Hear, Localize, and Reason: Spatially Aware Scene Understanding for Audio-Visual LLMs

Kim Sung-Bin[1] **Lee Jung-Mok**[2] **Jinwoo Jung**[2] **Oh Hyun-Bin**[1]
Hyeonggon Ryu[3] **David Harwath**[4] **Tae-Hyun Oh**[2]

[1]POSTECH [2]KAIST [3]Hankuk University of Foreign Studies
[4]The University of Texas at Austin

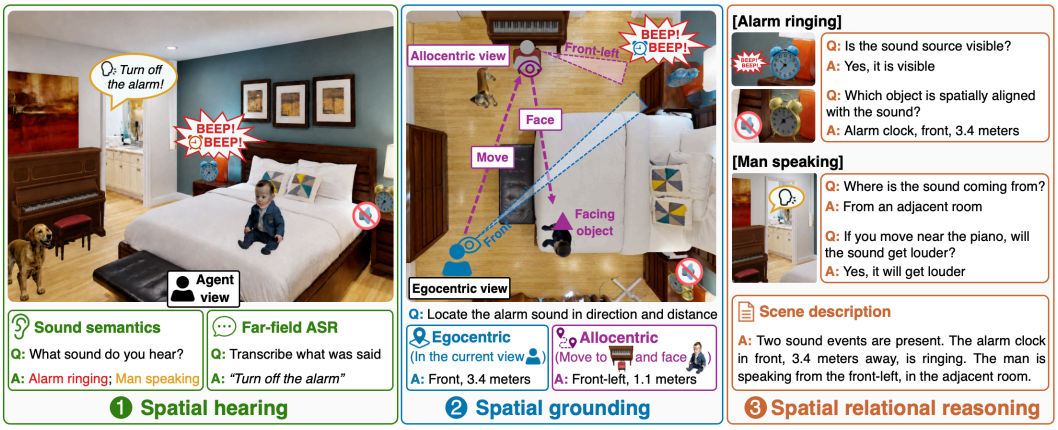


Figure 1: **Overview of HLR-AVSceneQA.** We introduce HLR-AVSceneQA, a comprehensive benchmark for spatial audio-visual scene understanding organized around a hear-localize-reason hierarchy. For a queried sound source, the model recognizes what is heard, localizes where it originates in egocentric and allocentric views, and uses the localized source together with visual scene evidence to reason about how it relates to the surrounding scene.

Abstract

Audio-visual large language models (AV-LLMs) have made strong progress in multimodal understanding, yet most treat audio as monaural semantic content, limiting their ability to reason about where sounds originate and how they relate to the visual scene. Recent spatial audio-visual studies address parts of this problem, but often focus on specific abilities, such as spatial correspondence or direction and distance reasoning. In this work, we introduce HLR-AVSceneQA, a comprehensive benchmark with a hear-localize-reason hierarchy that evaluates whether models can recognize what is heard, localize sound sources in egocentric and allocentric views, and use these localized sources to reason about how they relate to the visual scene. We further introduce HLR-LLM, which augments a strong audio-visual foundation model with a binaural spatial audio branch alongside its monaural semantic audio pathway. We train HLR-LLM with a curriculum that incorporates final-stage chain-of-thought supervision for multi-step spatial reasoning. Experiments show that existing AV-LLMs remain limited in spatial grounding and relational reasoning, while HLR-LLM substantially improves spatial audio-visual scene understanding without compromising semantic audio-visual perception. Code and benchmark data are available at https://hlrllm.github.io/.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Spatial audio and vision provide complementary cues for understanding a scene. We may hear someone calling from another room, infer that a ringing alarm clock is hidden below the bed by combining its direction with the visible layout, or decide where to move to get closer to a sound by relating auditory distance cues to nearby objects. Such everyday situations show that audio-visual scene understanding requires more than recognizing a sound event or matching it to a visible object. Accordingly, a spatially aware agent must infer where the sound originates, whether it is visible, why it may be hidden, and how its relation to the scene changes as the agent moves.

Recently, audio-visual large language models (AV-LLMs) [1, 2, 3, 4, 5] have shown strong progress in diverse multimodal understanding tasks, such as audio-visual question answering [6, 7], temporal grounding [8], and multimodal reasoning. However, most existing AV-LLMs treat audio primarily as monaural semantic content, focusing on *what* is heard rather than *where* it is heard. As a result, they struggle to reason about auditory sources in the physical scene, especially when a source is outside the camera view, occluded by objects, located in another room, or spatially ambiguous among multiple candidates. Spatial audio, such as binaural or multi-channel recordings, offers the direction and distance cues needed to overcome this limitation and support spatially grounded scene understanding.

Despite this growing need for spatially grounded audio-visual understanding, two key challenges remain. First, existing spatial audio-(visual) benchmarks typically focus on specific subproblems, such as audio-only [9, 10, 11] or speech-centric settings [12], egocentric or allocentric direction and distance reasoning [13], or precise RGB-D-based 3D grounding [14]. Existing evaluations therefore remain fragmented across modalities, assumptions, and task formats, making it difficult to assess a single AV-LLM's ability to jointly handle semantic understanding, spatial grounding, and relational reasoning. Second, although spatial audio-visual tasks require multi-step reasoning, many current training settings rely on direct-answer supervision [15, 14]. This gives models insufficient structure for decomposing complex spatial queries into interpretable intermediate steps that connect auditory semantics, spatial cues, and visual context.

In this work, we introduce **HLR-AVSceneQA**, a comprehensive benchmark for spatial audio-visual scene understanding. As shown in Fig. 1, HLR-AVSceneQA follows a hear-localize-reason hierarchy: spatial hearing evaluates what is heard, spatial grounding evaluates source localization in egocentric and allocentric views, and spatial relational reasoning evaluates how localized sound sources relate to the visual scene. By treating localized sound-source information not only as a prediction target but also as spatial evidence for downstream reasoning, HLR-AVSceneQA enables systematic evaluation of spatial grounding and scene-level reasoning. Unlike panoramic-input settings [15], HLR-AVSceneQA uses normal-field-of-view observations, requiring models to reason about visible, occluded, and off-screen sources from partial visual evidence. We construct the benchmark with a simulator-based pipeline [16], producing diverse, controllable spatial audio-visual scenes with reliable annotations.

We further introduce **HLR-LLM**, a spatially aware AV-LLM that extends audio-visual semantic understanding with explicit spatial audio modeling. Built on an existing AV-LLM [4] and augmented with a spatial audio branch, HLR-LLM improves spatial grounding and reasoning while preserving strong semantic perception. To address the multi-step nature of spatial audio-visual reasoning, we incorporate chain-of-thought [17, 18] supervision in the final stage of curriculum training, encouraging the model to connect auditory, spatial, and visual cues rather than directly predicting an answer.

Experiments show that existing AV-LLMs, including commercial models that rely on monaural audio, struggle with spatial grounding and reasoning in our benchmark. In contrast, HLR-LLM preserves strong semantic hearing ability, including sound event and speech recognition, while substantially improving spatial grounding and relational reasoning. Transfer experiments on real-world spatial audio-visual recordings [19] further suggest that pretraining on HLR-AVSceneQA provides useful spatial priors for adaptation to real-world data. Our contributions are summarized as follows:

- We introduce HLR-AVSceneQA, a comprehensive benchmark that evaluates spatial hearing and grounding both directly and as intermediate evidence for downstream spatial relational reasoning.
- We propose HLR-LLM, a spatially aware AV-LLM that augments a strong foundation model with a spatial audio pathway to improve grounding and reasoning while preserving semantic perception.
- We demonstrate that final-stage chain-of-thought curriculum training improves multi-step spatial reasoning by connecting auditory, spatial, and visual cues through intermediate reasoning steps.

Table 1: **Evaluation scope of spatial audio and audio-visual benchmarks.** We compare whether each benchmark explicitly evaluates key spatial audio-visual capabilities. A+V denotes audio-visual input; Multi-source denotes simultaneous sound sources; Far-field ASR and Sound event denote speech recognition from far-field audio and non-speech sound recognition; Ego./Allo. denote egocentric/allocentric reasoning; and Corr. denotes audio-visual correspondence. Cross-room, Hidden Source, Proximity, and AV Caption denote reasoning about across-room sources, invisible or occluded sources, nearest-object relations, and open-ended audio-visual captioning, respectively.

Benchmark	A+V	Multi-source	Far-field ASR	Sound event	Ego.	Allo.	AV Corr.	Cross-room	Hidden Source	Proximity	AV Caption
BAT [9]	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗
SpatialWhisper [12]	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗
Hear You Are [15]	✓	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗
SAVVY-Bench [13]	✓	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗
SpatialSceneQA [14]	✓	✓	✗	✗	✓	✗	✓	✗	✗	✗	✗
HLR-AVSceneQA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related work

Spatial audio understanding. Spatial audio provides geometric cues about sound sources beyond monaural semantic content. While monaural audio understanding mainly targets semantic tasks, such as automatic speech recognition [20, 21], sound event recognition [22, 23, 24], and audio captioning [25, 26, 27], spatial audio formats, such as binaural audio and first-order ambisonics (FOA), capture direction- and distance-related cues. Early work on spatial audio explores sound event localization and detection (SELD), where models jointly identify active sound events and estimate the direction of arrival [28, 29, 30]. More recent work connects spatial audio with language models for sound localization and reasoning [9, 10, 31], learns spatially aware audio-text representations through contrastive objectives [11, 32], and adapts FOA inputs for far-field speech understanding [12, 33]. These works highlight the value of spatial hearing, but remain largely audio-only or speech-centric, leaving open how spatial auditory cues should be grounded in and reasoned over a visual scene.

Spatial audio-visual benchmarks and tasks. General audio-visual benchmarks [34, 35, 36, 37, 38] evaluate broad multimodal abilities, but do not focus on spatial sound-source understanding. Recent benchmarks have begun to address specific spatial capabilities: SAVVY-Bench [13] evaluates dynamic egocentric/allocentric direction and distance reasoning, SpatialSceneQA [14] focuses on RGB-D and FOA-based 3D grounding, and Hear You Are [15] studies sound-object correspondence with panoramic images. As summarized in Table 1, these benchmarks cover complementary aspects of spatial audio-visual understanding, but focus on different subsets of capabilities. HLR-AVSceneQA instead provides a unified hear-localize-reason evaluation that spans spatial hearing, grounding, multi-source scenes, and scene-level relational reasoning, including whether localized sound-source information can support broader reasoning about the surrounding scene. It further uses normal-field-of-view egocentric observations, requiring models to reason about visible, off-screen, and hidden sound sources from partial visual observations.

Spatial audio-visual large language models (AV-LLMs). General AV-LLMs [1, 2, 3, 4, 5] support video understanding, temporal localization, and audio-visual question answering, but mostly rely on monaural audio. Spatially aware AV-LLMs add spatial auditory cues to visual inputs. Hear You Are LLM [15] uses spatial audio and panoramic images, SAVVY [13] builds dynamic spatial maps with a training-free pipeline, and JAEGER [14] integrates RGB-D observations with FOA audio for 3D grounding and reasoning. While promising, these systems are often tailored to specific tasks, such as direction/distance QA or precise 3D grounding, and do not fully inherit the broad semantic and reasoning abilities of general AV-LLMs. They are also commonly trained with direct-answer supervision despite the multi-step nature of spatial audio-visual reasoning [15, 14]. HLR-LLM addresses these gaps by extending a strong AV-LLM with a spatial audio branch and chain-of-thought supervision [17, 18], improving spatial reasoning while preserving general audio-visual capabilities.

3 HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark

We introduce HLR-AVSceneQA, a comprehensive benchmark for spatial audio-visual scene understanding. HLR-AVSceneQA follows a hear-localize-reason hierarchy that evaluates three capabilities: *spatial hearing* for recognizing what is heard, *spatial grounding* for localizing sound sources from

Table 2: **Representative QA examples in HLR-AVSceneQA.** Most tasks use controlled answer formats, such as multiple choice or discretized bins; simplified prompts and answers are shown for clarity. The final row shows one CoT supervision example used for rationale-augmented training.

Task family	Subtask	Representative example
Spatial hearing	Sound event	***Q*:** What sound do you hear? ***A*:** `Hair dryer drying`.
	Far-field ASR	***Q*:** Write the exact phrase spoken. ***A*:** `"Yes it is suppressed"`.
Egocentric grounding	Direction	***Q*:** From the current viewpoint, which direction is the sound coming from? ***A*:** `front-left`.
	Distance	***Q*:** From the current viewpoint, how far is the sound source? ***A*:** `2.5-3 m`.
Allocentric grounding	Direction	***Q*:** If you stand at the sofa and face the TV, which direction is the sound coming from? ***A*:** `back-left`.
	Distance	***Q*:** If you stand at the sofa and face the TV, how far is the sound source? ***A*:** `0-1.0 m`.
Spatial relational reasoning	Visibility	***Q*:** From the current viewpoint, is the sound source visible? ***A*:** `no`.
	Hidden-source	***Q*:** Given the sound, where is its source located? ***A*:** `hidden behind the armchair`.
	Correspondence	***Q*:** Which object is spatially aligned with the sound? ***A*:** `man in front-right`.
	Loudness change	***Q*:** If you move to the sofa, does the sound get louder or quieter? ***A*:** `quieter`.
	Proximity	***Q*:** Which object among the options is closest to the sound source? ***A*:** `armchair`.
	Scene captioning	***Q*:** Describe the sound event(s), including direction, visibility, and hidden location if applicable. ***A*:** `A hair dryer is drying from the front-left direction. It is not visible and is likely occluded by the armchair.`
CoT supervision example *provided for allocentric grounding and spatial relational reasoning; hidden-source example shown*		***Q*:** Given the sound, where is its source located? ***A*:** `The sound event is a hair dryer drying, arriving from the front-left direction and 2.7 meters away. The hair dryer is not visible because it is occluded by the armchair. Therefore, the source is hidden behind the armchair.`

egocentric and allocentric viewpoints, and *spatial relational reasoning* for understanding how localized sources relate to visible, hidden, nearby, or off-screen objects in the scene (Fig. 1). This hierarchy spans 12 tasks, as summarized in Table 2. HLR-AVSceneQA contains 175,061 training and 21,808 test QA pairs, including 53,279 training and 4,869 test pairs from scenes with multiple simultaneously active sound sources. We also provide 94,519 CoT-augmented training pairs for multi-step spatial reasoning. Because collecting real-world spatial audio-visual data with accurate annotations is challenging, we construct the benchmark with a two-stage simulator-based pipeline: Stage I synthesizes spatial audio-visual scenes with structured metadata, and Stage II generates QA pairs from this metadata.

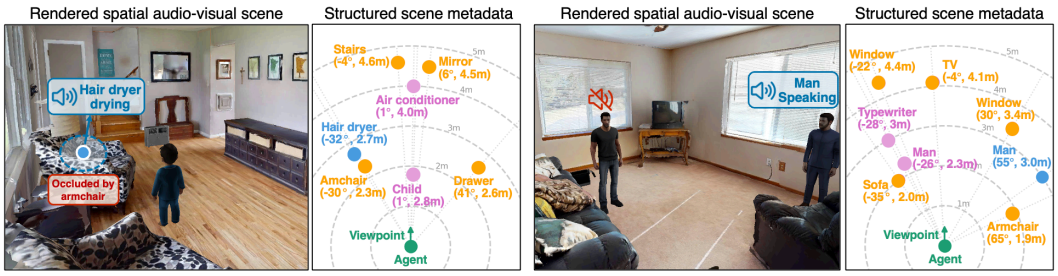
3.1 Stage I: Spatial audio-visual scene synthesis

We synthesize each sample in an indoor 3D scene from an agent's egocentric viewpoint, where the agent pose defines both the RGB camera and the acoustic receiver. We populate the scene with diverse candidate objects and designate up to two objects as active sound sources. Each sample includes a synchronized RGB observation, binaural audio, object layouts, source assignments, and geometric metadata for QA generation. Examples of rendered scenes and metadata are shown in Fig. 2.

Scene and acoustic simulator. We use SoundSpaces 2.0 [16], which renders geometry-aware room acoustics and spatializes dry monaural clips into binaural audio. For 3D environments, we use HM3DSem [39], which provides 216 indoor spaces and 3,100 rooms with dense object- and room-level semantics. These semantics support realistic rendering, layout-aware QA generation, and filtering of physically implausible placements. We use 145 scenes for training and 36 for testing.

Audio source curation. We use non-speech sounds from VGGSound [40], which contains over 200,000 in-the-wild clips across 310+ classes. Since some classes, such as *sea waves* or *people cheering*, are not naturally grounded to a single indoor object, we manually exclude them. We further use ImageBind-based text-audio similarity to filter out likely mislabeled clips [41]. For speech audio, we use LibriSpeech [42], which provides clean utterance-level transcripts for far-field speech recognition. We follow the official train/test splits of both VGGSound and LibriSpeech.

Visual object curation. We build a 3D mesh pool of sounding objects to insert into scenes. We collect available meshes from Objaverse-XL [43], and supplement categories with insufficient coverage



(a) Occluded sound source (b) Ambiguous sound source among same-category instances

Figure 2: **Examples of rendered scenes and metadata.** Our dataset covers diverse scenarios that require spatially aware audio-visual reasoning, including (a) sound sources occluded by objects and (b) ambiguous sound sources among same-category instances. Each sample is summarized with structured scene metadata. Yellow denotes existing scene objects, pink denotes inserted objects, blue denotes sounding objects, and green denotes the agent.

using off-the-shelf 2D/3D generation models [44, 45]. All assets are manually checked to remove low-quality meshes. We keep approximately 81 meshes per category for training and 35 for testing.

Spatial audio-visual scene assembly and structured metadata. Given a room from HM3DSem, we insert 2 to 6 candidate object meshes based on available free space and room scale. We activate up to two sound sources and attach them to selected objects. Some audio-object pairs are semantically matched, while others are intentionally *spatially aligned but semantically mismatched*. We further include sources that are fully visible, occluded by an object, placed in an adjacent room, or outside the agent's field of view, introducing realistic ambiguity and preventing reliance on category matching alone. The final RGB image is rendered from the agent viewpoint with a 90° horizontal field of view at 1280×720 resolution, paired with binaural audio. For each sample, we export structured metadata, including agent pose, object identities and 3D coordinates, source-object assignments, visibility state, occlusion type, and pairwise geometry, to generate QA pairs with accurate labels.

3.2 Stage II: Benchmark task design and QA construction

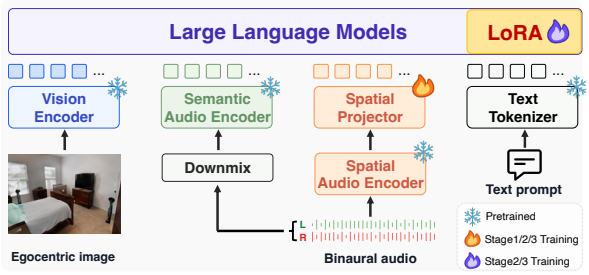
Given each rendered scene and its metadata, we generate QA pairs for three task families. To improve linguistic diversity while keeping evaluation stable, we vary question templates but preserve controlled answer formats. Most tasks use constrained outputs, such as multiple choice or discretized bins, while speech transcription and scene captioning are open-ended. Representative examples are shown in Table 2, and additional construction and quality-control details are provided in Appendix C.

Task family 1: Spatial hearing. This family evaluates semantic perception from spatial audio through ***sound event recognition*** and ***far-field speech recognition***, which identify audible event categories and transcribe spoken content as the foundation for grounding and reasoning.

Task family 2: Spatial grounding. This family evaluates sound-source localization in egocentric and allocentric frames, each with ***direction*** and ***distance*** estimation. Egocentric grounding localizes the source relative to the agent's current viewpoint, while allocentric grounding localizes it from a transformed viewpoint, defined by standing at a reference object and facing another object. This tests both source localization and viewpoint-aware spatial reasoning.

Task family 3: Spatial relational reasoning. This family extends source localization to scene-level reasoning by evaluating how localized sound sources relate to the visual scene. ***Visibility*** asks whether the source is visible from the current view, and ***correspondence*** asks which visible object is spatially aligned with the sound, where the answer may be semantically mismatched or `None`. ***Hidden-source localization*** identifies whether a non-visible source is hidden behind an object, in an adjacent room, or outside the field of view. ***Loudness change*** asks whether the source becomes louder or quieter after the agent moves to a target object; ***proximity reasoning*** asks which candidate object is nearest to the source; and ***scene captioning*** summarizes sound events, directions, and visibility states, including multiple emitters when present.

Chain-of-thought supervision. We further provide chain-of-thought (CoT) answer variants for allocentric grounding and spatial relational reasoning. The CoT rationales are deterministically generated from structured simulator metadata using task-specific templates, while GPT-4 [46] is



(a) Model architecture

Stage	Training Dataset → Objective	# Samples
1	Spatial audio-only QA [9] → Spatial audio alignment	514k
2	HLR-AVSceneQA → Spatial audio-visual instruction tuning	175k
3	HLR-AVSceneQA-CoT → Chain-of-thought tuning	94.5k

(b) Training curriculum

Figure 3: **Architecture and training curriculum of HLR-LLM.** (a) HLR-LLM augments a strong AV-LLM with a binaural spatial audio branch, enabling the language model to jointly attend to text, visual, downmixed monaural audio, and binaural spatial audio tokens. (b) The model is trained through a three-stage curriculum: spatial audio alignment, spatial audio-visual instruction tuning, and final-stage CoT tuning for structured spatial reasoning.

used only to paraphrase question wording. These rationales expose task-specific intermediate cues beyond the final answer, such as sound event identity, source location, visibility checks, distance comparisons, and how source direction and distance change under transformed viewpoints. This supervision encourages the model to connect auditory, spatial, and visual cues through explicit intermediate reasoning steps rather than directly predicting the final answer.

4 Method

We introduce HLR-LLM for spatial audio-visual scene understanding. Our goal is to enable spatial audio-visual grounding and reasoning with binaural spatial cues while retaining the semantic perception ability of a strong base AV-LLM. To this end, HLR-LLM introduces a spatial audio pathway to the base AV-LLM, where a spatial audio encoder and projector provide binaural spatial cues alongside visual context and monaural semantic audio. We train the model with a curriculum that incorporates chain-of-thought supervision to guide multi-step spatial reasoning.

4.1 HLR-LLM architecture

Figure 3 (a) shows the architecture of HLR-LLM. The model consists of the base visual and monaural audio encoders, a newly introduced spatial audio encoder, a lightweight spatial audio projector, and the language model. The monaural audio branch preserves semantic hearing abilities, such as sound event recognition and speech understanding, while the spatial audio branch provides direction- and distance-sensitive cues for grounding and reasoning. The language model integrates visual, monaural audio, spatial audio, and text tokens to generate task-specific responses.

Spatial audio tokenization. Given a binaural audio signal $b \in \mathbb{R}^{2\times T}$, we encode it with a pretrained spatial audio encoder f_{spa} [9] and map the resulting features into the language model embedding space using a trainable MLP projector p_{spa}: $Z_{\text{spa}} = p_{\text{spa}}(f_{\text{spa}}(b))$. The projected embeddings Z_{spa} serve as spatial audio tokens that provide acoustic localization cues to the LLM.

Integration with the AV-LLM. We build HLR-LLM on Qwen2.5-Omni [4], which supports text, image, and audio inputs. We introduce a new media token, `<spatial_audio>`, whose placeholder positions are replaced with the projected spatial audio tokens Z_{spa}. The resulting multimodal input sequence is $X = [E(q), Z_{\text{vis}}, Z_{\text{aud}}, Z_{\text{spa}}]$, where $E(q)$ denotes text embeddings, Z_{vis} visual tokens, Z_{aud} monaural audio tokens, and Z_{spa} spatial audio tokens. The language model generates the answer y conditioned on this sequence, *i.e.*, $p(y \mid X)$. This design preserves the base model's semantic audio-visual ability while explicitly injecting spatial acoustic cues.

4.2 Curriculum training for spatial audio-visual reasoning

Spatial audio-visual scene understanding requires low-level perception of what is heard and where it comes from, as well as high-level spatial reasoning over the scene. Figure 3 (b) summarizes the

three-stage curriculum used to train HLR-LLM, which progresses from spatial audio alignment to spatial audio-visual instruction tuning and final-stage CoT tuning.

Stage 1: Spatial audio alignment. We first pretrain the spatial audio projector using a spatial audio-only QA dataset [9]. This aligns frozen spatial audio encoder features with the LLM embedding space and initializes the model to verbalize spatial audio properties, such as sound event and location.

Stage 2: Spatial audio-visual instruction tuning. We then fine-tune the model on the training split of HLR-AVSceneQA. This stage teaches the model to jointly use visual tokens, monaural audio tokens, and spatial audio tokens across spatial hearing, grounding, and relational reasoning tasks. For example, recognizing a ringing sound can rely on the monaural audio branch, while determining whether the source is visible or hidden requires combining spatial cues with visual evidence.

Stage 3: Chain-of-thought tuning. Finally, we continue instruction tuning with CoT-formatted answers for multi-step spatial tasks. The model learns to produce a compact rationale that connects auditory cues, spatial estimates, and visual evidence before giving a normalized final answer. This encourages the model to decompose spatial audio-visual queries into interpretable reasoning steps rather than relying on direct answer prediction.

Training objective. At each stage, we minimize the standard auto-regressive cross-entropy loss over the target answer tokens. Stage 1 updates only the spatial audio projector, while Stages 2–3 update the spatial audio projector and LoRA parameters of the language model. All visual, monaural audio, and spatial audio encoders remain frozen throughout training.

4.3 Implementation details

We follow the audio preprocessing of BAT [9]. Each source clip is loudness-normalized and convolved with the room impulse response generated by SoundSpaces to produce binaural audio. The resulting two-channel waveform is sampled at 32 kHz. We use a window size of 1024, a hop size of 320, and 128 mel bins to compute spectrogram features. The spatial audio branch receives the binaural signal, while the monaural audio branch receives its channel-averaged version. For visual and monaural audio inputs, we use the default preprocessing of Qwen2.5-Omni without modification. For LoRA fine-tuning, we use rank 16, alpha 16, dropout 0.05, a learning rate of 2×10^{-4}, and weight decay of 0.01. We train for 2 epochs on four NVIDIA A6000 GPUs with a per-device batch size of 4.

5 Experiment

5.1 Experimental setup

Datasets and metrics. We primarily evaluate models on HLR-AVSceneQA for comprehensive spatial audio-visual scene understanding. Except for far-field ASR, which is evaluated by word error rate (WER), and scene captioning, which is evaluated with CIDEr-D [47], METEOR [48], and SPIDEr [49], all other tasks are evaluated using accuracy. We also construct QA formats on the STARSS23 [19] dataset to evaluate adaptation to real-world spatial audio-visual recordings, covering sound event recognition, egocentric direction and distance prediction, and visibility prediction. We restrict the evaluation to these tasks because STARSS23 does not provide the object- and scene-level annotations required to reliably define hidden-source, allocentric, and other relational reasoning tasks. All fine-tuned models are trained on the corresponding training split and evaluated on the test split. Unless otherwise noted, we report mean $\pm$ standard deviation over three runs.

Competing methods. We compare HLR-LLM with both zero-shot and fine-tuned audio-visual models. For zero-shot evaluation, we consider several open-source AV-LLMs [1, 2, 3, 4, 5] and commercial models [50, 51], which mainly use visual input and monaural audio. We also compare with SAVVY [13], which uses multichannel audio but focuses on spatial grounding tasks. For fine-tuned baselines, we initialize Hear You Are LLM [15] and Qwen2.5-Omni [4] from their respective public checkpoints and fine-tune them on the HLR-AVSceneQA training split. Although Hear You Are LLM is originally trained with panoramic images, we adapt it to the same normal-FOV inputs and HLR-AVSceneQA supervision by fine-tuning its visual pathway, projectors/adapters, and LLM while keeping the audio encoder frozen. We additionally fine-tune a dual-channel variant of Qwen2.5-Omni, which separately encodes the left and right binaural channels through the original audio pathway and feeds both audio token streams to the LLM without an explicit spatial audio branch.

Table 3: **Evaluation on spatial hearing and spatial grounding tasks.** Sound event recognition and grounding tasks are reported in accuracy (%), while speech recognition is reported in WER (%, lower is better). **V**, **M**, and **B** denote visual input, monaural audio, and binaural audio, respectively. $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$ denote independently encoded left- and right-channel audio inputs. **Best** and second-best mean results are shown in bold and underlined, respectively.

Model	Modality	Spatial Hearing		Egocentric Grounding		Allocentric Grounding	
		Sound (Acc. %)	Speech (WER ↓)	Direction	Distance	Direction	Distance
Random chance		10.00	-	12.50	10.00	12.50	10.00
Zero-shot models							
MiniCPM-o-v2.6 [1]	**V+M**	35.26±0.32	126.10±5.51	12.83±0.48	14.26±0.66	11.92±0.36	18.42±0.51
VideoLLaMA2 [2]	**V+M**	54.54±0.11	113.54±0.82	12.15±0.41	7.73±0.05	15.84±0.36	7.19±0.17
Video-SALMONN 2 [3]	**V+M**	32.20±0.13	**13.06**±0.02	17.00±0.22	7.62±0.07	17.55±0.20	6.98±0.12
Qwen2.5-Omni [4]	**V+M**	57.21±0.09	14.92±0.56	17.33±0.07	7.10±0.07	21.43±0.60	7.03±0.13
Qwen3-Omni [5]	**V+M**	55.84±0.03	15.27±1.29	15.64±0.07	11.20±0.06	18.09±0.21	9.17±0.26
SAVVY [13]	**V+M+B**	-	-	52.36±0.46	19.87±0.15	17.81±0.38	12.37±0.60
GPT-Realtime-Mini [50]	**V+M**	34.21±0.36	18.07±0.86	12.65±0.42	14.55±0.60	14.27±0.28	17.28±0.96
Gemini-2.5-Flash-Lite [51]	**V+M**	38.08±0.10	21.58±0.18	19.67±0.14	12.23±0.24	19.39±0.29	8.04±0.11
Fine-tuned models							
Hear You Are LLM [15]	**V+B**	67.67±0.37	148.00±1.13	44.65±0.42	22.63±0.54	21.53±0.19	21.57±0.33
Qwen2.5-Omni [4]	**V+M**	80.57±0.27	15.33±2.33	32.69±0.26	30.92±0.42	26.26±0.18	27.73±1.22
Dual-channel Qwen2.5-Omni	**V+$\mathbf{M}_{\mathrm{L}}$+$\mathbf{M}_{\mathrm{R}}$**	**84.52**±0.07	13.42±1.14	64.27±0.32	46.59±0.27	42.71±0.25	37.84±0.24
HLR-LLM (Ours)	**V+M+B**	81.03±0.11	14.43±0.54	**78.21**±0.34	**55.56**±0.27	**52.03**±0.26	**45.08**±0.85

Table 4: **Evaluation on spatial relational reasoning and scene captioning tasks.** Relational reasoning tasks are reported in accuracy (%). Scene captioning is evaluated with CIDEr-D, METEOR, and SPIDEr. **V**, **M**, and **B** denote visual input, monaural audio, and binaural audio, respectively. $\mathbf{M}_{\mathrm{L}}$ and $\mathbf{M}_{\mathrm{R}}$ denote independently encoded left- and right-channel audio inputs. **Best** and second-best mean results are shown in bold and underlined, respectively.

Model	Modality	Spatial Relational Reasoning (Acc. %)					Scene Captioning		
		Visibility	Correspondence	Hidden source	Loudness	Proximity	CIDEr-D	METEOR	SPIDEr
Random chance		50.00	9.49	20.00	50.00	28.68	-	-	-
Zero-shot models									
MiniCPM-o-v2.6 [1]	**V+M**	75.20±0.14	22.52±1.09	11.65±0.62	47.29±0.48	26.83±0.62	0.078±0.004	0.173±0.001	0.136±0.002
VideoLLaMA2 [2]	**V+M**	68.37±0.22	19.77±0.14	21.25±0.26	39.18±0.20	25.52±0.30	0.112±0.002	0.131±0.001	0.079±0.000
Video-SALMONN 2 [3]	**V+M**	75.56±0.00	7.97±0.22	22.00±0.31	51.89±0.00	30.10±0.05	0.075±0.001	0.094±0.001	0.074±0.000
Qwen2.5-Omni [4]	**V+M**	75.51±0.05	33.30±0.13	21.48±0.54	51.87±0.04	32.95±0.10	0.301±0.002	0.168±0.001	0.172±0.001
Qwen3-Omni [5]	**V+M**	73.90±0.30	26.18±0.28	23.58±0.36	51.29±0.85	32.06±0.39	0.344±0.004	0.203±0.000	0.260±0.001
GPT-Realtime-Mini [50]	**V+M**	69.46±0.61	12.84±0.25	20.10±0.42	50.13±0.99	26.03±0.44	0.052±0.002	0.237±0.001	0.151±0.002
Gemini-2.5-Flash-Lite [51]	**V+M**	49.81±0.32	20.29±0.05	15.36±0.17	48.27±0.57	30.89±0.34	0.570±0.010	0.209±0.002	0.218±0.003
Fine-tuned models									
Hear You Are LLM [15]	**V+B**	71.31±0.05	54.73±0.08	43.97±0.20	65.41±0.28	33.33±0.49	4.160±0.008	0.385±0.000	2.409±0.004
Qwen2.5-Omni [4]	**V+M**	62.72±0.10	55.73±0.07	51.24±0.58	68.84±0.32	47.11±0.27	4.796±0.014	0.420±0.000	2.751±0.007
Dual-channel Qwen2.5-Omni	**V+$\mathbf{M}_{\mathrm{L}}$+$\mathbf{M}_{\mathrm{R}}$**	77.30±0.12	69.07±0.15	60.12±0.43	72.59±0.31	55.64±0.36	**5.414**±0.011	**0.444**±0.001	3.070±0.005
HLR-LLM (Ours)	**V+M+B**	**80.51**±0.38	**76.19**±0.11	**67.23**±0.37	**80.87**±0.20	**61.50**±0.51	4.461±0.004	0.439±0.001	**3.109**±0.004

5.2 Main results on HLR-AVSceneQA

Tables 3 and 4 summarize the main results on HLR-AVSceneQA. In Table 3, zero-shot AV-LLMs and commercial models show reasonable spatial hearing ability, but perform poorly on direction and distance grounding, indicating that monaural audio with visual input is insufficient for spatial localization. Baselines with spatial audio improve some grounding tasks, with SAVVY [13] performing well on egocentric direction and Hear You Are LLM improving over zero-shot models, but they remain limited in broader grounding and reasoning. Fine-tuning Qwen2.5-Omni, our base model, substantially improves sound-event recognition and spatial grounding, yet accurate localization remains challenging without explicit binaural cues. Dual-channel Qwen2.5-Omni further improves localization by processing left and right channels separately through the original audio pathway, confirming the value of binaural spatial cues. HLR-LLM outperforms this baseline across all egocentric and allocentric grounding metrics, demonstrating the benefit of explicitly modeling binaural spatial information rather than exposing the LLM to separate left- and right-channel tokens.

A similar trend appears in Table 4. Zero-shot models perform reasonably well on visibility, which can often be inferred from visual evidence, but struggle with tasks requiring spatial alignment and relational reasoning. Fine-tuned baselines improve substantially, highlighting the importance of task-specific learning for spatial audio-visual understanding. HLR-LLM achieves the best performance across all spatial relational reasoning tasks and consistently outperforms the dual-channel Qwen2.5-Omni baseline, further supporting the benefit of explicit binaural modeling for scene-level reasoning.

Table 5: **Ablation study.** We ablate input modalities and training strategies to validate the contribution of each component. All metrics are reported in accuracy (%). **V**, **M**, and **B** denote visual input, monaural audio, and binaural audio, respectively. Results are mean ± standard deviation over three seeds. **Best** and second-best mean results are shown in bold and underlined, respectively. The full ablation is provided in Appendix B.

ID	Variant	Modality	Egocentric Grounding		Allocentric Grounding		Spatial Relational Reasoning		
			Direction	Distance	Direction	Distance	Correspondence	Hidden source	Proximity
Modality ablation									
(a)	Without vision	**M+B**	74.72 ± 0.36	50.64 ± 0.54	21.51 ± 0.41	26.63 ± 0.62	57.11 ± 0.11	66.62 ± 0.21	34.13 ± 0.12
(b)	Visual only	**V**	31.20 ± 0.14	20.33 ± 0.49	28.11 ± 0.36	22.96 ± 0.22	26.71 ± 0.02	24.16 ± 0.43	37.09 ± 0.55
(c)	Without binaural	**V+M**	32.69 ± 0.26	30.92 ± 0.42	26.26 ± 0.18	27.73 ± 1.22	55.73 ± 0.07	51.24 ± 0.58	47.11 ± 0.27
Training ablation									
(d)	Stage-2 training	**V+M+B**	77.92 ± 0.05	52.44 ± 0.37	$\underline{48.37} \pm 0.21$	38.83 ± 0.96	67.69 ± 0.11	66.52 ± 0.35	56.23 ± 0.43
(e)	Joint training	**V+M+B**	$\underline{78.00} \pm 0.15$	$\underline{53.04} \pm 0.17$	45.01 ± 0.59	$\underline{39.25} \pm 0.68$	$\underline{72.71} \pm 0.23$	$\mathbf{69.78} \pm 0.24$	$\underline{56.92} \pm 0.21$
(f)	HLR-LLM (Ours)	**V+M+B**	$\mathbf{78.21} \pm 0.34$	$\mathbf{55.56} \pm 0.27$	$\mathbf{52.03} \pm 0.26$	$\mathbf{45.08} \pm 0.85$	$\mathbf{76.19} \pm 0.11$	$\underline{67.23} \pm 0.37$	$\mathbf{61.50} \pm 0.51$

Table 6: **Real-world adaptation on STARSS23.** "Pretrain" denotes HLR-AVSceneQA pretraining before STARSS23 fine-tuning. "Sound" denotes sound event recognition; direction, distance, and visibility are egocentric. All metrics are accuracy (%). Results are mean ± standard deviation over three seeds. **Best** and second-best mean results are shown in bold and underlined, respectively.

ID	Model	Pretrain	Sound	Direction	Distance	Visibility
(a)	Hear You Are LLM [15]	–	38.66 ± 0.16	15.25 ± 0.57	$\underline{34.87} \pm 0.35$	62.35 ± 0.07
(b)	Hear You Are LLM [15]	✓	40.80 ± 0.76	$\underline{16.15} \pm 0.36$	26.93 ± 0.45	63.80 ± 0.48
(c)	HLR-LLM	–	$\underline{44.59} \pm 0.13$	16.13 ± 0.04	28.11 ± 0.04	$\underline{69.67} \pm 0.18$
(d)	HLR-LLM	✓	$\mathbf{53.93} \pm 0.14$	$\mathbf{29.92} \pm 0.13$	$\mathbf{35.13} \pm 0.07$	$\mathbf{71.68} \pm 0.09$

Compared with Hear You Are LLM, HLR-LLM achieves consistent gains across relational reasoning tasks, demonstrating stronger scene-level reasoning beyond source localization. For scene captioning, HLR-LLM achieves the best SPIDEr score, while dual-channel Qwen2.5-Omni obtains the best CIDEr-D and METEOR scores. Overall, the results show that explicit spatial audio modeling benefits not only source localization but also reasoning about sound sources in the visual scene.

5.3 Ablation study

Table 5 examines the contribution of each modality and training strategy. Removing vision (a) still retains egocentric direction and distance performance, since binaural audio provides direct spatial cues. However, it substantially reduces allocentric grounding and proximity reasoning, where visual object locations are required. Interestingly, the model without vision still performs relatively well on hidden-source reasoning, suggesting that binaural cues provide useful evidence for distinguishing non-visible source configurations. The visual-only model (b) performs poorly across most tasks, showing that vision alone is insufficient for reliable sound-source localization, especially when sources are off-screen or hidden. Removing binaural audio (c) also causes a large drop in grounding tasks, showing that monaural audio lacks the spatial cues needed for accurate localization. At the same time, its relatively stronger proximity performance suggests that visual context and semantic audio can still support some object-level relational reasoning.

The training ablation shows how each stage contributes to spatial reasoning. Stage-2 training (d), which includes Stages 1–2 but omits CoT tuning, already provides strong performance. Joint training with CoT (e), which mixes Stage-2 instruction data with CoT-formatted data, improves egocentric grounding, correspondence, and hidden-source reasoning, but does not consistently improve all grounding tasks. In contrast, our final model (f), which applies CoT tuning after Stage-2 instruction tuning, performs best across all grounding metrics and most relational reasoning tasks, with clear gains in allocentric grounding, correspondence, and proximity. These gains are most evident on tasks requiring multi-step spatial reasoning, such as transforming viewpoints or comparing object–source relations. Overall, the ablation indicates that binaural audio supports spatial localization, vision supports object-level grounding, and final-stage CoT tuning improves complex spatial reasoning.

5.4 Real-world adaptation

Table 6 evaluates real-world adaptation on STARSS23. We fine-tune each model on the STARSS23 training set and compare performance with and without HLR-AVSceneQA pretraining. Comparing (a)

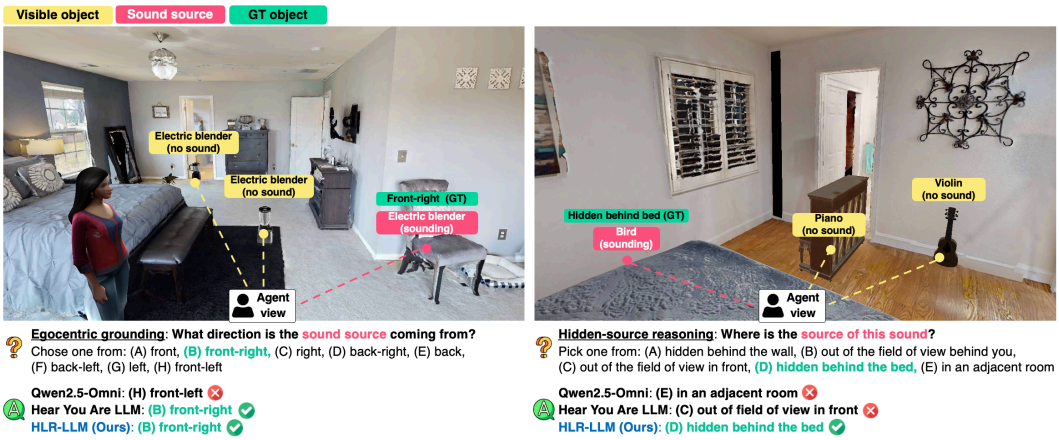


Figure 4: **Qualitative examples of spatial grounding.** HLR-LLM correctly identifies the sound direction in a scene with multiple similar objects (left) and localizes a non-visible sound source as hidden behind the bed (right), outperforming the comparison models in the hidden-source example.

vs. (b) and (c) vs. (d), HLR-AVSceneQA pretraining improves most evaluated tasks, suggesting that it provides useful spatial priors for adapting to real spatial audio-visual recordings. Among models with HLR-AVSceneQA pretraining, HLR-LLM (d) outperforms Hear You Are LLM (b) across all metrics, with particularly large gains in sound recognition and direction grounding, demonstrating stronger adaptation to real-world spatial audio-visual data.

5.5 Qualitative results

Figure 4 shows representative qualitative results for spatial grounding and relational reasoning. In the left example, multiple visually similar electric blenders are present, while the sound originates from the front-right direction. The monaural Qwen2.5-Omni predicts an incorrect direction, whereas models with binaural spatial input (Hear You Are LLM and HLR-LLM) correctly localize the source. The right example is more challenging because the sounding bird is not visible and must be localized using spatial audio and scene layout. While Qwen2.5-Omni and Hear You Are LLM incorrectly attribute the source to other non-visible regions, HLR-LLM correctly identifies it as hidden behind the bed. Together, these examples show that explicit spatial audio cues help HLR-LLM localize sound sources and ground them within visible or hidden regions of the scene. Further results and analyses across diverse spatial reasoning scenarios are provided in Appendix B.

6 Conclusion and discussion

We introduce HLR-AVSceneQA, a comprehensive benchmark for spatial audio-visual scene understanding organized around a hear-localize-reason hierarchy. By covering spatial hearing, egocentric and allocentric grounding, and spatial relational reasoning, HLR-AVSceneQA evaluates whether AV-LLMs can move beyond recognizing audio-visual content to connect what is heard, where it originates, and how the source relates to the surrounding scene. We also propose HLR-LLM, which augments a strong AV-LLM with a binaural spatial audio branch and a staged curriculum with final-stage CoT supervision, improving spatial reasoning while retaining strong semantic audio-visual perception. Real-world adaptation experiments on STARSS23 further show that HLR-AVSceneQA pretraining provides useful spatial priors for downstream adaptation. These results highlight the importance of explicit spatial audio modeling and structured training for spatially aware AV-LLMs.

Our work also has limitations. HLR-AVSceneQA is built with a simulator-based pipeline, which provides controllable annotations but cannot fully capture the acoustic complexity and visual diversity of the real world. The benchmark focuses on indoor scenes, leaving dynamic agents, moving sound sources, and richer embodied interaction for future work. Future research should extend spatial audio-visual benchmarks to more realistic and dynamic settings and develop AV-LLMs that can actively perceive, move, and reason in spatial environments.

Acknowledgments and disclosure of funding

Funding. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean 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 (33%); No. RS-2024-00358135, Corner Vision: Learning to Look Around the Corner through Multi-modal Signals (24%)), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2025-25443318, Physically-grounded Intelligence: A Dual Competency Approach to Embodied AGI through Constructing and Reasoning in the Real World (33%); No. RS-2026-25518317, Development of AI memory mechanism that reflects human cognitive principles (10%)), the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01; N10250156), and the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

Competing interests. The authors declare no competing interests.

References

[1] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 2, 3, 7, 8, 15

[2] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3, 7, 8, 15, 18

[3] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Video-SALMONN 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. 2, 3, 7, 8, 15, 18

[4] 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. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 2, 3, 6, 7, 8, 15, 16, 18

[5] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 2, 3, 7, 8, 15

[6] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. CAT: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[7] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[9] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. BAT: Learning to reason about spatial sounds with large language models. In *International Conference on Machine Learning (ICML)*, 2024. 2, 3, 6, 7, 15

[10] Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. OWL: Geometry-aware spatial reasoning for audio large language models. In *International Conference on Learning Representations (ICLR)*, 2026. 2, 3

[11] Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. Learning spatially-aware language and audio embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3

[12] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, et al. Can large language models understand spatial audio? In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2024. 2, 3

[13] Mingfei Chen, Zijun Cui, Xiulong Liu, Jinlin Xiang, Caleb Zheng, Jingyuan Li, and Eli Shlizerman. SAVVY: Spatial awareness via audio-visual LLMs through seeing and hearing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2, 3, 7, 8, 16

[14] Zhan Liu, Changli Tang, Yuxin Wang, Zhiyuan Zhu, Youjun Chen, Yiwen Shao, Tianzi Wang, Lei Ke, Zengrui Jin, and Chao Zhang. JAEGER: Joint 3D audio-visual grounding and reasoning in simulated physical environments. In *International Conference on Machine Learning (ICML)*, 2026. 2, 3

[15] Hyeonggon Ryu, Joon Son Chung, and David Harwath. Hear you are: Teaching LLMs spatial reasoning with vision and spatial sound. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 3, 7, 8, 9, 16

[16] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 4, 21

[17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[18] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. 2, 3

[19] Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, et al. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 7, 16

[20] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[21] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023. 3

[22] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017. 3

[23] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020. 3

[24] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2021. 3

[25] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 3

[26] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. 3

[27] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. Prefix tuning for automated audio captioning. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023. 3

[28] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 2019. 3

[29] Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2021. 3

[30] Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022. 3

[31] Oh Hyun-Bin, Kazuki Shimada, Yuhta Takida, Kim Sung-Bin, Toshimitsu Uesaka, Takashi Shibuya, Kyeongyoon Lee, Tae-Hyun Oh, and Yuki Mitsufuji. Spatio-temporal audio language modeling for dynamic sound sources. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026. 3

[32] Kentaro Seki, Yuki Okamoto, Kouei Yamaoka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. Spatial-CLAP: Learning spatially-aware audio–text embeddings for multi-source conditions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026. 3

[33] Jiamin Xie, Ju Lin, Yiteng Huang, Tyler Vuong, Zhaojiang Lin, Zhaojun Yang, Peng Su, Prashant Rawat, Sangeeta Srivastava, Ming Sun, et al. Thinking in directivity: Speech large language model for multi-talker directional speech recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2025. 3

[34] Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. AVH-Bench: A cross-modal hallucination benchmark for audio-visual large language models. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[35] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. AVTrustBench: Assessing and enhancing reliability and robustness in audio-visual LLMs. In *IEEE International Conference on Computer Vision (ICCV)*, 2025. 3

[36] Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Probing audio-visual reasoning in multimodal language models through the lens of audio. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026. 3

[37] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. AVQA: A dataset for audio-visual question answering on videos. In *ACM International Conference on Multimedia (MM)*, 2022. 3

[38] Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-Omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025. 3

[39] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-Matterport 3D semantics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4, 21

[40] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 4

[41] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One embedding space to bind them all. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[42] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015. 4

[43] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4, 21

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5, 21

[45] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 5, 21

[46] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 26

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7

[48] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005. 7

[49] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 7

[50] OpenAI. Realtime API documentation, 2025. Accessed: 2026-05-07. 7, 8, 15

[51] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 7, 8, 15

[52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 15

[53] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 15

[54] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001. 16

[55] Kourosh Saberi and David R Perrott. Minimum audible movement angles as a function of sound source trajectory. *The Journal of the Acoustical Society of America (JASA)*, 1990. 16

[56] Andrew J Kolarik, Brian CJ Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan. Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics*, 2016. 16

[57] Rachel Ege, A John Van Opstal, and Marc M Van Wanrooij. Accuracy-precision trade-off in human sound localisation. *Scientific Reports*, 2018. 21

Table of Contents

A Additional experimental details **15**

A.1 Implementation details 15

A.2 Details on comparison model experiments 15

A.3 STARSS23 preprocessing and QA construction 16

B Additional results and analysis **17**

B.1 Ablation study 17

B.2 Analysis on the STARSS23 dataset 18

B.3 Qualitative results 18

C Details on HLR-AVSceneQA **21**

C.1 Scene initialization, agent placement, and scene generation 21

C.2 Task definitions and details 23

C.3 Chain-of-thought supervision 26

C.4 Quality control 28

C.5 Dataset statistics 28

D Broader impact **29**

E LLM usage **29**

F Licenses and terms of existing assets **30**

A Additional experimental details

A.1 Implementation details

HLR-LLM is built on Qwen2.5-Omni with an additional binaural spatial audio pathway. For the spatial branch, we use a pretrained Spatial-AST encoder following BAT [9]. Spatial-AST produces audio tokens together with class tokens trained for semantic classification, direction prediction, and distance prediction. We use these tokens as spatial audio features and map them into the language embedding space with a lightweight projector. The projector is an MLP with a linear–GELU–linear structure, which has been shown to be effective for extending LLMs to multimodal inputs [52, 53].

Training follows the three-stage curriculum described in the main paper. In Stage 1, we train only the spatial alignment modules, including the spatial audio resampler and projector, while freezing the language model, visual encoder, monaural audio encoder, and spatial audio encoder. In Stages 2–3, we LoRA-finetune the language model and continue updating the spatial alignment modules while keeping the encoders frozen. We apply LoRA to all linear layers with rank 16, alpha 16, and dropout 0.05, and use LoRA+ with a larger learning-rate ratio for the LoRA B matrix. We train with bfloat16 precision, FlashAttention, learning rate 2×10^{-4}, warmup ratio 0.05, weight decay 0.01, and gradient clipping at 1.0. Training is performed for 2 epochs on four NVIDIA A6000 GPUs with per-device batch size 4 and gradient accumulation 4.

A.2 Details on comparison model experiments

Since spatial AV-LLMs are still limited, we first evaluate existing AV-LLMs that accept visual input and monaural audio [1, 2, 3, 4, 5, 50, 51]. For these models, the binaural audio is converted to

monaural during preprocessing. For SAVVY [13], we follow the official code to process our inputs and evaluate only the applicable spatial grounding tasks, matching the scope of the original paper.

For fairer comparisons, we additionally fine-tune three baselines using the same direct-answer HLR-AVSceneQA training split.

Hear You Are LLM. We fine-tune Hear You Are LLM [15], the closest existing trainable model that uses both visual input and binaural audio for spatial audio-visual reasoning. Although it was originally trained with panoramic observations, HLR-AVSceneQA is defined under normal-field-of-view (normal-FOV) inputs, where sources may be visible, occluded, off-screen, or located in another room. We therefore adapt Hear You Are LLM to the same 1280×720 normal-FOV images, 32-kHz binaural audio, QA prompts, answer targets, and evaluation protocol used for HLR-LLM. Its SigLIP image encoder can directly process the normal-FOV images as patch tokens without architectural modification. During fine-tuning, we apply LoRA to the image encoder (rank 16, alpha 32) and update the projectors/adapters and LLM, while keeping the audio encoder frozen. We retain the original Hear You Are architecture and do not add the spatial audio branch used in HLR-LLM. Accordingly, this comparison evaluates two adapted spatial AV-LLMs under matched data and evaluation conditions, rather than providing a fully architecture-matched comparison.

Qwen2.5-Omni. We fine-tune Qwen2.5-Omni [4], our monaural-audio base model, on HLR-AVSceneQA. This provides a controlled baseline for evaluating spatial reasoning without explicit binaural spatial modeling.

Dual-channel Qwen2.5-Omni. We additionally fine-tune a dual-channel variant of Qwen2.5-Omni to test whether simply providing both binaural channels is sufficient for spatial reasoning. Instead of downmixing the binaural input to mono, the left and right channels are separately encoded through the original monaural audio pathway, and both resulting audio token streams are provided to the LLM. Unlike HLR-LLM, this variant does not introduce a dedicated spatial audio encoder or spatial projector. Thus, it isolates the benefit of explicit spatial-audio modeling from simply exposing the base model to both audio channels.

A.3 STARSS23 preprocessing and QA construction

STARSS23 [19] is an audio-visual dataset of real-world spatial recordings with synchronized video, first-order ambisonic (FOA) audio, and frame-level spatio-temporal annotations of sound events. We use STARSS23 as a real-world benchmark complementary to our simulator-generated HLR-AVSceneQA. Starting from the official STARSS23 development split, we follow the official train/test partition and divide each recording into non-overlapping 10-second clips. Clips without a corresponding video stream are removed, resulting in 1,317 train clips and 1,116 test clips. For each clip, we extract a rectilinear egocentric image from the midpoint video frame. We also convert the FOA audio into binaural audio using the CIPIC subject-048 HRTF convention [54]. This makes the real-world recordings compatible with the binaural audio interface used by our spatial audio-visual models while retaining the original STARSS23 annotations.

We convert STARSS23 azimuth annotations into our egocentric convention, where 0° is front, 90° is right, 180° is back, and 270° is left. We construct QA pairs for sound-event recognition, egocentric direction prediction, egocentric distance prediction, and visibility prediction. We distinguish static and dynamic sources using angular and distance variation of active annotation frames, and use only static sources for spatial QA. Following psychoacoustic studies of auditory motion and distance discrimination [55, 56], a source is treated as static when its angular variation is within 5° and its relative distance variation is within 20%.

For each retained static source, we compute representative azimuth and distance values from its active annotation frames and discretize them into the same egocentric bins used in HLR-AVSceneQA. We remove ambiguous QA items, including cases where frame-level annotations cross bin boundaries or multiple same-class sources make the queried source unclear. Since STARSS23 does not provide object-level semantic masks, visibility is defined geometrically by whether the representative source direction falls inside the canonical rectilinear camera field of view.

Scope of real-world tasks. We restrict the real-world evaluation to tasks that can be reliably derived from STARSS23 annotations. STARSS23 does not provide the dense object- and scene-level annotations required for HLR-style hidden-source or allocentric grounding, including object geometry,

Table S1: **Ablation study on spatial hearing and spatial grounding tasks.** Sound event recognition and grounding tasks are reported in accuracy (%), while speech recognition is reported in WER (%, lower is better). **V**, **M**, and **B** denote visual input, monaural audio, and binaural audio, respectively. Results are mean ± standard deviation over three seeds. **Best** and second-best mean results are shown in bold and underlined, respectively.

ID	Variant	Modality	Spatial Hearing		Egocentric Grounding		Allocentric Grounding	
			Sound (Acc. %)	Speech (WER ↓)	Direction	Distance	Direction	Distance
Modality ablation								
(a)	Without vision	**M+B**	78.44 ± 0.11	14.62 ± 2.81	74.72 ± 0.36	50.64 ± 0.54	21.51 ± 0.41	26.63 ± 0.62
(b)	Visual only	**V**	29.16 ± 0.04	250.14 ± 15.37	31.20 ± 0.14	20.33 ± 0.49	28.11 ± 0.36	22.96 ± 0.22
(c)	Without binaural	**V+M**	80.57 ± 0.27	15.33 ± 2.33	32.69 ± 0.26	30.92 ± 0.42	26.26 ± 0.18	27.73 ± 1.22
(d)	Without monaural	**V+B**	80.81 ± 0.03	361.05 ± 4.00	77.87 ± 0.14	47.78 ± 0.49	50.23 ± 0.41	37.95 ± 0.24
Data scale ablation (Stage-2 training)								
(e)	10% data	**V+M+B**	75.77 ± 0.26	**10.65** ± 0.18	75.39 ± 0.07	48.15 ± 0.15	29.31 ± 0.11	29.10 ± 0.71
(f)	50% data	**V+M+B**	**81.09** ± 0.13	13.97 ± 0.98	76.22 ± 0.05	**55.88** ± 0.19	49.73 ± 0.36	41.20 ± 0.33
(g)	100% data	**V+M+B**	**81.09** ± 0.28	11.06 ± 0.11	77.92 ± 0.05	52.44 ± 0.37	48.37 ± 0.21	38.83 ± 0.96
Training ablation (100% data; compare (g)–(i))								
(h)	Joint training	**V+M+B**	79.09 ± 0.03	12.19 ± 0.43	78.00 ± 0.15	53.04 ± 0.17	45.01 ± 0.59	39.25 ± 0.68
(i)	HLR-LLM (Ours)	**V+M+B**	81.03 ± 0.11	14.43 ± 0.54	**78.21** ± 0.34	55.56 ± 0.27	**52.03** ± 0.26	**45.08** ± 0.85

Table S2: **Ablation study on spatial relational reasoning and scene captioning tasks.** Relational reasoning tasks are reported in accuracy (%). Scene captioning is evaluated with CIDEr-D, METEOR, and SPIDEr. **V**, **M**, and **B** denote visual input, monaural audio, and binaural audio, respectively. Results are mean ± standard deviation over three seeds. **Best** and second-best mean results are shown in bold and underlined, respectively.

ID	Variant	Modality	Spatial Relational Reasoning (Acc. %)					Scene Captioning		
			Visibility	Correspondence	Hidden source	Loudness	Proximity	CIDEr-D	METEOR	SPIDEr
Modality ablation										
(a)	Without vision	**M+B**	65.43 ± 0.42	57.11 ± 0.11	66.62 ± 0.21	69.37 ± 0.14	34.13 ± 0.12	5.390 ± 0.014	0.441 ± 0.001	3.070 ± 0.007
(b)	Visual only	**V**	53.21 ± 0.10	26.71 ± 0.02	24.16 ± 0.43	52.97 ± 0.34	37.09 ± 0.55	1.757 ± 0.006	0.316 ± 0.001	1.171 ± 0.002
(c)	Without binaural	**V+M**	62.72 ± 0.10	55.73 ± 0.07	51.24 ± 0.58	68.84 ± 0.32	47.11 ± 0.27	4.796 ± 0.014	0.420 ± 0.000	2.751 ± 0.007
(d)	Without monaural	**V+B**	75.67 ± 0.07	66.58 ± 0.09	66.48 ± 0.10	77.64 ± 0.14	55.23 ± 0.62	5.475 ± 0.016	0.440 ± 0.001	3.091 ± 0.008
Data scale ablation (Stage-2 training)										
(e)	10% data	**V+M+B**	71.92 ± 0.29	65.72 ± 0.24	63.53 ± 0.44	74.75 ± 0.22	46.42 ± 0.66	4.946 ± 0.013	0.424 ± 0.001	2.842 ± 0.008
(f)	50% data	**V+M+B**	76.74 ± 0.20	66.82 ± 0.16	66.17 ± 0.48	78.27 ± 0.14	55.54 ± 0.60	5.612 ± 0.017	0.448 ± 0.001	3.168 ± 0.009
(g)	100% data	**V+M+B**	77.62 ± 0.07	67.69 ± 0.11	66.52 ± 0.35	78.49 ± 0.36	56.23 ± 0.43	**5.664** ± 0.010	**0.450** ± 0.000	**3.208** ± 0.005
Training ablation (100% data; compare (g)–(i))										
(h)	Joint training	**V+M+B**	79.68 ± 0.12	72.71 ± 0.23	**69.78** ± 0.24	79.36 ± 0.21	56.92 ± 0.21	5.485 ± 0.000	0.439 ± 0.001	3.100 ± 0.001
(i)	HLR-LLM (Ours)	**V+M+B**	**80.51** ± 0.38	**76.19** ± 0.11	67.23 ± 0.37	**80.87** ± 0.20	**61.50** ± 0.51	4.461 ± 0.004	0.439 ± 0.001	3.109 ± 0.004

occlusion state, room layout, and reference-object information. We therefore avoid constructing uncertain pseudo-labels for these unsupported tasks.

B Additional results and analysis

B.1 Ablation study

Tables S1 and S2 provide the full ablation results on HLR-AVSceneQA. The modality ablation shows that binaural audio is essential for spatial grounding. Removing the binaural pathway in (c) substantially degrades egocentric and allocentric grounding as well as spatial relational reasoning. The visual-only variant (b) also performs poorly across these tasks. In contrast, removing monaural audio in (d) has a smaller effect on spatial grounding, but severely degrades far-field speech recognition, highlighting the role of the monaural pathway in preserving semantic audio understanding.

The data-scale ablation in (e)–(g) further disentangles the effect of training data from explicit spatial audio modeling. Even with only 10% of the training data, (e) substantially outperforms the 100%-data model without binaural audio (c) on egocentric grounding and most relational reasoning tasks. Increasing the training data generally improves spatial reasoning, particularly relational reasoning and scene captioning, although some grounding tasks exhibit non-monotonic variation.

The training ablation compares Stage-2 training (g), joint training (h), and our final staged curriculum with CoT tuning (i). Final-stage CoT tuning improves both egocentric and allocentric grounding

Table S3: **Zero-shot evaluation on STARSS23.** We evaluate direct transfer to real-world spatial audio-visual recordings without any STARSS23 fine-tuning. Sound denotes sound event recognition; direction, distance, and visibility are egocentric. All metrics are reported in accuracy (%).

Model	Sound	Direction	Distance	Visibility
VideoLLaMA2 [2]	34.52 ± 0.31	12.61 ± 0.16	4.91 ± 0.17	61.80 ± 0.31
Video-SALMONN 2 [3]	35.30 ± 0.11	14.36 ± 0.10	4.80 ± 0.29	66.99 ± 0.00
Qwen2.5-Omni [4]	$\mathbf{43.22} \pm 0.79$	14.66 ± 0.76	3.62 ± 0.17	67.61 ± 0.25
HLR-LLM (Ours)	33.99 ± 0.34	$\mathbf{17.48} \pm 1.17$	$\mathbf{19.76} \pm 3.42$	$\mathbf{68.61} \pm 0.09$

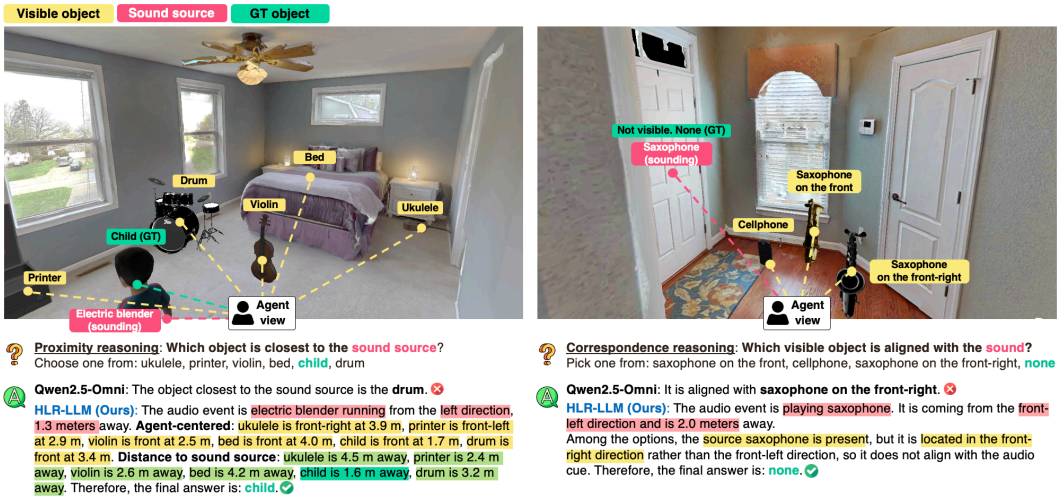


Figure S1: **Qualitative chain-of-thought reasoning examples from HLR-AVSceneQA.** We compare Qwen2.5-Omni with HLR-LLM on proximity and audio-visual correspondence tasks. HLR-LLM generates explicit CoT rationales that localize the sound source, relate it to visible objects, and derive the correct answer.

and achieves the strongest performance on most relational reasoning tasks, with particularly clear gains in correspondence and proximity. The CoT supervision explicitly exposes intermediate steps, such as identifying the sound event, estimating the source location, transforming viewpoints, and comparing source–object relations, which helps the model decompose complex spatial queries before predicting the final answer. We also observe that CoT tuning degrades scene captioning metrics and far-field ASR, suggesting a trade-off between structured multi-step reasoning and tasks that primarily require direct semantic prediction. Overall, we select (i), which uses all modalities and final-stage CoT tuning, as our final model.

B.2 Analysis on the STARSS23 dataset

Table S3 evaluates direct transfer from HLR-AVSceneQA to real-world STARSS23 recordings without target-domain fine-tuning. Despite being trained only on simulated scenes, HLR-LLM outperforms existing AV-LLMs on direction, distance, and visibility prediction, with particularly notable gains in distance estimation. Although sound recognition remains less competitive, these results suggest that spatial grounding capabilities learned from HLR-AVSceneQA can transfer to real-world recordings without target-domain supervision.

B.3 Qualitative results

We present qualitative chain-of-thought (CoT) examples from HLR-AVSceneQA in Figure S1. The examples focus on spatial relational reasoning, where HLR-LLM explicitly decomposes the problem into sound-source localization and subsequent reasoning over visible objects. In the proximity example, Qwen2.5-Omni selects the drum, whereas HLR-LLM's CoT first estimates the sound-source location and candidate-object positions, compares their distances to the source, and correctly identifies the child as the closest object. In the correspondence example, Qwen2.5-Omni incorrectly

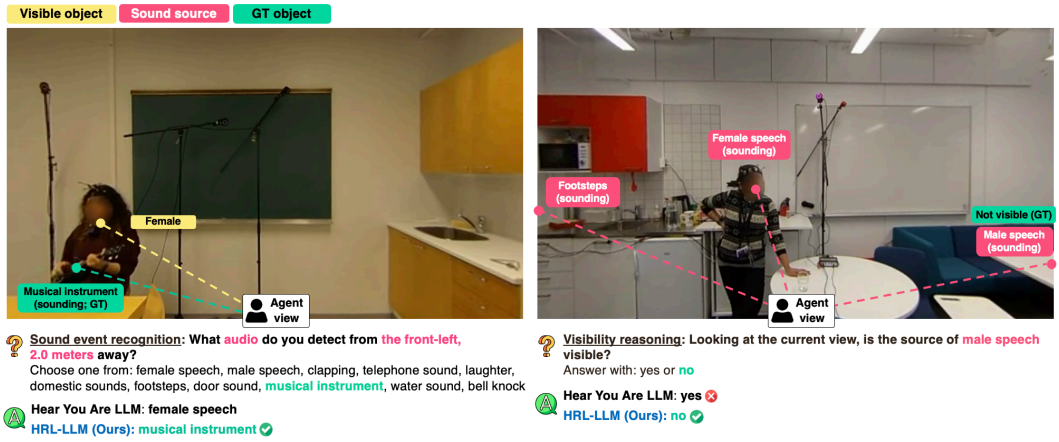


Figure S2: **Qualitative examples from the STARSS23 real-world evaluation.** We show two representative task types, sound-event recognition and visibility prediction, together with the midpoint video frame, annotated sound-source locations, ground-truth answers, and model predictions.

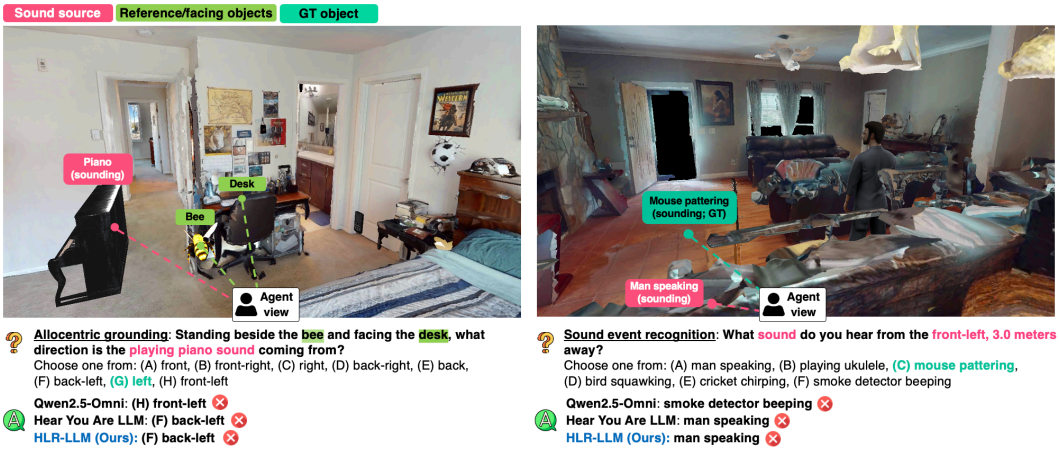


Figure S3: **Failure cases.** We show representative failure cases where models make fine-grained spatial errors, such as confusing back-left with left, or struggle with multi-source scenes in which sound sources are spatially close.

selects a visible saxophone. HLR-LLM's CoT instead localizes the sounding saxophone to the front-left, compares this position with the visible candidates, and correctly concludes that none of them is spatially aligned with the sound. These examples illustrate how CoT supervision enables HLR-LLM to explicitly connect sound-source localization with downstream scene-level relational reasoning.

Figure S2 shows representative STARSS23 examples for sound-event recognition and visibility prediction. HLR-LLM correctly recognizes the target event in multi-source audio and determines source visibility by combining spatial audio with visual evidence. These examples further demonstrate that the learned spatial audio-visual reasoning transfers to real-world recordings.

Failure cases. We also provide failure cases in Fig. S3. Most errors are caused by fine-grained spatial ambiguity. For example, the model may confuse adjacent direction labels, such as `back-right` vs. `right` or `front-left` vs. `left`, especially when the source lies near a bin boundary. Errors also occur when several candidate objects are spatially clustered, making it difficult to associate the localized sound with the correct object. In multi-source scenes, failures often arise when two sources are close to each other and the non-target source is louder. In such cases, overlapping acoustic cues can dominate the prediction, causing the model to localize or reason about the wrong source.



Figure S4: **Counterfactual analysis with conflicting audio-visual cues.** The image is kept fixed with the dog visible at its original location, while the binaural source position is changed. HLR-LLM follows the counterfactual spatial audio and correspondingly updates its localization and downstream loudness reasoning.

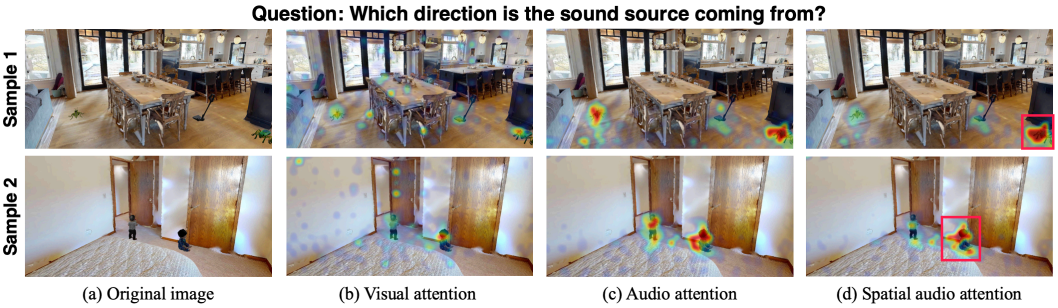


Figure S5: **Modality-specific attention visualization.** From left to right: input image, visual-to-visual, audio-to-visual, and spatial-audio-to-visual attention from the 18th transformer layer. Spatial-audio tokens attend more selectively to the ground-truth sound-source regions (red boxes).

Counterfactual analysis. We further conduct a counterfactual analysis to examine whether HLR-LLM relies on spatial audio cues when visual and auditory evidence conflict. We keep the rendered image fixed, where the dog remains visible at its original location, and replace the binaural audio with dog barking rendered from different spatial positions. As shown in Fig. S4, HLR-LLM follows the counterfactual spatial audio in both cases. In the left case, it localizes the barking to the front-left at 0.9 m and infers that moving to the radiator increases the source distance, resulting in a quieter prediction. In the right case, it instead localizes the barking to the front at 4.1 m and predicts that moving to the step brings the agent closer to the source, resulting in a louder prediction, even though the visible dog remains at its original front-left location. These results provide evidence that HLR-LLM uses spatial audio for localization and subsequent spatial reasoning rather than relying solely on the visible source position.

Modality-specific attention analysis. To examine how modality-specific representations interact with visual content, we extract self-attention weights from the 18th transformer layer of HLR-LLM. For each modality, we use its corresponding tokens as queries and the visual tokens as keys, averaging the attention weights across all query tokens and attention heads. We visualize three modality-to-visual attention maps: visual-to-visual using Z_{vis}, audio-to-visual using Z_{aud}, and spatial-audio-to-visual using Z_{spa}. The visual-to-visual map reflects regions emphasized by the visual representation, while the latter two indicate which image regions are attended to by the conventional audio and spatial-audio representations, respectively. As shown in Fig. S5, spatial-audio tokens attend more selectively to the ground-truth sound-source regions than conventional audio tokens, providing qualitative evidence that the spatial pathway captures localization-relevant visual associations.

C Details on HLR-AVSceneQA

We provide implementation-level details on HLR-AVSceneQA, including the dataset construction pipeline, task characteristics, question and answer formats, quality-control procedures, and additional details that were not introduced in the main paper.

C.1 Scene initialization, agent placement, and scene generation

Simulator setup. We generate each sample in Habitat-Sim with acoustic rendering using the SoundSpaces simulator [16]. The agent is initialized with synchronized RGB, semantic, depth, panoramic RGB, and binaural audio sensors. The RGB sensor is a pinhole camera used as the main visual input. The semantic and depth sensors are used for visibility checking and geometric supervision, while the panoramic RGB sensor is used for visualization and debugging. The binaural audio sensor is co-located with the visual agent and uses a two-channel receiver at a 32 kHz sampling rate with material-aware indirect propagation. We set the sensor height to 150 cm and the camera field of view to 90°.

We select valid rooms from HM3D scenes [39] by checking that the semantic region contains a floor, a valid room-level region of interest, and sufficient navigable positions. In the standard setting, we place the agent near a room corner and orient it toward the room center, which increases the chance that inserted objects and room semantics are visible. Since this placement may bias sound sources toward the front view, we also sample agent poses closer to the room center. This makes it possible to construct scenes in which one reference object remains visible while another object or sound source falls outside the camera view.

Coordinate system and azimuth definition. We use Habitat coordinates for camera, agent, and source placement, and convert positions into metric distances when constructing QA labels. For azimuth, we annotate angles around the horizontal plane, ranging from 0 to 360°. In our setup, 0° indicates the front direction, 90° indicates the right direction, 180° indicates the back direction, and 270° indicates the left direction relative to the agent. We do not consider elevation in this manuscript, as human spatial-audio localization is generally known to be more accurate and robust in azimuth than in elevation [57].

For both direction and distance, we discretize continuous spatial values into bins to make training and evaluation more stable. For azimuth, we use 8 bins, each covering 45°. Angles increase clockwise from straight ahead: `front` corresponds to 337.5°–22.5°, `right` to 67.5°–112.5°, `back` to 157.5°–202.5°, and `left` to 247.5°–292.5°. For distance, we use 10 bins, with a 1-meter interval for 0–1 m, 0.5-meter intervals from 1 to 5 m, and a final bin for distances beyond 5 m. The full bin definitions are shown in Table S4. This discretization reduces ambiguity from small geometric differences and facilitates stable training and evaluation.

3D object meshes and scene insertion. We insert 3D object meshes into HM3D rooms to synthesize controlled layouts containing both sounding and non-sounding objects. The object pool is built from curated meshes, including Objaverse-XL assets [43], as well as generated meshes for sound categories whose object coverage is insufficient. For generated sounding-object meshes, we first use Stable Diffusion 2 [44] for text-to-image synthesis, and then use TRELLIS [45] for image-to-3D mesh generation. The size and brightness of each object category are manually determined to ensure plausible rendering in indoor scenes. Examples are shown in Fig. S6.

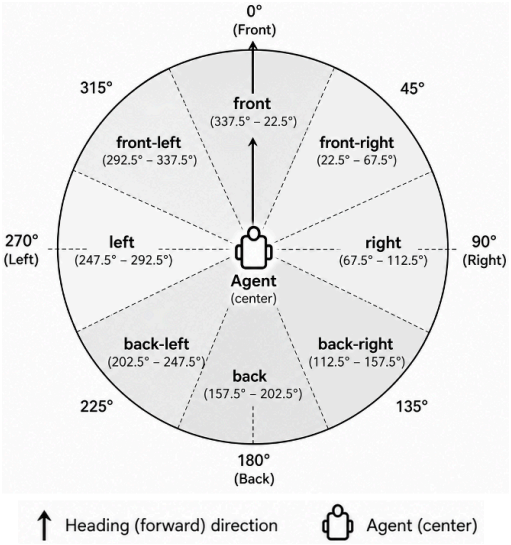
When inserting objects into a room, we reject placements that collide with room geometry, overlap existing objects above a horizontal IoU threshold, or violate minimum agent-to-object or object-to-object distance constraints. These filtering steps help ensure that inserted objects are physically plausible and visually distinguishable.

Binaural audio generation. For each active source, we render binaural audio using SoundSpaces 2.0. Let $x_j[t]$ be the dry monaural waveform assigned to source j, located at $\mathbf{s}_j$, and let the agent receiver be at position $\mathbf{r}$ with heading θ. SoundSpaces produces a channel-wise binaural room impulse response:

$$h_{m,j}[\tau] = \mathcal{R}_m(\tau; \mathbf{s}_j, \mathbf{r}, \theta, \mathcal{E}), \qquad m \in \{L, R\}, \tag{1}$$

Table S4: **Azimuth and distance bins used in HLR-AVSceneQA.** Azimuth angles increase clockwise from the agent's forward direction.

Type	Bin	Range
Azimuth	front	337.5°–22.5°
	front-right	22.5°–67.5°
	right	67.5°–112.5°
	back-right	112.5°–157.5°
	back	157.5°–202.5°
	back-left	202.5°–247.5°
	left	247.5°–292.5°
	front-left	292.5°–337.5°
Distance	A	0–1.0 m
	B	1.0–1.5 m
	C	1.5–2.0 m
	D	2.0–2.5 m
	E	2.5–3.0 m
	F	3.0–3.5 m
	G	3.5–4.0 m
	H	4.0–4.5 m
	I	4.5–5.0 m
	J	> 5.0 m



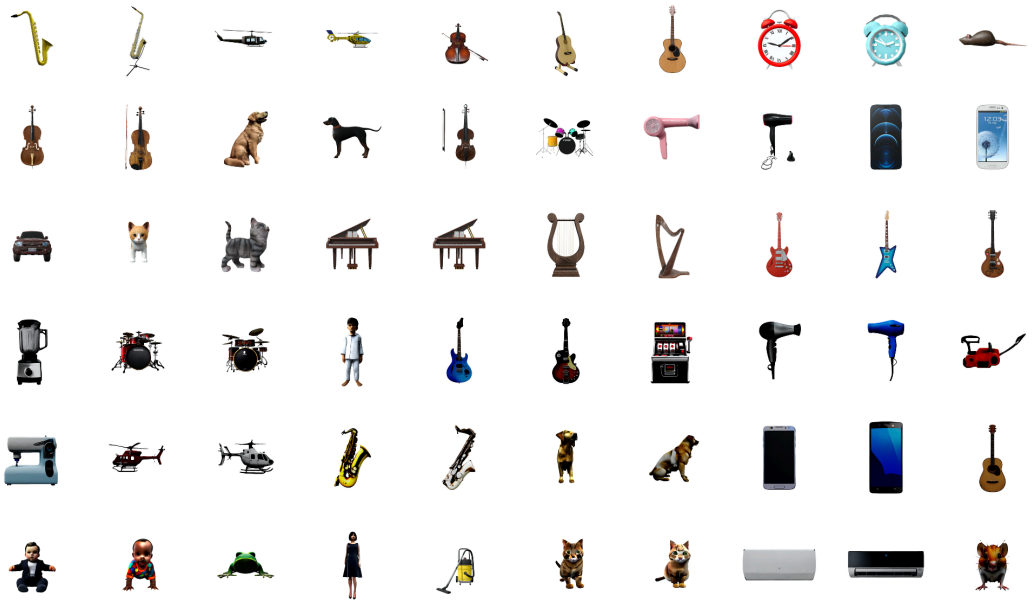


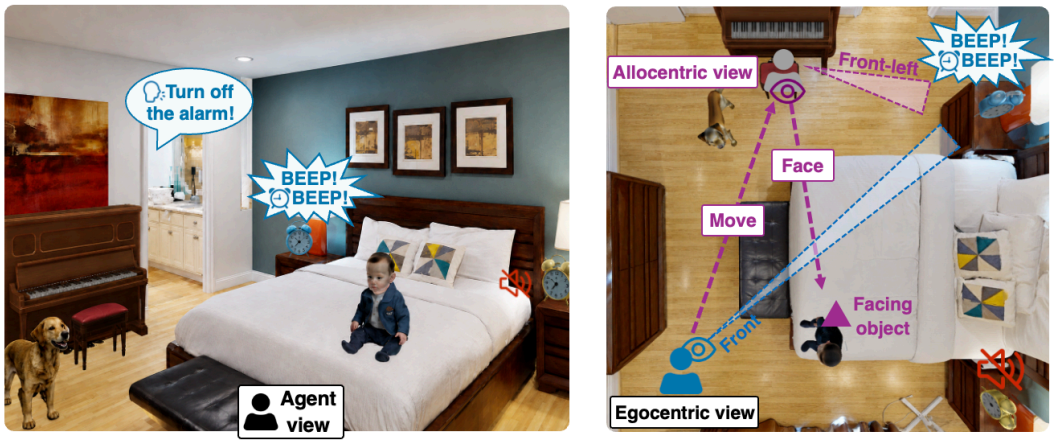
Figure S6: **Examples of 3D object meshes.** The first three rows show meshes retrieved from Objaverse-XL, while the last three rows show meshes generated by generative models.

where $\mathcal{E}$ denotes the scene geometry and material configuration. The binaural signal is then obtained by convolving the normalized dry waveform $\tilde{x}_j$ with the corresponding room impulse response:

$$y_{m,j}[t] = (h_{m,j} * \tilde{x}_j)[t] = \sum_{\tau=0}^{T_h-1} h_{m,j}[\tau]\tilde{x}_j[t-\tau]. \tag{2}$$

For multiple active sources, the rendered signals are summed:

$$y_m[t] = \sum_{j=1}^{N} y_{m,j}[t], \qquad m \in \{L, R\}. \tag{3}$$



(a) Normal filed-of-view (model input) (b) Bird's-eye-view

Figure S7: **Example illustration.** This example data sample is used to facilitate understanding of our task definitions and details.

The final stereo waveform $\mathbf{y}[t] = [y_L[t], y_R[t]]$ is saved at a 32 kHz sampling rate.

Spatial reasoning scenarios. We intentionally include several scenarios that require audio-visual spatial reasoning beyond semantic matching. *Occluded sound sources* place the active source behind a visible object or behind a wall, so the source may project into the camera view while having no visible semantic pixels. *Ambiguous sound sources among same-category instances* include multiple objects of the same category, requiring the model to determine which instance is spatially aligned with the sound. *Semantically mismatched but spatially aligned cases* attach a sound event to an object that does not semantically match the event, preventing models from relying only on category correspondence. *Adjacent-room sound sources* move the source into a neighboring room connected through the scene geometry, creating cases where the source is not visible but can still be heard from a distant location.

C.2 Task definitions and details

Here we provide additional details on the task definitions, their descriptions, and the intention of each task, using the example in Fig. S7. In this scenario, Fig. S7 (a) shows the agent's view, which is the model input with a normal field of view. Several instances are inserted, including a dog, a piano, a baby, and two alarm clocks. There are two sound events: one is a man speaking, saying "Turn off the alarm," coming from the agent's front-left direction and originating from an adjacent room. The other is an alarm ringing from the blue alarm clock, coming from the front direction, 3.4 meters away. Based on this example, we describe the goal and intended evaluation behavior of each task.

Furthermore, to improve linguistic diversity and reduce overfitting to fixed question formats, we generate paraphrased templates for each task. We create approximately 15 paraphrases per task and show three representative prompts below. The full set of templates will be included in the dataset release.

Spatial hearing: sound event recognition. This task evaluates semantic audio perception. The model is asked to identify the audible sound event from a set of candidate sound categories. A total of 10 candidate sound categories are provided as options. The target sound category is always included, while distractors are sampled from semantically related categories, sound categories associated with scene objects, and random categories. When multiple sources are present, the prompt may additionally specify the source direction and distance to indicate which source should be recognized.

- ***Q:*** *What sound do you hear from the front, 3.4 meters away?*
- ***A:*** *alarm clock ringing.*
- ***Paraphrase 1:*** *Which sound is coming from the front, 3.4 meters away?*
- ***Paraphrase 2:*** *Identify the sound event from the front direction at 3.4 meters.*

Spatial hearing: far-field automatic speech recognition. This task evaluates far-field automatic speech recognition (ASR) from spatialized audio. Since speech may contain important information that the agent should understand, the model is required to recognize speech as part of spatial hearing. Because speech can originate from both nearby and distant locations, far-field ASR is more challenging than clean or near-field ASR. We ask the model to return only the exact quoted transcription, without any additional explanation.

- ***Q:*** *Transcribe what the person said. Output only the transcription inside quotes.*
- ***A:*** *"Turn off the alarm."*
- ***Paraphrase 1:*** *Write the exact phrase spoken. Return only the quoted transcription.*
- ***Paraphrase 2:*** *Provide the exact spoken words inside quotes.*

Spatial grounding: egocentric direction prediction. This task evaluates egocentric sound-source direction estimation. The answer is computed from the vector pointing from the agent to the sound source, with its azimuth measured relative to the agent's current forward direction. We ask the model to answer using a discretized 8-bin azimuth label as shown in Table S4. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *From this camera perspective, which direction bin does the source of alarm clock ringing fall into?*
- ***A:*** *front.*
- ***Paraphrase 1:*** *Based on the current view, which direction is the alarm clock ringing coming from?*
- ***Paraphrase 2:*** *From this viewpoint, identify the direction bin for the source of alarm clock ringing.*

Spatial grounding: egocentric distance prediction. This task evaluates egocentric source-distance estimation. The answer is computed as the horizontal distance between the agent and the active sound source, and is then converted into a discrete distance bin, as shown in Table S4. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *From this camera perspective, what is the distance to the source of alarm clock ringing?*
- ***A:*** *F.*
- ***Paraphrase 1:*** *Based on the current view, how far away is the source of alarm clock ringing?*
- ***Paraphrase 2:*** *From this viewpoint, select the distance bin for the source of alarm clock ringing.*

Spatial grounding: allocentric direction prediction. This task evaluates direction reasoning from a different viewpoint. The prompt asks the model to imagine standing at a reference object and facing another object, as shown in Fig. S7 (b). The source direction is then recomputed in this new coordinate frame. It is intended to evaluate whether the model can reason from a different perspective and understand the 3D spatial structure of the current scene. The task is generated only when valid visible reference and facing objects are available. We avoid prompts where the reference and facing objects have identical final category names, which would make the instruction ambiguous. As in the egocentric case, the answer is a discretized 8-bin azimuth label. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *If you stand by the piano facing the baby, which direction is the source of alarm clock ringing coming from?*
- ***A:*** *front-left.*
- ***Paraphrase 1:*** *Standing next to the piano and facing the baby, which direction does the alarm clock ringing come from?*
- ***Paraphrase 2:*** *From a position at the piano facing the baby, identify the direction bin of the alarm clock ringing.*

Spatial grounding: allocentric distance prediction. This task evaluates distance reasoning from a different viewpoint. Similar to allocentric direction prediction, the prompt asks the model to imagine standing at a reference object and facing another object. The source distance is then recomputed from the reference object's position. This task evaluates whether the model can reason about source distance from a different location in the scene. We avoid prompts where the reference and facing objects have identical final category names, which would make the instruction ambiguous. As in the

egocentric case, the answer is a discretized 10-bin distance label. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *If you stand by the piano facing the baby, what is the distance to the source of alarm clock ringing?*
- ***A:*** *B.*
- ***Paraphrase 1:*** *Standing next to the piano and facing the baby, how far away is the alarm clock ringing?*
- ***Paraphrase 2:*** *From a position at the piano facing the baby, select the distance bin for the source of alarm clock ringing.*

Spatial relational reasoning: visibility. This task evaluates whether the model can determine if the sound source is visible from the current viewpoint. This task becomes difficult when the scene contains the same or similar categories, but the actual sound comes from outside the current viewpoint. For example, one may incorrectly assume that the man speaking sound in Fig. S7 comes from the baby, since the baby is also human-related. However, with sufficient spatial audio perception, the model should distinguish that the baby is not making the sound and that the sound source is not visible. The ground truth is computed from semantic rendering: a source is considered visible if the number of semantic pixels corresponding to the source object exceeds a configured threshold. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *From the current camera perspective, is the source of man speaking visible?*
- ***A:*** *no.*
- ***Paraphrase 1:*** *Based on the current view, can you see the source of man speaking?*
- ***Paraphrase 2:*** *From this viewpoint, is the man speaking source in view?*

Spatial relational reasoning: correspondence. This task evaluates audio-visual spatial correspondence. The model is asked to identify which visible object is spatially aligned with the sound source. The correct answer may be semantically different from the sound category, since the task focuses on spatial alignment rather than category matching. This task becomes challenging when multiple visible objects exist, as in Fig. S7. Without spatial audio understanding, the model may confuse whether the yellow alarm clock is ringing. If no visible object corresponds to the sound source, the answer is `None`. When multiple visible inserted objects share the same category, we disambiguate them using orientation labels, such as *"alarm clock on the front-right."* If multiple sources are present, we specify the queried sound source.

- ***Q:*** *From the current camera perspective, which object is spatially aligned with the sound of alarm clock ringing?*
- ***A:*** *alarm clock in front.*
- ***Paraphrase 1:*** *Which visible object corresponds to where the alarm clock ringing comes from?*
- ***Paraphrase 2:*** *Which object best matches the spatial origin of the alarm clock ringing?*

Spatial relational reasoning: hidden-source localization. This task evaluates reasoning about non-visible sound sources. It is asked when the source is not reliably visible in the current image. The answer includes five options: hidden behind the wall, hidden behind a specific object, in an adjacent room, out of the field of view in front, and out of the field of view behind the agent. This task goes one step beyond visibility prediction. The model must reason where the source may be if it is not visible. It should decide whether the sound comes from a distant place, such as another room, or whether it should be in the field of view but is occluded by an object, which requires accurate visual perception. The ground-truth label is determined using visibility, projection, adjacent-room metadata, and occluder information. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *You hear a man speaking. Where is the source of this sound?*
- ***A:*** *in an adjacent room.*
- ***Paraphrase 1:*** *Given the sound of man speaking, where is its source?*
- ***Paraphrase 2:*** *A man speaking is audible. Where could the source be situated?*

Spatial relational reasoning: loudness change. This task evaluates relative acoustic reasoning under hypothetical movement. The prompt asks whether the sound would become louder or quieter if the agent moved to a specified object. Here, louder or quieter is defined as a controlled relative movement judgment based on whether the source becomes closer or farther, rather than an estimate of absolute perceived loudness. This assesses whether the model can decide whether a movement would bring the agent closer to or farther from the sound source. It also goes one step beyond allocentric distance prediction. The model must estimate the source distance from the changed location and compare it with the current source distance. The label is computed by comparing the current agent-to-source distance with the candidate-object-to-source distance. If the candidate object is closer to the source than the current agent position, the answer is `louder`; otherwise, it is `quieter`. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *If you move to the piano, will the sound of man speaking become louder or quieter?*
- ***A:*** *louder.*
- ***Paraphrase 1:*** *When at the piano, will the sound of man speaking be louder or quieter?*
- ***Paraphrase 2:*** *Upon moving to the piano, does the man speaking become louder or quieter?*

Spatial relational reasoning: proximity reasoning. This task goes one step beyond loudness change and evaluates source-to-object spatial reasoning. The model is given a set of candidate objects and must select the one closest to the sound source. This requires comparing the source distance to each candidate object, similar to evaluating loudness change for each object, and then ranking the candidates to select the closest one. The ground truth is computed using the Euclidean distance from the source to each candidate object. As with other distance-based relational tasks, we exclude adjacent-room sources to avoid cases where direct Euclidean geometry does not reflect the effective acoustic propagation path. Tied cases are skipped to avoid ambiguous supervision. If multiple sources are present, we specify the queried sound source.

- ***Q:*** *Which object is closest to the source of alarm clock ringing?*
- ***A:*** *bed.*
- ***Paraphrase 1:*** *To which object is the alarm clock ringing nearest?*
- ***Paraphrase 2:*** *Which object is at the shortest distance from the source of alarm clock ringing?*

Spatial reasoning: scene captioning. This task evaluates compositional spatial audio-visual scene description. The model is asked to describe each sound source using complete sentences and include the sound event, direction, visibility, and hidden-source location when applicable. The prompt explicitly instructs the model not to describe unrelated scene content. When multiple sources are present, the model must provide a separate description for each source.

- ***Q:*** *Describe only the sound source or sound sources.*
- ***A:*** *The sound event is alarm clock ringing. It is coming from the front direction. The sound source is visible in the scene.*
- ***Paraphrase 1:*** *Summarize only the sound source information.*
- ***Paraphrase 2:*** *For each source, state the sound event, direction, visibility, and hidden location if applicable.*

C.3 Chain-of-thought supervision

We additionally construct chain-of-thought-style supervision to encourage the model to connect audio perception, visual grounding, and spatial reasoning before producing the final answer. Both the final answers and CoT rationales are deterministically derived from structured simulator metadata through task-specific templates rather than free-form generation. GPT-4 [46] is used only to paraphrase question wording for linguistic diversity. This is especially important for tasks that cannot be solved from a single cue alone, such as allocentric reasoning, hidden-source localization, loudness change, and proximity reasoning. In these tasks, the model must first identify the sound event, estimate its direction or distance, relate it to visible or non-visible objects, and then convert this intermediate spatial evidence into the requested answer. We keep the reasoning brief and require the model to end with a normalized final-answer line.

Spatial grounding: allocentric direction prediction.

- ***Q:*** *If you are located at the cabinet facing the door frame, determine which direction the source of woman singing is coming from. Reason briefly, then end with exactly one line: "Final answer: <8bin_direction_label>".*
- ***A:*** *The audio event woman singing is heard from the front direction at 1.2 m.*
  The cabinet is positioned front-left at a distance of 2.6 m from the current location.
  After moving to cabinet while maintaining the same orientation, the sound source appears at right with a distance of 2.0 m.
  From this position, door frame is found in the front-right direction at 3.5 m.
  Rotate right to face door frame, establishing the new forward axis.
  In this new frame, the sound source lies front-right.
  Final answer: front-right

Spatial grounding: allocentric distance prediction.

- ***Q:*** *Standing next to the cabinet and facing the typewriter, what is the distance to the sound source? Reason briefly, then end with exactly one line: "Final answer: <distance_bin>".*
- ***A:*** *The sound event is playing acoustic guitar, initially at 1.0 m in the right direction.*
  The reference object cabinet is positioned front-right at a distance of 1.7 m.
  After moving to cabinet, the sound source is approximately 1.2 m away.
  This measurement places it within the 1-1.5 m interval.
  Final answer: B

Spatial relational reasoning: visibility.

- ***Q:*** *From this perspective, is the sound source in sight? Answer with [yes, no]. Reason briefly, then end with exactly one line: "Final answer: <yes_or_no>".*
- ***A:*** *The audio event is playing clarinet.*
  It is coming from the back direction and is 1.0 meters away.
  In that direction, the clarinet is not present, so the sound source cannot be seen. Thus, the source is not visible.
  Final answer: no

Spatial relational reasoning: correspondence.

- ***Q:*** *From the camera's perspective, which object is at the location of the sound of mouse squeaking? Pick one option; select None if no object in view is producing sound. Reason briefly, then end with exactly one line: "Final answer: <object_or_None>".*
- ***A:*** *The audio event is mouse squeaking.*
  It is coming from the left direction and is 1.6 meters away.
  Among the options, no visible source in the left direction matches the sound.
  Final answer: None

Spatial relational reasoning: hidden-source localization.

- ***Q:*** *The sound of woman speaking is present. Where is it originating from? Choose one from: (A)–(E). Reason briefly, then end with exactly one line: "Final answer: <hidden_source_option>".*
- ***A:*** *The audio event is woman speaking.*
  Although it comes from the right direction, the source lies outside the current field of view in that direction.
  Therefore, the source is out of the field of view in front.
  Final answer: out of the field of view in front

Spatial relational reasoning: loudness change.

- ***Q:*** *By moving to pillow, does the sound become louder or quieter? Choose one from: [louder, quieter]. Reason briefly, then end with exactly one line: "Final answer: <louder_or_quieter>".*
- ***A:*** *The audio event is hair dryer drying.*
  It originates from the front direction and is 4.0 meters away.

Once you move to pillow, the source is 2.8 m away, making you closer to it than before.
Therefore, the sound becomes louder.
Final answer: louder

Spatial relational reasoning: proximity reasoning.

- ***Q:*** *Which object is closest to the sound source? Select one from: [bed, vacuum cleaner, chair, air conditioning]. Reason briefly, then end with exactly one line: "Final answer: <object_label>".*
- ***A:*** *The audio event is man speaking.*
  It originates from the back-right direction and is 0.7 meters away.
  Agent-centered candidate layout: bed is front-right at 1.6 m from the agent, vacuum cleaner is front-right at 3.0 m from the agent, chair is front at 3.4 m from the agent, and air conditioning is front at 2.0 m from the agent.
  Comparing source-to-candidate distances gives: bed is 1.4 m away from the sound source, vacuum cleaner is 3.1 m away from the sound source, chair is 3.7 m away from the sound source, and air conditioning is 2.4 m away from the sound source.
  Therefore, the nearest object is bed.
  Final answer: bed

C.4 Quality control

Automatic quality control. We apply several filtering and retry rules to avoid noisy supervision. First, rooms are skipped if they do not have a valid floor, contain too few insertion slots, or lack enough valid navigable points. Second, object placements are rejected if they collide with room geometry, overlap existing objects above the IoU threshold, violate minimum distance constraints, or fail semantic visibility checks. Third, generated 3D meshes and assets from Objaverse-XL are manually inspected by human annotators. We retain only sufficiently high-quality objects, filtering out low-quality meshes and implausible categories. We also find that, in some cases, sounding objects, reference objects, facing objects, or candidate objects used in task options can be too small to be clearly visible. We therefore apply a pixel-count threshold and remove QA pairs that ask about objects that are hardly visible. Finally, proximity questions with tied nearest objects are skipped.

Human verification. We additionally conduct a human verification study on 150 samples, covering 1,144 QA pairs, with three human annotators. Annotators verify whether the target sound is audible, whether its approximate spatial location can be identified, whether candidate and reference objects mentioned in the questions are observable, and whether the annotated source visibility and hidden-source location are consistent with the audio-visual scene. As shown in Table S5, 1,133 of 1,144 QA pairs pass human verification, corresponding to an overall pass rate of 99.0%. All spatial-hearing and egocentric-grounding questions achieve a 100% pass rate, while the remaining task categories achieve pass rates between 96.8% and 100%. These results provide human validation that our metadata- and template-based construction pipeline produces reliable spatial audio-visual QA annotations across the benchmark tasks.

C.5 Dataset statistics

Table S6 reports the number of generated QA pairs for the training and test splits, including the number of CoT supervision examples for each task and the breakdown of QA pairs constructed from single-source and multi-source scenes. Table S7 further characterizes the benchmark composition by reporting the distributions of hidden-source and visibility cases, together with train–test overlap statistics. The hidden-source test set covers diverse non-visible configurations, including off-screen, adjacent-room, wall-occluded, and object-occluded sources. The visibility task contains 53.6% visible and 46.4% non-visible cases in training, and 30.7% visible and 69.3% non-visible cases in the test set. Importantly, the training and test splits use disjoint environments and rooms, as well as disjoint object mesh instances, while all test sound and object categories are represented in the training split. Thus, the benchmark evaluates generalization to unseen environments, scene layouts, and object instances while maintaining consistent semantic task coverage.

Fig. S8 and Fig. S9 further visualize the distributions of distance and direction bins for egocentric and allocentric settings in both the training and test sets. Finally, Fig. S10 and Fig. S11 show the top 30 inserted object categories and active sound-source categories in the training and test sets.

Table S5: **Human verification of HLR-AVSceneQA annotations.** Three human annotators verify 1,144 QA pairs from 150 sampled scenes. We report the number of passed and failed QA pairs and the corresponding pass rate for each task.

Task family	Task	Total	Passed	Failed	Pass rate
Spatial hearing	Sound event recognition	112	112	0	100.0%
	Far-field ASR	29	29	0	100.0%
Egocentric grounding	Direction	98	98	0	100.0%
	Distance	98	98	0	100.0%
Allocentric grounding	Direction	148	146	2	98.6%
	Distance	148	148	0	100.0%
Relational reasoning	Visibility	107	105	2	98.1%
	Hidden source	66	65	1	98.5%
	Loudness change	111	109	2	98.2%
	Proximity	126	122	4	96.8%
	Correspondence	101	101	0	100.0%
Overall	–	**1,144**	**1,133**	**11**	**99.0%**

Table S6: **Dataset statistics by task split.** Each entry reports the number of QA pairs. For Train and Test, Single and Multi are computed from the source-count annotations. For Train-CoT, Single and Multi report the number of CoT QA pairs in each source-count group after task-specific filtering.

Task family	Task	Train			Train-CoT			Test		
		Single	Multi	Total	Single	Multi	Total	Single	Multi	Total
Spatial hearing	Sound event	14,794	7,054	21,848	–	–	–	1,498	772	2,270
	Far-field ASR	2,708	867	3,575	–	–	–	380	128	508
Spatial grounding	Egocentric direction	13,770	6,500	20,270	–	–	–	1,442	632	2,074
	Egocentric distance	13,770	6,500	20,270	–	–	–	1,442	632	2,074
	Allocentric direction	8,410	3,909	12,319	8,086	3,735	11,821	1,454	461	1,915
	Allocentric distance	8,410	3,909	12,319	8,086	3,735	11,821	1,454	461	1,915
Spatial relational reasoning	Visibility	11,277	5,350	16,627	11,277	5,350	16,627	1,644	410	2,054
	Correspondence	14,794	7,054	21,848	14,794	7,054	21,848	2,066	268	2,334
	Hidden-source	5,173	2,693	7,866	5,173	1,321	6,494	1,094	382	1,476
	Loudness change	12,542	5,860	18,402	12,542	5,860	18,402	1,164	291	1,455
	Proximity	5,166	2,391	7,557	5,166	2,340	7,506	720	248	968
	Scene captioning	10,968	1,192	12,160	–	–	–	2,581	184	2,765
Total	–	**121,782**	**53,279**	**175,061**	**65,124**	**29,395**	**94,519**	**16,939**	**4,869**	**21,808**

D Broader impact

HLR-AVSceneQA and HLR-LLM aim to support spatially aware audio-visual understanding for embodied AI, assistive agents, robotics, and accessibility applications. The ability to recognize what sound is present, localize where it originates, and reason about how it relates to the current scene is especially important for agents that navigate indoor environments and assist in daily situations. For example, such systems may help localize alarms, identify speakers, or reason about events outside the camera view. Our benchmark is simulator-based and does not contain private real-world recordings. However, future deployment of spatial audio-visual systems should follow strict privacy, consent, and data-minimization practices, especially when operating in real homes or shared indoor spaces.

E LLM usage

We use LLMs in three ways. First, the proposed HLR-LLM is built on a large language model as the reasoning module that integrates unified inputs, including visual, monaural audio, and binaural spatial audio tokens. Second, we use an LLM to paraphrase question templates to improve linguistic diversity and reduce the risk of models learning fixed template patterns. Finally, we use an LLM to assist with editing and grammar checking, table refinement (Table S4), figure generation and refinement (Fig. S8, S9, S10, S11), and development of the data-viewer web page.

Table S7: **Additional dataset statistics and train–test overlap.** We report label distributions for hidden-source and visibility tasks, followed by statistics on the train and test splits. Percentages are computed within each corresponding task split. Train and test environments, rooms, and object mesh instances are disjoint, while test semantic categories are covered by the training split.

Statistic	Train	Test	Train–Test Overlap
Hidden-source distribution			
Off-screen (front)	2,681 (34.1%)	408 (27.6%)	–
Off-screen (back)	2,752 (35.0%)	518 (35.1%)	–
Adjacent room	959 (12.2%)	237 (16.1%)	–
Hidden behind wall	418 (5.3%)	123 (8.3%)	–
Hidden behind object	1,056 (13.4%)	190 (12.9%)	–
Total	**7,866**	**1,476**	–
Visibility distribution			
Visible	8,912 (53.6%)	631 (30.7%)	–
Not visible	7,715 (46.4%)	1,423 (69.3%)	–
Total	**16,627**	**2,054**	–
Split statistics and overlap			
Environments	145	36	0
Rooms	1,028	241	0
Sound categories	108	79	79
Object categories	37	37	37
Object mesh instances	3,018	1,299	0

F Licenses and terms of existing assets

We use existing datasets, simulators, models, and 3D assets only under their stated licenses and terms of use.

- **SoundSpaces 2.0**: used for spatial acoustic rendering. SoundSpaces is released under the Creative Commons Attribution 4.0 International license. Matterport3D-based task data and trained models follow the Matterport3D terms and CC BY-NC-SA 3.0 US license.
- **HM3DSem / HM3D**: used as the source of indoor 3D environments and semantic scene annotations. We use the dataset under its academic research terms and respect the restrictions of the underlying Matterport3D assets.
- **VGGSound**: used for non-speech sound events. We use the official split and retain only object-grounded categories suitable for indoor spatial reasoning. Copyright remains with the original video owners.
- **LibriSpeech**: used for speech audio and far-field ASR supervision. LibriSpeech is based on public-domain audiobooks and is distributed under a Creative Commons license.
- **Objaverse-XL**: used as a source of 3D object meshes. The dataset as a whole is licensed under ODC-By v1.0, while individual objects have their own licenses. We manually check and retain only assets compatible with our release setting.
- **Stable Diffusion 2**: used to generate canonical object images. The code is released under the MIT license, and model weights are released under the CreativeML Open RAIL++-M license.
- **TRELLIS**: used for image-to-3D mesh generation. TRELLIS models and code are released under the MIT license, with some submodules under their own licenses.
- **Comparison models**: MiniCPM-o-v2.6, VideoLLaMA2, Video-SALMONN 2, Qwen2.5-Omni, Qwen3-Omni, SAVVY, Hear You Are LLM, GPT-Realtime, and Gemini-2.5-Flash-Lite are used only for evaluation or baseline comparison under their respective model licenses, APIs, and terms of use.

All generated 3D assets are synthetic and are not intended to reproduce real people, private environments, or sensitive locations. We will release only assets and annotations that are compatible with the applicable licenses and redistribution terms.

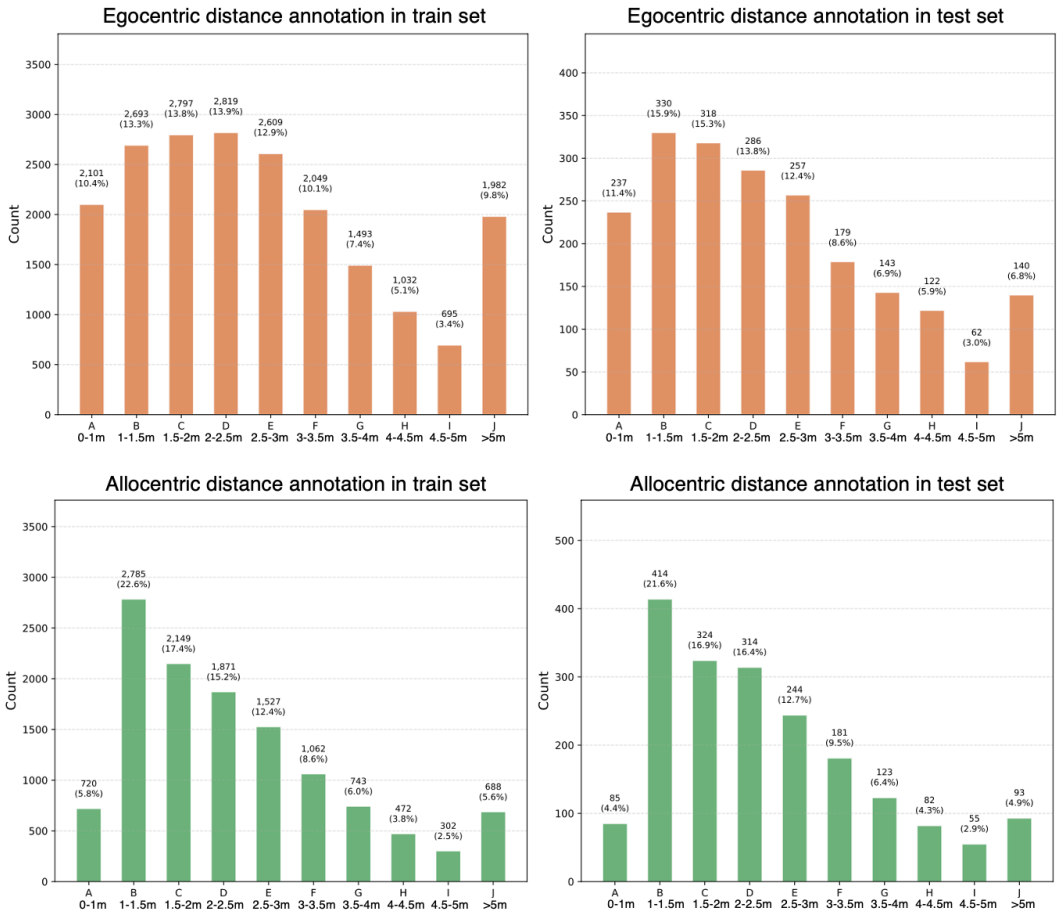


Figure S8: **Distance annotation statistics.** We show histograms of distance-bin annotations for egocentric and allocentric perspectives in both the training and test sets.

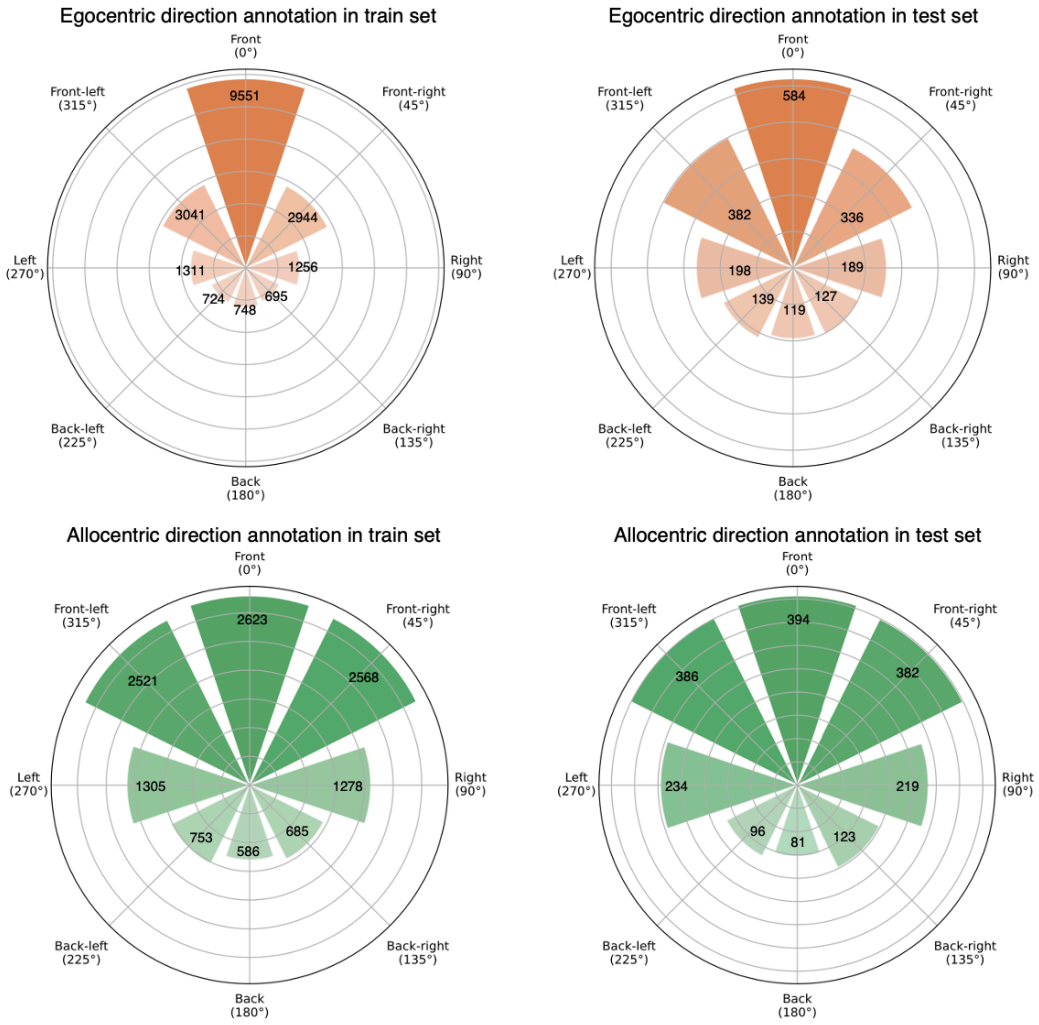


Figure S9: **Direction annotation statistics.** We show log-scaled distributions of direction-bin annotations for egocentric and allocentric perspectives in the training and test sets.

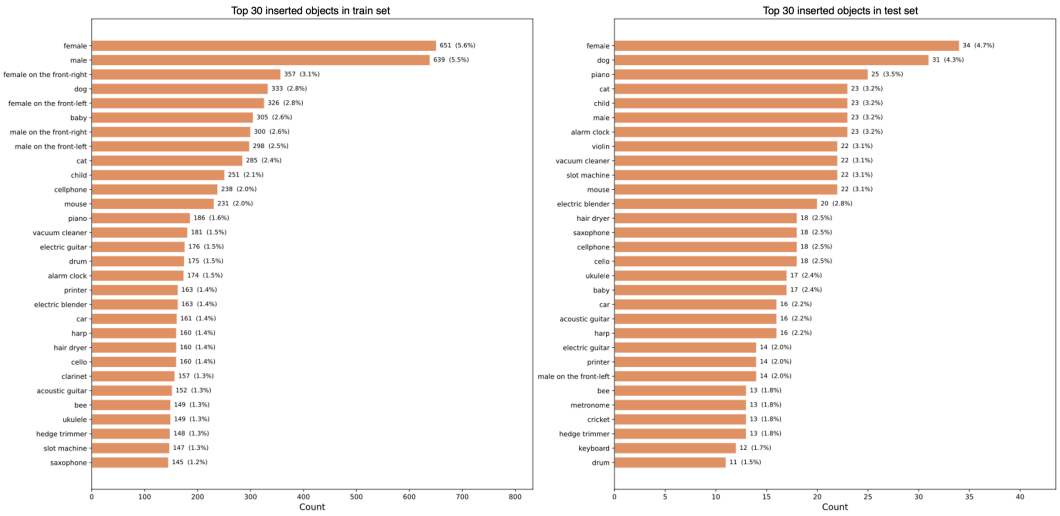


Figure S10: **Top 30 inserted object categories.** We show the 30 most frequent object categories inserted into training and test scenes.

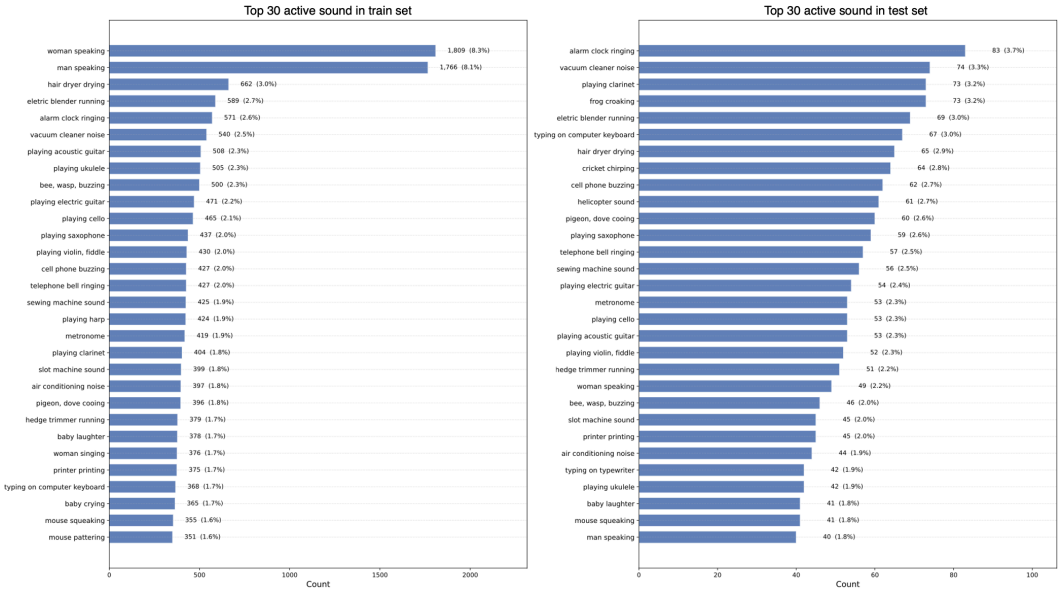


Figure S11: **Top 30 active sound categories.** We show the 30 most frequent sound categories used as active sound sources in training and test scenes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Section 1, "Introduction", clearly state the paper's main contributions: HLR-AVSceneQA, a hear-localize-reason benchmark for spatial audio-visual scene understanding, and HLR-LLM, a spatially aware AV-LLM with a binaural spatial audio branch and curriculum training. These claims are supported by Section 3, "HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark", Section 4, "Method", and Section 5, "Experiment".

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6, "Conclusion and discussion", discusses limitations, including that HLR-AVSceneQA is simulator-based, does not fully capture real-world acoustic complexity and visual diversity, and focuses on indoor scenes without dynamic agents, moving sound sources, or richer embodied interactions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. Its contributions are primarily benchmark construction, model design, curriculum training, and empirical evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3, "HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark" describes the benchmark construction pipeline, Section 4, "Method" describes the model architecture and curriculum training, and Section 5, "Experiment" describes datasets, metrics, and baselines. Appendix A, "Additional experimental details" provides implementation details, baseline configurations, and STARSS23 preprocessing, while Appendix C, "Details on HLR-AVSceneQA" further describes benchmark generation, QA construction, and quality control.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The project page linked in the abstract provides access to the released code and benchmark data. Section 3, "HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark", Section 4, "Method", Section 5, "Experiment", Appendix A, "Additional experimental details", Appendix B, "Additional results and analysis", and Appendix C, "Details on HLR-AVSceneQA" provide benchmark construction, implementation, and evaluation details needed to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3, "HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark" specifies benchmark generation, training/test scene splits, task design, and QA construction, while Section 4, "Method", gives preprocessing and fine-tuning hyperparameters. Section 5, "Experiment", and Appendix A, "Additional experimental details" provide datasets, metrics, comparison models, STARSS23 preprocessing, LoRA settings, training precision, batch size, gradient accumulation, and compute setup.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean $\pm$ standard deviation over three independent runs for the main benchmark results, ablations, and real-world adaptation experiments. Section 5, "Experiment" specifies this evaluation protocol, and the corresponding tables report the resulting standard deviations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4, "Method", states that training is performed on four NVIDIA A6000 GPUs with a per-device batch size of 4. Appendix A, "Additional experimental details" further specifies bfloat16 precision, FlashAttention, LoRA settings, training for 2 epochs, per-device batch size 4, and gradient accumulation 4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Appendix D, "Broader impact", discusses intended beneficial applications and privacy/consent considerations for future real-world deployment. Appendix F, "Licenses and terms of existing assets", further states that existing datasets, simulators, models, and 3D assets are used under their licenses and terms of use.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix D, "Broader impact", discusses positive impacts for embodied AI, assistive agents, and robotics. It also discusses negative or sensitive deployment considerations, especially privacy, consent, and data minimization for future real-world use in homes or shared indoor spaces.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Appendix D, "Broader impact", states that the benchmark is simulator-based and does not contain private real-world recordings. Appendix F, "Licenses and terms of existing assets", states that generated 3D assets are synthetic, are not intended to reproduce real people, private environments, or sensitive locations, and that only assets and annotations compatible with applicable licenses and redistribution terms will be released.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix F, "Licenses and terms of existing assets", lists the licenses or usage terms for SoundSpaces 2.0, HM3DSem/HM3D, VGGSound, LibriSpeech, Objaverse-XL, Stable Diffusion 2, TRELLIS, and comparison models. The paper also cites the corresponding original works throughout the main text and references.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Section 3, "HLR-AVSceneQA: Spatial audio-visual scene understanding benchmark", documents the new benchmark, task hierarchy, dataset size, generation pipeline, and QA construction. Appendix C, "Details on HLR-AVSceneQA", further documents scene generation, task definitions, QA formats, quality-control procedures, statistics, and examples. The project page linked in the abstract provides access to the released benchmark assets and accompanying documentation.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The human verification was conducted internally by the research team as an annotation-quality check and did not involve recruited human subjects or crowdsourced workers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or human-subject studies requiring participant risk disclosure or IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: Appendix E, "LLM usage", explicitly describes LLM usage in three ways: HLR-LLM uses an LLM as the reasoning module, an LLM is used to paraphrase question templates, and an LLM is used for editing, grammar checking, table/figure refinement, figure generation, and the data viewer webpage.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.