

Spatio-Temporal Audio Language Modeling for Dynamic Sound Sources

Oh Hyun-Bin^{1*} Kazuki Shimada² Yuhta Takida² Kim Sung-Bin¹ Toshimitsu Uesaka²
Takashi Shibuya² Kyeongyoon Lee⁴ Tae-Hyun Oh^{5†} Yuki Mitsufuji^{2,3†}

¹POSTECH ²Sony AI ³Sony Group Corporation ⁴Sungkyunkwan University ⁵KAIST

Abstract

Sound events have semantic identities, locations, and trajectories, yet existing audio-language models typically encode clips as global event content. Conversely, sound event localization models track source directions over time but offer limited semantic coverage for language reasoning. To bridge this gap, we introduce ST-AudioQA, a spatio-temporal audio QA dataset and benchmark built from first-order ambisonic (FOA) renderings of static and moving sound sources. Each scene provides source identity, activity, direction, distance, and motion metadata, enabling dense trajectory supervision and QA about what is sounding, where it is, how it moves, and how sources relate. We further introduce ST-Audio Encoder, which jointly learns event semantics and time-resolved source trajectories, and ST-AudioLM, which exposes these audio tokens to an LLM for spatio-temporal audio QA. Experiments show that this representation improves the balance between semantic preservation and localization and yields stronger reasoning than static spatial and localization-oriented baselines. Beyond the controlled benchmark, ST-AudioLM shows zero-shot transfer to real recordings, with further gains after light adaptation. Project page: <https://st-audiolm.github.io/>.

1 Introduction

Audio-language models have made rapid progress in recognizing, describing, and reasoning about sound events (Deshmukh et al., 2023; Gong et al., 2024; Tang et al., 2024a; Chu et al., 2023; Kong et al., 2024). Yet most systems still treat an audio clip primarily as event labels or a caption to be generated. This abstraction misses a basic property of everyday hearing: sound events unfold in space and time. Humans use auditory location and motion to navigate, monitor danger, and decide how to act;

a looming vehicle, receding footsteps, or speech fixed in place while another source passes by can change the interpretation of the same acoustic scene (Carlini et al., 2024; Neuhoff, 2001; Seifritz et al., 2002). In such scenes, a sound is not merely a semantic category, such as *siren* or *speech*, but also an entity with a location, distance, and trajectory. An audio-language model should therefore answer not only what is present, but also where each source is, whether it moves, how its distance changes, and how its motion relates to other sources.

Recent spatial audio-language work shows that large language models (LLMs) can use multi-channel cues to reason about rendered sound scenes (Zheng et al., 2024; Tang et al., 2024b; Wilkinghoff and Tan, 2026; Jiang et al., 2026; Biswas et al., 2026; Sakshi et al., 2025; Zhu et al., 2026), while emerging benchmarks begin to consider dynamic audio (Liu et al., 2026; Sridhar et al., 2026). However, most models still focus on static or clip-level spatial attributes without explicitly exposing time-resolved source trajectories. Meanwhile, sound event localization and detection (SELD) models estimate time-varying directions and activities (Adavanne et al., 2018; Shimada et al., 2022; Hu et al., 2025). Yet their constrained sound event taxonomies limit their use as broad-vocabulary semantic encoders for open-ended audio-language reasoning over AudioSet-scale event vocabularies (Gemmeke et al., 2017). Together, these limitations reveal a gap between two lines of work: audio-language models with broad semantics but weak temporal localization, and localization models with strong spatial inductive bias but limited language-level semantic coverage.

To bridge this gap, we contribute a dataset, an encoder, and an audio-language model. First, we construct ST-AudioQA, a controlled benchmark of static and moving sound sources rendered in first-order ambisonics (FOA). We use FOA for its listener-centered full-sphere spatial coverage (Ger-

*Work done during an internship at Sony AI.

†Co-corresponding authors.

zon, 1973; Daniel, 2003) and alignment with established SELD direction-estimation practice (Adavanne et al., 2018), complementing prior binaural spatial audio-language work (Zheng et al., 2024). Unlike event-only audio QA or static spatial QA, ST-AudioQA is designed to evaluate binding between sound identity, time, and spatial state: it asks not only *what* and *where*, but also *where at a given time*, *how did it move*, and *which source changed*. The rendered spatio-temporal audio provides both dense trajectory supervision and structured source-level metadata for generating controlled QA, enabling training of both the spatio-temporal encoder and audio LLM. Second, we introduce ST-Audio Encoder, a time-resolved spatio-temporal FOA audio encoder built on audio spectrogram transformers (Gong et al., 2021) that predicts AudioSet semantics together with activity, direction, and distance trajectories. Third, we connect ST-Audio Encoder to a LLM (Walsh et al., 2025) via audio tokens, yielding ST-AudioLM, and train it with a curriculum organized as A) single-source perception, B) two-source perception and grounding, and C) compositional source-time-space reasoning.

Experiments show that spatio-temporal audio QA requires more than recognizing events or localizing sources in isolation. Encoders adapted from static spatial audio models (Zheng et al., 2024) struggle to track motion, while localization-oriented SELD models (Hu et al., 2025) lose event information needed for language questions. ST-Audio Encoder addresses this tradeoff by learning representations that retain sound-event identity while tracking source direction and distance over time. When used as audio tokens for an LLM, these representations outperform baselines based on temporalized static-encoder features or localization-oriented dynamic features across single-source perception, two-source grounding, and compositional motion questions. Evaluations on real recordings demonstrate zero-shot transfer and gains from light adaptation (Shimada et al., 2023); ST-AudioLM also leads the zero-shot comparison on a real-recording spatial-reasoning benchmark (Liu et al., 2026). Together, these results suggest that spatio-temporal audio-language reasoning benefits from binding *what is sounding* to *where it is over time*. Our contributions are summarized as follows:

- We propose ST-AudioQA, a controlled spatio-temporal audio QA benchmark for static and moving sources, covering event recognition,

time-anchored localization, motion, source grounding, and compositional reasoning.

- We introduce ST-Audio Encoder, a time-resolved FOA spatial audio encoder that preserves event semantics while predicting source direction and distance trajectories.
- We build ST-AudioLM, which equips an LLM with semantic and trajectory-aware FOA audio tokens, improving spatio-temporal perception, grounding, and compositional reasoning.

2 Related Work

Sound source localization models. Sound source localization and SELD studies provide a complementary line of work focused on estimating source activity and direction over time (Grumiaux et al., 2022; Adavanne et al., 2018; Cao et al., 2021). Formulations such as Multi-ACCD OA, which represents active sound events as Cartesian direction vectors, have been effective for localizing overlapping sources (Shimada et al., 2022). Recent extensions further incorporate source distance estimation (Krause et al., 2024; Diaz-Guerra et al., 2024) and exploit pretrained models, such as PSELDNets pretrained on large synthetic SELD data (Hu et al., 2025). These models explicitly address the spatial and temporal localization problem that audio-language models often overlook. However, their objectives are typically framed around localization and constrained sound event taxonomies. This makes them effective for estimating source trajectories, but less suitable as general audio interfaces for LLMs, where broad semantic understanding and language-facing representations are essential.

Audio-language models for spatial reasoning. Audio-language models connect pretrained audio or speech encoders (Gong et al., 2021; Elizalde et al., 2023; Chen et al., 2023; Radford et al., 2023) to LLMs for captioning, question answering, and instruction following (Deshmukh et al., 2023; Gong et al., 2024; Tang et al., 2024a; Chu et al., 2023; Kong et al., 2024). Recent systems further broaden general audio understanding through scaled open models, time-aware representations, and diffusion language backbones (Chu et al., 2024; Ghosh et al., 2025; Yang et al., 2026; Zhou et al., 2026). Although these models improve language-level reasoning over everyday sounds using large-scale audio and audio-text supervision (Gemmeke et al., 2017; Kim et al., 2019; Drossos et al., 2020; Wu

et al., 2023), they typically encode audio as global semantic content, leaving spatio-temporal source state implicit. Recent spatial audio-language models begin to address this gap. BAT introduces a spatial audio encoder for binaural audio and enables LLMs to answer controlled questions about sound source semantics and locations (Zheng et al., 2024). Subsequent work extends auditory LLMs to multichannel spatial tasks such as far-field speech recognition (Tang et al., 2024b), improves spatial audio representations (Wilkinghoff and Tan, 2026), estimates source and room descriptors (Jiang et al., 2026), or adds geometry-aware and plug-in spatial modules for spatial audio-language reasoning (Biswas et al., 2026; Sakshi et al., 2025). Spatial-Omni (Zhu et al., 2026) integrates FOA encoding into a LLM for spatial audio understanding. However, most spatial audio-language models still focus on static or clip-level spatial reasoning without explicit time-resolved source trajectories.

Recent work has begun to address dynamic audio reasoning: STAR-Bench includes a broad real-recording evaluation (Liu et al., 2026), while Sridhar et al. present dynamic spatial audio QA with rendered binaural scenes and BAT-style encoder adaptation (Sridhar et al., 2026). In contrast, ST-AudioQA provides controlled trajectory supervision and a hierarchy of source-state perception, grounding, and reasoning tasks, while ST-Audio Encoder produces dedicated semantic and trajectory tokens directly from FOA. Together with ST-AudioLM, this enables audio-language reasoning grounded in time-varying source states.

3 ST-AudioQA: Spatio-Temporal Audio QA Benchmark

Existing spatial audio QA benchmarks (Zheng et al., 2024; Biswas et al., 2026) reduce spatial reasoning to static source locations, whereas real acoustic scenes require tracking sources as they move. We therefore formulate spatial audio QA as a spatio-temporal grounding problem: each event should be tied to where it is over time, whether it moves, and how its direction and distance change. ST-AudioQA adopts the source trajectory, rather than a clip-level position, as the unit of supervision.

To build this benchmark, we render FOA audio and derive controlled short-answer QA pairs from structured source metadata. For each sound source indexed by k , we define its state at time frame t as

$$s_k(t) = \{a_k(t), \mathbf{d}_k(t), r_k(t), y_k\}, \quad (1)$$

where $a_k(t)$ denotes binary source activity (active vs. inactive), $\mathbf{d}_k(t)$ is the listener-centered direction, $r_k(t)$ is metric distance, and y_k is the audio label set associated with the clip. Static sources are represented as zero-motion trajectories, giving a unified metadata schema for static and moving sources. Sec. 3.1 describes how we generate spatio-temporal FOA audio, and Sec. 3.2 describes how these source states are converted into QA pairs. Additional details on audio rendering and QA generation are provided in Appendix E.

3.1 Spatio-Temporal FOA Audio Generation

Spatial audio simulation. We use SoundSpaces 2.0 (Chen et al., 2022) with Matterport3D (Chang et al., 2017) scene meshes to simulate room acoustics, following prior spatial audio-language work (Zheng et al., 2024). For each room, we sample listener and source anchors from navigable same-room candidates, then query SoundSpaces for FOA room impulse responses (RIRs) that encode source-to-listener direction, distance, and reflections. This produces 10-second, 32 kHz FOA clips with controllable source trajectories and room-scale reverberation.

Static and dynamic sources. We represent both static and moving sources with a common 40-bin temporal schema. Each 10-second clip is divided into 40 time bins, corresponding to one spatial state every 0.25 seconds. For a moving source, the listener is fixed and the source follows a same-room trajectory between valid source anchors. For a static source, the trajectory is constant, *i.e.*, the same source position is used for all time bins. To render motion, we query a per-bin FOA RIR and crossfade adjacent rendered segments, obtaining time-varying renderings of a sound source. We retain only trajectories that pass audibility and RIR-stability checks; additional filtering and rendering details are given in Appendix E.1.

Sound events and mixtures. We use AudioSet clips (Gemmeke et al., 2017) as monaural event waveforms and spatialize each waveform by rendering it through the SoundSpaces FOA RIRs. AudioSet provides the event content and label set, while SoundSpaces provides the spatial propagation and trajectory annotations. A rendered scene contains either one source ($K = 1$) or two sources ($K = 2$). Each source is rendered independently into FOA audio, and two-source scenes are mixed by summing the rendered FOA

(a) Perception and grounding with primitive QA families		
Primitive QA family	A: Single-source perception	B: Two-source perception and grounding
Event semantics	<i>Detect or verify the event content of one source</i> Cases: S, D Q: What sound events are present? A: dog bark	<i>Detect or verify the event content of two sources</i> Cases: SS, SD, DD Q: What sound events are present? A: dog bark; clapping
Movement status	<i>Decide whether one source moves</i> Cases: S, D Q: Is the sound moving or stationary? A: moving	<i>Decide whether a named source moves</i> Cases: SS, SD, DD Q: Is the clapping moving or stationary? A: moving
Spatial location	<i>Localize one source over time</i> Cases: S, D Q: What is the azimuth of the sound at the end? A: back-right	<i>Localize a named source over time</i> Cases: SS, SD, DD Q: What is the distance of the clapping at the end? A: 2.0--2.5 m
Spatial change	<i>Estimate beginning-to-end change</i> Cases: D Q: Does the sound get closer or farther overall? A: closer	<i>Estimate change for a named source</i> Cases: SD, DD Q: Does the clapping move clockwise or counterclockwise? A: clockwise
Source grounding	<i>Not instantiated for one-source clips</i> Cases: – Q: Not asked A: Not applicable	<i>Select the event matching a spatial or motion attribute</i> Cases: SS, SD, DD Q: Which sound moves clockwise, dog bark or clapping? A: clapping
(b) Compositional source-time-space reasoning		
Compositional QA family	Composed primitive families	C: Compositional example
Temporal relation change	{Event semantics, Spatial location, Spatial change}	<i>Ask whether a relation changes from beginning to end</i> Cases: SD, DD Q: Does dog bark start closer than clapping and end farther away? A: no
Movement-conditioned spatial relation	{Movement status, Spatial location}	<i>Compare the moving and stationary sounds at a time anchor</i> Cases: SD Q: At the middle, is the moving sound higher than the stationary sound? A: no
Cross-source trajectory relation	{Event semantics, Spatial change}	<i>Compare how the two source trajectories change</i> Cases: SD, DD Q: Does dog bark change distance more than clapping? A: yes

Table 1: **ST-AudioQA hierarchy.** (a) Primitive QA families instantiated for single-source perception and two-source perception/grounding. (b) Compositional reasoning integrates these primitives into source-time-space relation questions over two trajectories. S/D denote static/dynamic one-source clips; SS/SD/DD denote the three two-source motion cases. Controlled answers include event labels, yes/no, spatial bins, and movement labels.

source signals channel-wise. Single-source scenes contain either one static or one moving source; two-source scenes cover static–static (SS), static–dynamic (SD), and dynamic–dynamic (DD) motion configurations. The resulting metadata provides time-varying source activity, direction, distance, and event labels. These structured source states are used both for encoder supervision and for generating the QA pairs described next.

3.2 Spatio-Temporal QA Generation

Using the structured source state $s_k(t)$, we instantiate controlled short-answer QA pairs from the rendered scene metadata. The questions are designed to probe event identity, time-localized spatial state, movement status, spatial change, and relations between source trajectories. Table 1 summarizes the hierarchy of question families and types.

Primitive QA families. We define five primitive QA families over the source state $s_k(t)$, corresponding to Table 1(a):

- *Event semantics* covers event detection and verification, asking what sound events are present or whether a queried event is audible.
- *Movement status* decides whether a source is moving or stationary.
- *Spatial location* localizes a source at a beginning, middle, or end time anchor, using listener-centered azimuth, elevation, and distance bins.
- *Spatial change* estimates how a source changes from the beginning to the end of the clip, such as clockwise versus counterclockwise, upward versus downward, or closer versus farther.
- *Source grounding* selects the sound event that matches a given spatial or motion attribute in two-source scenes.

QA types. The QA types form a progression from perceiving a single source, to grounding sources in mixtures, to reasoning over relations between trajectories. Type A, *single-source perception*, asks whether the model can recover the event identity, location, movement status, and spatial change of one source trajectory, testing basic source-state understanding. Type B, *two-source perception and grounding*, places the same source-level cues in a mixture: the model must answer about a named event or select the event matching a spatial or motion attribute, thereby binding each event identity to the correct trajectory. Type C, *compositional source-time-space reasoning*, composes the primitive families into two-source relation questions, requiring the model to compare grounded source states across event identity, time, space, and motion (see Table 1(b)).

4 Method

Figure 1 illustrates ST-AudioLM, which connects a time-resolved FOA audio encoder to an LLM for spatio-temporal audio QA. We first describe ST-Audio Encoder, which learns event semantics together with source activity, direction, and distance trajectories (Sec. 4.1). We then describe the audio-token interface and LLM training procedure used to turn the encoder representations into language-model inputs for perception, grounding, and compositional reasoning (Sec. 4.2).

4.1 ST-Audio Encoder: Spatio-Temporal FOA Audio Encoder

FOA frontend and static heads. ST-Audio Encoder uses an AST-style Transformer (Gong et al., 2021) backbone over FOA spatial audio features. Given a 10-second FOA waveform, we compute log-mel features from the four ambisonic channels (W, X, Y, Z) and concatenate three acoustic intensity-vector features (I_x, I_y, I_z), yielding a seven-channel time-frequency input, as commonly used for FOA SELD features (Politis et al., 2020). A 3×3 convolutional frontend fuses these channels before patch embedding, and the resulting patch sequence is processed by the Transformer. To learn clip-level semantic and static spatial summaries, we prepend three learned global tokens: a semantic token, a direction-of-arrival (DoA) token, and a distance token. The semantic token supports multi-label AudioSet prediction, while the DoA and distance tokens supervise static localization heads for

clip-level DoA and source-listener distance. This gives us a static FOA perception encoder, which we next turn into a time-resolved trajectory encoder.

Dynamic trajectory heads. The global tokens summarize clip-level semantics and static spatial state, but moving-source localization requires time-resolved evidence. We therefore convert the final Transformer patch sequence into temporal trajectory representations. Specifically, we reshape the final patch sequence into a grid $\mathbf{Z} \in \mathbb{R}^{T' \times F' \times C}$, where T' and F' are the temporal and frequency dimensions after patch embedding, and C is the hidden dimension. We then average over frequency to obtain a temporal representation:

$$\mathbf{e}_t = \frac{1}{F'} \sum_{f=1}^{F'} \mathbf{z}_{t,f}. \quad (2)$$

The sequence $\{\mathbf{e}_t\}$ is linearly interpolated to 40 bins for each 10-second clip, corresponding to one representation every 250 ms. Lightweight temporal self-attention layers contextualize these bin-level representations before prediction. Per-bin dynamic heads then predict source activity, a 3D direction vector, and source-listener log-distance. Static-source examples are represented as constant trajectories over time, allowing the same heads to train on both stationary and moving sources.

Encoder training. We train ST-Audio Encoder in three stages (see Fig. 1(a)). Stage 1 trains the semantic token for multi-label AudioSet prediction on static FOA renderings. Stage 2 adds clip-level static localization by supervising the DoA and distance tokens, yielding a static FOA encoder with both event and spatial objectives. Stage 3 converts this encoder into a time-resolved trajectory encoder: the dynamic heads are trained on moving-source trajectories, while static examples are replayed as constant trajectories over time. To retain event recognition during this dynamic stage, we keep the semantic head active and distill semantic logits from a frozen Stage-1 semantic teacher.

4.2 ST-AudioLM: Spatio-Temporal Audio Reasoning with LLM

Spatio-temporal audio token interface. After training ST-Audio Encoder, we remove all prediction heads and freeze the encoder. For each audio clip, we use one semantic token and 40 temporal trajectory representations from the encoder as audio tokens, forming a 41-token audio sequence.

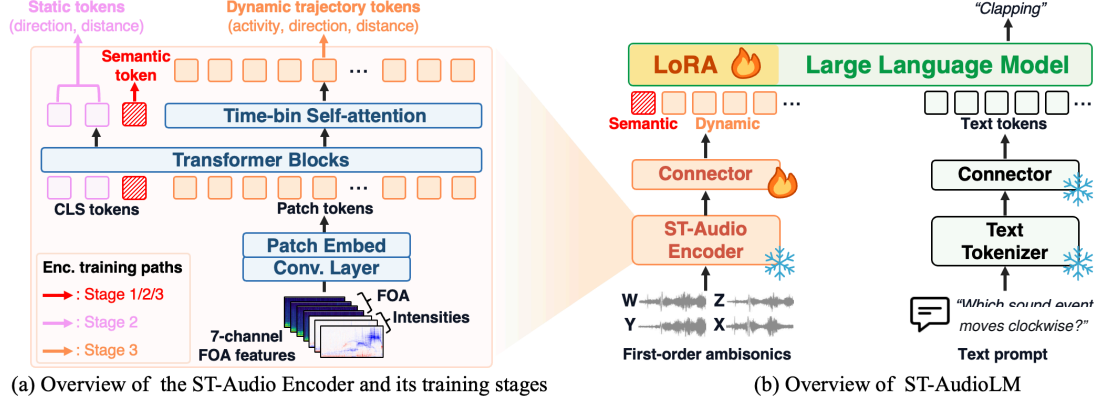


Figure 1: **Overview of ST-Audio Encoder and ST-AudioLM.** (a) ST-Audio Encoder converts FOA waveforms into seven-channel FOA features and uses an AST-style Transformer (Gong et al., 2021) to produce semantic, static localization, and dynamic trajectory tokens, supervised by the corresponding objectives in three-stage encoder training. (b) For QA finetuning, the frozen ST-Audio Encoder provides one semantic token and dynamic trajectory tokens as audio tokens. A trainable two-layer MLP connector maps them into the LLM embedding space, and only the connector and LoRA adapters are trained on the staged spatio-temporal QA curriculum.

The semantic token provides a clip-level event representation, while the temporal trajectory representations preserve the temporal order of spatial information. The global DoA and distance tokens are used only for encoder supervision and are not included in the LLM input. A trainable two-layer MLP connector projects each audio token into the LLM embedding space. Thus, the LLM receives both a semantic representation and a temporally ordered sequence of trajectory representations, rather than a single pooled audio vector (see Fig. 1(b)).

LLM finetuning. During QA training, we freeze both ST-Audio Encoder and the base LLM weights, and update only the two-layer MLP connector and LoRA adapters. We train these modules with a staged spatio-temporal QA curriculum, with training-pool sizes summarized in Table 2. Type A trains single-source perception, requiring the model to recognize event, location, and motion from the 41 audio tokens. Type B adds two-source perception and grounding, where questions refer to named events or ask the model to select the source matching a spatial or motion attribute. Type C adds compositional relation questions over grounded sources, combining event identity, time, location, motion, and source-to-source comparisons.

5 Experiments

We evaluate our approach by tracing the path from spatio-temporal audio representation to QA. Encoder-level experiments examine whether ST-Audio Encoder captures both event identity and time-varying source state. Audio-language experi-

Stage	Question Type	# Tr. Samples	Percentages
I	A	216K	29.5%
II	A, B	516K	70.3%
III	A, B, C	734K	100%

Table 2: **Staged QA curriculum for LLM finetuning.** Training samples are cumulative over the listed question types: A is single-source perception, B adds two-source perception and grounding, and C adds compositional reasoning. Percentages are relative to the Stage-III pool.

ments then assess ST-AudioLM on ST-AudioQA, measuring single-source perception, two-source grounding, and compositional source-time-space reasoning. Sec. 5.1 describes the setup, metrics, and baselines; Sec. 5.2 presents the main analyses.

5.1 Experimental Setup

Implementation details. Our encoder adapts the Spatial-AST backbone used in BAT (Zheng et al., 2024) to the FOA frontend in Sec. 4.1. Stages 1–2 train this static FOA encoder on room-rendered AudioSet clips for multi-label event recognition and clip-level DoA/distance localization. Stage 3 initializes ST-Audio Encoder from the Stage-2 checkpoint and learns time-resolved source trajectories through two training phases on single-source moving clips; Appendix A.4 provides stage-wise training details. For audio-language training, we use OLMo2-7B-Instruct (Walsh et al., 2025), freeze the audio encoder and base LLM, and train only the MLP connector and LoRA adapters. We use LoRA rank 16 and $\alpha = 32$; Type A uses connector and LoRA learning rates of 2×10^{-4} and 10^{-4} , while Types B and C use 10^{-4} and 5×10^{-5} , respectively.

Encoder baselines. Table 3 compares static and dynamic encoder baselines. For static evaluation, *Spatial-AST* is the binaural BAT encoder (Zheng et al., 2024), and *Spatial-AST-FOA* is our Stage-2 static FOA encoder before trajectory training. For dynamic evaluation, *Intensity-based DOA estimator* converts the FOA intensity vector into time-bin directions without learning. *Spatial-AST-FOA (temporal crop)* applies the trained static FOA encoder to short waveform crops centered at each trajectory time bin, predicting bin-level DoA and distance independently. *PSELDNets-mACCD OA + semantic/dist. heads* extends a dynamic PSELDNets architecture (Hu et al., 2025) to 355 AudioSet labels and source-listener distance prediction.

Audio-language baselines. Tables 4 and 5 compare ST-AudioLM with QA-trained baselines, and zero-shot evaluations of general audio-language models (Chu et al., 2024; Ghosh et al., 2025; Yang et al., 2026; Zhou et al., 2026) and the FOA-based Spatial-Omni (Zhu et al., 2026). Using each model’s native or closest supported input (separate binaural channels for Qwen2-Audio, mono for the other general models, and FOA for Spatial-Omni), these comparisons test whether general audio-language or spatial-audio modeling suffices without trajectory-aware training. *BAT* uses the same QA pairs with corresponding binaural renderings, serving as a static binaural baseline. For controlled FOA comparisons, *PSELDNets-mACCD OA + OLMo2* connects the dynamic PSELDNets-based encoder from Table 3 to OLMo2, testing localization-oriented dynamic features for QA. *Spatial-AST-FOA + OLMo2* uses the frozen Stage-2 static FOA encoder, reshaping its patch tokens into 40 temporal tokens and appending the semantic token. This matches ST-AudioLM’s 41-token interface but lacks ST-Audio Encoder’s trajectory-specific temporal self-attention, dynamic heads, and trajectory supervision.

Metrics. Encoder evaluation reports AudioSet mAP, DoA MAE as 3D angular error in degrees, Dist. Acc over 0.5 m static distance bins, and Dist. MAE in meters for dynamic clips. Traj. Acc@20° counts a clip as correct only when all active time bins are localized within 20°, measuring full-path rather than average localization. For QA, event detection uses mAP, event verification uses yes/no accuracy, and the other families use answer accuracy; Appendix B details aggregation and parsing.

(a) Static source evaluation

Model	Input	Semantic	Spatial loc.		
		mAP ↑	DoA MAE ↓	Dist. MAE ↓	Acc ↑
Spatial-AST (Zheng et al., 2024)	Binaural	50.1	18.0°		67.2
Spatial-AST-FOA	FOA	47.7	19.2°		74.1

(b) Dynamic source evaluation

Model	Input	Semantic	Spatial loc.			Tracking
		mAP ↑	DoA MAE ↓	Dist. MAE ↓	Traj. Acc@20° ↑	
Intensity-based DOA estimator	FOA	–	23.9°	–		33.0
Spatial-AST-FOA (temporal crop)	FOA	45.2	41.4°	0.64 m		–
PSELDNets-mACCD OA + sem./dist. heads	FOA	29.7	13.7°	0.38 m		61.4
ST-Audio Encoder (ours)	FOA	62.8	13.8°	0.32 m		62.3

Table 3: **Encoder evaluation for static perception and dynamic tracking.** (a) reports static event recognition, DoA estimation, and distance-bin prediction; the binaural Spatial-AST row is only a reference because FOA uses different renderings and input features. (b) evaluates semantic prediction, DoA/distance errors, and whole-trajectory tracking on single-source moving clips.

5.2 Experimental Results and Analyses

We analyze the results from encoder representation to audio-language reasoning, examining (i) semantic-preserving trajectory modeling, (ii) single- and two-source perception and grounding, (iii) compositional source-time-space reasoning, (iv) transfer to real recordings, and (v) key representation and training choices. Appendix D.1 provides detailed STARSS23 transfer results.

Can an audio encoder model source trajectories while preserving event semantics? Table 3 evaluates static spatial-semantic perception and dynamic moving-source tracking. Table 3(a) includes binaural Spatial-AST only as a reference point because its input format and renderings differ from *Spatial-AST-FOA*, while *Spatial-AST-FOA* shows that our Stage-2 FOA encoder provides a strong static spatial-semantic representation before trajectory training. In dynamic scenes, *Spatial-AST-FOA (temporal crop)* shows that applying this static encoder over temporal windows is insufficient for trajectory tracking. *PSELDNets-mACCD OA + semantic/dist. heads* is a strong localization-oriented baseline, obtaining the best DoA error and competitive trajectory accuracy but substantially lower AudioSet mAP. In contrast, ST-Audio Encoder achieves the best semantic mAP and distance error while remaining competitive with PSELDNets on dynamic localization. This suggests that its main benefit is a joint representation that preserves broad event semantics while exposing time-resolved spatial cues for downstream language modeling.

Type	Model	A: Single-source perception							B: Two-source perception and grounding										
		Semantic		Spatial loc.		Spatial chg.		Move	Semantic		Spatial loc.		Spatial chg.		Move	Source grounding			
		mAP	Y/N	DoA	Dist.	Δ DoA	Δ Dist.	Acc.	mAP	Y/N	Src. DoA	Src. Dist.	Src. Δ DoA	Src. Δ Dist.	Src. Move	Loc.	Chg.	Move	
–	Random chance	0.8	50.0	22.9	10.0	50.0	50.0	50.0	1.5	50.0	22.9	10.0	50.0	50.0	50.0	25.0	25.0	25.0	
Zero-shot	Qwen2-Audio (Chu et al., 2024)	8.8	76.3	15.9	0.0	51.2	34.3	37.2	4.7	69.1	20.5	0.0	48.8	36.3	36.9	3.6	19.4	4.7	
	Audio Flamingo 3 (Ghosh et al., 2025)	17.8	82.3	19.3	8.0	35.2	39.1	43.1	7.8	67.6	18.4	3.4	37.2	34.9	40.8	0.1	0.4	0.1	
	MOSS-Audio-8B (Yang et al., 2026)	6.5	79.0	19.2	21.0	17.7	48.6	51.2	3.7	62.1	19.2	19.5	22.0	49.2	50.6	13.4	5.8	1.2	
	DIFFA-2 (Zhou et al., 2026)	3.8	75.2	17.8	17.1	32.1	45.3	51.1	2.8	60.6	18.2	16.7	32.6	48.6	49.1	0.2	0.1	0.2	
	Spatial-Omni (Zhu et al., 2026)	3.6	68.6	19.3	8.6	18.2	34.0	9.1	3.2	44.4	24.3	8.1	18.5	29.9	4.2	8.1	6.4	8.6	
QA-trained	BAT (Zheng et al., 2024)	1.0	64.6	48.5	29.8	49.7	52.8	99.4	1.8	60.5	35.6	21.0	49.8	50.4	64.7	39.4	37.4	57.0	
	PSELDNets-mACCDQA + OLMo2	9.1	89.9	72.0	39.7	51.3	62.6	73.4	5.4	76.1	51.2	28.2	50.5	58.3	56.1	46.1	32.8	30.1	
	Spatial-AST-FOA + OLMo2	25.3	93.3	71.4	44.9	78.4	82.4	99.6	12.6	83.2	51.0	30.9	65.2	68.8	73.9	48.8	40.2	70.0	
	ST-AudioLM (ours)	27.6	93.7	81.6	51.5	91.1	87.8	99.8	14.3	83.8	54.3	32.8	71.6	70.0	66.2	49.7	47.4	59.3	

Table 4: **Audio-language QA results on ST-AudioQA.** All values are on a 0–100 scale; higher is better. Type A evaluates single-source perception; Type B evaluates two-source perception and grounding. Semantic columns report event-detection mAP and event-verification accuracy; Spatial loc. reports DoA/distance accuracy, Spatial chg. reports beginning-to-end change accuracy, and Move reports moving/stationary accuracy. For Type B, “Src.” denotes questions about the named sound event, and source grounding asks the model to select the event matching a spatial or motion attribute. Zero-shot models use their native or closest supported input format. Qwen2-Audio and BAT use binaural audio rendered by SoundSpaces from the same scene and source metadata.

Type	Model	C: Compositional source-time-space reasoning			
		Temp. rel.	Move-spat.	Traj. rel.	Avg.
–	Random chance	50.0	50.0	50.0	50.0
Zero-shot	Qwen2-Audio (Chu et al., 2024)	51.6	35.9	45.6	44.4
	Audio Flamingo 3 (Ghosh et al., 2025)	49.8	33.1	49.8	44.2
	MOSS-Audio-8B (Yang et al., 2026)	50.4	33.1	45.6	43.0
	DIFFA-2 (Zhou et al., 2026)	49.7	33.4	45.1	42.7
	Spatial-Omni (Zhu et al., 2026)	49.9	28.2	46.8	41.6
QA-trained	BAT (Zheng et al., 2024)	76.3	51.6	55.3	61.1
	PSELDNets-mACCDQA + OLMo2	86.3	50.3	51.4	62.7
	Spatial-AST-FOA + OLMo2	80.4	55.2	54.3	63.3
	ST-AudioLM (ours)	86.0	55.8	60.6	67.5

Table 5: **Compositional QA results.** All values are controlled-answer accuracy on a 0–100 scale; higher is better. Temp. rel., Move-spat., and Traj. rel. denote temporal relation change, movement-conditioned spatial relation, and cross-source trajectory relation. Avg. is computed over the reported Type-C families.

Do trajectory-aware audio tokens improve single-source perception? Table 4 evaluates Type-A single-source perception, where the model must recognize the event and recover the source’s spatial state over time. Under zero-shot evaluation, general audio-language models retain some event-semantic ability but remain weak on localization, spatial change, and motion. Among QA-trained baselines, *BAT* remains weak in event detection and spatial-change reasoning, while *PSELDNets-mACCDQA + OLMo2* is limited by weak event semantics. The strongest overall baseline, *Spatial-AST-FOA + OLMo2*, uses the same 41-token interface as ST-AudioLM but derives its tokens from a static FOA encoder. Compared with this, ST-AudioLM improves semantic recognition and spatial perception, with the clearest gains on trajectory-sensitive questions such as Δ DoA. This suggests that trajectory-supervised encoder representations make the audio tokens more informative for perceiving time-varying source states.

Do trajectory-aware tokens remain useful for grounding in mixtures? Type B in Table 4 moves from single-source perception to two-source mixtures, where the model must answer about a named event or select the event matching a spatial or motion attribute. The static FOA baseline already performs strongly, showing that the 41-token interface provides a useful basis for mixture QA. ST-AudioLM further improves semantic recognition, source-conditioned localization, and spatial-change grounding. This suggests that trajectory-aware tokens remain useful beyond single-source perception, especially when spatial questions must be grounded to a queried sound event in a mixture.

How well do trajectory-aware tokens support compositional reasoning? Table 5 evaluates Type-C questions that combine event semantics, motion, spatial change, and cross-source comparison. These zero-shot baselines remain below chance on average, showing that general audio-language modeling and FOA spatial modeling alone are insufficient for Type-C reasoning. ST-AudioLM achieves the best average accuracy, with the clearest gain on cross-source trajectory relations and performance comparable to the strongest localization-oriented baseline on temporal relation change. However, movement-conditioned spatial relations remain only modestly above chance for all models, indicating that jointly resolving motion and spatial relations is still challenging. Overall, these results suggest that trajectory-aware tokens help compositional source-time-space reasoning, while relation-level reasoning over grounded sources remains open.

(a) STARSS23 transfer and adaptation						(b) STAR-Bench zero-shot evaluation					
Setting	Type A	Type B	Type C			Model	Input	Static loc.	Relation	Trajectory	Overall
	Overall	Overall	Move-spat.	Traj. rel.	Overall	Random chance	–	33.3	33.3	33.3	33.3
Zero-shot	39.9	18.3	54.0	49.8	51.4	Qwen2-Audio (Chu et al., 2024)	2ch	42.2	22.2	13.3	37.9
Light adapt.	59.7	59.4	56.4	67.8	63.5	Audio Flamingo 3 (Ghosh et al., 2025)	Mono	36.6	52.8	26.7	37.4
Real mixture	67.1	59.1	44.2	62.6	55.7	Spatial-Omni (Zhu et al., 2026)	FOA	46.2	41.7	10.0	42.8
						ST-AudioLM (ours)	FOA	47.9	63.9	60.0	50.4

Table 6: **Real-recording evaluation.** (a) STARSS23 transfer (Shimada et al., 2023). Overall for each QA type is accuracy over all questions in its evaluation pool after answer normalization (chance for Type C: 50.0). *Light adapt.* updates only the connector/LoRA after synthetic training; *Real mixture* adds family-stratified STARSS23 QA amounting to approximately 5% of the synthetic QA count at each curriculum stage (Appendix D.2). (b) Zero-shot STAR-Bench evaluation (Liu et al., 2026) using each model’s native or closest compatible input; none is trained on STAR-Bench. Overall reports micro-averaged accuracy. Values are percentages; higher is better.

Setting	Semantic	Spatial loc.		Tracking
	mAP ↑	DoA MAE ↓	Dist. MAE ↓	Traj. Acc@20° ↑
(A) Semantic Distillation				
Without	51.0	12.2	0.30	73.0
With (ours)	62.8	13.8	0.32	62.3
(B) Temporal Resolution				
10 bins	58.5	14.1	0.31	65.4
20 bins	58.3	14.0	0.31	64.4
40 bins (ours)	62.8	13.8	0.32	62.3
(C) FOA Input Features				
FOA log-mel only	56.7	71.5	0.48	1.6
+ IPD	56.9	20.1	0.36	51.9
+ Intensity vector (ours)	62.8	13.8	0.32	62.3

Table 7: **Encoder design ablations.** (A) semantic distillation, (B) temporal resolution, and (C) FOA input features. IPD denotes inter-channel phase difference. Variants share the Stage-3 training setup within each block. DoA/Dist. MAE are in degrees/meters; other metrics are percentages.

Does source-time-space modeling transfer to real recordings? Table 6(a) evaluates transfer to STARSS23 recordings (Shimada et al., 2023). The zero-shot model retains semantic and some single-source perception abilities, while lightweight connector/LoRA adaptation substantially improves multi-source grounding and Type-C reasoning. Mixing a small real-QA set into the synthetic curriculum yields the strongest Type-A transfer and improves Type-B grounding over zero-shot, although explicit adaptation remains strongest on Type-C. Table 6(b) further evaluates all models zero-shot on the real-recording subset of STAR-Bench (Liu et al., 2026). ST-AudioLM achieves the best result in every category and overall, with its clearest advantage on trajectory questions, providing evidence that its trajectory-aware representation transfers beyond synthetic rendering.

What enables semantic-preserving spatio-temporal audio-language modeling? Tables 3, 4, and 5 compare static and trajectory-trained representations at the encoder level and under the same LLM protocol, while Table 7 isolates

semantic distillation, temporal resolution, and FOA frontend features. Distillation improves semantic preservation despite a localization trade-off, and the distilled encoder still exceeds PSELDNets in trajectory accuracy. Among the tested resolutions, 40 bins give the best semantic mAP and DoA error while keeping distance and trajectory tracking competitive. Magnitude-only FOA severely degrades dynamic localization; IPD partially recovers it, but intensity-vector features perform best overall. Appendix C shows that the A→B→C curriculum progressively develops the three task types and better preserves Type-A/B than joint training under the same epoch budget.

6 Conclusion

We present spatio-temporal audio-language modeling for dynamic sound sources. To this end, we introduce ST-AudioQA, a controlled, hierarchical source-time-space QA benchmark; ST-Audio Encoder, a time-resolved FOA encoder that preserves broad event semantics while modeling source trajectories; and ST-AudioLM, which exposes semantic and trajectory-aware audio tokens to an LLM. Experiments show that these representations support dynamic localization and improve single-source perception, multi-source grounding, and compositional reasoning over static spatial and localization-oriented baselines. ST-AudioLM also shows zero-shot transfer to real recordings, with further benefits from target-domain supervision. Taken together, our results show that reasoning over moving sound sources requires more than static spatial understanding or localization alone; it calls for representations that bind broad event semantics to time-resolved source trajectories.

7 Limitations

This work uses controlled synthetic rendering to obtain precise trajectory labels and systematic QA, but it cannot cover the full range of real acoustic conditions, microphone responses, source behaviors, noise, or outdoor environments. We therefore assess transfer through zero-shot inference, light adaptation, and mixed real/synthetic QA training on STARSS23 (Shimada et al., 2023), together with zero-shot evaluation on STAR-Bench (Liu et al., 2026). Although these results demonstrate transfer beyond synthetic rendering, broader real-world evaluation remains necessary.

The current benchmark also simplifies dynamic scenes. Our dynamic rendering uses per-bin stationary FOA RIRs with crossfades. In our motion regime, the omitted Doppler shift is negligible; Appendix E.1 quantifies this approximation, while high-speed pass-by events and continuously changing scattering paths remain outside our scope. Sources follow controlled 10-second trajectories, and many questions focus on beginning-to-end spatial change rather than arbitrary non-monotonic motion. Movement-conditioned spatial-relation reasoning also remains an open challenge because it requires joint motion identification, source grounding, localization, and relation comparison. The main setting is limited to one- and two-source scenes, so denser mixtures, same-class source ambiguity, source entrances and exits, pauses, acceleration, and curved paths remain future directions.

8 Ethical Considerations

The system may inherit biases from AudioSet labels and the simulated scene distribution. We do not annotate or infer speaker identity or other personal attributes, but speech-related and other audio labels may encode demographic or contextual bias. To support transparency and reproducibility, we will release the generation scripts and metadata filters. We avoid treating synthetic benchmark performance as a substitute for real-world robustness.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea government (MSIT) in 2026 (Grant No. RS-2024-00451947, Project Title: Interaction Framework for Connecting Users to Generative Experiences via Time-Space Expanded Multimedia: Centering on

Social Communication Disorders (50%); No. RS-2024-00358135, Corner Vision: Learning to Look Around the Corner through Multi-modal Signals (50%)), the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01), and the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

References

- Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. 2018. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*.
- Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2026. OWL: Geometry-aware spatial reasoning for audio large language models. In *ICLR*.
- Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D. Plumbley. 2021. An improved event-independent network for polyphonic sound event localization and detection. In *ICASSP*.
- Alessandro Carlini, Camille Bordeau, and Maxime Ambard. 2024. Auditory localization: A comprehensive practical review. *Frontiers in Psychology*, 15:1408073.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*.
- Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. 2022. SoundSpaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS Datasets and Benchmarks*.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xi-angzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*.
- 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.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. *Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models*. Preprint, arXiv:2311.07919.

- J r me Daniel. 2003. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *AES*.
- Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *NeurIPS*.
- 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 DCASE 2024 challenge. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.
- Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP*.
- Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. CLAP: Learning audio concepts from natural language supervision. In *ICASSP*.
- Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*.
- Michael A. Gerzon. 1973. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*.
- Sreyan Ghosh, Arushi Goel, 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. In *NeurIPS*.
- Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio spectrogram transformer. In *Inter-speech*.
- Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. In *ICLR*.
- Pierre-Amaury Grumiaux, Srđan Kitic, Laurent Girin, and Alexandre Gu rin. 2022. A survey of sound source localization with deep learning methods. *Journal of the Acoustical Society of America*.
- Jinbo Hu, Yin Cao, Ming Wu, Fang Kang, Feiran Yang, Wenwu Wang, Mark D. Plumbley, and Jun Yang. 2025. PSELDNets: Pre-trained neural networks on a large-scale synthetic dataset for sound event localization and detection. *IEEE Transactions on Audio, Speech and Language Processing*.
- Xilin Jiang, Sebastian Braun, and Hannes Gamper. 2026. Sci-Phi: A large language model spatial audio descriptor. *IEEE Open Journal of Signal Processing*.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *NAACL-HLT*.
- 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. In *ICML*.
- Daniel Aleksander Krause, Archontis Politis, and Anamaria Mesaros. 2024. Sound event detection and localization with distance estimation. In *EUSIPCO*.
- Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, Leilei Sun, Dahua Lin, and Jiaqi Wang. 2026. STAR-Bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. In *ICLR*.
- John G. Neuhoff. 2001. An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13(2):87–110.
- Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. 2020. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. In *Workshop on Detection and Classification of Acoustic Scenes and Events*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.
- S. Sakshi, Vaibhavi Lokegaonkar, Neil Zhang, Ramani Duraiswami, Sreyan Ghosh, Dinesh Manocha, and Lie Lu. 2025. [SPUR: A plug-and-play framework for integrating spatial audio understanding and reasoning into large audio-language models](#). *Preprint*, arXiv:2511.06606.
- Erich Seifritz, John G. Neuhoff, Deniz Bilecen, Klaus Scheffler, Henrietta Mustovic, Hartmut Schachinger, Raffaele Elefante, and Francesco Di Salle. 2002. Neural processing of auditory looming in the human brain. *Current Biology*, 12(24):2147–2151.
- Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji. 2022. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *ICASSP*.
- 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. In *NeurIPS Datasets and Benchmarks*.

- Arvind Krishna Sridhar, Yinyi Guo, and Erik Visser. 2026. [Spatial audio question answering and reasoning on dynamic source movements](#). *Preprint*, arXiv:2602.16334.
- 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 *ICLR*.
- 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? In *Interspeech*.
- Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafford, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, and 23 others. 2025. 2 OLMo 2 Furious (COLM’s Version). In *COLM*.
- Kevin Wilkinghoff and Zheng-Hua Tan. 2026. DSpAST: Disentangled representations for spatial audio reasoning with large language models. In *ICASSP*.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*.
- Chen Yang, Chufan Yu, Hanfu Chen, Jie Zhu, Jingqi Chen, Ke Chen, Wenxuan Wang, Yang Wang, Yaozhou Jiang, Yi Jiang, Zhengyuan Lin, Ziqi Chen, Zhaoye Fei, Chenghao Liu, Donghua Yu, Jun Zhan, Kang Yu, Kexin Huang, Liwei Fan, and 11 others. 2026. [MOSS-Audio technical report](#). *Preprint*, arXiv:2606.01802.
- 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. In *ICML*.
- Jiaming Zhou, Xuxin Cheng, Shiwan Zhao, Yuhang Jia, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. 2026. DIFFA-2: A practical diffusion large language model for general audio understanding. In *Findings of ACL*.
- Zhiyuan Zhu, Yixuan Chen, Yiwen Shao, Wenxiang Guo, Changhao Pan, Yu Zhang, Yuxiang Wang, Wei Liu, Houhua Zhang, Chengkuan Zeng, Wenbo Cheng, Yunxi Liu, Rui Yang, Steve Yves, Liefeng Bo, and Zhou Zhao. 2026. [Spatial-Omni: Spatial audio understanding integration in multimodal LLMs via FOA encoding](#). *Preprint*, arXiv:2606.10738.

Contents

A Model and Training Details

- A.1 ST-Audio Encoder Architecture
- A.2 Audio Tokens for the LLM
- A.3 Baseline Implementations
- A.4 Training Configuration

B Evaluation Details

C Curriculum versus Joint Training

D Real-Data Transfer

- D.1 STARSS23 Transfer QA Construction
- D.2 STARSS23 Transfer and Adaptation
- D.3 STAR-Bench Zero-Shot Evaluation

E Data Generation Details

- E.1 FOA Scene and Dynamic Audio Rendering
- E.2 QA Generation Rules and Templates
- E.3 ST-AudioQA Statistics

A Model and Training Details

We provide additional implementation details for ST-Audio Encoder, its interface with the language model, the baseline systems, and the training configurations used in our experiments.

A.1 ST-Audio Encoder Architecture

Input features. The encoder operates on a seven-channel time-frequency input. The first four channels are log-mel features computed from the first-order ambisonic signal. The remaining three channels are acoustic intensity-vector features derived from the same FOA signal. We stack the mel and intensity features along the channel dimension and pass them through a 3×3 convolutional frontend before AST-style patch embedding. This frontend fuses spectral, ambisonic, and intensity cues before transformer processing. We sample the waveforms at 32 kHz and compute a 1,024-point STFT using a 1,024-sample Hann window and a 320-sample hop. The resulting power spectra are projected onto 128 mel bands spanning 50 Hz to 14 kHz and compressed to decibels with an 80 dB range. For $q \in \{X, Y, Z\}$, the intensity feature is the mel projection of $\text{Re}(W^*q)/(|W|^2 + (|X|^2 + |Y|^2 + |Z|^2)/3 + \epsilon)$.

Backbone configuration. The AST backbone uses 768-dimensional embeddings, 12 Transformer blocks, 12 attention heads, an MLP ratio of 4, and 16×16 patches with stride 16. The frontend feature map is resized to 1,024 time frames before patch embedding.

Static spatial pretraining. We first train the FOA encoder with a Spatial-AST-style objective (Zheng et al., 2024). This stage uses global representations to predict AudioSet event semantics, and static direction and distance. The semantic representation is trained with AudioSet multi-label supervision, while the direction and distance representations are trained with static localization losses. This provides a spatial-semantic initialization before introducing moving-source trajectory supervision.

Dynamic trajectory modeling. For dynamic trajectory modeling, we reuse the transformer patch sequence from the pretrained encoder instead of adding a separate audio frontend. The post-transformer patch sequence is reshaped into a time-frequency grid, pooled over frequency, and represented as 40 temporally ordered states, matching the 40 spatial states used in the renderer. We apply a self-attention temporal modeling block over these states. The dynamic heads predict source activity, 3D direction, and log-distance for each time bin. During dynamic training, we retain the global semantic representation and use semantic supervision and distillation to reduce degradation in event recognition. The temporal module consists of two self-attention blocks with eight heads and an MLP ratio of 4. Its prediction heads output one activity logit, a normalized 3D direction vector, and scalar log-distance per time bin.

Static examples as constant trajectories. Static clips are represented as constant trajectories. The same direction and distance target is repeated across all 40 time bins, so a stationary source becomes a zero-motion special case of the dynamic formulation. This allows the temporal heads to learn moving and non-moving sources under one representation.

Training objectives. For the final trajectory-training phase, the per-example dynamic loss is

$$\begin{aligned} \mathcal{L}_{\text{dyn}} = & \mathcal{L}_{\text{sem}} + \mathcal{L}_{\text{dir}} + 0.3\mathcal{L}_{\text{dist}} \\ & + 0.05\mathcal{L}_{\Delta\text{dist}} + 0.05\mathcal{L}_{\Delta\text{dir}} + 0.1\mathcal{L}_{\text{act}}. \end{aligned} \quad (3)$$

Here, $\mathcal{L}_{\text{sem}}$ and $\mathcal{L}_{\text{act}}$ are binary cross-entropy losses for clip-level AudioSet labels and per-bin source activity, respectively. $\mathcal{L}_{\text{dir}}$ is the cosine loss between predicted and target 3D direction vectors, and $\mathcal{L}_{\text{dist}}$ is the Smooth L1 loss on log-distance. $\mathcal{L}_{\Delta\text{dir}}$ and $\mathcal{L}_{\Delta\text{dist}}$ apply Smooth L1 loss between predicted and target first-order differences of direc-

tion and log-distance across adjacent time bins, respectively. The full objective additionally includes $20\mathcal{L}_{KD}$, which distills semantic logits from the frozen Stage-1 teacher with temperature 1, and $0.05\mathcal{L}_{replay}$, which applies the trajectory losses to replayed static clips represented as constant trajectories.

A.2 Spatio-Temporal Audio Tokens for the LLM

During QA finetuning, the encoder prediction heads are removed and the frozen encoder is used as a representation extractor. The language model receives the semantic global representation and the 40 temporally ordered trajectory representations. The static direction and distance global representations are used for encoder training, but are not passed to the LLM.

The text prompt contains a single audio placeholder. At the embedding level, this placeholder is replaced by the 41-representation audio block. A trainable connector maps each audio representation to the LLM hidden size, and the projected block is inserted into the prompt embedding sequence. The two-layer connector maps each 768-dimensional encoder representation to 4,096 dimensions and then applies a second 4,096-dimensional projection, with GELU activation and no dropout. The connector is trained jointly with LoRA adapters on the QA curriculum, while the audio encoder and the base LLM remain frozen.

A.3 Baseline Implementations

Encoder baselines. We compare ST-Audio Encoder against three classes of encoder baselines: a non-learned localization method, a static spatial-audio encoder, and a localization-focused dynamic encoder.

Intensity-based DOA estimator. This baseline is a non-learned method with no trainable parameters. It estimates source direction from FOA active-intensity cues computed from the rendered FOA waveform. Because it does not contain an event classifier, distance head, or language-facing representation, it is used only as a physics-based localization reference.

Spatial-AST-FOA temporal crop. We build a Spatial-AST encoder for static FOA input. The architecture follows (Zheng et al., 2024), with the frontend modified to accept FOA features. For dynamic evaluation, the moving-source clip is evaluated through temporal crops, and the static local-

ization heads are applied to each crop. This gives the baseline local temporal evidence while keeping the model architecture static. It does not include the 40-bin temporal modeling block, dynamic trajectory heads, or moving-source trajectory supervision used by ST-Audio Encoder.

PSELDNets-mACCDOA with semantic and distance heads. We adapt the PSELDNets mACCDOA architecture as a localization-focused dynamic baseline. The original model is designed around frame-wise sound event localization with a smaller event taxonomy, so we extend it with a 355-class AudioSet semantic head and an additional distance head. For the encoder evaluation, we use ground-truth-class track selection: localization is evaluated on the output associated with the ground-truth event class. This focuses the comparison on localization and representation quality rather than on an additional class-track selection problem.

Audio-language baselines. We compare ST-AudioLM with several audio-language baselines that differ in the audio representation supplied to the language model. The QA-trained baselines use ST-AudioQA supervision, and all models are evaluated with the same answer parser.

BAT. The BAT baseline (Zheng et al., 2024) is a representative prior spatial-audio LLM designed for static spatial sources. We evaluate it under our QA protocol using the same QA pairs and scene/source metadata as ST-AudioLM, but render the audio in the binaural format expected by BAT. We train it with the same QA supervision to compare against a prior static spatial-audio language modeling pipeline.

PSELDNets-mACCDOA + OLMo2. This baseline connects the adapted PSELDNets-mACCDOA (Hu et al., 2025) encoder to the same OLMo2 language model (Walsh et al., 2025). We extract the shared final HTSAT feature map rather than final SELD logits, average it over frequency and linearly resample it to 40 temporal tokens, and append its global average as one clip token. We then train the same connector and LoRA adapters used by ST-AudioLM. This tests whether localization-focused dynamic features are sufficient for source-time-space QA when attached to the same LLM training pipeline.

Spatial-AST-FOA + OLMo2. This baseline connects the frozen static FOA encoder to OLMo2. To match the audio-token count of ST-AudioLM,

we reshape the post-transformer patch sequence into a time-frequency grid, average over frequency, resample the resulting temporal sequence to 40 positions, and concatenate the semantic global representation. Thus, the LLM receives 41 audio representations, as in ST-AudioLM. The key difference is that these temporal representations come from a static encoder and are not trained with dynamic trajectory heads or moving-source supervision.

Zero-shot audio-language models. We evaluate Qwen2-Audio (Chu et al., 2024), Audio Flamingo 3 (Ghosh et al., 2025), MOSS-Audio-8B (Yang et al., 2026), and DIFFA-2 (Zhou et al., 2026) from their released instruction checkpoints without task-specific finetuning. Qwen2-Audio receives SoundSpaces-rendered left and right binaural channels as separate audio inputs; the other three receive the omnidirectional FOA W channel as 16-kHz mono. All use the same QA questions, valid-answer format, and answer normalization. The three additional models generate at most 24 tokens; Audio Flamingo 3 and MOSS-Audio-8B use greedy decoding, while DIFFA-2 uses 16 diffusion steps.

Spatial-Omni. We evaluate the released SO-7B and SO-Encoder checkpoints (Zhu et al., 2026) without ST-AudioQA adaptation. Its native preprocessing receives 10-second, 16-kHz four-channel FOA audio, and generation uses the same exact-answer prompts and scoring protocol.

A.4 Training Configuration

The trajectory-learning stage of ST-Audio Encoder (Stage 3) proceeds in two phases. We first train the encoder on 20K single-source moving clips for 50 epochs with a learning rate of 1×10^{-5} . We then continue training on 100K reflected moving-source clips for 20 epochs with a learning rate of 3×10^{-5} . In this final phase, we apply semantic supervision and distillation from the frozen Stage-1 semantic teacher, while replaying static examples as constant trajectories. Both phases use AdamW with weight decay 5×10^{-4} , and all dynamic examples use 40 time bins per 10-second clip.

For audio-language training, we use OLMo2-7B-Instruct as the base language model. The audio encoder and the base LLM weights are frozen. We train a two-layer MLP connector together with LoRA adapters. The Type-A, Type-B, and Type-C curriculum stages are each trained for two epochs. The LoRA adapters use rank 16, $\alpha = 32$, and dropout 0.05. Type A uses connector and LoRA

learning rates of 2×10^{-4} and 10^{-4} , while Types B and C use 10^{-4} and 5×10^{-5} , respectively. The OLMo2-based QA-trained baselines use the same stage-wise connector and LoRA schedule.

B Evaluation Details

All models are evaluated with the same answer parser and the same family-level aggregation.

Family-level aggregation. Detailed scores are computed separately for each question family. Overall is accuracy over all questions in the corresponding evaluation pool after answer normalization. For Type B, source-to-attribute questions contribute to the corresponding attribute column, such as spatial location, spatial change, or movement status. In contrast, attribute-to-source questions are evaluated under source grounding, since the model must select which source satisfies a queried attribute. For Type C, Avg. is the unweighted average over the reported compositional reasoning families.

Event metrics. Event-detection questions are evaluated as multi-label mAP. For ST-AudioQA, the positive labels are the accepted AudioSet labels attached to the queried source or source mixture; for two-source clips, the positive set is the union of both sources’ retained labels. We lowercase each generated answer, remove wrapper text and surrounding punctuation, and retain its first output line. We then represent the answer and every canonical event label by TF-IDF-weighted token counts and use their cosine similarity as the score for that label. Because this scoring is token based rather than exact-set matching, multi-label answers may use commas, semicolons, slashes, or conjunctions without requiring a fixed separator. Comma-separated AudioSet display-name synonyms are treated as aliases when mapping target metadata to canonical class indices; generated variants are scored by their token similarity to the canonical label rather than by a separate exact-match synonym table. We compute AP separately for every class with at least one positive example and report the macro-average over these classes. Event-verification questions ask whether a specific event is present and are scored separately as binary yes/no accuracy. For STARSS23, positives are the annotated STARSS23 event classes, so STARSS23 mAP is computed within its 13-class taxonomy and is not directly comparable to the 355-class AudioSet mAP.

Stage	Training data	Type-A eval	Type-B eval	Type-C eval
I	A	76.5	32.7	13.7
II	A+B	76.8	53.0	10.2
III	A+B+C	76.3	52.6	67.5

Table 8: **Stage-wise QA curriculum evaluation.** The model after each stage is evaluated on the same held-out pools. Type-A/B values report accuracy over all questions in their respective pools after answer normalization, while Type C is the unweighted average over its three compositional families.

Direction aggregation. In the QA tables, DoA is the arithmetic mean of the azimuth- and elevation-question accuracies, while Δ DoA averages the corresponding beginning-to-end azimuth- and elevation-change accuracies.

Closed-answer parsing and random baselines.

All non-event-detection questions use finite answer spaces. We lowercase model outputs, strip surrounding punctuation, and map the result to the answer vocabulary of the corresponding family; unmapped outputs are counted as incorrect. Random baselines use the same family aggregation. For closed-answer questions, the random predictor samples uniformly from the valid answer space of each question. For event mAP, random chance is the macro-average class prevalence over classes with positives in the corresponding evaluation split, corresponding to chance-level AP under random scoring.

C Curriculum versus Joint Training

Table 8 shows the progression across the QA curriculum: Stage I establishes single-source perception, Stage II substantially improves two-source grounding, and Stage III enables compositional reasoning while largely preserving Type-A/B performance.

Table 9 compares the staged QA curriculum with a joint A+B+C training run under the same total number of QA-training epochs. The joint run trains on the combined QA pool from the beginning, whereas the curriculum introduces single-source perception, two-source grounding, and compositional reasoning sequentially. This ablation tests whether staged training better preserves the perception and grounding skills learned in Types A and B while adding Type-C compositional reasoning.

Joint training reaches comparable Type-C accuracy, and is slightly higher on the Type-C average.

However, under the same epoch budget, the staged curriculum consistently retains stronger Type-A and Type-B semantic, spatial-location, and spatial-change performance. This indicates that staged training better preserves earlier source-state perception and grounding while adding compositional reasoning. We therefore use the staged curriculum for the main model and report joint training as an ablation.

D Real-Data Transfer

We evaluate real-world transfer in two complementary settings: transfer and lightweight adaptation on STARSS23-derived QA (Shimada et al., 2023), and zero-shot evaluation on STAR-Bench (Liu et al., 2026).

D.1 STARSS23 Transfer QA Construction

We construct a transfer benchmark from the original STARSS23 (Shimada et al., 2023) frame-level annotations. Each 10-second clip is paired with its FOA waveform and annotations for event class, source track, azimuth, elevation, and distance. We define a physical source as a (class, track) pair within a clip. For source-grounded questions, a class name is used as a referent only when that class appears in a single track; repeated class instances are treated as ambiguous for source selection but remain valid for event detection and verification.

As summarized in Table 10, the STARSS23 transfer set contains Type A, Type B, and Type C questions satisfying the stated validity criteria. Type A is generated from clips with exactly one annotated source and evaluates event semantics, event verification, movement status, spatial location, and spatial change. Type B is generated from clips with two or more annotated sources and evaluates the same perceptual attributes together with source grounding. For Type C, we retain movement-conditioned spatial relation and cross-source trajectory relation, the two compositional families supported by the real annotations. We omit temporal relation change because the same validity criteria yield insufficient balanced instances.

Temporal anchors follow the same linguistic convention as ST-AudioQA: questions use *beginning*, *middle*, and *end* without exposing numeric timestamps. For label generation, we use the nearest active annotation around the queried anchor and instantiate a question only when the relevant source is active. Movement labels are computed from

Training strategy	Epoch	A: Single-source perception								B: Two-source perception and grounding								C: Compositional reasoning								
		Semantic		Spatial loc.		Spatial chg.		Move		Semantic		Spatial loc.		Spatial chg.		Move		Source grounding		Temp. rel.	Move-spat.	Traj. rel.	Avg.			
		mAP	Y/N	DoA	Dist.	Δ DoA	Δ Dist.	Acc.	mAP	Y/N	Src.	DoA	Src.	Dist.	Src.	Δ DoA	Src.	Δ Dist.	Src.	Move	Loc.	Chg.	Move	Acc.	Acc.	Acc.
Curriculum A+B+C	6 (2+2+2)	27.6	93.7	81.6	51.5	91.1	87.8	99.8	14.3	83.8	54.3	32.8	71.6	70.0	66.2	49.7	47.4	59.3	86.0	55.8	60.6	67.5				
Joint A+B+C	6	25.0	93.5	79.8	48.7	87.7	85.4	99.7	12.9	82.5	52.0	30.6	69.5	66.8	67.7	48.2	47.4	59.0	85.0	57.6	61.0	67.9				

Table 9: **Curriculum versus joint QA training.** All values are on a 0–100 scale. The curriculum checkpoint follows the staged schedule used in the main experiments and reports the same Type-C checkpoint as Table 5. The joint checkpoint is trained from scratch on the combined A+B+C pool. Type-C Avg. is computed over the three reported Type-C families.

QA type	Train QA	Test QA	Question families
Type A	4,003	1,775	event, movement, location, change
Type B	49,163	12,328	event, movement, location, change, grounding
Type C	1,075	436	movement–spatial and trajectory relations
Total	54,241	14,539	joint Type A/B/C

Table 10: **STARSS23-derived transfer QA split.** Type C includes the two compositional families that remain reliable after filtering the real annotation distribution.

the first and last active annotations of a source. A source is considered moving if its azimuth change is at least 10 degrees, elevation change is at least 10 degrees, distance change is at least 0.5 m, or relative distance change is at least 20%.

D.2 STARSS23 Transfer and Adaptation

We aim to measure whether the synthetic source-time-space representation transfers to real recordings and how different uses of real QA affect transfer. *Zero-shot* directly evaluates the ST-AudioLM checkpoint trained on ST-AudioQA without STARSS23 supervision. *Adapted* fine-tunes only the connector and LoRA adapters for two epochs on the corresponding STARSS23-derived QA training pools after the synthetic curriculum, keeping ST-Audio Encoder frozen. *Mixture* instead adds STARSS23-derived QA during each curriculum stage, with the number of real questions set to approximately 5% of that stage’s synthetic QA count. This gives 10,822, 25,822, and 41,422 real questions in Stages 1, 2, and 3, respectively. Using seed 0, we sample by question family: Stages 1–2 use replacement when a family does not meet its quota, while Stage 3 samples without replacement and redistributes unused family quotas across the remaining pool.

Table 11 provides a fine-grained Type-A/B breakdown complementing the overall and Type-C

results in Table 6(a). The zero-shot model transfers event recognition, but source-time-space grounding remains difficult without real-data supervision. Light adaptation substantially improves Type-B localization, movement, and source grounding. The real-data mixture yields the strongest Type-A overall result and matches light adaptation on Type-B overall (59.1 vs. 59.4), with complementary strengths across individual metrics. As shown in Table 6(a), light adaptation remains stronger on Type-C compositional reasoning.

D.3 STAR-Bench Zero-Shot Evaluation

We use the 123 STARSS23-derived items of STAR-Bench (Liu et al., 2026) whose released clips are at most 10 seconds: 101 single-source static-localization, 12 multi-source spatial-relation, and 10 dynamic trajectory-tracking items. We align each released clip to its original STARSS23 recording by short-time envelope correlation and zero-pad the corresponding FOA segment to 10 seconds. Each three-option question is evaluated under all three cyclic option orders, yielding 369 trials and reducing answer-position bias. Models answer with the exact option text and use their native or closest compatible interface: FOA for FOA-capable models, separate binaural channels for Qwen2-Audio, and mono for Audio Flamingo 3. All models are evaluated zero-shot, and no STAR-Bench data is used to train ST-AudioLM.

As reported in Table 6(b), ST-AudioLM achieves the best static-localization, relation, trajectory, and overall results. Together with STARSS23 transfer, this zero-shot result provides complementary evidence that the trajectory-aware representation generalizes beyond the synthetic rendering pipeline.

Setting	Model	A: Single-source perception							B: Multi-source perception and grounding								
		Semantic		Spatial loc.		Spatial chg.		Move	Semantic		Spatial loc.		Spatial chg.		Move	Source grd.	
		mAP	Y/N	DoA	Dist.	Δ DoA	Δ Dist.	Acc.	mAP	Y/N	DoA	Dist.	Δ DoA	Δ Dist.	Acc.	DoA	Dist.
–	Random chance	12.50	50.00	22.92	12.50	50.00	50.00	50.00	23.22	50.00	22.92	12.50	50.00	50.00	18.25	17.74	17.68
Zero-shot	ST-AudioLM	49.47	83.67	26.51	25.46	40.91	40.00	88.67	42.96	72.24	19.74	11.79	51.19	54.29	1.49	12.07	8.18
Light adapt.	ST-AudioLM + STARSS23	73.52	83.67	59.97	32.55	50.00	30.00	92.67	59.81	87.47	53.46	33.93	51.19	55.71	69.78	60.77	67.49
Real mixture	ST-AudioLM + 5% STARSS23	71.81	94.00	68.24	37.53	63.64	30.00	93.33	67.48	88.67	56.88	32.91	48.81	52.86	73.51	57.89	64.72

Table 11: **Detailed STARSS23 Type-A/B results.** All values are on a 0–100 scale and higher is better. Semantic mAP is computed over the 13 STARSS23 event classes. DoA averages azimuth and elevation accuracy. Source grd. evaluates Type-B attribute-conditioned source selection. *Real mixture* adds a family-stratified STARSS23 QA set amounting to approximately 5% of the synthetic QA count at each curriculum stage; exact counts and sampling are described above.

E Data Generation Details

E.1 FOA Scene and Dynamic Audio Rendering

Listener and source placement. We sample listener and source anchors from precomputed navigable locations in each Matterport3D scene (Chang et al., 2017). Before RIR rendering, we enforce a same-room constraint on each listener–source pair and discard pairs that cross room boundaries, preventing them from entering either encoder training or QA generation. For a moving source, the listener remains fixed and the source trajectory is constructed from valid source anchors in the same room. For a static source, we use a constant trajectory by repeating the same source position across all time bins.

Time-varying rendering. SoundSpaces (Chen et al., 2022) provides FOA RIRs for fixed source–listener configurations. We therefore approximate a moving source as a sequence of short stationary renderings. Each 10-second clip is divided into $N = 40$ time bins, with $\Delta = 10/N = 0.25$ seconds. Let $\tau_j = (j + \frac{1}{2})\Delta$ denote the center of bin j , and let $\mathbf{p}_k(\tau_j)$ be the 3D position of source k at that time. For each bin, SoundSpaces is queried for the FOA RIR $h_{k,m,j}$ from source k to ambisonic channel m at position $\mathbf{p}_k(\tau_j)$.

Given a monaural event waveform x_k , we render the m -th FOA channel of source k by convolving x_k with the per-bin RIRs and crossfading the rendered outputs:

$$\tilde{x}_{k,m}(t) = \sum_{j=0}^{N-1} w_j(t) (h_{k,m,j} * x_k)(t), \quad (4)$$

where $w_j(t)$ are 50 ms crossfade weights satisfying $\sum_j w_j(t) = 1$ for all t . This crossfaded-RIR strategy follows prior moving-source SELD

rendering (Politis et al., 2020). The maximum radial source velocity in our trajectory design is $v_r = 0.28$ m/s, yielding a relative Doppler shift of $|v_r|/c \approx 8.2 \times 10^{-4}$. Thus, the omitted frequency shift is below 0.1% in our motion regime and is unlikely to dominate direction/distance trajectory QA; high-speed pass-by events remain outside our scope. The rendered source signal is $\tilde{\mathbf{x}}_k(t) = \{\tilde{x}_{k,m}(t)\}_m$ over all FOA channels. For two-source scenes, source stems are rendered independently and summed channel-wise:

$$\tilde{\mathbf{x}}^{\text{scene}}(t) = \sum_{k=1}^K \tilde{\mathbf{x}}_k(t). \quad (5)$$

Trajectory filtering. We retain render-valid trajectories with finite, nonsilent RIRs. A source is dynamic when its active beginning and end anchors differ by at least 10° in azimuth or elevation, or 0.5 m in distance; a static source is formed by freezing a randomly selected beginning, middle, or end state. Two-source scenes use distinct source-audio, trajectory, and canonical event identifiers under the same scene, room, and listener, with mean and minimum inter-source angular separations of at least 35° and 10° , respectively.

Activity-aware QA labeling. We generate QA labels only when the queried source is audible at the requested time anchor. To determine this, we compute a source-level activity mask before mixing sources. This is important for two-source clips: at the same time anchor, one source may be active and valid for a question, while the other may be inactive and therefore skipped. This prevents generating questions whose answers would depend on inaudible or near-silent source segments.

Metadata for QA generation. For each rendered clip, the manifest stores all information needed

Azimuth label	Center angle	Angular sector
Front (F)	0°	$[-22.5^\circ, 22.5^\circ)$
Front-right (FR)	45°	$[22.5^\circ, 67.5^\circ)$
Right (R)	90°	$[67.5^\circ, 112.5^\circ)$
Back-right (BR)	135°	$[112.5^\circ, 157.5^\circ)$
Back (B)	180°	$[157.5^\circ, 202.5^\circ)$
Back-left (BL)	225°	$[202.5^\circ, 247.5^\circ)$
Left (L)	270°	$[247.5^\circ, 292.5^\circ)$
Front-left (FL)	315°	$[292.5^\circ, 337.5^\circ)$

Table 12: **Azimuth-sector quantization.** Azimuth is quantized into eight listener-centered 45-degree sectors. Angles are measured clockwise on the horizontal plane, with front at 0°.

to generate and verify QA labels. This includes the scene identifier, listener pose, source trajectories, source-to-listener directions and distances, source activity masks, AudioSet labels, and canonical source labels. It also stores quantized spatial states at the beginning, middle, and end of the clip, which are used to generate time-anchored localization questions. Finally, beginning-to-end spatial deltas are stored to determine whether spatial-change questions are valid.

E.2 QA Generation Rules and Templates

Answer quantization. We quantize continuous source states into listener-centered azimuth sectors, elevation categories, distance bins, and spatial-change labels. For azimuth, we use eight listener-centered 45-degree sectors, as summarized in Table 12. Elevation labels are defined as *above*, *level*, and *below*, with a $\pm 10^\circ$ band for *level*. Distance labels use ten 0.5 m bins, from 0.5–1.0 m to 5.0–5.5 m. For spatial-change questions, we instantiate a question only when the beginning-to-end change exceeds the corresponding threshold. Questions below these thresholds are skipped to avoid noisy or perceptually ambiguous labels.

Event labels and source referents. AudioSet clips can contain multiple valid labels. For event detection, the accepted answer set is the full retained AudioSet label set; for two-source event detection, the accepted set is the union of the two sources’ label sets. For source-specific questions, however, each source also has one canonical event label used as the linguistic referent. This keeps questions such as “the clapping” well-defined while still evaluating event recognition with multi-label targets.

Source-grounding validity. Two-source source-grounding questions use a label-choice answer

space: the two canonical source labels, *both*, and *neither*. The generator instantiates a source-grounding question only when the queried attribute is valid for the relevant source state and the answer is well-defined under this four-option space. This is especially important for spatial attributes, since both sources may occupy the same sector or neither source may match the queried sector.

Type-C validity. Type C is generated from two-source clips after all primitive labels have been computed. Temporal relation-change questions require the compared relation to be valid at both the beginning and end anchors. Movement-conditioned spatial relations require a valid movement-status referent, such as a moving source and a stationary source, before the spatial comparison is formed. Cross-source trajectory relations require both trajectories to have valid beginning-to-end spatial-change labels.

E.3 ST-AudioQA Statistics

Dataset splits. We partition the eligible dynamic-source pool by base trajectory identifier into 80%/10%/10% train/validation/test pools before constructing single- and two-source clips. Balanced sampling then produces 40,000/4,000/10,000 clips: half single-source, equally split between static and dynamic, and half two-source, equally split among SS, SD, DS, and DD motion cases; SD and DS constitute the mixed static–dynamic configuration reported in the main text. No source-audio or base-trajectory identifier is shared across splits.

Table 13 summarizes the number of generated QA pairs for each task family across the training, validation, and test splits. The table separates single-source and two-source QA pairs, where Type A covers single-source perception QA, Type B covers two-source perception and grounding QA, and Type C covers compositional source-time-space reasoning QA. The primitive QA families are grouped into event semantics, movement status, spatial location, spatial change, and source grounding. Type C further introduces three compositional reasoning families: temporal relation change, movement-conditioned spatial relation, and cross-source trajectory relation.

Figures 2–6 provide additional annotation statistics. Figure 2 shows the distribution of motion cases for Type A, Type B, and Type C across splits,

highlighting the balance between static and dynamic source configurations. Figure 3 visualizes the answer distributions for spatial-change questions along azimuth, elevation, and distance axes. Figures 4, 5, and 6 show the distributions of azimuth, elevation, and distance annotations for static and dynamic sources.

Table 13: **Dataset statistics by task split.** Each entry reports the number of QA pairs. Single and Two denote the number of single-source and two-source QA pairs, respectively. Type A corresponds to single-source perception QA, Type B to two-source perception and grounding QA, and Type C to compositional source-time-space reasoning QA. Primitive QA families are grouped into five high-level tasks, and compositional reasoning families into three relation tasks.

Task family	Task	Train			Validation			Test		
		Single	Two	Total	Single	Two	Total	Single	Two	Total
Type A: Single-source perception QA	Event semantics	59,988	0	59,988	6,000	0	6,000	14,988	0	14,988
	Movement status	19,996	0	19,996	2,000	0	2,000	4,996	0	4,996
	Spatial location	116,100	0	116,100	11,574	0	11,574	29,115	0	29,115
	Spatial change	20,353	0	20,353	2,034	0	2,034	5,082	0	5,082
Type B: Two-source perception and grounding QA	Event semantics	0	50,000	50,000	0	5,000	5,000	0	12,500	12,500
	Movement status	0	20,000	20,000	0	2,000	2,000	0	5,000	5,000
	Spatial location	0	74,400	74,400	0	7,440	7,440	0	18,600	18,600
	Spatial change	0	37,980	37,980	0	3,764	3,764	0	9,496	9,496
Type C: Compositional source-time-space reasoning QA	Source grounding	0	117,620	117,620	0	11,762	11,762	0	29,378	29,378
	Temporal relation change	0	78,000	78,000	0	7,800	7,800	0	19,500	19,500
	Movement-conditioned spatial relation	0	78,000	78,000	0	7,800	7,800	0	19,500	19,500
Total	Cross-source trajectory relation	0	61,965	61,965	0	6,291	6,291	0	15,679	15,679
	–	216,437	517,965	734,402	21,608	51,857	73,465	54,181	129,653	183,834

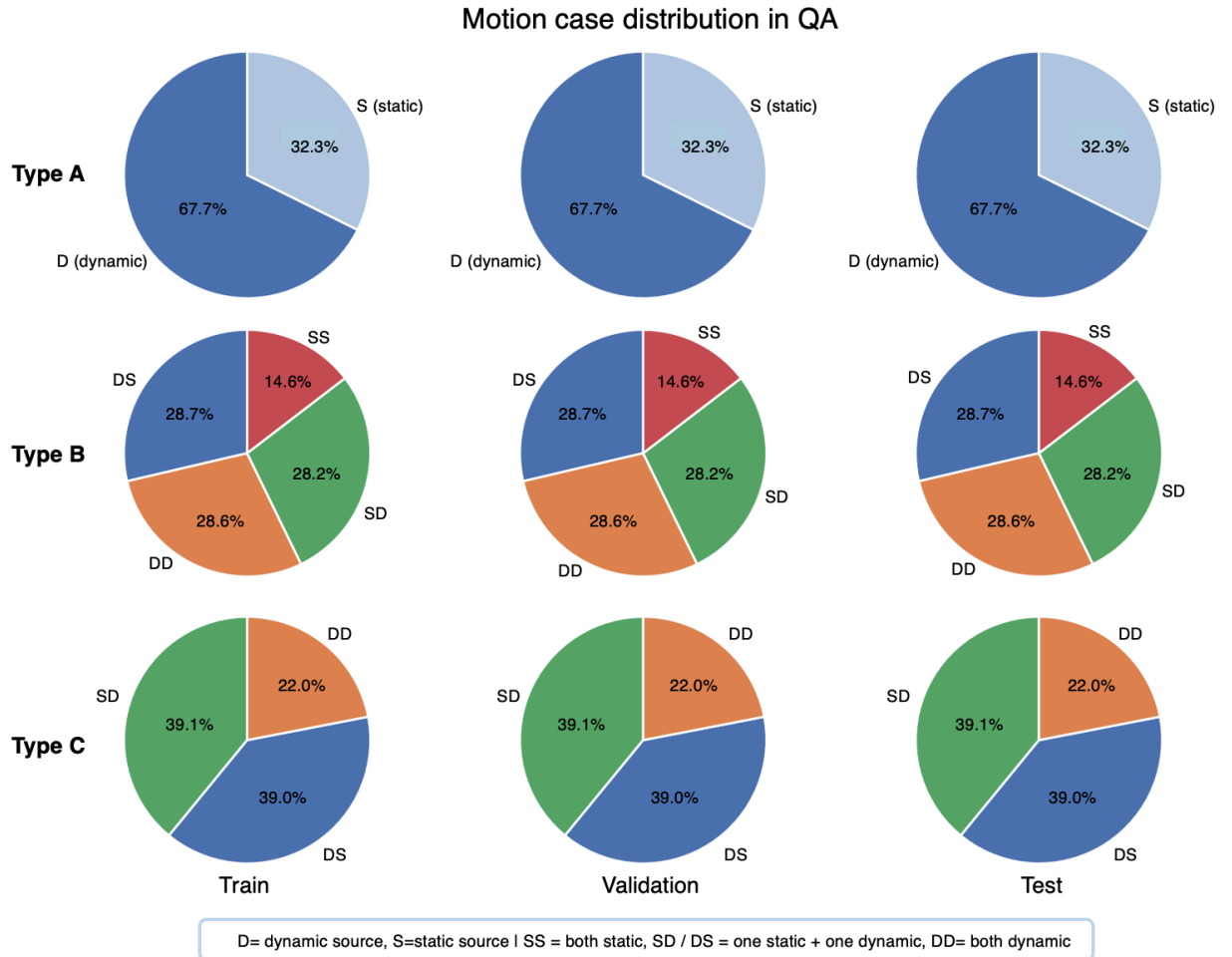


Figure 2: **Motion-case statistics.** We show distributions of motion cases for Type A, Type B, and Type C QA examples across the training, validation, and test sets.

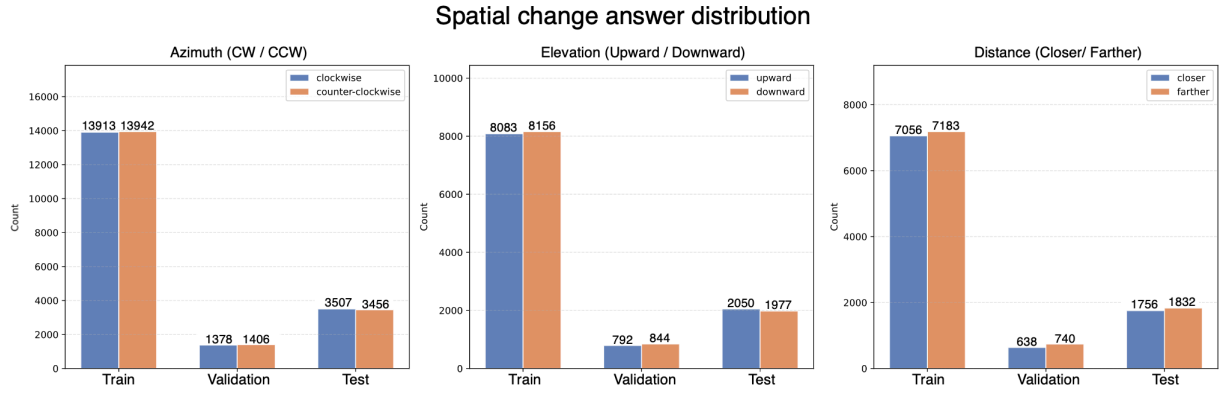


Figure 3: **Spatial-change annotation statistics.** We show distributions of spatial-change annotations along azimuth, elevation, and distance axes across the training, validation, and test sets.

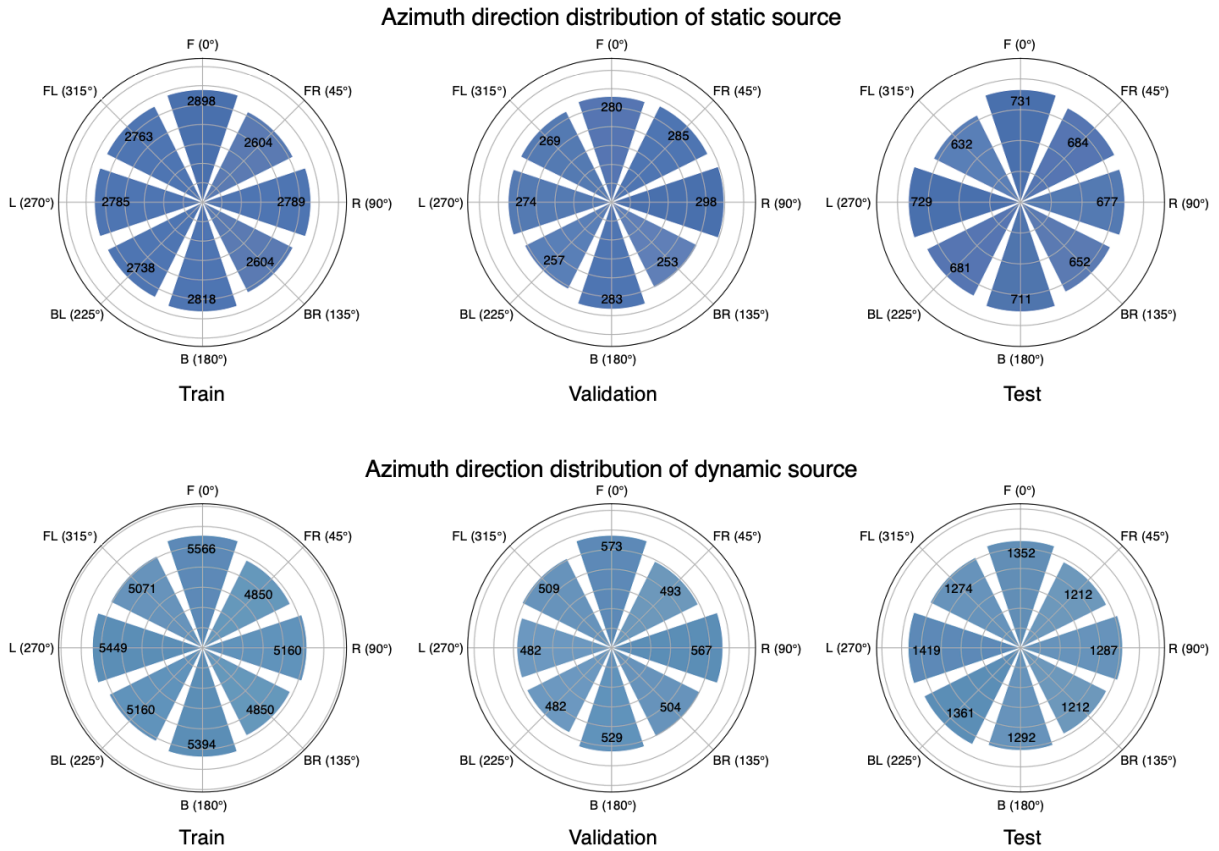


Figure 4: **Azimuth annotation statistics.** We show polar histograms of azimuth-direction annotations for static and dynamic sources across the training, validation, and test sets.

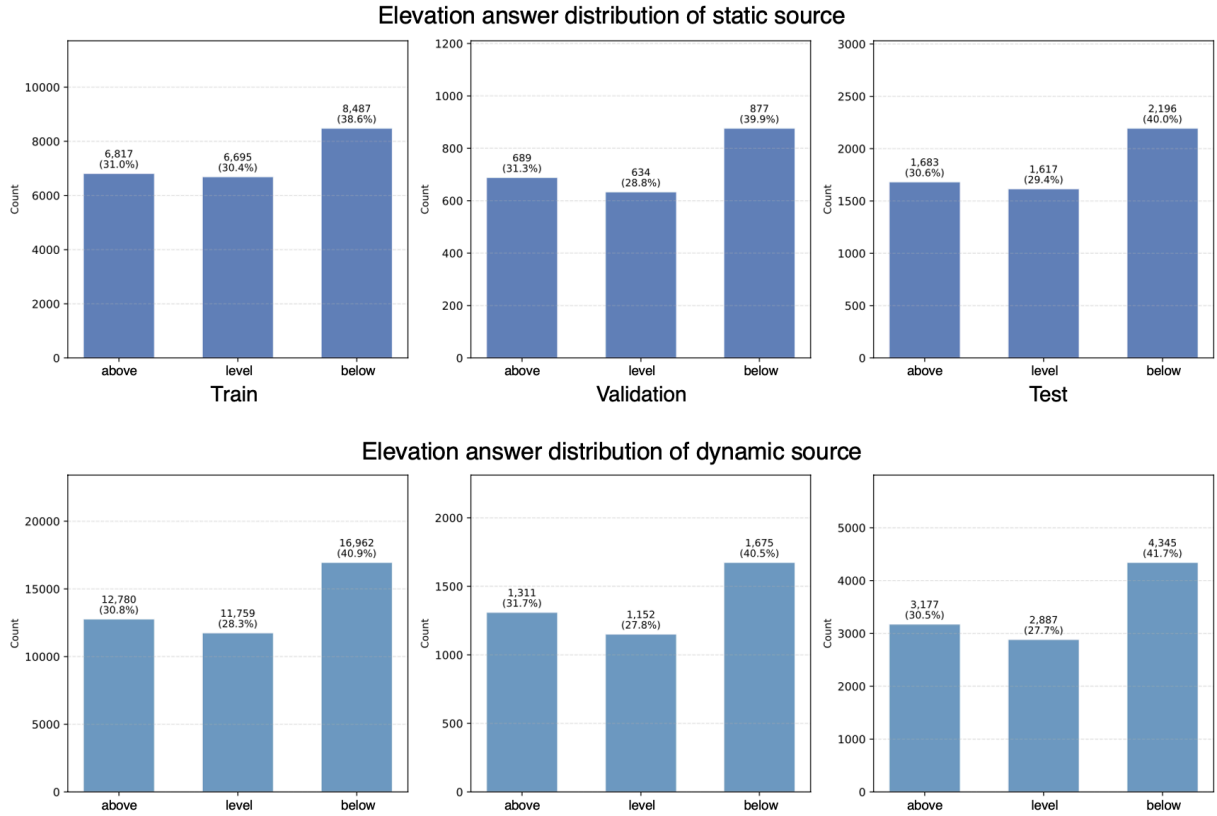


Figure 5: **Elevation annotation statistics.** We show distributions of elevation annotations for static and dynamic sources across the training, validation, and test sets.

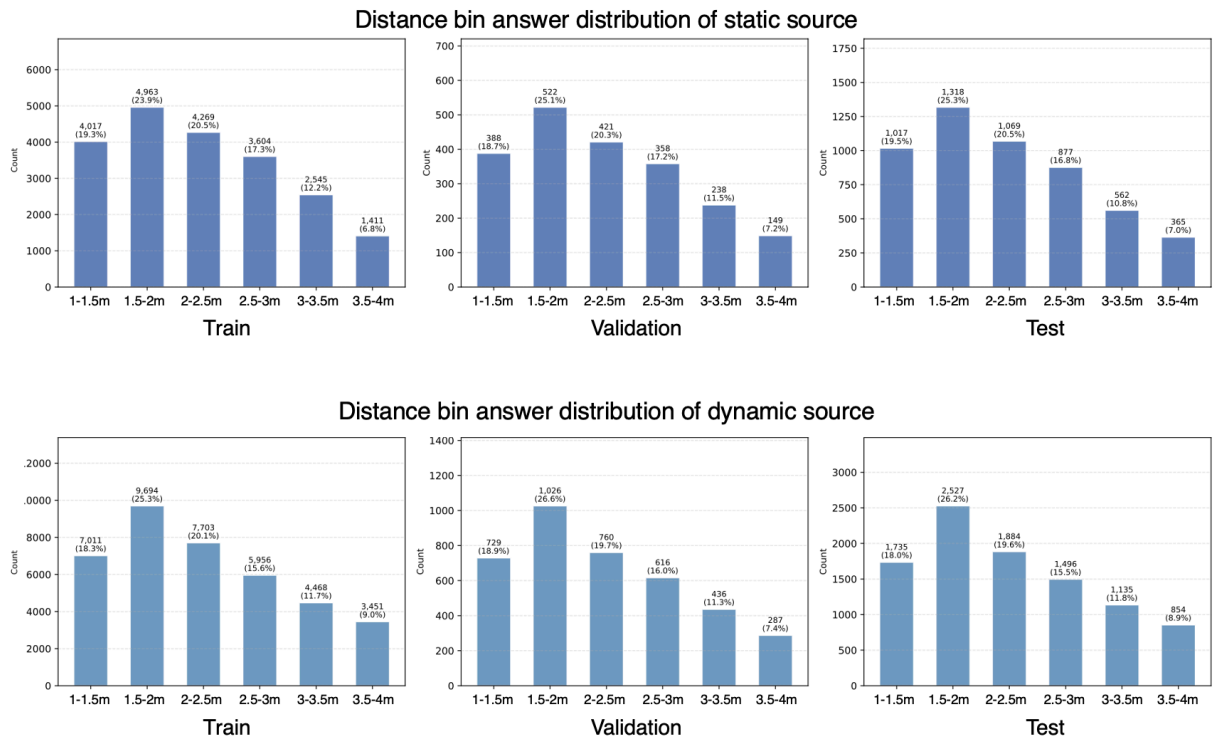


Figure 6: **Distance annotation statistics.** We show histograms of distance-bin annotations for static and dynamic sources across the training, validation, and test sets.