used to improve the clarity and readability of the manuscript. The authors reviewed all AI-assisted content and assume full responsibility for the final questions and manuscript.

ETHICS STATEMENT

Our dataset and benchmark are derived from publicly available datasets. We acknowledge that limitations or biases in the source datasets may propagate to our benchmark. We will comply with the applicable terms of use of the source datasets when releasing our data and benchmark.

REPRODUCIBILITY STATEMENT

Appendix A details the construction rules for the dataset and benchmark; Appendix B provides the experimental settings and implementation details. We will publicly release the dataset and benchmark upon acceptance.

REFERENCES

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.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

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.

Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Yang Zheng, Jingyuan Li, and Eli Shlizerman. Savvy: Spatial awareness via audio-visual llms through seeing and hearing. *Advances in Neural Information Processing Systems*, 38:118999–119038, 2026.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. *Advances in Neural Information Processing Systems*, 37:33505–33537, 2024.

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.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

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. *arXiv preprint arXiv:2503.03983*, 2025.

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

Yuan Gong, Hongyin Luo, Alexander Liu, Leonid Karlinsky, and James R Glass. Listen, think, and understand. In *International Conference on Learning Representations*, volume 2024, pp. 18516–18545, 2024.

Google DeepMind. Gemini 3.1 pro: Model card, February 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-pro/`.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 35634–35670, 2026.

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. *arXiv preprint arXiv:2402.01831*, 2024.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5622–5632. IEEE, 2025.

Bin Lin, Bo Zhao, Boyong Wu, Chao Yan, Chen Wu, Cheng Yi, Chengyuan Yao, Daijiao Liu, Fei Tian, Feng Tian, et al. Stepaudio 2.5 technical report. *arXiv preprint arXiv:2605.23463*, 2026.

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.

Hyeonggon Ryu, Joon Son Chung, and David Harwath. Hear you are: Teaching llms spatial reasoning with vision and spatial sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 38606–38615, 2026.

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*, 2025.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36:72931–72957, 2023.

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. Enhancing temporal understanding in audio question answering for large audio language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pp. 1026–1035, 2025a.

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. Spatial audio motion understanding and reasoning. *arXiv preprint arXiv:2509.14666*, 2025b.

Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. Spatial audio question answering and reasoning on dynamic source movements. *arXiv preprint arXiv:2602.16334*, 2026.

Parthasaarathy Sudarsanam and Archontis Politis. Towards spatial audio understanding via question answering. *arXiv preprint arXiv:2507.09195*, 2025.

Luoyi Sun, Xiao Zhou, Zeqian Li, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spotsound: Enhancing large audio-language models with fine-grained temporal grounding. *arXiv preprint arXiv:2604.13023*, 2026.

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, et al. Can large language models understand spatial audio? *arXiv preprint arXiv:2406.07914*, 2024b.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Hualei Wang, Yiming Li, Shuo Ma, Hong Liu, and Xiangdong Wang. Listening between the frames: Bridging temporal gaps in large audio-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26233–26241, 2026.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

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

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

Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, et al. Moss-audio technical report. *arXiv preprint arXiv:2606.01802*, 2026a.

Chiao-An Yang, Ryo Hachiuma, Sifei Liu, Subhashree Radhakrishnan, Raymond A Yeh, Yu-Chiang Frank Wang, and Min-Hung Chen. 4d-rgpt: Toward region-level 4d understanding via perceptual distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31042–31053, 2026b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026c.

Masahiro Yasuda, Yasunori Ohishi, and Shoichiro Saito. Echo-aware adaptation of sound event localization and detection in unknown environments. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 226–230. IEEE, 2022.

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.

Goksenin Yuksel, Marcel van Gerven, and Kiki van der Heijden. Gram: Spatial general-purpose audio representations for real-world environments. *arXiv preprint arXiv:2602.03307*, 2026.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8995–9006. IEEE, 2025.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 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. *arXiv preprint arXiv:2402.01591*, 2024.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to Reason in 4D: Dynamic Spatial Understanding for Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9637–9646, June 2026.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. Vlm4d: Towards spatiotemporal awareness in vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8600–8612. IEEE, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 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.

In the appendix, we provide the dataset and benchmark construction details and statistics in Appendix A, and additional implementation details and more experiment results in Appendix B.

A Construction of the Training Data and Benchmark

This section describes the synthesis rules, data splits, dataset statistics, and representative examples for the spatial-audio question-answering data. The dataset consists of two components: spatial-perception questions constructed from real first-order Ambisonics (FOA) recordings in STARSS23, and room-acoustics questions generated by rendering anechoic FOA-MEIR signals in shoebox rooms. All questions are multiple-choice. The options and correct answers are computed from annotations or rendering metadata rather than generated by a language model. Question templates are paraphrased while preserving their semantics and numerical values.

A.1 Data Construction

A.1.1 STARSS23

STARSS23 provides FOA recordings annotated with source azimuth, elevation, and distance. Annotations are sampled at 0.1 s intervals. Each frame records the sound class, source identifier within that class, and the source azimuth, elevation, and distance in the listener-centered coordinate system. An azimuth of 0° denotes the front; positive and negative values denote the left and right, respectively; and $\pm 180^\circ$ denotes the rear. An elevation of 0° corresponds to the ear plane, with positive values indicating positions above it. Distances are converted from centimeters in the original annotations to meters.

Because the original recordings vary in duration, we first divide them into short clips while preserving event integrity whenever possible. Frame-level annotations are then converted into source-centric metadata, from which multiple-choice questions are generated.

Recordings no longer than 60 s are retained in full. Longer recordings are segmented from left to right at approximately 45 s. The final segment is retained in full if it does not exceed 60 s. A sound event is defined as a continuous active interval associated with a particular sound class and source identifier. Events no longer than 45 s are kept intact whenever possible. If a planned boundary falls within such an event, it is moved to the nearer endpoint so that the event remains on one side of the boundary. If this adjustment would produce a segment longer than 60 s, the recording is instead split near 45 s. When multiple short events cross a candidate boundary, we select a boundary that cuts the fewest events while remaining close to 45 s. Continuous events longer than 45 s, such as music or domestic background sounds, may be split. A trailing segment shorter than 5 s is merged into the preceding segment, and segment boundaries are rebalanced where possible to obtain durations between 30 and 60 s.

The CSV annotations of each segment are then converted into JSON metadata organized by sound class and source entity. Different source identifiers within the same class are treated as distinct entities. Consecutive active frames are merged into active intervals. A frame-level trajectory is retained if the source position changes within the segment; otherwise, only one static position is stored. When multiple entities of the same class must be distinguished, they are named *first*, *second*, and *third* according to their order of appearance within the current segment. Entity indices from the original unsegmented recording are not retained.

Questions are generated only while the target entity is active. Comparison and interval questions are generated only when the candidates are sufficiently distinguishable: distance differences must be at least approximately 0.1 m, and azimuth or elevation differences must be at least approximately 15°. Distractors in fine-grained questions are also separated by comparable margins. If a valid and sufficiently separable question cannot be formed for an event or entity pair, that task is skipped. At most one question of each task type is generated per event, entity pair, or segment, preventing the 0.1 s frame rate from artificially inflating the dataset.

Coarse-grained questions explicitly define their bins. Distance is divided into *near* (< 1.5 m), *mid-range* (1.5–4.0 m), and *far* (> 4.0 m). Azimuth is divided into four 90° sectors: front, left, back, and right. Elevation is divided into above, approximately level with, and below the ear plane, using

thresholds of approximately $\pm 10^\circ$. Fine-grained distance questions use either intervals approximately 0.5 m wide or point values rounded to 0.1 m. Fine-grained azimuth and elevation questions use intervals containing the ground-truth value or rounded point values.

Relative Position questions are asked when two sources are simultaneously active. They cover Relative Distance, Relative Azimuth, Relative Elevation, Same Side, and Same Quadrant. Multi-source questions cover Source Count, Extreme Source, Azimuth Ordering, and Spatial Identification. Motion questions compare source positions at the beginning and end of a trajectory (Motion Trajectory) or predict the source sector one second into the future (Motion Prediction). Spatio-temporal Localization questions ask when a source is closest (Closest Moment), when it starts moving (Motion Onset), when it occupies a specified direction (Directional Timing), or whether a source on the left or right begins first (Onset Comparison).

Hallucination questions cover Source Existence, Directional Existence, and Temporal Existence. Negative examples use either classes absent from the segment, inactive times, or directions not occupied by the queried source. Ego-rotation questions assume that the listener turns 90° left, 90° right, or 180°, and ask for Rotated Azimuth under the same sector definitions.

A.1.2 FOA-MEIR AND ROOM SIMULATION

FOA-MEIR provides anechoic sound events and measured room impulse responses. We use only its anechoic sound library. Room reverberation is rendered as FOA audio using SoundSpaces 2.0 in rectangular shoebox rooms. The library contains 12 classes with 20 independent recordings per class, for a total of 240 recordings: organ, piano, toy_train, gun_shot, metallophone, bicycle_bell, security_buzzer, shaker, hand_clap, wood_block, shaking_bell, and hit_drum.

Room volumes are divided into four non-overlapping ranges. Within each range, we independently sample the room volume, ceiling height, and horizontal aspect ratio, from which the room dimensions are derived. Minimum and maximum side-length constraints are imposed to avoid unrealistically flat or elongated rooms. The listener is placed near floor level at an ear height of approximately 1.5 m, with a randomly sampled position separated from the room boundaries by a safety margin. Minimum distances are enforced between the listener and sources and between source pairs. Source heights are approximately aligned with the listener's ear plane. Moving sources are linearly interpolated between start and end positions in the horizontal plane, with a minimum displacement sufficient to produce perceptible motion.

Each single-room scene is 30 s long and contains two or three event classes. One anechoic recording is sampled from each class, except in approximately one-third of the cases, in which two recordings are sampled. Events may overlap or be replayed, and each scene contains at least one moving source. We generate 2,000 training scenes and 500 test scenes, balanced across volume ranges A–D.

Room-comparison scenes concatenate two 15 s recordings into one 30 s example. Each half contains one or two event classes, and source motion is optional. Comparison pairs are sampled either from the same range among B, C, and D or from adjacent ranges A–B, B–C, and C–D. Pairs separated by two or more ranges are excluded to avoid trivial comparisons. For same-range pairs, the room-volume ratio is approximately 1.2–1.5. We generate 600 training pairs and 150 test pairs, balanced across comparison types. No test room shares the same length–width–height signature.

All Room Acoustics answers are computed from the rendered room volume. Each single-room scene produces two questions. The first is a Volume Range question that asks which of the four fixed volume ranges contains the room. The second is a Volume Estimation question that asks for its approximate volume: the correct option is the true volume, while the three distractors are sampled from the same range with relative errors of approximately 10%–20%. This avoids both cross-range shortcuts and nearly indistinguishable options. Comparison questions are Size Comparison questions that ask whether the first segment (0–15 s) or the second segment (15–30 s) was rendered in the larger room.

A.1.3 VALIDATION RULES

After generation, each clip's annotations are reloaded, and the gold answer is recomputed using the same coordinate system, activity criteria, and separability thresholds as in generation. The result is then compared with the stored answer.

For STARSS23, all entities and trajectories in the clip are reconstructed. The queried time must fall within the target entity's active interval, with a nearby trajectory point available. Comparison questions are rejected if the distance difference is below approximately 0.1 m or the angular difference is below 15°. A Same Side question is invalid if the sources do not lie on the queried axis. Each question must have two to five distinct options, with a valid answer label. Ordinals such as first and second require at least two same-class entities in the current clip and cannot be inherited from the original recording. For hallucination questions, entity existence, activity, and directional sector are recomputed from the annotations rather than inferred from the stored polarity. Questions without a reconstructable gold answer are marked as unverified.

Room questions are validated against the metadata recorded during rendering. The volume must equal length times width times height and fall within its assigned category (A–D). Comparison scenes must contain exactly two segments, and their recorded ordering must agree with their computed volumes. For the Volume Range task, the answer and selected interval must match the metadata category. For the Volume Estimation task, the answer must be the option closest to the true volume, with all four options in the same category. For the Size Comparison task, the answer must identify the larger room, and the rooms must belong to the same or adjacent categories.

Finally, paraphrasing must not change the options, answer label, or question type, and numerical bins in coarse-grained questions must be preserved.

A.1.4 Training and Benchmark Splits

All splits strictly follow the original train/test partitions of the datasets. Neither test recordings nor test anechoic signals are used for training. For STARSS23, `dev-train-sony` and `dev-train-tau` are used for training, while `dev-test-sony` and `dev-test-tau` are used for evaluation. For FOA-MEIR, recordings indexed 0–15 within each class are assigned to training, and recordings indexed 16–19 are assigned to testing. Thus, no anechoic recording is shared across the two splits. The synthesized room scenes follow the same partition.

Spatial-perception questions are generated from all eligible STARSS23 segments, and room-acoustics questions are generated from all eligible single-room and comparison scenes. To reduce answer imbalance, we consider each question type separately. When one answer occurs more frequently than all others, we randomly downsample questions with that answer until its count matches that of the second-most frequent answer. The training data combine the resulting STARSS23 and room-acoustics question pools. After stratifying by task and holding out 200 questions for validation, the final training set contains 22,472 question-answer pairs, as shown in Fig. 4(a).

The benchmark is not the full test question pool. We apply the same answer-balancing procedure within each question type. We additionally balance the number of questions across question types by randomly downsampling types with disproportionately large pools. Questions are sampled from the test split with balancing across sound classes and task types. The resulting benchmark contains 3,045 question-answer pairs, covering all 269 STARSS23 test segments and a subset of the FOA-MEIR test scenes.

In addition, we randomly sample 10k examples from SO-Dataset for joint training. The training data distribution, shown in Fig. 4(b), is largely consistent with that of the original dataset. Moreover, there is no overlap between the training and test data.

A.2 Dataset Statistics

A.2.1 Task Taxonomy and Training Distribution

The dataset contains eight task categories.

- **Absolute Position**: Distance Estimation and Coarse Distance for fine- and coarse-grained distance; Azimuth Estimation and Coarse Azimuth for fine- and coarse-grained azimuth; and Elevation Estimation and Coarse Elevation for fine- and coarse-grained elevation.
- **Relative Position**: Relative Distance for identifying the closer source; Relative Azimuth for left–right relations; Relative Elevation for vertical relations; Same Side for determining

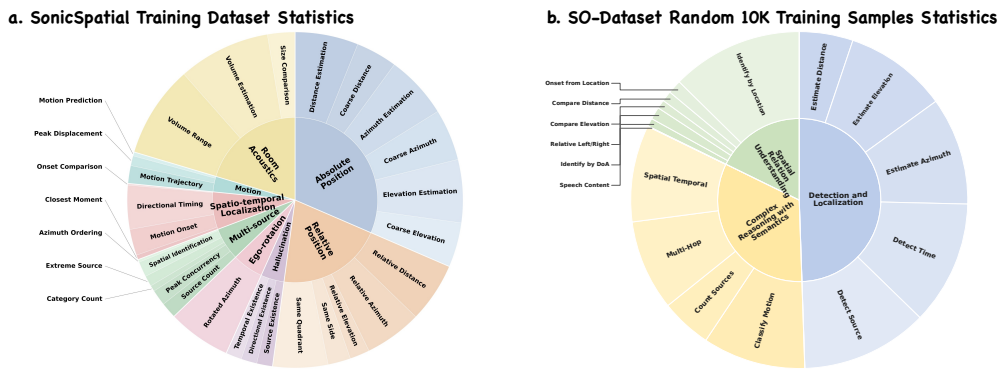


Figure 4: Statistics of Datasets.

whether two sources are on the same side of the listener; and Same Quadrant for determining whether they occupy the same quadrant.

- **Multi-source**: Source Count for the total number of source entities; Peak Concurrency for the maximum number of simultaneously active sources; Category Count for the number of entities in a given class; Extreme Source for identifying the nearest or farthest source; Spatial Identification for inferring a sound category from its direction; and Azimuth Ordering for clockwise source ordering from the front.
- **Spatio-temporal Localization**: Closest Moment for identifying when a source is closest; Motion Onset for identifying motion onset; Directional Timing for identifying when a source occupies a specified direction; and Onset Comparison for comparing the onset times of sources on the left and right.
- **Motion**: Motion Trajectory for classifying motion during an active interval; Peak Displacement for identifying the source with the largest azimuthal displacement; and Motion Prediction for predicting the source sector one second ahead.
- **Hallucination**: Source Existence for determining whether a sound class occurs; Directional Existence for determining whether it occurs in a specified direction; and Temporal Existence for determining whether it occurs at a specified time.
- **Ego-rotation**: Rotated Azimuth for inferring source azimuth after the listener rotates.
- **Room Acoustics**: Volume Range for room-volume classification; Volume Estimation for numerical volume estimation; and Size Comparison for comparing the sizes of two rooms.

A.2.2 Audio Duration

The training data cover 2,944 unique audio clips, including 350 STARSS23 segments and 2,594 FOA-MEIR scenes. The benchmark covers 576 clips, including 269 STARSS23 segments and 307 FOA-MEIR scenes. All FOA-MEIR single-room and comparison scenes are 30 s long. STARSS23 segments range from 30 to 60 s. Their mean durations are 42.5 s in the training split and 43.0 s in the benchmark, with most segments falling between 40 and 50 s. Because FOA-MEIR scenes constitute the majority of the data, the overall mean audio durations are 31.5 s for training and 36.1 s for the benchmark.

A.2.3 Number of Sound Events per Clip

In STARSS23, events are counted as distinct entities. In FOA-MEIR, they are counted as the anechoic sources inserted into a scene; for comparison scenes, sources in both halves are included. The training and benchmark clips contain an average of 3.56 and 3.75 events, respectively. The similar event densities indicate that the benchmark is neither substantially simpler nor more crowded than the training data.

B EXPERIMENT DETAILS

We use the open-source MOSS-Audio-8B-Instruct audio-language model as the backbone and integrate a pretrained GRAM-Ambisonics spatial encoder. The model is jointly optimized in a single stage. The MOSS audio encoder and the GRAM backbone remain frozen. We train the Spatial Selection Module (SSM) and the LoRA parameters in the language model. All other parameters remain frozen. The LoRA modules use a rank of 16, a scaling factor of 32, and a dropout rate of 0.05, with no bias parameters trained. The language-model output head remains frozen.

B.1 TRAINING DATA AND OPTIMIZATION.

We train exclusively on four-channel first-order Ambisonics (FOA) multiple-choice examples. The training and validation sets contain $32,472$ and 200 examples, respectively. We use AdamW with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 0.01. The learning rate follows a cosine schedule with linear warm-up over the first 3% of the total training steps. Gradients are clipped to a maximum norm of 1.0. Training uses bfloat16 mixed precision and gradient checkpointing. The per-device batch size is 16, without gradient accumulation, and the model is trained for three epochs.

B.2 INPUT REPRESENTATION AND ARCHITECTURE

The model takes a four-channel FOA waveform as input. The semantic branch uses the W channel, resampled to $16\,\mathrm{kHz}$, and extracts 128-dimensional log-Mel features with a hop size of 160 samples. These features are processed by the MOSS audio encoder, which has a hidden dimension of $1{,}280$ and approximately 12.5 audio tokens per second. An adapter with a hidden dimension of $8{,}192$ then projects the tokens into the $4{,}096$-dimensional language-model space. The spatial branch resamples the waveform to $32\,\mathrm{kHz}$ and extracts seven-channel, 128-dimensional log-Mel and acoustic-intensity-vector features at $100\,\mathrm{Hz}$. The GRAM encoder converts these features into 768-dimensional spatial tokens, which are mapped to $4{,}096$ dimensions through a two-layer linear projection with a GELU activation. The SSM is dimensionally aligned with the language model, with a hidden dimension of $4{,}096$, and 32 attention heads. It comprises two Q-Former layers, in which learnable queries first interact with the question text and then attend to the GRAM spatial tokens. The SSM produces a fixed set of 32 geometry tokens, each of dimension $4{,}096$, which are concatenated after the audio tokens and passed to the language model. Cross-entropy loss is computed only over the answer tokens, with the option letter as the sole supervision target.

B.3 CONVERSATION TEMPLATE

We follow the MOSS-Audio chat template. The user message begins with audio boundary markers and placeholder tokens corresponding to the audio sequence, followed by 32 geometry placeholders and the question text. Each question is formatted as a question stem, a list of options labeled A, B, C, and so forth, and an output constraint. The assistant target is the corresponding single option letter, such as `A`.

Prompt for Spatial Audio Question Answering

System: You are a professional spatial audio understanding expert. Please listen to the audio carefully, answer the question, and select the correct option.

User: {audio_start}{audio_tokens}{geometry_tokens}{audio_end}

{question} Choose the best answer from the following options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}

Answer with the option letter only.

Assistant: {answer_letter}