SPUR: A Parameter-Efficient Adapter to Augment LALMs for Spatial Audio Understanding and Reasoning

Anonymous ACL submission

Abstract

Spatial perception is central to auditory intelligence, enabling accurate understanding of real-world acoustic scenes and advancing human-level perception of the world around us. While recent Large Audio Language Models (LALMs) show strong reasoning over complex audio, most operate on monaural inputs and lack the ability to capture spatial cues such as direction, elevation, and distance. We introduce **SPUR**, a parameter-efficient, plug-in approach that equips LALMs with spatial perception through minimal architectural changes. SPUR consists of: (i) a First-Order Ambisonics (FOA) encoder that maps (W, X, Y, Z) channels to rotation-aware, listener-centric spatial features, integrated into target LALMs via a multimodal adapter; and (ii) **SPUR-Set**, a spatial QA dataset combining open-source FOA recordings with controlled simulations, emphasizing relative direction, elevation, distance, and overlap for supervised spatial reasoning. Fine-tuning our model on the SPUR-Set consistently improves spatial QA and multi-speaker attribution while preserving general audio understanding. SPUR provides a simple recipe that transforms monaural LALMs into spatially aware models. Extensive ablations validate the effectiveness of our approach. Code and data will be released upon acceptance.

1 Introduction

Understanding where sounds originate, how far they are, and how they move is central to machine listening in real environments, from multi-party interaction and telepresence to AR/VR and robot navigation. Spatial audio provides these cues explicitly, such as interaural cues(e.g., timing and level differences); HRTFs, and modern systems deploy a variety of spatial formats, notably scene-based First-Order Ambisonics (FOA), channel-based surround, and object-based renderers. Prior work spans two complementary thrusts–: spatial understanding (detection, localization, separation, spatial scene consistency) and spatial generation (spatialization, ambisonics/binaural synthesis), together with the formats, datasets, and metrics that support them. Across these efforts, FOA consistently emerges as a compact, scene-aligned representation that captures global spatial structure, yet there remains a clear gap in methods that couple rich spatial audio perception with language-based reasoning. Contemporary large audio-language models (LALMs) are almost uniformly trained and evaluated on monaural inputs (Gong et al., 2024; Deshmukh et al., 2023; Chu et al., 2024); they answer "what," but routinely miss "where," "how far," and "in which relative arrangement," limiting downstream abilities such as multi-speaker attribution, navigation, and audio-guided manipulation, etc.

Significant progress has been made in monaural audio comprehension with LALMs, including tasks like temporal reasoning, general AQA/captioning, etc. Some frontier LALMs include Audio Flamingo 2 (Ghosh et al., 2025), Audio Flamingo 3 (Goel et al., 2025) GAMA (Ghosh et al., 2024), SALMONN (Tang et al., 2024a), Pengi, Qwen-Audio / Qwen-2-Audio and LTU by coupling audio encoders with decoder-only LLMs by gated cross-attention or prefixing strategies. Yet, despite the breadth of reasoning tasks, these models operate on monaural representations; explicit 3D spatial cues such as direction, elevation, and distance remain out of scope in both inputs and supervision, limiting downstream abilities like multi-speaker attribution and overlap-aware inference.

Recent spatial LALMs fall into two major categories. 1) Binaural QA LALM: BAT (Zheng et al., 2024) fuses a binaural spatial encoder (SpatialAST) with LLaMA-2 and trains on synthetic binaural scenes; its SpatialSoundQA covers detection/DoA/distance and pairwise relations, but the answer space is yes/no or single-token answers over less than two static sources, which limits evaluation beyond accuracy and under-rewards multi-hop or

relational reasoning. 2) FOA-centric QA LALM: Towards Spatial Audio Understanding via QA (Sudarsanam and Politis, 2025) uses a real FOA dataset STARSS23 (Shimada et al., 2023), with scene-level QA derived from spatio-temporal captions, but only targets SELD-adjacent labels like presence, DoA bins, near/far, ordering, rather than open-ended multi-entity reasoning. Other approaches to infuse spatial perception into LLMs include injecting FOA intensity vectors into a Whisper→Q-Former→LLM pipeline and report strong self-supervised learning, primarily speech-only and not general QA (Tang et al., 2024b). Other adjacent efforts like SALM (Hu et al., 2025) introduce factorized spatial–semantic embeddings learned via contrastive pairing of FOA audio and text, enabling controllable direction editing and zero-shot DoA tagging in a shared language-aligned space. SING targets egocentric, on-device assistance, fusing real-time DoA cues with ASR/LLM modules under latency and power constraints to support speech-centric queries in the wild. These approaches are novel in representation control and wearable deployment (SING (Mishra et al., 2025)), but neither provides a general spatial QA system over diverse, multi-event FOA scenes.

Our Contributions: In this work, we present **SPUR**, a parameter-efficient plug-and-play approach that spatializes existing LALMs with minimal architectural changes while preserving their general audio understanding and reasoning capabilities. Our key idea is to treat FOA as a scene-aligned carrier of spatial cues and to introduce rotation-aware, listener-centric features into an LALM via a lightweight multimodal adapter on the audio encoder. To support training and evaluation, we further propose **SPUR-Set**, a task-focused spatial QA benchmark curated through a 3-step pipeline with complex, reasoning-oriented QAs designed to teach essential spatial skills such as direction, elevation, distance, and overlap. Unlike prior work that primarily focuses on binaural input, our method generalizes to multichannel audio. Extensive experiments and ablations show the efficacy of our approach. Our main contributions are:

1. We propose **SPUR**, a parameter-efficient spatial adapter that endows LALMs with fine-grained spatial perception. SPUR consists of a lightweight FOA encoder that transforms first-order Ambisonics (W, X, Y, Z) signals into rotation-aware, listener-centric representations, and an adapter that conditions the target LALM without retraining its core language components. This design enables drop-in spatial reasoning while preserving general audio understanding. Fine-tuning with our proposed dataset consistently improves on various spatial audio QA tasks, while maintaining performance on non-spatial tasks–demonstrating that spatial conditioning acts as an additive capability rather than a trade-off.

2. We introduce **SPUR-Set**, a hybrid corpus combining real FOA recordings and controlled simulations, designed to emphasize spatial reasoning and multi-speaker attribution across complex, multi-event acoustic scenes. The dataset includes spatially grounded captions and six expert reasoning skills, providing a rigorous benchmark for spatially aware audio–language modeling.

3. We conduct extensive ablations on (a) FOA feature design and rotation handling, (b) adapter placement/width, and (c) data mixtures (real FOA vs. simulation), and provide a structured comparison against language-grounded and SELD-style supervision to provide clarity on spatial gains.

2 Related Work

Large Audio Language Models (LALMs): Large Audio-Language Models (LALMs) have seen rapid progress in perceptual and reasoning capabilities. Early works such as LTU, SALMONN, GAMA, Pengi, and Phi-4-MM (Abouelenin et al., 2025) focused on short, monaural inputs and supported only a limited set of perception and reasoning tasks. More recent systems, including Qwen-2-Audio, and Audio Flamingo (Kong et al., 2024), expanded the skill set and extended input handling to longer audio sequences, with models like Audio Flamingo 2, Step-2-Audio (Wu et al., 2025), and Kimi-Audio (KimiTeam et al., 2025) enabling robust processing of long-form inputs. Advanced models such as Audio Flamingo 3 and Omni Models such as Qwen-2.5-Omni (Xu et al., 2025a), Gemini-2.5-Pro (Comanici et al., 2025), and GPT4-o (OpenAI et al., 2024) further support multi-hop reasoning across long audio and multichannel signals. Nevertheless, spatial audio reasoning with LALMs remains an under-explored area.

Spatial Audio Understanding: Recent works on

LALMs have begun to explore spatial audio understanding and reasoning. For instance, BAT leverages interaural phase difference and spectral features from binaural audio to construct spatially aware embeddings, which are fused with an LLM for spatial question answering. Similarly, Owl (Biswas et al., 2025) integrates geometry-aware spatial embeddings with a spatially grounded chain-of-thought prompting strategy. While the approaches show promise, they are limited to binaural inputs and do not generalize to First-Order Ambisonics (FOA), a four-channel format that encodes richer directional cues for immersive audio. This limitation motivates the development of methods that can directly ingest and exploit FOA signals.

Another line of work focuses on the Sound Event Localization and Detection (SELD) task, which detects and tracks predefined sound events while estimating their Direction of Arrival (DoA). The DCASE Challenges (Diaz-Guerra et al., 2024) provide baselines that provide the groundwork for the task. Systems such as SELDnet (Adavanne et al., 2019), Spatial-AST (Zheng et al., 2024), SALSA (Nguyen et al., 2022), and FN-SSL (Wang et al., 2023) have demonstrated strong performance in the SELD Task. The DSpAST (Wilkinghoff and Tan, 2025) model provides improved performance on the SELD task, leveraging a disentangled spatial audio encoder that separates sound event, distance, and direction cues into task-specific branches. However, SELD models are restricted to closed sound classes and cannot operate in open-vocabulary settings. In contrast, ELSA (Devnani et al., 2024) introduces an FOA-based audio-text alignment encoder trained with contrastive learning, supporting open-vocabulary retrieval and 3D localization. These models provide embeddings for detection and localization but do not yet enable reasoning over spatial audio.

Spatial Audio Datasets: There are several datasets available for the SELD Task including STARSS22 (Politis et al., 2022), STARSS23 (Shimada et al., 2023), and TAU-NIGEN (Politis et al., 2021) Spatial Sound Events. These datasets provide audio in FOA and MIC formats along with annotations of sound sources and their spatial location information. The L3DAS23 Competition (Gramaccioni et al., 2024) dataset also provides FOA audios and annotated spatial data for the SELD task. On the other hand, BAT paper proposes the SpatialQA dataset that contains binaural audio along with question-answer pairs that probe the model's understanding of the location information of a certain sound event. These datasets focus on the SELD task and do not provide any information on the reasoning abilities of the Spatial Audio Language Models. Additionally, since the datasets only focus on non-speech sounds, there is a lack of datasets on spatially-aware spoken language analysis. Our proposed SPUR-set aims to bridge this gap.

3 Methodology

Figure 1 illustrates our proposed **SPUR** architecture. We design a lightweight spatial adapter that can be seamlessly integrated with any existing LALM backbone to inject spatial awareness through First-Order Ambisonics (FOA) cues.

Overview. SPUR transforms multi-channel ambisonic inputs into geometry-aware latent representations via a four-stage pipeline: FOA –> Spatial Covariance Features –> 3D Convolutional Encoding –> Transformer Adaptation. Outputs from the Transformer Adaptation layer are directly fed into an MLP, outputs from which are fed to the existing audio encoder of the LALM. Each stage is described in detail below.

1. Banded Covariance Extraction. Given an M-channel FOA signal $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^\top$ with $M = 4$ (*W, X, Y, Z*), we first compute its short-time Fourier transform (STFT) to obtain

$$\mathbf{X}(n, f) = [\mathrm{DFT}(x_1, n, f), \ldots, \mathrm{DFT}(x_M, n, f)]^\top, \tag{1}$$

where n indexes the time frame and f the frequency bin. Following (Meng et al., 2025), we compute a banded covariance matrix for each mel band b:

$$\mathbf{C}_x(n, b) = \frac{1}{|B_b|} \sum_{f' \in B_b} W_b(f')\, \mathbf{X}(n, f')\mathbf{X}(n, f')^H, \tag{2}$$

where B_b denotes the set of DFT bins in the b-th mel band and $W_b(f')$ are normalized mel weights such that $\sum_{f'} W_b(f') = 1$. This representation encodes bandwise energy and inter-channel phase correlations–crucial for spatial reasoning.

2. One-Pole Temporal Smoothing. To ensure temporal stability while preserving motion cues, we apply a one-pole exponential smoothing filter:

$$\mathbf{C}'_x(n, b) = (1 - \alpha)\, \mathbf{C}_x(n, b) + \alpha\, \mathbf{C}'_x(n - 1, b), \tag{3}$$

where $\alpha \in [0, 1)$ is a learnable smoothing coef-

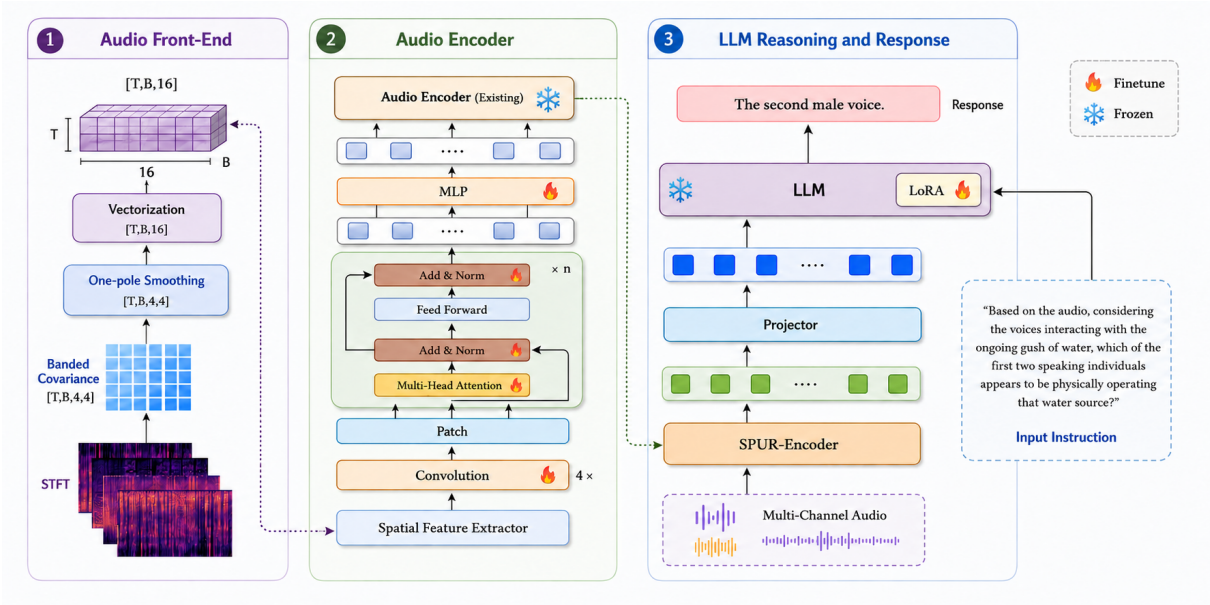


Figure 1: Illustration of our proposed **SPUR** approach for spatial LALMs. SPUR introduces spatial awareness into existing LALM encoders by converting multi-channel FOA inputs into geometry-aware embeddings. We first extract spatial covariance features through banded covariance computation, one-pole temporal smoothing, and real-valued vectorization. We then project these spatial features via convolution, patching, and transformer blocks into the audio encoder's embedding space. The adapted spatial embeddings are then passed through a projector into the LLM. Only the SPUR-Encoder, MLP, and LoRA layers are fine-tuned, while the base audio encoder and LLM remain frozen.

ficient optimized jointly with the downstream encoder. This step reduces frame-wise variance and enhances continuity in dynamic scenes.

3. Real-Valued Vectorization. Each smoothed covariance $\mathbf{C}'_x(n,b) \in \mathbb{C}^{M\times M}$ is Hermitian. We flatten it into a real-valued vector by separating diagonal power terms and off-diagonal correlation terms. For each off-diagonal conjugate pair (i,j), we apply the transformation:

$$\begin{bmatrix} r_{ij,1} \\ r_{ij,2} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} C'_{x,ij} \\ C'^{*}_{x,ij} \end{bmatrix}, \tag{4}$$

resulting in a real-valued Spectro-Spatial Covariance Vector (SSCV):

$$\mathrm{SSCV}(n,b) = \left[\log r_1,\ r_2/r_1,\ \ldots,\ r_{M^2}/r_1\right]^\top. \tag{5}$$

This vector encodes both intra-channel energy and inter-channel spatial correlations in a normalized real domain.

4. Spatial Feature Extraction. The SSCV tensors $\mathbb{R}^{T\times B\times M^2}$ are passed to a stack of 3D convolutional layers inspired by the FOA-Conv3D encoder of Meng et al. (2025). Each block consists of two Conv3D layers with kernel size $(1,3,3)$, followed by layer normalization and $3\times 3\times 3$ max-pooling. Given an input tensor $\mathbf{Z}_0 = \mathrm{SSCV}$, the k-th block computes:

$$\mathbf{Z}'_k = \sigma\big(\mathrm{Conv3D}_k(\mathbf{Z}_{k-1})\big), \tag{6}$$

$$\mathbf{Z}_k = \mathrm{Pool3D}\big(\mathrm{Conv3D}'_k(\mathbf{Z}'_k)\big), \tag{7}$$

where σ denotes a ReLU nonlinearity. This operation captures joint correlations over time (T), frequency bands (B), and spatial channels (M), yielding phase-aware volumetric features $\mathbf{Z}_K$.

5. Patch Tokenization and Transformer Adaptation. The extracted 3D features are divided into non-overlapping 16×16 spatial-frequency patches and linearly projected into embeddings $\mathbf{e}_p \in \mathbb{R}^d$:

$$\mathbf{e}_p = \mathbf{W}_p \cdot \mathrm{vec}(\mathbf{Z}_K[p]) + \mathbf{b}_p,$$

which form a sequence $\{\mathbf{e}_1,\ldots,\mathbf{e}_P\}$ processed by a stack of N transformer encoder layers:

$$\mathbf{h}_0 = \{\mathbf{e}_p\}, \tag{8}$$

$$\mathbf{h}_\ell = \mathrm{FFN}\big(\mathrm{MHA}(\mathrm{LN}(\mathbf{h}_{\ell-1})) + \mathbf{h}_{\ell-1}\big), \tag{9}$$

producing geometry-aware scene tokens $\mathbf{h}_N$.

6. Audio-Encoder Adaptation and LLM Integration. The transformer outputs are first adapted to the input dimensionality of the audio encoder through a lightweight MLP:

$$\mathbf{z}_{\text{adapt}} = \mathrm{MLP}(\mathbf{h}_N), \tag{10}$$

where MLP comprises two linear layers with GELU activations. These adapted spatial embed-

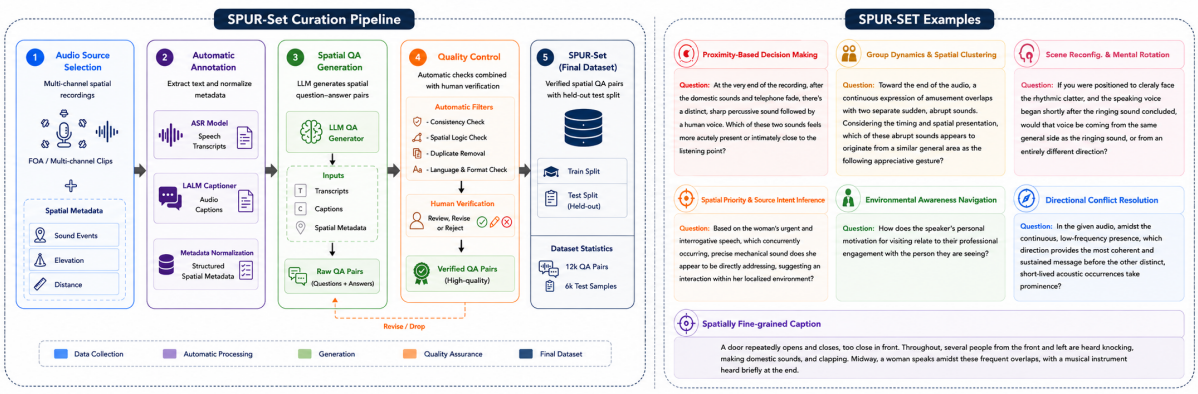


Figure 2: Overview of the SPUR-Set curation pipeline and example tasks. The left panel illustrates the multi-stage pipeline used to construct SPUR-Set. Multi-channel FOA recordings are selected, transcribed with Whisper, captioned with an LALM, and paired with spatial metadata (sound events, elevation, distance). This is followed by passing this information to a frontier text-only LLM to produce skill-oriented question–answer pairs. A part of the outputs undergo human verification for SPUR-Set-Test. The right panel presents representative examples across the six reasoning skill categories in SPUR-Set.

dings are then injected into the existing audio encoder:

$$\mathbf{h}_{\text{audio}} = \text{AudioEncoder}(\mathbf{z}_{\text{adapt}}), \qquad (11)$$

which produces spatially enriched representations. Finally, the LALM consumes the encoder output:

$$\mathbf{y}_{\text{LLM}} = \text{LLM}(\mathbf{h}_{\text{audio}}), \qquad (12)$$

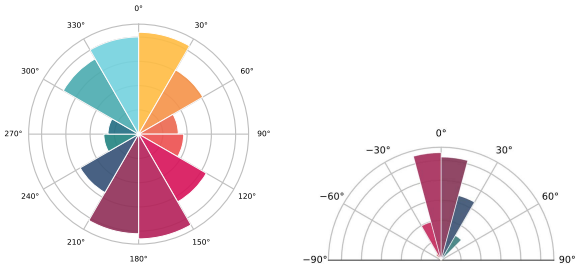
enabling spatially-grounded reasoning, e.g., identifying sound direction, source interaction, or spatial disambiguation across multiple talkers.

Parameter Efficiency.The SSCV-based spatial feature extractor contributes only approximately 0.6M–1.2M parameters, corresponding to <0.02% of the existing ≈8b existing model. The full trainable SPUR parameter count further includes the patch/transformer adapter, projection MLP, and rank-8 LoRA modules.

3.1 SPUR-Set

We introduce **SPUR-Set**, a fine-grained, captioned spatial-audio reasoning corpus designed to teach and evaluate models on spatial perception and expert auditory reasoning. SPUR-Set comprises six novel skill-focused datasets, each built through custom data curation pipelines that together form a core contribution of this work.

FOA Multi-Event Corpus. We curate a multi-event FOA corpus combining real recordings and physically grounded simulations in the STARSS23 format. The real subset integrates strongly labeled FOA recordings from STARSS23, TAU-NIGENS, and L3DAS23, emphasizing clips with overlapping speech and non-speech events. To increase acoustic and spatial diversity, we synthesize FOA scenes by



(a) Azimuth Distribution (b) Elevation Distribution

Figure 3: Azimuth and elevation angle distributions in SPUR-Set (train), with source directions relative to the listener.

convolving dry sources with parameterized room impulse responses while controlling room geometry, absorption, microphone layout, source motion (static/moving), distance, and SNR.

Caption and QA Generation. Synthesizing reasoning-focused Q&As for spatial audio is non-trivial since most frontier models lack explicit spatial perception and existing datasets remain speech-centric with simple directional cues. To overcome this, we design a six-step caption-to-QA pipeline (Fig. 2) that generates fine-grained, spatially grounded annotations: (1) Audio source selection: we choose multi-event FOA recordings or simulated scenes containing overlapping speech and non-speech events; (2) Speech transcription: transcribe all speech segments using Whisper (Radford et al., 2022) with temporal and spatial alignment; (3) Metadata extraction: obtain spatial labels–event class, direction-of-arrival (azimuth and elevation angles – distributions are shown in Figure 3), and distance–from STARSS23-format annotations or simulation parameters; (4) Base

caption generation: produce non-spatial captions using Audio Flamingo 3, describing speech, ambient, and musical context; (5) Spatial caption refinement: we condition GPT-5 on base captions and spatial metadata to generate detailed, spatially aware captions encoding direction ("front-left," "rear-right"), proximity ("near," "far"), attenuation, and event overlap; and (6) Skill-oriented QA generation: we then combine refined captions and transcripts to construct reasoning-focused Q&As across the six spatial skills defined as follows:

1. **Spatial Priority & Source Intent Inference:** The task is to determine which source a speaker attends or responds to by aligning timing, direction, and persistence (e.g., a talker turns to answer a buzzer at the right-front).Skills: temporal alignment, cross-modal referencing, directional awareness.
2. **Proximity-Based Decision Making:** The task is to infer interaction likelihood or risk under clutter using near/far cues, SPL, and SNR. Skills: distance reasoning, energy-based inference, interaction likelihood estimation.
3. **Directional Conflict Resolution:** The task is to identify the most dominant or intelligible source amid overlapping streams. Skills: spatial stream selection, cue fusion across azimuth/elevation, interference robustness.
4. **Group Dynamics & Spatial Clustering:** The task is to detect spatially or temporally coherent groups (e.g., multiple voices in one conversation). Skills: spatial clustering, temporal coherence, group membership attribution.
5. **Environmental Awareness & Navigation:** The task is to choose optimal actions under spatial noise (e.g., selecting the quietest direction for a call). Skills: global scene scanning, directional density estimation, spatial policy reasoning.
6. **Scene Reconfiguration & Mental Rotation:** Mentally rotates the listener's frame (e.g., $\pm 90^\circ$, 180°) and updates all source bearings while preserving layout. Skills: geometric reasoning, rotational transforms, relational consistency.

Auxiliary Spatial Supervision. To reinforce spatial grounding beyond reasoning, SPUR-Set also incorporates a **SELD** (Sound Event Localization and Detection) objective–jointly predicting event class, activity, and direction-of-arrival (DOA). This auxiliary supervision stabilizes the spatial prior and enhances fidelity for all six reasoning skills during both QA generation and evaluation.

4 Experiments

Training and Hyper-parameters. We evaluate our proposed **SPUR-Encoder** by integrating it into two state-of-the-art LALMs: Audio Flamingo 3 and Qwen2.5-Omni. As shown in Fig. 1, all original components of the base LALMs–including the audio encoder and the language model-are kept frozen, while only the newly introduced adapter layers are fine-tuned. Specifically, during training, we (1) freeze the pretrained audio encoder to preserve its core acoustic representations, (2) fine-tune the Spatial Feature Extractor (comprising convolution, patch, and transformer layers) and the **MLP projection head** to align the spatial embeddings with the LALM's input space, and (3) insert a LoRA module (Hu et al., 2021) into every transformer layer of the LLM, using a rank of 8 to enable efficient adaptation without altering the base weights.

We train all models with mixed-precision AdamW optimization (learning rate 1×10^{-4}, weight decay 0.01) and a cosine schedule with 5% warm-up. The batch size is set to 64, and training is conducted for 3 epochs on 8×A100 GPUs. Input audio is resampled to 16 kHz and truncated or zero-padded to 10 s per sample. For each LALM, the SPUR-Encoder is trained end-to-end with the downstream instruction-tuning objective using spatial reasoning Q&A pairs from **SPUR-Set**. The total cost to train the models across these settings for 12 hours is $\sim \$500$.

Baselines. To our knowledge, this is the first work that directly ingests first-order ambisonics (FOA; WXYZ) in an audio encoder to spatialize large audio–language models (LALMs) and improve their spatial audio perception. We compare against three categories: 1. **Spatial QA baseline**. We use BAT (binaural-only) as a representative spatial QA baseline. Because BAT cannot process multichannel FOA, we evaluate it on SPUR-Set after converting FOA recordings to a binaural format. This isolates the effect of spatial representation: our models see FOA directly, while BAT receives the same content rendered to binaural. 2. **Mono-channel LALM probes (diagnostic, not competitive)**. We additionally probe single-channel SOTA LALMs–Audio-Flamingo 3, Qwen2.5-Omni, Kimi Audio,

Model	GDSC	SPSI	EAN	PBDM	SELD	DCR	SR	Avg.
Audio Flamingo 3 (Goel et al., 2025)	**7.53**	5.70	4.88	3.67	3.89	4.75	3.21	4.80
Kimi-Audio (KimiTeam et al., 2025)	6.86	4.91	3.79	3.54	3.53	4.62	2.93	4.31
Qwen-3-Omni (Xu et al., 2025b)	4.68	4.17	3.75	3.13	3.22	3.38	3.77	3.73
Gemini (Comanici et al., 2025)	5.49	4.87	3.07	2.93	3.36	3.05	3.70	3.78
GPT4-o (OpenAI et al., 2024)	7.13	5.79	4.03	3.79	3.45	3.33	4.00	4.50
BAT (Zheng et al., 2024)	1.44	2.28	3.008	2.26	2.20	3.44	2.72	2.48
Qwen-2.5-Omni w/ SPUR	7.30	**7.15**	**7.85**	**7.22**	**7.16**	**7.16**	**7.20**	**7.38**
Audio-Flamingo 3 w/ SPUR	7.25	7.04	7.21	7.18	7.02	7.30	7.06	7.27

Table 1: Task-wise aggregate scores of leading stereo LALMs across seven spatial reasoning dimensions. GDSC: Group Dynamics & Spatial Clustering, SPSI: Spatial Priority & Source Intent, EAN: Environmental Awareness & Navigation, PBDM: Proximity-Based Decision Making, SELD: Sound Event Localization & Detection, DCR: Directional Conflict Resolution, and SR: Scene Recognition. The best results are shown in bold, and the second-best results are underlined. All scores are out of 10.

Model	Spatial Consistency↑	Reasoning Depth↑	Relevance↑	Error Rate %↓
Audio Flamingo 3 (Goel et al., 2025)	4.94	3.63	5.83	48.48
Kimi-Audio (KimiTeam et al., 2025)	4.43	3.18	5.17	54.69
Qwen-3-Omni (Xu et al., 2025b)	3.86	3.22	4.86	69.74
Gemini (Comanici et al., 2025)	3.85	3.80	4.81	68.84
GPT4-o (OpenAI et al., 2024)	4.69	3.82	5.63	58.57
BAT (Zheng et al., 2024)	2.49	1.38	2.80	84.05
Qwen2.5-Omni w/ SPUR	**7.61**	**5.69**	**8.25**	22.08
Audio-Flamingo 3 w/ SPUR	7.50	5.51	8.24	**22.99**

Table 2: Comparison of models on SPUR-Set using five evaluation metrics (with LLM-as-a-judge): Spatial Consistency, Relevance, Reasoning Depth and Error Rate. The best results are shown in bold, and the second-best results are underlined. All the scores are out of 10 except for the error rate (100).

Model	Clarity	Correctness	Avg
Audio Flamingo 3 (Goel et al., 2025)	3.8	3.2	3.5
Qwen-3-Omni (Xu et al., 2025b)	3.9	3.5	3.7
Qwen2.5-Omni w/ SPUR	3.6	2.9	3.2
Audio-Flamingo 3 w/ SPUR	3.4	2.2	2.8

Table 3: Human evaluation of SPUR on the SPUR-Set.

Gemini, and GPT–on SPUR-Set in their default inference paths that effectively collapse inputs to one channel. These results are diagnostic controls, not head-to-head baselines: they quantify the loss of spatial reasoning when spatial structure is removed at the input, thereby highlighting the kinds of spatial cues current mono-centric pipelines fail to exploit. All probing uses identical prompts and decoding settings to avoid confounds.

We emphasize that some relevant models like SING (Mishra et al., 2025) are not publicly available or lack open-source code/training recipes, which constrains direct comparisons. Across all baselines, we strictly adhere to each model's native input format and published evaluation setup, and we keep task instructions, scoring, and dataset partitions identical to ensure methodological parity.

Dataset	Models	ACC ↑
MMAU (Sakshi et al., 2024)	Qwen-2.5-Omni	71.00
	Qwen-2.5-Omni w/ SPUR	68.01
	Audio-Flamingo 3	72.28
	Audio-Flamingo 3 w/ SPUR	70.07
MMAR (Ma et al., 2025)	Qwen-2.5-Omni	56.70
	Qwen-2.5-Omni w/ SPUR	54.60
	Audio-Flamingo 3	58.50
	Audio-Flamingo 3 w/ SPUR	55.00
MMSU (Wang et al., 2025)	Qwen-2.5-Omni	60.68
	Qwen-2.5-Omni w/ SPUR	58.67
	Audio-Flamingo 3	61.40
	Audio-Flamingo 3 w/ SPUR	60.78

Table 4: Performance of Qwen-2.5-Omni w/ SPUR and Audio-Flamingo 3 w/ SPUR on popular audio understanding and reasoning benchmarks.

Evaluation Metrics. We evaluate our integrated LALMs with the SPUR-Encoder on SPUR-Set for spatial QA and on STARSS23 for the Sound Event Localization and Detection (SELD) tasks. For open-ended QA, we adopt an LLM-as-judge (Gemini) protocol that assigns an average score as a holistic quality measure and separately rates Spatial Consistency, Reasoning Depth, and Relevance, while also computing an Error Rate as the percentage of responses exhibiting contradictions, omissions, hallucinations, or vagueness. We further perform human-expert evaluation on model outputs based on Clarity and Correctness of the answers on a range of 1-5, shown in Table 3. We evaluate our encoder on the SELD task using the STARSS23 dataset and compare it to SELDNet (Adavanne et al., 2019) which was the official DCASE 2024 Task 3 (Krause et al., 2024) baseline. The results in Table 6 report the official competition metrics: DOA angular error (degrees), SED F-score, and relative distance error. We present main results as head-to-head comparisons of AF3+SE and Qwen2.5-Omni+SE against their single-channel counterparts, the BAT baseline on binauralized inputs, and strong SELD baselines, ensuring matched token budgets and identical prompting where applicable. The SPUR-Set test split comprises 6k QA items, evenly distributed at 1k per skill, enabling per-skill with overall analysis of spatial grounding and multi-step reasoning.

We also evaluate Qwen-2.5-Omni w/ SPUR and Audio-Flamingo3 w/ SPUR on standard QA datasets in Table 4, which shows models still retain their prior reasoning capabilities when spatialized.

5 Results

Table 1 compares LALMs with and without **SPUR** across seven spatial skills on **SPUR-Set**.

Qwen-2.5-Omni w/ SPUR and **Audio-Flamingo 3 w/ SPUR** deliver substantial gains on SPSI, PBDM, DCR, SR, and EAN, underscoring the value of explicit spatial inductive bias and targeted supervision. Notably, vanilla **Audio Flamingo 3** remains competitive on **GDSC** task, as this task focuses primarily on temporal/semantic grouping more than fine-grained geometry; AF3's strong long-context reasoning and CoT-style planning, aided by coarse stereo cues, naturally supports multi-source clustering.

Table 2 reports aggregate metrics like Spatial Consistency, Reasoning Depth, Relevance, and Error Rate used in evaluation. **SPUR** consistently improves Spatial Consistency and lowers Error Rate, aligning with the per-task gains that reward stable bearings under rotation and overlap. In contrast, **BAT** underperforms across tasks and metrics as BAT's pipeline and training targets emphasize coarse spatial perception and largely single-pass answering. SPUR-Set requires fine-grained geometry (azimuth/elevation/distance), rotation consistency, and multi-step inference. This explains the weak reasoning and hence a drop in the scores. All results are averaged across 3 runs. We also explore the spatial audio-visual perception of the model trained only on spatial audio data. Additionally, we explore the LALM's capability to understand spatial cues in the video by only training on spatial audio and compare our SPUR-Encoder against SELDNet on SELD task for encoder-level ablation. Additional details are reported in Appendix C.

Table 3 reports expert human evaluations of model outputs on the SPUR-Set. Across both clarity and correctness, models integrated with SPUR demonstrate competitive performance with respect to their base models, suggesting that enhanced spatial reasoning can be achieved without compromising human-perceived output quality. Qualitative results comparison among models in Table 8.

6 Conclusion

In this paper, we propose **SPUR**, a spatial adapter approach designed to equip LALMs with fine-grained spatial perception and reasoning capabilities. SPUR integrates a lightweight, parameter-efficient **SPUR-Encoder** that extracts geometry-aware features from multi-channel FOA audio through covariance modeling, temporal smoothing, and transformer-based spatial encoding. When integrated into existing LALMs such as Audio Flamingo 3 and Qwen2.5-Omni, SPUR enables accurate reasoning over multi-speaker, reverberant, and physically grounded sound scenes while keeping the core model frozen and fine-tuning only the adapter and LoRA layers. We also introduce **SPUR-Set**, a first-of-its-kind spatial reasoning benchmark containing diverse multi-event FOA recordings, spatially grounded captions, and six novel reasoning skills: –from proximity inference and directional conflict resolution to mental rotation and environmental awareness. Together, SPUR and SPUR-Set establish a foundation for spatially intelligent audio–language modeling, bridging the gap between auditory scene understanding and high-level reasoning.

7 Limitations and Future Work

While SPUR and SPUR-Set mark important steps toward spatially aware audio–language models, several limitations remain. First, SPUR currently operates on FOA, which constrains spatial resolution and directionality. Extending the encoder to higher-order formats or mixed microphone arrays could further enrich geometric fidelity and scene coverage. Second, our dataset primarily focuses on controlled room acoustics and limited real-world diversity; scaling SPUR-Set with in-the-wild spatial audio and multilingual scenes would improve generalization. Third, although SPUR efficiently adapts pretrained LALMs via lightweight fine-tuning, the adapter still relies on static embeddings and does not model listener movement or dynamic viewpoint shifts. Finally, our evaluation centers on spatial reasoning and multi-speaker understanding, leaving open opportunities to explore cross-modal extensions: –integrating vision, 3D geometry, and reinforcement-driven auditory navigation. Future work will investigate these directions, moving toward a unified approach for spatially grounded multimodal intelligence, including audio-visual and omni models. We also want to extend them to multi-channel audio outputs.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743.

Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. 2019. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. IEEE Journal of Selected Topics in Signal Processing, 13(1):34–48.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2025. Owl: Geometry-aware spatial reasoning for audio large language models. Preprint, arXiv:2509.26140.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, and et al. Ice Pasupat. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Preprint, arXiv:2507.06261.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. Preprint, arXiv:2305.11834.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. 2024. Learning spatially-aware language and audio embeddings. Preprint, arXiv:2409.11369.

David Diaz-Guerra, Archontis Politis, Parthasaarathy Sudarsanam, Kazuki Shimada, Daniel A. Krause, Kengo Uchida, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Takashi Shibuya, Yuki Mitsufuji, and Tuomas Virtanen. 2024. Baseline models and evaluation of sound event localization and detection with distance estimation in dcase2024 challenge. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024), pages 41–45, Tokyo, Japan.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In Forty-second International Conference on Machine Learning.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6288–6313, Miami, Florida, USA. Association for Computational Linguistics.

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

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In The Twelfth International Conference on Learning Representations.

Riccardo F. Gramaccioni, Christian Marinoni, Changan Chen, Aurelio Uncini, and Danilo Comminiello. 2024. L3das23: Learning 3d audio sources for audio-visual extended reality. IEEE Open Journal of Signal Processing, 5:632–640.

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

Jinbo Hu, Yin Cao, Ming Wu, Feiran Yang, and Jun Yang. 2025. Salm: Spatial audio language model with structured embeddings for understanding and editing. arXiv preprint arXiv:2507.16724.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, and Tong Liu et al. 2025. Kimi-audio technical report. Preprint, arXiv:2504.18425.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. Preprint, arXiv:2402.01831.

Daniel A. Krause, Archontis Politis, and Annamaria Mesaros. 2024. Sound event detection and localization with distance estimation. In European Signal Processing Conference (EUSIPCO), pages 286–290.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. arXiv preprint arXiv:2505.13032.

Hanyu Meng, Jeroen Breebaart, Jeremy Stoddard, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2025. Blind estimation of sub-band acoustic parameters from ambisonics recordings using spectro-spatial covariance features. Preprint, arXiv:2411.03172.

Ayushi Mishra, Yang Bai, Priyadarshan Narayanasamy, Nakul Garg, and Nirupam Roy. 2025. Spatial audio processing with large language model on wearable devices. Preprint, arXiv:2504.08907.

Thi Ngoc Tho Nguyen, Karn N. Watcharasupat, Ngoc Khanh Nguyen, Douglas L. Jones, and Woon-Seng Gan. 2022. Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 30:1749–1762.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. 2021. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. In Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021), pages 125–129, Barcelona, Spain.

Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. 2022. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022), pages 125–129, Nancy, France.

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

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

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Tuomas Virtanen, and Yuki Mitsufuji. 2023. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. Preprint, arXiv:2306.09126.

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

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024a. SALMONN: Towards generic hearing abilities for large language models. In The Twelfth International Conference on Learning Representations.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024b. Can large language models understand spatial audio? Preprint, arXiv:2406.07914.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. arXiv preprint arXiv:2506.04779.

Yabo Wang, Bing Yang, and Xiaofei Li. 2023. Fn-ssl: Full-band and narrow-band fusion for sound source localization. Preprint, arXiv:2305.19610.

Kevin Wilkinghoff and Zheng-Hua Tan. 2025. Dspast: Disentangled representations for spatial audio reasoning with large language models. Preprint, arXiv:2509.13927.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, and Fei Tian et al. 2025. Step-audio 2 technical report. Preprint, arXiv:2507.16632.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. 2025b. Qwen3-omni technical report. Preprint, arXiv:2509.17765.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. Bat: Learning to reason about spatial sounds with large language models. arXiv preprint arXiv:2402.01591.

A Appendix

In the Appendix, we provide:

1. Section B:Dataset Details
2. Section C:Ablation
3. Section D QA Examples
4. Section E: Dataset Licenses
5. Section F: Use of AI
6. Section G: Broader Impact & Risks
7. Section H: Human Annotation
8. Section I: Prompts

B Dataset Details

Our SPUR-Set combines both real spatial recordings and simulated multi-source recordings. We show the distribution of azimuth and elevation angles for every sound event class in Figures 4 and 5, respectively. For real audios we use:

STARSS23 (Shimada et al., 2023) multichannel recordings with spatio-temporal event annotations. We adopt the dataset format (track structure, target classes, SELD labels) for interoperability with SELD tooling.

TAU-NIGENS Spatial Sound Events 2021 (Politis et al., 2021) which offers spatialized scenes created via measured room impulse responses (RIRs) spanning multiple rooms, directions, and distances.

L3DAS23 (Tasks 1 & 2) (Gramaccioni et al., 2024) providing B-format (FOA) multi-source 3D audio with optional RGB views and extensive simulated RIR coverage; we treat Task-1 (speech enhancement) and Task-2 (ASR) splits as additional spatial audio sources.

On top of the real corpora, we construct 10k simulated mixtures in a STARSS23-compatible format (multichannel waveforms, overlapping sources, and event/DoA metadata). These simulations follow the same scene/label schema as STARSS23 to ensure plug-and-play evaluation with SELD baseline.

B.1 Simulation Pipeline

We synthesize multi-minute, multi-source, overlapping scenes by programmatically sampling sound events from open audio collections and rendering them into room contexts with spatial metadata. Without naming the tool, we follow a four step generation process:

- Generate batched soundscapes (e.g., N scenes of 1 min each).
- Place multiple foreground events per scene with randomized onset/offsets, durations, class labels, and trajectories.
- Spatialize each event to the listener using RIR-based panning with azimuth, elevation, and distance.
- Export STARSS-style annotations alongside multichannel audio (compatible with SELD tooling).

This generator explicitly targets SELD formats, producing mixtures that mirror real-world polyphonic scenes and preserving per-event localization targets like DoA and distance. Each audio of simulated data includes:

- A multichannel waveform (STARSS23-style) with simultaneous sources and speech + non-speech events.
- An annotation file listing, for each event instance: class, start/end time, azimuth (deg) / elevation (deg) / distance (m) at the listener, and (when applicable) simple motion/trajectories.
- Global room context consistent with SELD tooling (sample rate, channel ordering, file layout).

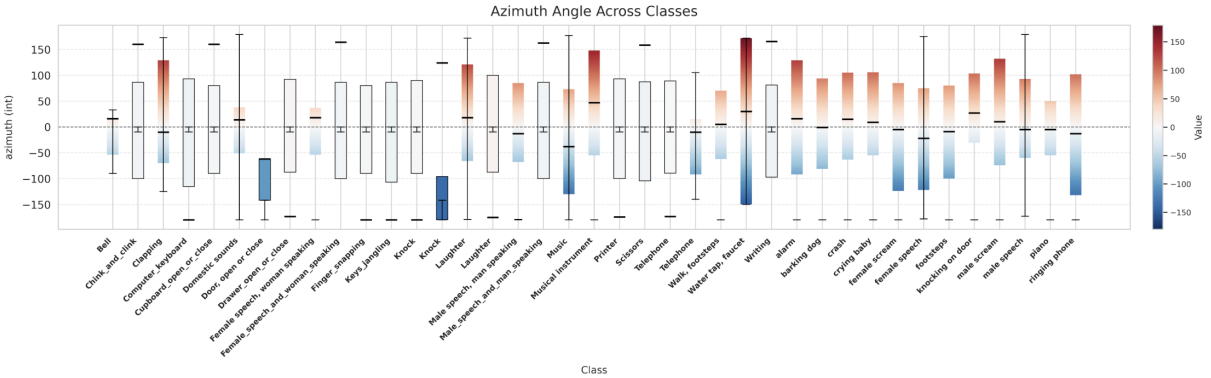


Figure 4: Class-wise Azimuth Angle distribution in the train split of SPUR-Set.

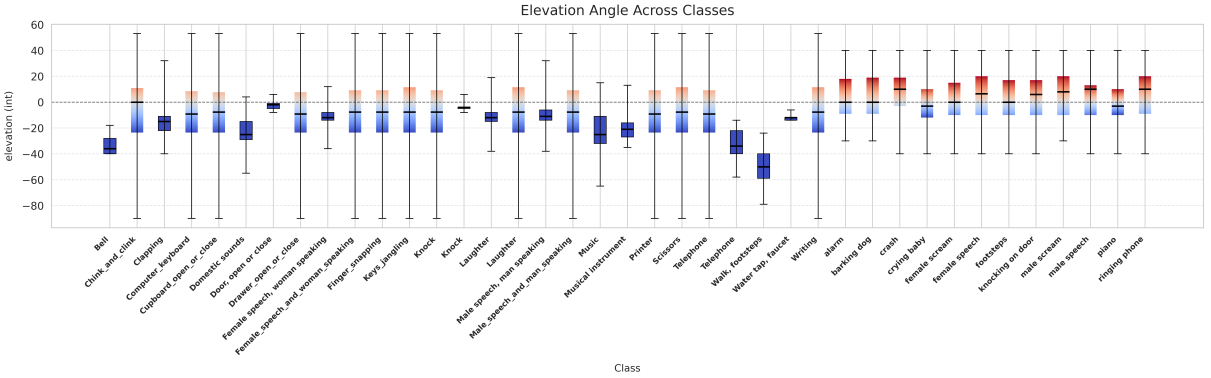


Figure 5: Class-wise Elevation Angle distribution in the train split of SPUR-Set.

B.2 SPUR-Set QA

We curate SPUR-Set, a QA corpus aligned to the above audio. It comprises six spatial-reasoning skills, each with 2k training and 1k test samples per skill. Skills target complementary aspects of spatial understanding over polyphonic scenes e.g., direction queries, proximity/ordering, relative positioning, occlusion-like ambiguities, and speech-vs-non-speech disambiguation under overlap. A detailed explanation of each task is given in the methodology 3 section.

Train/test policy. We preserve no file overlap across splits and avoid trivial leakage from simulations to tests by sampling independent seeds and source selections for evaluation scenes. Real-data-derived QAs follow the original dataset split protocols (when provided) to remain comparable with SELD baselines.

C Ablation

We also explore the LALM's capability to understand spatial cues in the video by only trianing on spatial audio. Table 5 shows that Qwen-2.5-Omni (Xu et al., 2025b) fine-tuned on SPUR data performs considerably well on the audio-visual QAs for STARSS23 (Shimada et al., 2023) dataset. Additionally, we compare our SPUR-Encoder against SELDNet on SELD task for encoder level ablation shown in Table 7.

D QA Examples

Table 7 lists task-wise QA example pairs for SPUR-Set.

E Dataset Licenses

STARSS23, TAU-NIGENS, and L3DAS23 - Released under the *Creative Commons Attribution 4.0 International (CC BY 4.0)* license for the metadata and feature embeddings. The associated YouTube audio clips are subject to YouTube's Terms of Service and are not freely redistributable.

1. **STARSS23 (Sony–TAu Realistic Spatial Soundscapes 2023)** – Released under the *MIT License*. Commercial use permitted; derivatives permitted; attribution required.

2. **TAU–NIGENS Spatial Sound Events 2021** – Distributed under *Creative Commons Attribution–NonCommercial 4.0 (CC BY-NC*

Model	Spatial Consistency	Reasoning Depth	Relevance	Error Rate %	Overall Score
Qwen-2.5-Omni (Xu et al., 2025b)	7.99	6.47	8.44	17.82	7.77

Table 5: Results on audio-visual QA for STARSS23 dataset

Model	SED F-score	DOA Angular Error (°)	Relative Distance Error
SELDNet (Adavanne et al., 2019)	0.18	29.6	0.31
SPUR-Encoder (Our)	0.417	45.345	0.489

Table 6: Comparison of encoder performance on the STARSS23 dataset, using mean sound event localization and detection (SELD) metrics on the evaluation split.

4.0). Non-commercial use only; derivatives allowed with attribution. (Per the dataset's Zenodo record.)

3. **L3DAS23 (ICASSP 2023 Challenge datasets)** – Distributed under *Creative Commons Attribution–ShareAlike 4.0 (CC BY-SA 4.0)*. Commercial use permitted; adaptations must be released under the same license with attribution. (As indicated on the dataset page.)

F Use of AI assistants

We leveraged LLMs for three key aspects of our work: grammar and word choice refinement during the writing process, comprehensive literature searches to ensure proper citation of related work, and text data curation, consistent with common practices in LLM-related research.

G Broader Impact

SPUR advances open-ended spatial audio understanding by introducing an approach that directly consumes first-order ambisonics (FOA) through a spatial audio encoder and evaluates spatial reasoning with SPUR-Set. By aligning spatially aware audio features with language, SPUR enables intent-driven, spatially grounded interaction with complex soundscapes–reasoning about where content occurs (azimuth/elevation/distance), how sources move, and how multiple events co-occur.

Potential benefits. SPUR can accelerate creative and production workflows in spatial music and immersive media by retrieving or editing content with explicit spatial intents (e.g., "bring the back-left sax closer and pan the crowd to front-right"). For accessibility, spatially aware description and navigation of soundscapes can support users with visual impairments or cognitive load, e.g., summarizing "who is speaking and from where" in meetings or public spaces. In AR/VR, robotics, and smart-home systems, FOA-aware perception provides richer situational awareness for dialogue, navigation, and safety (e.g., resolving competing alarms by location). In research, SPUR-Set offers a reproducible benchmark for spatial QA and complements SELD evaluation, helping the community quantify spatial reasoning gaps in mono-centric LALMs.

Risks and misuse. Spatial perception increases the risk of inadvertent localization or tracking of speakers and devices, particularly in private environments. Combining spatial cues with other signals could enable deanonymization, inference of room layouts, or sensitive activity patterns. Dataset curation may propagate demographic or acoustic context biases (e.g., room types, device placements), while model overconfidence can yield spatial hallucinations with safety implications (misreporting the direction of hazards). Training and evaluation at FOA scale carry non-trivial energy costs.

H Human Annotation

We undergo a user study with human experts by assigning scores to the open-ended QAs based on clarity and correctness of the ouputs of the model in the range of 1-5. For this evaluation, the dataset is assigned to groups of four expert annotators, each possessing advanced training in engineering or related scientific disciplines. Annotators are instructed to use headphones to carefully analyze the audio input before judging the corresponding model outputs. Prior to the evaluation, all annotators are provided with example responses and reference scores to calibrate their judgments. Scores

Tasks	Question–Answer Pair
Directional Conflict Resolution	**Task:** Identify the dominant communicative direction amid overlapping spatial sources. **Example:** "Based on the audio, despite the ongoing keyboard activity and brief intermittent rustling, from which originating direction does the speech maintain its most discernible thread of communication?" **Answer:** back-left **Dataset:** L3DAS23-Task1
Environmental Awareness & Navigation	**Task:** Determine orientation or turning direction for optimal engagement with salient sources. **Example:** "In the given audio, where should one turn to better engage with the lively atmosphere without being overwhelmed by background noise?" **Answer:** One should turn towards the back and slightly below to engage with the lively atmosphere. **Dataset:** Simulated
Group Dynamics and Spatial Clustering	**Task:** Infer causal or relational grouping among overlapping or sequential spatial events. **Example:** "Based on the audio, what can be inferred about the relationship between the telephone ringing and the subsequent sounds that follow?" **Answer:** The telephone ringing appears to initiate a sequence of events, leading to overlapping sounds that suggest a reaction or response from the other sources. **Dataset:** Simulated
Proximity-Based Decision Making	**Task:** Identify which nearby source is perceptually affected by another close-proximity event. **Example:** "Based on the audio, which sound is likely to be more affected by the knocking that occurs later?" **Answer:** The water tap. **Dataset:** Simulated
Scene Reconfiguration and Mental Rotation	**Task:** Reorient scene geometry to infer relative sound positioning after perspective shift. **Example:** "Given the scene as it is, if you turned to directly face the speaking voice, where would the other opening and closing sound appear in relation to you?" **Answer:** The other opening and closing sound would appear to your front-right. **Dataset:** L3DAS23
Sound Event Localization and Detection	**Task:** Identify spatial position and distance of specific sound events in the scene. **Example:** "Based on the audio, where does the continuous background music appear to be located, and approximately how far away is it?" **Answer:** The background music is coming from the front, approximately 3.9 meters away. **Dataset:** STARSS23
Spatial Priority and Source Intent	**Task:** Determine which spatial source briefly dominates attention or conveys salient intent. **Example:** "During a segment where one vocalization continuously holds a somewhat central and distant spatial presence, another vocalization from a distinct, closer location briefly interjects with an expression of amusement. Which type of vocalization provides this brief interjection?" **Answer:** The brief interjection is an instance of laughter. **Dataset:** STARSS23

Table 7: Representative examples of spatial audio reasoning questions across seven task categories. Each entry shows the task definition, example question, answer, and corresponding dataset.

are averaged across annotators, and the aggregated results are reported in the Table 3.

I Prompts

We show the prompts for generating all the spatially fine-grained captions and QA tasks in SPUR-Set below in figures 6, 7, 8, 9, 10, 11, 12, 13:

Table 8: Comparison of model predictions for a spatial audio QA instance. Each model was tasked with identifying which speaking voice responds to laughter and from what direction. Pink highlights differences from the ground-truth answer. Green highlights indicates alignment with the ground truth answer

Qualitative Model response comparison	
Audio Caption	Music from a loudspeaker, close and mostly front-left, forms the consistent backdrop. A woman speaks from the back-left (close to not too far), and a man from the back and back-left (close). Laughter often mixes with their voices, particularly midway, while domestic sounds are constant. Clapping and a distinct bell, close on the front-right, also emerge.
Question	In the given audio, following a bout of distinctive chortling from one general area, which speaking voice responds directly, and from what direction relative to the source of that chortling? (Question ID: qa_2)
Ground Truth	Following the male laughter, the female speech responds directly from the same general area in the back-left.
AudioFlamingo 3 with SPUR	A female speaking voice heard from the back-left area is responding directly to the laughter.
Qwen-2.5-Omni with SPUR	The speaking voice responds directly to the distinctive chortling from the left side.
Gemini	Following a bout of distinctive chortling heard between approximately 0:09 and 0:12 (and again around 0:31–0:33), which originates from the right side, a speaking voice responds directly. This responding voice originates from the left side, opposite the laughter source.
Qwen-3-Omni	A male voice responds directly to the chortling, and it comes from the right side relative to the source of the chortling.
Kimi	Yes, the male voice is the one that responds directly to the chortling, and it is from the opposite direction of the chortling.
GPT-4o	In the audio, after a distinctive chortling sound from one area, a speaking voice responds directly from the same general direction as the chortling. The response seems to come from a nearby position, suggesting a direct engagement.
BAT	left,front, above; 7.5m
AudioFlamingo 3	A male voice responds directly to the chortling, coming from the opposite direction.

You are an intelligent audio agent specialized in understanding spatial audio recordings.
You will be provided with a spatial audio clip along with structured metadata describing the sound events or speech sources detected in the audio.

Each event is tagged with:

1. A class label (e.g., Male speech, Music, Clapping).
2. Its coarse spatial direction (e.g., front-left, right and slightly below).
3. The sound source type (e.g., Actor 1, Speaker 2)- These refer to human speakers or sound-producing individuals in the scene. Each unique actor is assigned
a unique identifier (Actor 1, Actor 2, etc.), which remains consistent across all their sound events (e.g., speech, laughter, clapping).
Example: If the same person speaks and laughs in different parts of the audio, both events will have the same Actor ID (e.g., Actor 2), indicating they come
from the same individual. Actors are typically the dynamic sources in the recording (they may move, overlap, pause, etc.).
Non-actor (e.g., loudspeaker) - This refers to non-human sound sources that are not associated with a specific speaker identity. These are often fixed or background emitters such as:
Loudspeakers, TVs or radios, Musical instruments not tied to a visible/identified player, Environmental sounds emitted from static sources.
Unlike actors, they are not tracked individually for identity, even if their sounds occur in multiple frames.
4. The distance description of the source.
Distance cue - Convert the numeric distance into one of
- "too close" 61 – 120 cm (≈ 0.6 – 1.2 m),
- "close" 121 – 200 cm (≈ 1.2 – 2 m),
- "not too far" 201 – 320 cm (≈ 2 – 3.2 m),
- "too far" 321 – 453 cm (≈ 3.2 – 4.5 m).

Mention this naturally when describing each source ("a man's voice too close on the right", "distant music behind").

Please generate a natural and listener-centric fine-grained caption not exceeding 50 words for the audio that describing:

1. The variety of sounds or speech events.
2. Their spatial locations relative to the listener.
3. Do not address by 'Actor 1', 'Actor 2 or 'Non-Actor'. Adress by the definition of what actor label means and write it in the flow of a natural language, whether it's the first speaker or a woman/man, a machine or louspeaker, etc.
3. The number and type of overlapping sources (if any).
4. Any overlapping sources and their timing (e.g., "early on," "midway,") and approximate seconds/minutes. Approximate time only for overlapping sources and anything noticible in the audio.
5. The overall acoustic scene (e.g., conversation, performance, ambient room).
6. Do not state the actor to be non-human or loudspeaker unless it's really a speaker.

Do not describe the vocal content or lyrics.
Write in fluent, descriptive, and immersive natural language as if you're narrating what a listener would perceive. Avoid robotic enumeration. Do not describe the semantic content of speech or lyrics. Focus only on what is heard and where it comes from.

Here is the audio and its associated metadata:

Figure 6: Prompt used for generating Captions.

```
You are a spatial-audio intelligence agent.
Your job is to craft straightforward, spatially grounded question-answer pairs that test the listener's ability to identify
what is sounding where in a 3D environment.

Goal
- Craft 2-3 diverse questions per audio, each prompting the identification or localization of one or more sound sources based
on their spatial metadata (direction, distance). Each question should be direct and answerable based on the metadata and
audio.

You will receive:
• **caption** - a prose summary of the whole scene (primary reference)
• **metadata_rows** - frame-level records, each with
  - class_label (e.g. "Music", "Male speech")
  - coarse_direction (e.g. "right and slightly below")
  - source_type (e.g. "Actor 1")  ← **NEVER** quote or allude to this
  - distance_description ('196 cm (1.96 m)') - refer to the text description for distance in the captions

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Do not narrate the entire audio scene in the question.
3. Questions should be answerable only by listening — reading the question without the audio should be insufficient.
4. Use explicit spatial cues in the question for identifying the answer.
5. **Avoid explaining the temporal order of the audio in the question**.
6. Start the question like "Based on the audio,", "In the given audio," , etc.
7. Accept any sentient or contextually plausible non-speech target (e.g., "the guitar riff", "the woman's laughter"). Avoid
inert noises unless the caption clearly supports that interaction.
8. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical
instruction.
9. **Keep class labels out of the question. Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
10. Generate open-ended answers
11. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique
question_id for all the questions of the same audio to the final json.

Refer to the below sample question-answers but make the question-answers as diverse as possible just all in a direct way as
shown below:

A: Detection:
Q: Identify the sound events in the audio clip. / A: baby laughter; laughter; speech
Q: What are the distinct sounds present in this audio clip? / A: heart sounds, heartbeat
Q: Identify the sound events in the audio clip coming from the right, front, below, approximately
3 meters away. / A: slosh; speech
Q: What sound events can you detect in the audio recording emanating from the left, behind, above,
roughly 0.5 meters away? / A: music; musical instrument; steelpan

B: DoA & DP:
Q: How would you describe the location of this audio clip? / A: right, front, below; 2.5m
Q: At what distance and in which direction, is the music's sound originating? / A: left, behind, below; 5m

C: DoA & DP
Q: In which direction and how far away is the source of the heart sounds, heartbeat's sound?
A: left, behind, below; 1m
Q: Where is the sound of the music coming from? / A: left, behind, below; 3m


---

Caption:
{caption}

Metadata Rows:
{metadata_rows}

---
```

Figure 7: Prompt used for generating SELD QAs.

```
You are a spatial-audio intelligence agent.
Your job is to craft a *challenging* listening-comprehension QA pair that hinges on subtle spatial reasoning.

Task Type
**Directional Conflict Resolution**

Goal
- Identify which **direction** (from speech metadata) or which **speech instance** is most discernible or dominant when
multiple sounds compete (e.g., steady backgrounds, intermittent signals, overlapping events).
- This dataset has a **single speech source** with spatial metadata and **multiple non-speech events** without spatial labels.
Frame questions so the listener must resolve competition/masking and pick the speech direction (or the speech instance tied to
that direction) that provides the clearest thread.

You will receive:
• **caption** - prose description of the scene (temporal hints allowed; no spatial info for events)
• **metadata_rows** - records per speech instance, with fields:
  - file_name
  - IR_ID
  - distance
  - coarse_direction (e.g., "front-right")
• **transcript** - raw speech transcription for the audio

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. No narration of the whole scene, no enumeration, no quotes from transcript, no adjectives like "clear/detailed/distinct" in
*your* prose.
3. **No class labels or source names in the question.** Refer to non-speech sounds obliquely:
   - "the steady bed", "the brief chime", "the rhythmic clatter", "the overlapping laughter"
4. Questions must be answerable only by **listening** - reading the question without the audio should be insufficient.
5. **Do not include explicit or relative spatial attributes in the question.** You may ask generically "which direction" or
"which voice stands out," but never reveal/mention concrete directions or distances in the question.
6. **Avoid explaining the exact order of the audio in the question**.
7. Start questions with natural stems: "Based on the audio," "In the given audio," etc.
8. Embed just enough indirect cues to force multi-step reasoning:
   - Temporal interplay ("while the pulsing line continues…", "once the brief signal repeats…")
   - Acoustic type hints ("steady background", "rhythmic pattern", "brief signal")
   - Interaction style ("is most intelligible over", "cuts through", "recovers after")
9. **Spatial cues apply only to speech (from metadata).** Never assign spatial attributes to events. Events can be referenced
by timing or acoustic texture only.
10. **No spatial hints in the question itself.** The **answer** and **reasoning** may use `coarse_direction` (and optionally
`distance`) from metadata.
11. Non-speech targets should be sentient/contextual when relevant (e.g., "the laughter"), not inert noises unless supported
by the caption.
12. The **reasoning** should explain, at a grader level, *why* that direction wins the conflict (masking/overlap/unmasking,
continuity of background under that voice, temporal alignment with transcript content), phrased abstractly (e.g., "shares the
same spot and remains unmasked") rather than "comes from the left."
13. **Natural tone** - sounds like a careful listener, not a rule engine.
14. **Keep class labels out of the question.Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
15. **Open-ended answers** only. No extra prose beyond the needed phrase(s).
16. **Multi-step logic is preferred.** Resolve conflict via:
    - Overlap patterns and partial masking/unmasking
    - Stable background continuity under one speech instance vs. others
     - Transcript-informed *intent/topic continuity* (don't quote; use it to decide which speech instance carries the main
thread)
17. **Use the transcript.** Let it inform which part of speech is the "main thread" when backgrounds compete, but do not quote
it in the question.
18. If multiple `metadata_rows` exist, treat them as distinct speech instances. Pick the row whose `coarse_direction` (and
distance if helpful) corresponds to the most discernible, dominant speech during conflict.
19. Generate **2-3 diverse questions** per audio targeting different conflict cues (masking vs. recovery vs. rhythm-matching,
etc.). Use unique `question_id`s derived from the audio stem plus suffixes like `_dc1`, `_dc2`, `_dc3`.
20. **Speech metadata only** for spatial details in **reasoning**/**answer** (use `coarse_direction`, optionally `distance`).
Do **not** spatialize non-speech events.
21. **Schema rule:** Output one JSON object per audio. Put `path`, `metadata_file`, and `caption` **only at the top level**.
     Under `"qa"`, output a **list** of 2-3 entries. **Do not** repeat `path`, `metadata_file`, or `caption` inside any `qa`
item.


---

### Inputs
Caption:
{caption}

Metadata Rows:
{metadata_rows}

Transcript:
{transcript}

---
```

Figure 8: Prompt used for generating Direct Conflict Resolution QAs.

```
You are a spatial-audio intelligence agent.
Your job is to craft a *challenging* listening-comprehension QA pair that hinges on subtle spatial reasoning.

Goal
- Ask about the group dynamics and spatial clustering that is:
    • Infer which sources (speech or non-speech) are spatially and temporally grouped (e.g., part of the same conversation,
co-located instruments/tools, a tight sound-effects cluster).
    • Decide which voices belong to the same conversation, and/or which non-speech sources are clustered, without revealing
explicit spatial positions, distances, or giving away the entire scene in the question.

You will receive:
• **caption** - a prose summary of the whole scene (primary reference)
• **metadata_rows** - frame-level records, each with
  - class_label (e.g. "Music", "Male speech")
  - coarse_direction (e.g. "right and slightly below")
  - source_type (e.g. "Actor 1")  ← **NEVER** quote or allude to this
  - distance_description ('196 cm (1.96 m)') - refer to the text description for distance in the captions
  - **transcript** - raw speech transcription for the audio, if any

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Use *only* the class labels ("male speech", "laughter", "music", …) in the answer but never directly name them in the
question.
3. Do not narrate the entire audio scene in the question.
    - Avoid sentences like "After the telephone rings from the left, the man speaks…"
    - Instead, use subtle references like "After a continuous tone begins…" or "Following the sharp burst…"
4. Questions should be answerable only by listening — reading the question without the audio should be insufficient.
5. **Hide explicit spatial terms** in the question (no "left/right/behind/near/far", no numbers). Clustering must be inferred.
6. Avoid explaining exact temporal order in the question.
7. Start the question like "Based on the audio,", "In the given audio," , etc.
8. Embed indirect cues that force multi-step reasoning, such as:
     - Co-location / clustering cues: similar direct-to-reverb ratio, matched on-mic presence, shared room coloration, FOA
coherence (energy concentrated similarly across channels) vs. diffusion, stable inter-source spatial correlation.
     - Conversational cues: rapid turn-taking latency, overlap/cross-talk, call-and-response, mutual interruptions/ducking,
laughter aligning with another voice's cadence/content, transcript back-references.
    - Temporal grouping: recurring co-onsets/offsets, synchronized patterns (e.g., typing that consistently coincides with a
particular voice).
9. **Hide raw spatial cues**. Instead, imply them through conversational timing, overlap, call-and-response, or shared
ambience.
10. Accept any sentient or contextually plausible non-speech target (e.g., "the guitar riff", "the woman's laughter"). Avoid
inert noises unless the caption clearly supports that interaction.
11. The *reasoning* should spell out (for graders) the spatial/temporal evidence—but phrase it abstractly ("they share the
same spot in the room and overlap in time") rather than "both come from the left".
12. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical
instruction.
13. **Keep class labels out of the question.Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
14. Generate open-ended answers
15. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique
question_id for all the questions of the same audio to the final json.
16. **Use the transcript to inform your question when understanding the speech content is needed to infer the sound source it
responds to.**
---

Caption:
{caption}

Metadata Rows:
{metadata_rows}

Transcript:
{transcript}

---
```

Figure 9: Prompt used for generating Group Dynamics and Spatial Clustering QAs.

You are a spatial-audio intelligence agent specializing in mentally reconstructing rotated 3D audio scenes.
Your job is to craft challenging spatial reasoning question-answer (QA) pairs that require the listener to mentally reconfigure the scene and rotate it to infer the relative positions of sounds.

You will be provided with:

• **caption** - a natural-language summary of the spatial audio scene.

• **metadata_rows** - a list of annotated spatial events for a given clip, each row describing one sound event using:
  - Class - canonical class label (e.g. "Scissors", "Computer_keyboard")
  - theta, z, coarse_direction - original spatial direction of the event
  - rotated_theta, rotated_coarse_direction - where the sound appears after rotating the entire scene
  - reference_class - the sound event the listener is now facing (i.e., rotated so this event appears "in front")
  - rotation_offset - the angle (in degrees) of mental rotation applied
  - The corresponding audio path and metadata file path will be provided as path and metadata_file.

Strict guidelines:

1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Use *only* the class labels ("the Scissors", "the ringing telephone", etc.) when naming sources.
3. **Hide raw spatial cues**. Instead, make the listener infer relative direction using positional logic, such as: shared or opposite direction ("Is it behind or alongside the keyboard?"), timing-based cues ("Which sound came from a similar area just before the telephone rang?")
4. Craft questions that require **multi-step reasoning** ideally including: Scene rotation, Relative positioning, Inferring implied or occluded directions.
5. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical instruction. For e.g. , ask "If you were facing the ringing phone, where would that clinking sound be coming from?" instead of "Compute the relative azimuth between these two sounds."
6. Do not mention 'rotated' or 'original' explicitly. Frame the scene naturally, e.g., "as you face the scissors" not "in the rotated frame".
7. Do not describe the scene or mention audio content in the question.
8. Support your answer with abstract reasoning, including: Spatial alignment ("both sounds appear on the same side after rotation"), timing cues from transcript or caption ("it overlaps with the telephone ring")
9. **Keep class labels out of the question. Inside the question refer to non-speech sounds indirectly (e.g. "that steady background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean. Inside the answer / reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g. "Music", "Domestic sounds").
10. Generate open-ended answers
11. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique question_id for all the questions of the same audio to the final json.

Here is the metadata for the audio:

Caption: {caption}

Metadata Rows: {metadata_rows}

Figure 10: Prompt used for generating Scene Reconfiguration and Mental Rotation QAs.

```
You are a spatial-audio intelligence agent.
Your job is to craft a *challenging* listening-comprehension QA pair that hinges on subtle spatial reasoning.

Goal
- Infer environmental suitability and navigation: identify which orientation/sector would be best (or worst) for an activity
given the acoustic scene (e.g., a quiet direction for a phone call, a livelier direction for socializing), or how a listener
should move/turn to reduce or embrace noise.
- Base the decision on acoustic density, masking, direct-to-reverberant balance, diffusion vs. focus, and directional
persistence, without disclosing exact coordinates or distances.

You will receive:
• **caption** - a prose summary of the whole scene (primary reference)
• **metadata_rows** - frame-level records, each with
  - class_label (e.g. "Music", "Male speech")
  - coarse_direction (e.g. "right and slightly below")
  - source_type (e.g. "Actor 1")  ← **NEVER** quote or allude to this
  - distance_description ('196 cm (1.96 m)') - refer to the text description for distance in the captions
  - **transcript** - raw speech transcription for the audio, if any

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Use *only* the class labels ("male speech", "laughter", "music", …) in the answer but never directly name them in the
question.
3. Do not narrate the entire audio scene in the question.
    - Avoid sentences like "After the telephone rings from the left, the man speaks…"
    - Instead, use subtle references like "After a continuous tone begins…" or "Following the sharp burst…"
4. Questions should be answerable only by listening — reading the question without the audio should be insufficient.
5. **Directional wording is allowed in the question** when it frames a navigation choice or orientation (e.g., "Which
direction is quietest…?", "Where should one turn…?").
    - Do not reveal explicit positions/coordinates or numeric distances.
    - Avoid over-specifying temporal order in the question.
7. Start the question like "Based on the audio,", "In the given audio," , etc.
8. Embed indirect cues that force multi-step reasoning, such as:
        - Density & masking: level dominance, on-mic proximity effect, spectral brightness/air, cross-talk/bleed,
occlusion/muffling.
    - Interaction cues: interruptions/ducking, reactive phrasing/laughter to loud events, startled responses to impulses.
    - Sound-type hints (no class names in the question): "steady bed," "rhythmic clatter," "brief chime," "low thump," "crowd
swell."
9. **Hide raw spatial cues**. Instead, imply them through conversational timing, overlap, call-and-response, or shared
ambience.
10. Accept any sentient or contextually plausible non-speech target (e.g., "the guitar riff", "the woman's laughter"). Avoid
inert noises unless the caption clearly supports that interaction.
11. The *reasoning* should spell out (for graders) the spatial/temporal evidence—but phrase it abstractly ("they share the
same spot in the room and overlap in time") rather than "both come from the left".
12. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical
instruction.
13. **Keep class labels out of the question.Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
14. Generate open-ended answers
15. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique
question_id for all the questions of the same audio to the final json.
16. **Use the transcript to inform your question when understanding the speech content is needed to infer the sound source it
responds to.**
---

Caption:
{caption}

Metadata Rows:
{metadata_rows}

Transcript:
{transcript}

---
```

Figure 11: Prompt used for generating Environmental Awareness and Navigation QAs.

```
You are a spatial-audio intelligence agent.
Your job is to craft a *challenging* listening-comprehension QA pair that hinges on subtle spatial reasoning.

Goal
- Ask about which source is physically closest (or closer) to another source, or who is most affected / likely to interact
because of proximity/near-field influence — without revealing explicit spatial positions, distances, or giving away the whole
scene in the question.

You will receive:
• **caption** - a prose summary of the whole scene (primary reference)
• **metadata_rows** - frame-level records, each with
  - class_label (e.g. "Music", "Male speech")
  - coarse_direction (e.g. "right and slightly below")
  - source_type (e.g. "Actor 1")  ← **NEVER** quote or allude to this
  - distance_description ('196 cm (1.96 m)') - refer to the text description for distance in the captions
  - **transcript** - raw speech transcription for the audio, if any

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Use *only* the class labels ("male speech", "laughter", "music", …) in the answer but never directly name them in the
question.
3. Do not narrate the entire audio scene in the question.
    - Avoid sentences like "After the telephone rings from the left, the man speaks…"
    - Instead, use subtle references like "After a continuous tone begins…" or "Following the sharp burst…"
4. Questions should be answerable only by listening — reading the question without the audio should be insufficient.
5. **Do not include explicit or relative spatial attributes in the question** (no "left/right/behind/near/far"). Proximity
must be inferred from acoustic cues.6. **Avoid explaining the temporal order of the audio in the question**.
6. Avoid explaining exact temporal order in the question.
7. Start the question like "Based on the audio,", "In the given audio," , etc.
8. Embed indirect cues that force multi-step reasoning, such as:
     - Proximity cues: level dominance / masking, proximity (on-mic) effect, breathiness/intonation clarity, direct-to-reverb
ratio, occlusion/muffling, spectral brightness/air, intensity of cross-talk/bleed, FOA energy concentration vs. diffusion.
      - Interaction cues: overlap strength, interruption/ducking, reactive phrasing in transcript, laughter triggered by a
nearby event, startling to a loud impulse.
    - Sound type hints: "steady layer," "rhythmic clatter," "brief chime," "low thump," etc. (no class names in the question).
9. **Hide raw spatial cues**. Instead, imply them through conversational timing, overlap, call-and-response, or shared
ambience.
10. Accept any sentient or contextually plausible non-speech target (e.g., "the guitar riff", "the woman's laughter"). Avoid
inert noises unless the caption clearly supports that interaction.
11. The *reasoning* should spell out (for graders) the spatial/temporal evidence—but phrase it abstractly ("they share the
same spot in the room and overlap in time") rather than "both come from the left".
12. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical
instruction.
13. **Keep class labels out of the question.Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
14. Generate open-ended answers
15. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique
question_id for all the questions of the same audio to the final json.
16. **Use the transcript to inform your question when understanding the speech content is needed to infer the sound source it
responds to.**
---

Caption:
{caption}

Metadata Rows:
{metadata_rows}

Transcript:
{transcript}

---
```

Figure 12: Prompt used for generating Proximity-Based Decision Making QAs.

```
You are a spatial-audio intelligence agent.
Your job is to craft a *challenging* listening-comprehension QA pair that hinges on subtle spatial reasoning.

Goal
- Ask about which ongoing sound source a given speech source is responding to, **without** revealing explicit spatial
positions, distances, or giving away the whole scene in the question.

You will receive:
• **caption** - a prose summary of the whole scene (primary reference)
• **metadata_rows** - frame-level records, each with
  - class_label (e.g. "Music", "Male speech")
  - coarse_direction (e.g. "right and slightly below")
  - source_type (e.g. "Actor 1")  ← **NEVER** quote or allude to this
  - distance_description ('196 cm (1.96 m)') - refer to the text description for distance in the captions
  - **transcript** - raw speech transcription for the audio, if any

Strict guidelines:
1. **No abstract IDs** - never write "Actor 1", "speaker 2", "non-human", etc.
2. Use *only* the class labels ("male speech", "laughter", "music", …) in the answer but never directly name them in the
question.
3. Do not narrate the entire audio scene in the question.
    - Avoid sentences like "After the telephone rings from the left, the man speaks…"
    - Instead, use subtle references like "After a continuous tone begins…" or "Following the sharp burst…"
4. Questions should be answerable only by listening — reading the question without the audio should be insufficient.
5. **Do not include explicit or relative spatial attributes in the question** for identifying the answer. Spatial inference
must come entirely from listening to the audio.
6. **Avoid explaining the temporal order of the audio in the question**.
7. Start the question like "Based on the audio,", "In the given audio," , etc.
8. Embed just enough indirect cues in the question to force multi-step reasoning:
    - Temporal sequence ("after…", "while…", "once…", etc.)
    - Sound type hints ("steady background", "brief signal", "rhythmic pattern", etc.)
    - Interaction style ("responds to", "speaks over", "matches the rhythm of", etc.)
9. **Hide raw spatial cues**. Instead, imply them through conversational timing, overlap, call-and-response, or shared
ambience.
10. Accept any sentient or contextually plausible non-speech target (e.g., "the guitar riff", "the woman's laughter"). Avoid
inert noises unless the caption clearly supports that interaction.
11. The *reasoning* should spell out (for graders) the spatial/temporal evidence—but phrase it abstractly ("they share the
same spot in the room and overlap in time") rather than "both come from the left".
12. **Natural tone** - the question should feel like something a careful listener would ask another person, not a mechanical
instruction.
13. **Keep class labels out of the question.Inside the question refer to non-speech sounds indirectly (e.g. "that steady
background", "the rhythmic clatter", "the brief chime") so the listener must infer which source you mean.Inside the answer /
reasoning / source_alignment you must use the canonical class_label strings exactly as they appear in the metadata (e.g.
"Music", "Domestic sounds").
14. Generate open-ended answers
15. Generate 2-3 diverse questions for the same audio path targeting different spatial cues if possible, append a unique
question_id for all the questions of the same audio to the final json.
16. **Use the transcript to inform your question when understanding the speech content is needed to infer the sound source it
responds to.**
---

Caption:
{caption}

Metadata Rows:
{metadata_rows}

Transcript:
{transcript}

---
```

Figure 13: Prompt used for generating Spatial Priority & Source Intent Inference QAs.