OMNIECHO: AUDIO-VISUAL SPATIAL UNDERSTANDING FOR OMNI-MODAL EMBODIED AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans can effortlessly localize the direction of a sound source and integrate it with visual cues for reasoning, yet this remains challenging for embodied agents. In particular, it is still unclear how to effectively evaluate and model spatial audio understanding in embodied settings. To address this gap, we introduce **OmniEchoBench**, a unified benchmark for spatial audio-visual perception and audio-vision-language navigation. OmniEchoBench comprises six tasks over 197 real-world spatial audio-visual scenes, 2,972 question-answer pairs, and 900 navigation samples with first-order ambisonics (FOA) audio collected from 30 real-world environments. To enable scalable training supervision, we develop a controllable rendering pipeline for spatial audio. It preserves geometric consistency among sound sources, visual observations, and agent trajectories. Building on this, we propose **OmniEcho**, a spatially aware omni-modal model. It introduces an FOA spatial encoder alongside a pretrained semantic audio pathway. Extensive experiments show that OmniEcho achieves state-of-the-art performance on spatial audio-visual perception. For our sound-guided navigation, OmniEcho reaches a performance level close to that of traditional vision-language navigation. These results demonstrate that spatial audio can serve as a valuable signal for embodied scene reasoning and navigation, while also highlighting fine-grained spatial localization and distance estimation as important open challenges.

1 INTRODUCTION

Embodied agents rely on rich multimodal signals to perceive, reason, and act. Recent advances in omni-modal understanding have substantially improved the ability to jointly process vision, language, and audio (Deshmukh et al., 2026; Liu et al., 2026; Xu et al., 2025b; Liu et al., 2025; Ye et al., 2026). Meanwhile, progress in vision-language navigation (VLN) has enabled agents to reason about their surroundings and plan actions from language instructions (Wei et al., 2026; Wu et al., 2025; Wang et al., 2026; Wei et al., 2025). However, spatial audio, a common and important information source to humans, remains largely underexplored. Humans can effortlessly follow spatial auditory cues to localize target objects even if they are occluded, and to search for target objects beyond the current field of view. For embodied agents, such an ability is equally important. Sounds can provide information about events outside the camera's view, reveal the direction of targets, and guide agents toward their locations (Yang et al., 2024; Chen et al., 2020a; Gan et al., 2020).

In this work, we address spatial audio understanding for omni-modal embodied agents. The agent receives first-order ambisonics (FOA) spatial audio together with visual observations and, when applicable, language instructions. It then reasons about the surrounding environment to answer spatial questions or navigate toward a target. There arise three critical challenges in this context: (1) Collecting real-world spatial-audio data is particularly costly. The received signal jointly depends on source and listener poses, scene geometry, reflections, and reverberation, which must be faithfully captured while source trajectories and listener motion are accurately annotated. (2) Simulation can provide large-scale training data. However, effective training data should not only include rendered audio, but also maintain physical consistency with visual observations and moving agents. (3) Existing pre-trained models already excel in processing omni-modal signals. How to effectively integrate spatial audio into these pre-trained models remains an open question.

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

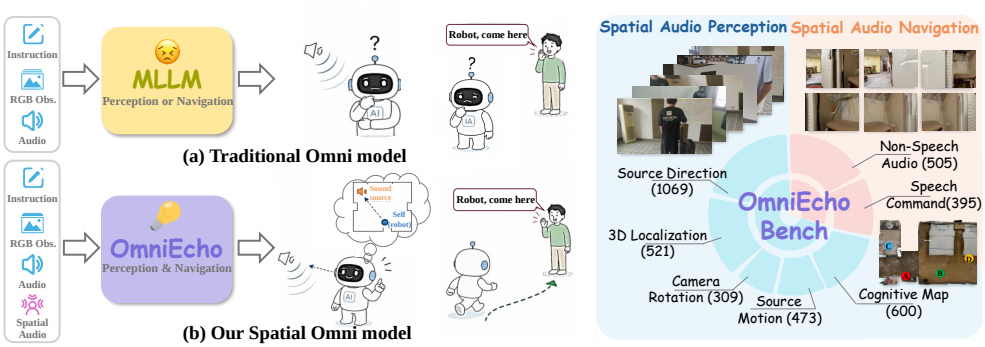


Figure 1: Overview of OmniEcho and OmniEchoBench. Left: OmniEcho extends a traditional omni model with spatial sound for perception and navigation. Right: OmniEchoBench benchmarks spatial perception and spatial audio navigation through six tasks.

To address these challenges, we first introduce **OmniEchoBench**, a unified benchmark for spatial audio-visual perception and audio-vision-language navigation, collected from real-world scenes. We also develop a scalable spatial-audio data construction pipeline. For perception, OmniEchoBench covers audio-visual scenarios involving visible, occluded, and dynamically changing sound sources, requiring agents to reason about source direction, 3D location, source motion, camera rotation, and cognitive maps (Tolman, 1948). For navigation, we collect spatial audio from dense positions in 30 real-world scenes and construct sound-guided navigation trajectories, together with a Habitat-based simulation platform for closed-loop evaluation (Savva et al., 2019). Moreover, to provide scalable supervision for model development, we synthesize **large-scale training data** through controllable spatial audio rendering. For spatial perception, we construct audio-visual scenes with manually specified sound-source configurations and trajectories, and render them into FOA audio to generate spatial question-answering data. For navigation, we augment existing VLN trajectories (Anderson et al., 2018; Ku et al., 2020; Wang et al., 2025; 2023) by placing sound sources at target locations and simulating acoustic effects such as distance attenuation and room reverberation.

Building on this, we propose **OmniEcho**, a spatially aware omni-modal foundation model built on Qwen3-Omni (Xu et al., 2025b). We design an FOA encoder and introduce a three-stage training procedure to align semantic and spatial information. By preserving the model's pretrained semantic audio pathway while introducing a spatial pathway, OmniEcho is able to jointly understand auditory semantics, spatial information, and environmental observations. We evaluate both spatial question answering and sound-guided navigation. Experiments show that existing omni-modal models struggle on our benchmark, whereas OmniEcho achieves state-of-the-art performance on spatial perception tasks. Although our sound-guided navigation task is more challenging than traditional text-guided VLN, OmniEcho achieves performance close to that of VLN methods. These results highlight the effectiveness of spatial audio guidance and the applicability of our task formulation.

Our main contributions are summarized as follows:

- To our knowledge, we are *the first* to address spatial audio understanding for omni-modal embodied agents, capable of spatial audio-visual QA and sound-guided navigation.
- We introduce **OmniEchoBench**, a benchmark consisting of spatial audio perception and navigation, with human annotations in real-world scenes. We further construct large-scale training data through spatial audio simulation, providing scalable training supervision.
- We present **OmniEcho**, a foundation model that can understand spatial audio and support both spatial audio perception and navigation. Extensive experiments demonstrate its strong performance on OmniEchoBench, while revealing fundamental limitations of existing omni-modal models in spatial audio reasoning.

2 Related Work

Traditional Vision-Language Navigation. Vision-and-language navigation (VLN) requires an embodied agent to follow natural-language instructions by grounding route descriptions, landmarks, spatial relations, and action choices in visual observations and navigation history. Since Room-to-

Room (R2R) introduced instruction-guided navigation in real indoor environments using the Matterport3D Simulator (Anderson et al., 2018), VLN has evolved from short-horizon route following toward more realistic and diagnostic navigation settings. Recent work has built fine-grained benchmarks for instruction understanding (Wang et al., 2024c) and studied online robustness through test-time adaptation (Gao et al., 2023) and obstructed environments (Hong et al., 2024). Other studies have improved reasoning or memory through causal learning, adaptive history filtering, language-based perceptual representations, explicit LLM reasoning, and bidirectional cross-modal modeling (Wang et al., 2024a; He et al., 2024; Pan et al., 2024; Zhou et al., 2024; Wu et al., 2025). Despite these advances, representative VLN models remain fundamentally vision-text driven: their decisions mainly depend on RGB/RGB-D observations, textual instructions, object or landmark descriptions, navigation histories, or language-converted scene representations. They do not treat spatial acoustic cues, such as sound-source direction, distance, and spatial layout, as native evidence for navigation. We instead introduce a navigation model with spatial audio understanding capability.

Audio-Visual Navigation. Audio-visual navigation studies embodied agents that use auditory and visual observations to reach sound-emitting targets in 3D environments. Look, Listen, and Act first formalized audio-visual embodied navigation as shortest-path planning toward a sound source from egocentric audio-visual inputs (Gan et al., 2020). SoundSpaces and SoundSpaces 2.0 then established widely used benchmarks by adding geometry-based acoustic rendering to Habitat over Matterport3D and Replica. However, these benchmarks rely on rendered acoustics and do not cover spoken-instruction navigation (Chen et al., 2020a; 2022). Subsequent work improved waypoint selection, acoustic memory, semantic sound handling, and robustness to moving or noisy sources (Chen et al., 2020b; 2021; Younes et al., 2023; Chen et al., 2023; Shi et al., 2025). Other studies explored language-mediated interaction or planning for audio-visual navigation, including AVLEN, CAVEN, AVLMaps, and RILA (Paul et al., 2022; Liu et al., 2024; Huang et al., 2023; Yang et al., 2024). However, these studies do not evaluate navigation in scenarios with real-world recorded spatial audio. In contrast, we introduce a unified benchmark for simultaneously evaluating spatial audio perception and navigation, and we provide a navigation simulator built upon real-world collected FOA audio, thereby offering more realistic and reliable acoustic evidence for navigation.

Spatial Audio Understanding. Recent work has improved spatial acoustic modeling from three complementary angles. One line develops encoders that use binaural phase differences, intensity cues, GCC-PHAT features, microphone geometry, or task-disentangled branches, improving localization, distance estimation, and spatial reasoning (Zheng et al., 2024; Wilkinghoff & Tan, 2026; Dementyev et al., 2026). A second line aligns spatial audio with language, either through contrastive learning over open-vocabulary text or through structured embeddings that separate semantic and spatial factors for retrieval, understanding, and editing (Devnani et al., 2024; Hu et al., 2025). A third line connects spatial audio representations to audio-language models for natural-language description and reasoning about direction, distance, reverberation, and multi-source relations (Sakshi et al., 2025; Jiang et al., 2026; Biswas et al., 2026; You et al., 2026; Liu et al., 2026). Overall, these works demonstrate that spatial audio can be encoded, aligned with language, and used for language-based spatial reasoning. Recent work has begun to equip MLLMs with spatial audio understanding. The closest related work is Spatial-Omni (Zhu et al., 2026), which explores spatial audio understanding. However, it does not incorporate visual frames during training and does not address sound-guided navigation. In contrast, our benchmark uses real-world FOA recordings and evaluates both spatial audio-visual understanding and sound-guided navigation.

3 BENCHMARK AND DATASETS

Overview. This section presents the benchmark and dataset constructed in this work, including OmniEchoBench-QA for evaluating spatial-audio understanding and OmniEchoBench-Nav for assessing sound-guided navigation. It further describes the training data synthesis pipeline, which is built upon a unified FOA spatial-audio rendering framework to generate spatial audiovisual question-answering data, audio-only data, and sound-guided navigation data.

3.1 OMNIECHOBENCH-QA

Design principle. As shown in Figure 2, we introduce OmniEchoBench-QA to evaluate whether an omni-modal model can perceive and reason about sound-source geometry from egocentric observa-

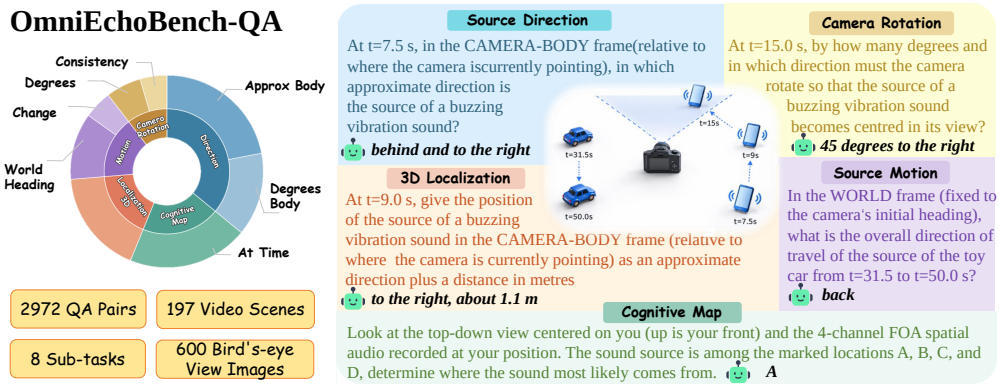


Figure 2: Overview of OmniEchoBench-QA, which comprises 2,972 QA pairs across 197 scenes and 600 bird's-eye view images, assessing a model's understanding of spatial audio.

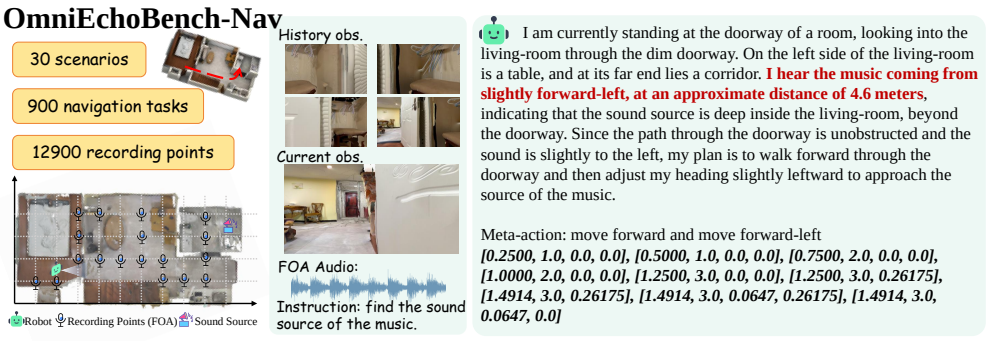


Figure 3: Overview of OmniEchoBench-Nav. Collected across 30 real-world scenes, the benchmark contains 900 navigation tasks, where agents use historical frames, the current view, FOA audio, and an instruction to navigate toward the target.

tions. This benchmark emphasizes sound sources that lie outside the current visual field or dynamically enter and leave the field of view over time. Unlike synthetic setups, all data are collected from real human performances. We first author a set of scripts (*screenplays*) that specify the acoustic events, the actors' motion, and their spatio-temporal relation to the camera (see the accompanying screenplay sheet), deliberately emphasizing off-screen and boundary-crossing sound sources so that the visual stream alone is insufficient for answering the questions. Human actors then perform and record each script, capturing, in a single synchronized take, a first-person main-view video, a 360° panoramic video, and a four-channel first-order Ambisonics (FOA, ACN/SN3D) audio track. The panoramic video is used *only* during annotation; at test time, the model is given a single-view video (or a still image) together with the FOA audio, matching a realistic embodied-perception setting.

Annotation and question construction. Using the panoramic recording as the ground-truth reference, annotators label the real-time position of every active sound source at 2 Hz, yielding a dense spatio-temporal trajectory for each source relative to the recording device. Then we apply rule-based sampling to instantiate questions in several typed categories, automatically deriving the ground-truth answers. The benchmark comprises two complementary subsets and 2,972 QA pairs in total. The first subset (2,372 QA) probes fine-grained temporal spatial reasoning over the recorded videos: *source direction* (1,069), *3D localization* (521), *source motion* (473), and *camera rotation* (309). The second subset (600 QA) targets audio-driven source localization from a bird's-eye view. In this setting, the model is presented with an ego-centered top-down map (with "up" aligned to its forward direction) together with the co-located FOA clip collected in navigation scenes. The model must then select the most plausible sound-source location from four candidates marked A, B, C, D. This setup more directly isolates the model's ability to integrate spatial hearing with a cognitive map. Additional information can be found in the Appendix A.2.1.

3.2 OmniEchoBench-Nav

Design and Annotation. As shown in Figure 3, we introduce OmniEchoBench-Nav, a real-world indoor navigation benchmark that jointly supports *spatial-audio* and *language-instruction* naviga-

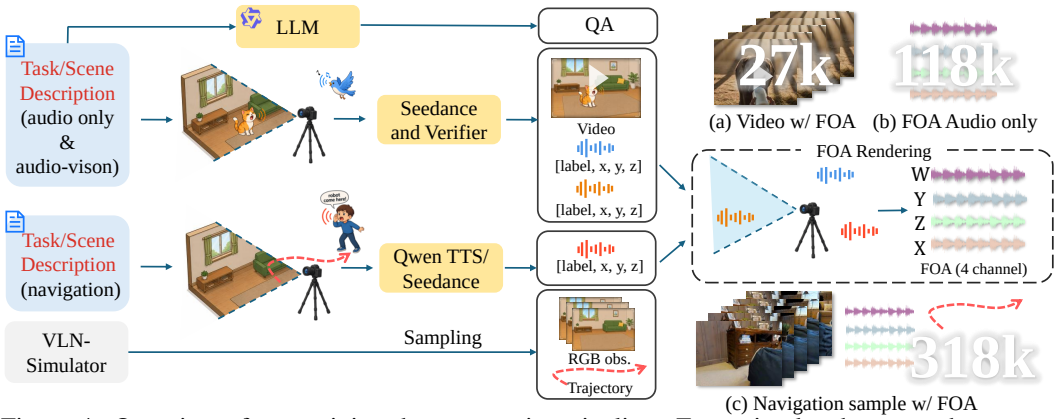


Figure 4: Overview of our training data generation pipeline. From simulated scenes, the system generates QA pairs, videos and sound event annotations and renders mono audio into FOA data.

tion over the same set of scenes and targets. The benchmark comprises 30 real-scanned indoor environments, distributed over single-room (10) and multi-room cross-room layouts (20, spanning two to four connected rooms), each provided as a textured mesh (`.glb`) together with a top-down floor render. Each scene contains 30 navigation samples, giving 900 samples in total. In each sample, a single sound source is placed at a known 3D position. Its signal is captured at a dense grid of receiver positions with known coordinates, producing 12,900 FOA recordings (4-channel, 48 kHz) that together form a spatially dense acoustic field of the environment. Sources cover a broad range of everyday audio, split into non-speech environmental events (e.g., mechanical, alarm, doorbell, animal, footsteps; 505 samples) and spoken commands (e.g., help requests, calls; 395 samples). Each source is annotated with its source device, room, semantic description, and a natural-language description of the navigation target. At test time, the agent is provided with the spatial audio recorded at the receiver position nearest to its current location, with directional remapping performed according to its current orientation. Based on this perceived spatial audio, the embodied agent must infer the location of the target sound source and navigate toward it. The Appendix A.2.2 describes the real-world scenes and acoustic-field collection process in greater detail.

3.3 Spatial Audiovisual Simulation for Data Synthesis

We build all spatial perception and navigation data on top of a unified spatial-audio rendering framework. Each sounding source is represented by a time-varying 3D trajectory, while the listener is co-located with the camera. We adopt a right-handed listener frame with the camera at the origin facing $+y$; azimuth $\theta \in [-180°, 180°]$ is defined with $0°$ straight ahead, $+90°$ to the left, and $-90°$ to the right. Given the source trajectories and listener pose, we encode each monophonic source stem into FOA (4 channels, ACN/SN3D) using real spherical harmonics (Nachbar et al., 2011; Zotter & Frank, 2019):

$$\mathbf{a}(t) = \sum_i \frac{g_i}{r_i(t)} \mathbf{Y}\big(\theta_i(t), \phi_i(t)\big) s_i(t), \tag{1}$$

where $s_i(t)$, $r_i(t)$, and g_i denote the source stem, source–listener distance, and source gain, respectively. Source directions are transformed into the current listener frame to maintain spatial alignment with the visual observations, with optional distance attenuation and room effects. This unified formulation is shared by both spatial audiovisual perception and sound-guided navigation.

Figure 4 illustrates two training-data pipelines. For spatial audiovisual QA, we generate dynamic scenes with Seedance (Seedance et al., 2026) and combine their visual content with trajectory-grounded sound events, from which FOA audio and QA pairs are synthesized. We further synthesize audio-only training data to support FOA encoder pretraining and QA. For navigation, we augment existing VLN trajectories with destination-conditioned sound events and render the corresponding FOA observations along sampled trajectories. The datasets provide synchronized visual observations, spatial audio, and geometry-grounded supervision for both perception and navigation. Further details of the data generation and annotation procedures are provided in the Appendix A.1.

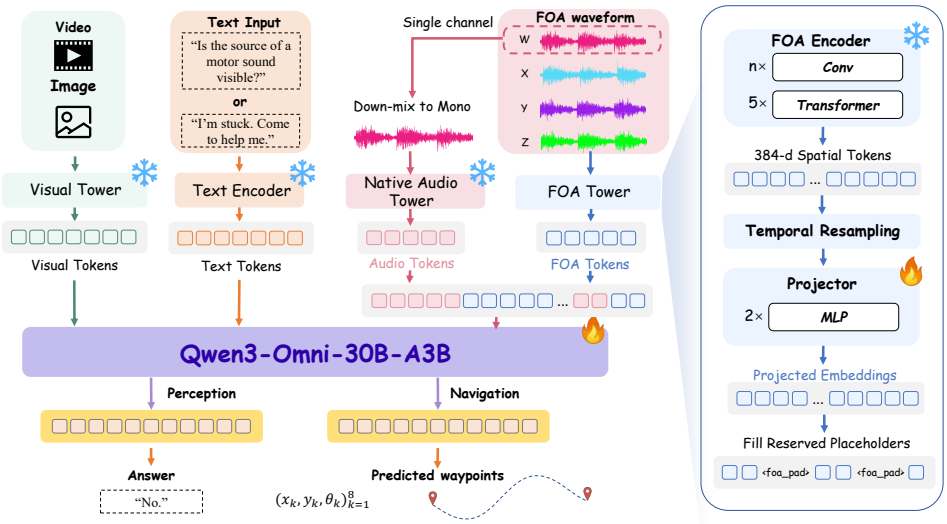


Figure 5: Overview of our unified spatial omni model built on Qwen3-Omni-Thinker. The pretrained FOA encoder extracts spatial tokens; these tokens are projected into the language embedding space by a trainable projector, and inserted after the corresponding semantic audio tokens.

4 METHOD

4.1 TASK FORMULATION

We unify spatial audio-visual perception and navigation within a single embodied-agent framework. For perception tasks, the model takes FOA audio A, a video sequence V or a single image I, and a question q, and predicts an answer a. For navigation tasks, we follow the setting of InternNav (Wei et al., 2026). At each decision step, the model receives a language instruction L, T frames captured by a forward-facing camera, and the FOA spatial audio A_t within the current temporal window, and predicts a trajectory over the next $K = 8$ time steps. We uniformly sample $T - 1$ keyframes from past observations and combine them with the current frame as visual input. This formulation enables unified spatial audio-visual perception and navigation within a single framework.

4.2 SPATIAL AUDIO ENCODING

Stage 1: Semantic alignment of the FOA encoder. We first pretrain a lightweight FOA encoder f_θ that maps an FOA clip into a temporal sequence of d-dimensional tokens ($d = 384$). Each clip is converted into a 5-channel input map $\mathbf{X} \in \mathbb{R}^{5\times128\times T}$ comprising the log-mel spectrogram of the omnidirectional (W) channel together with the three active-intensity components (i_x, i_y, i_z) and the diffuseness δ, all in the ACN/SN3D convention. The encoder is trained to align its representations with a frozen CLIP text encoder (Radford et al., 2021) using a SigLIP objective (Zhai et al., 2023), thereby establishing an open-vocabulary sound-semantic space. Training uses a 100k-clip corpus of synthetic FOA scenes containing 1–4 sound sources with diverse spatial configurations.

Stage 2: Query-conditioned spatial localization. Starting from the semantically aligned encoder, we further fine-tune it for query-conditioned sound-source localization. Given a text query specifying a target source, a cross-attention localization head attends to the encoder tokens and predicts the source's azimuth, elevation, and distance. The model is optimized with a localization loss together with the Stage-1 semantic objective to preserve the learned sound semantics.

Stage 3: Integration into the Omni backbone. As shown in Figure 5, we graft the frozen FOA encoder $f_{\theta^\star}$ into the Qwen3-Omni model (Xu et al., 2025b). Each FOA clip is processed along two complementary paths: (i) its W channel is down-mixed to mono and fed to the original, *frozen* audio tower to preserve the model's native semantic audio tokens; and (ii) the same clip is passed through $f_{\theta^\star}$ to obtain spatial tokens, which are temporally resampled to the 7 Hz audio-token grid and mapped into the language embedding space by a trainable projector. The projected spatial tokens occupy reserved placeholder embedding rows (`<foa_pad>`) and are inserted immediately after the corresponding audio tokens in temporal order. Only the LLM parameters Ψ and the projector ϕ are

Table 1: **Results on our OmniEchoBench-QA**. Our OmniEcho performs strongly in both audio-only and audio-visual settings, outperforming most evaluated models in overall accuracy and across the individual spatial reasoning tasks. 'A-single' denotes single-channel audio, 'A-foa' denotes four-channel FOA audio, and 'V' denotes visual input.

Method	Base Model	Setting	Camera Rotation	Source Direction	3D Localization	Source Motion	Cognitive Map	Overall
Random	-	-	33.6	12.5	-	23.8	25.0	-
Audio only								
Qwen3-Omni-30B-A3B (Xu et al., 2025b)	Qwen3-Omni-30B-A3B	A-single	22.7	18.4	8.3	22.0	-	17.5
Qwen2.5-Omni-7B (Xu et al., 2025a)	Qwen2.5-Omni-7B	A-single	22.7	8.3	4.6	14.6		10.6
GLM4-Voice-9B (Zeng et al., 2024)	GLM4-9B	A-single	20.4	9.5	5.4	20.3	-	12.2
SO-7B (Zhu et al., 2026)	Qwen2.5-Omni-7B	A-foa	21.4	7.1	0.0	20.7	-	10.1
OmniEcho (Ours)	Qwen3-Omni-30B-A3B	A-foa	**29.1**	**22.1**	**14.0**	**24.9**	-	**21.8**
Audio & Vision								
Qwen3-Omni-30B-A3B (Xu et al., 2025b)	Qwen3-Omni-30B-A3B	A-single + V	32.7	16.3	7.7	20.9	22.3	18.5
Qwen2.5-Omni-7B (Xu et al., 2025a)	Qwen2.5-Omni-7B	A-single + V	24.9	8.0	5.4	14.0	20.0	12.6
SO-7B (Zhu et al., 2026)	Qwen2.5-Omni-7B	A-foa + V	21.4	5.2	0.0	17.1	22.8	11.4
OmniEcho (Ours)	Qwen3-Omni-30B-A3B	A-foa + V	**41.7**	**22.6**	**14.2**	**24.7**	**47.5**	**28.5**

Table 2: **Results on our OmniEchoBench-Nav.** Our benchmark is designed to evaluate sound-guided navigation, and meanwhile, additionally provides text instruction for traditional text-guided VLN methods. Even if audio guidance is more challenging than text instruction, our method achieves close performance as previous VLN methods.

Model	Base Model	Guidance	Val-Unseen				
			SR↑	SPL↑	NE↓	OS↑	TL
Seq2Seq (Krantz et al., 2020)	End-to-end	Text Instruction	11.3	9.4	4.14	53.00	5.51
CMA (Krantz et al., 2020)	End-to-end	Text Instruction	9.8	8.5	4.34	48.56	5.00
InternVLA-N1 (Intern Robotics, 2025)	Qwen2.5-VL-7B	Text Instruction	17.8	16.7	2.87	63.89	4.67
Soundspaces (Chen et al., 2020a)	End-to-end	Binaural audio	5.4	3.9	5.32	25.67	4.16
OmniEcho	Qwen3-Omni-30B-A3B	FOA audio	16.2	11.5	4.09	52.89	7.02

updated, while the FOA encoder, audio tower, and visual tower remain frozen. Additional encoder architecture and optimization details for all three training stages are included in the Appendix A.3.

5 Experiments

5.1 Setup

Implementation Details. We build our model upon the pretrained Qwen3-Omni-30B-A3B (Xu et al., 2025b). Specifically, the model is trained on spatial audio-visual QA together with navigation-oriented supervision, resulting in 363,193 training examples in total. The navigation subset consists of 139,142 examples from ScaleVLN (Wang et al., 2023), 92,418 from RxR (Ku et al., 2020), 31,577 from R2R (Anderson et al., 2018), and 28,969 from VLN-PE (Wang et al., 2025). Stage 3 training is conducted on 32 NVIDIA A100 GPUs and requires approximately four days.

Benchmarks. For spatial audio-visual perception, we report results on OmniEchoBench-QA. For navigation, we evaluate on OmniEchoBench-Nav as well as the VLN-CE R2R val unseen split (Anderson et al., 2018; Krantz et al., 2020). More details are documented in the Appendix A.4.

5.2 Main results

5.2.1 Spatial Audio-Visual Question Answering Results

Table 1 reports the results on OmniEchoBench-QA. Under both the audio-only and audio-visual settings, OmniEcho outperforms evaluated models overall and across individual tasks. In the audio-only setting, OmniEcho surpasses the Qwen3-Omni-30B-A3B (Xu et al., 2025b) backbone which has only single-channel audio encoder, demonstrating the effectiveness of our FOA encoder. OmniEcho significantly outperforms SO-7B (Zhu et al., 2026) on our real-world collected benchmark. Although SO-7B claims to employ spatial-audio training, it lacks visual supervision during training

Table 3: **Performance on VLN-CE R2R with simulated spatial audio.** Although sound-guided navigation is more challenging than text-guided VLN, our method achieves comparable performance to previous baselines, demonstrating the potential of our model and dataset for future research.

Model	Base Model	Guidance	Val-Unseen				
			SR↑	SPL↑	NE↓	OS↑	TL
Seq2Seq (Krantz et al., 2020)	End-to-end	Text Instruction	21.0	-	7.81	28.0	8.39
CMA (Krantz et al., 2020)	End-to-end	Text Instruction	22.0	20.0	8.17	28.0	7.42
VLN-R1 (Qi et al., 2025)	Qwen2-VL-7B (Wang et al., 2024b)	Text Instruction	30.2	21.8	7.0	41.2	10.0
InternVLA-N1 (Intern Robotics, 2025)	Qwen2.5-VL-7B (Bai et al., 2025)	Text Instruction	50.6	43.3	4.73	56.7	-
NAViLA-SAGE (Miao et al., 2026)	navila-siglip-llama3-8b-v1.5-pretrain	Text Instruction	38.0	36.0	-	51.0	-
OmniEcho	Qwen3-Omni-30B-A3B (Xu et al., 2025b)	Audio Instruction	22.2	16.3	7.31	24.5	5.54

and still suffers from a sim-to-real gap. These findings suggest that existing methods remain limited in their robustness for spatial-audio understanding in real-world scenarios.

Comparing OmniEcho across the two input settings, removing visual input leaves its performance on Source Direction, 3D Localization, and Source Motion largely unchanged, indicating that the model can perform these forms of spatial reasoning primarily from FOA audio. Camera Rotation, however, degrades substantially without vision, suggesting that this task depends more strongly on visual evidence of viewpoint changes. This pattern is consistent with the FOA-encoder ablation, where explicit FOA features benefit source-centric spatial tasks but provide limited gains for camera-rotation reasoning. Together, these results suggest that FOA primarily captures sound-source geometry, while visual observations remain important for camera-centric reasoning.

We acknowledge that training with synthetically generated simulation data introduces bias. We align the distribution of simulated sound-source states with that of real-world states. Empirically, the simulated data has been shown to endow the model with a fundamental capability for spatial audio perception. In addition, the experiments in Appendix A.5.1 suggest that, in the future, the training distribution could potentially be corrected using a small amount of real spatial-audio data.

5.2.2 Spatial Audio Navigation Results

Baseline Selection. SoundSpaces-based models, AVLEN, and RILA (Chen et al., 2020a; Paul et al., 2022; Yang et al., 2024) are originally designed for simulated acoustic observations and do not support FOA input. To provide an informative comparison, we include the open-source SoundSpaces model and convert the FOA audio into binaural audio. Nevertheless, some navigation instructions in OmniEchoBench-Nav are conveyed through the spatial audio itself, requiring the agent not only to pursue a sound-emitting target but also to perform joint semantic understanding and spatial grounding. Therefore, the task formulations of these methods cannot be fully aligned with ours.

Results on OmniEchoBench. Table 2 reports the results on the OmniEchoBench-Nav. Following text-guided models, we adopt the agent-stopping setting, in which each model autonomously decides when to terminate. OmniEcho achieves 16.2% SR and 11.5% SPL, substantially outperforming the conventional text-guided baselines Seq2Seq and CMA. Compared with InternVLA-N1 (Intern Robotics, 2025), OmniEcho is lower by 1.6 points in SR and 5.2 points in SPL. These results indicate that, without route-level textual guidance or goal images, sound-guided navigation in real-world environments can still approach the performance of some traditional text-guided baselines. In contrast, SoundSpaces with binaural audio performs markedly worse in real spatial-audio settings and struggles to determine when to stop autonomously. For OmniEcho and other spatial-audio navigation models, efficient navigation and reliable stopping remain significant challenges.

Results on VLN-CE R2R. Table 3 further reports results on the simulated VLN-CE R2R Val-Unseen split (Krantz et al., 2020). We again use agent stopping for the cross-model comparison. Our OmniEcho achieves an SR of 22.2%, slightly exceeding Seq2Seq and CMA (Krantz et al., 2020), although its SPL of 16.3% is below CMA (20.0%). Its NE of 7.31 is better than those of Seq2Seq and CMA, and its trajectory length of 5.54 is the shortest among methods with reported TL. Nevertheless, OmniEcho remains behind VLN-R1 (Qi et al., 2025), NAViLA-SAGE (Miao et al., 2026), and InternVLA-N1 (Intern Robotics, 2025) in SR and SPL. Overall, spatial-audio guidance is competitive with earlier text-guided VLN baselines in success rate, but a clear gap remains relative to stronger text-guided systems, especially in path efficiency. We provide additional analyses of navigation failure modes and the effect of ground-truth sound direction in Appendix A.5.2 and A.5.3.

Table 4: **Ablation of the native audio encoder, FOA encoder, and Stage 3 training strategy on OmniEchoBench-QA.** The best overall performance is obtained by combining both encoders and freezing the FOA encoder during Stage 3.

Native audio encoder	FOA encoder	FOA encoder (stage 3)	Camera Rotation	Source textbfDirection	3D Localization	Source Motion	Cognitive Map	Overall
✓	✓	training	37.5	21.6	13.2	19.5	45.2	26.2
✓	✗	frozen	**43.7**	16.2	**14.2**	16.3	21.8	19.8
✗	✓	frozen	40.1	18.9	13.6	21.0	47.2	26.2
✓	✓	frozen	41.7	**22.6**	**14.2**	**24.7**	**47.5**	**28.5**

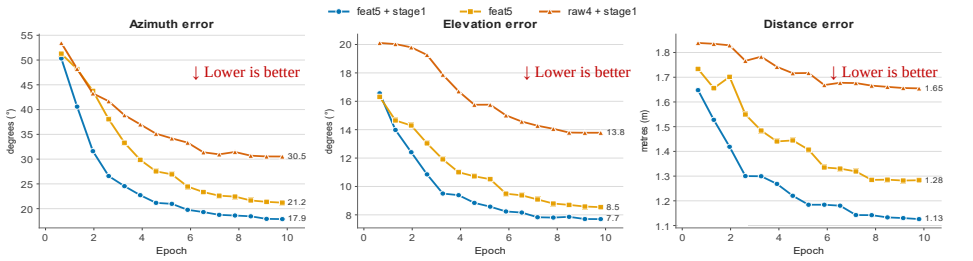


Figure 6: **Ablation on input representation and Stage-1 pretraining.** feat5+stage1 achieves the best azimuth, elevation, and distance errors, outperforming raw4+stage1 and feat5 (w/o stage 1 pretraining), showing that feature-based input is more effective than raw FOA waveform input and that Stage-1 pretraining further improves convergence and final accuracy.

5.3 Ablation Study

Ablation on native audio encoder and FOA encoder. Table 4 ablates three designs in Qwen3-Omni-30B-A3B (Xu et al., 2025b): introducing the FOA encoder, retaining the native audio encoder, and freezing the FOA encoder during Stage 3. The full configuration achieves the best Overall score of 28.5. Adding the frozen FOA encoder to the native-audio-only baseline improves Overall and most subtasks, confirming that it provides essential spatial cues beyond those captured by the native audio pathway. The slight decrease on Camera Rotation suggests that extra visual observations reduce reliance on spatial audio cues. Removing the native audio encoder while retaining the frozen FOA encoder reduces Overall from 28.5 to 26.2, showing that native semantic audio tokens remain complementary to the spatial representation. Finally, updating rather than freezing the FOA encoder during Stage 3 degrades every subtask, suggesting that freezing better preserves the spatial representations from the first two stages and avoids interference during multimodal instruction tuning.

Ablation on input representation and Stage-1 pretraining. As shown in Figure 6, feat5+stage1 achieves the best performance, with 17.9° azimuth error / 7.70° elevation error / 1.13 distance error, substantially outperforming raw4+stage1 (30.5° / 13.8° / 1.65). This suggests that, compared with directly modeling the raw FOA waveform, the pre-extracted 5-channel acoustic features provide stronger localization cues. Moreover, comparing feat5 with feat5+stage1 shows that removing Stage-1 semantic alignment pretraining degrades performance, indicating that such initialization improves optimization and helps the model converge to a better solution. Further tests show that simply adding an FOA encoder, without Stage 1 and Stage 2 training, does not bring clear gains over the variant without the FOA encoder. This result highlights the necessity of dedicated pretraining to avoid interference during multimodal instruction tuning that jointly mixes planning and QA tasks.

6 Conclusion

We introduce **OmniEchoBench**, a real-world benchmark unified for spatial audio-visual perception and sound-guided navigation, together with a controllable spatial audio rendering pipeline for data generation. We further propose **OmniEcho**, an omni-modal model that integrates an FOA encoder with a pretrained semantic audio pathway to jointly capture auditory semantics and spatial information. Experiments demonstrate that our model achieves state-of-the-art performance in spatial audio perception and navigation. Also, the performance of sound-guided navigation is close to that of traditional text-guided VLN methods. These results highlight the value of spatial audio for embodied scene reasoning and navigation, while identifying fine-grained spatial localization and distance estimation as key remaining challenges. We hope this work inspires future research on embodied agents that jointly reason about *what* they perceive and *where* it originates.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets. Specifically, generative AI assisted in constructing the simulated spatial audio–visual and sound-guided navigation training data, including the generation of synthetic scene descriptions, audio-visual media, sound events, and associated annotations. The corresponding data-generation and validation procedures are reported in the main text and Appendix A.1.

We did not use generative AI tools for the other tasks with required disclosure, including developing the core scientific contributions, designing the research methodology or experiments, conducting experimental analysis, selecting or verifying references, or making scientific claims; the remaining required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools to aid or polish writing, including language editing, grammar correction, readability improvements, and suggestions of alternative wording. We also used generative AI tools for research ideation or execution, specifically to assist with writing code. We have reviewed all AI-assisted work. All AI-assisted text was carefully edited and approved by the authors; AI-assisted code was inspected and tested for correctness; synthetic data were checked using the validation procedures described in the paper; and all technical content, claims, citations, and reported results were manually verified. We take responsibility for the final content of this work, including text, claims, code, data, annotations, and other artifacts produced with the aid of generative AI.

REFERENCES

Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

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

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *European conference on computer vision*, pp. 17–36. Springer, 2020a.

Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622*, 2020b.

Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15511–15520. IEEE, 2021.

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

Jinyu Chen, Wenguan Wang, Si Liu, Hongsheng Li, and Yi Yang. Omnidirectional information gathering for knowledge transfer-based audio-visual navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10959–10969, 2023. doi: 10.1109/ICCV51070.2023.01009.

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

Amala Sanjay Deshmukh, Kateryna Chumachenko, Tuomas Rintamaki, Matthieu Le, Tyler Poon, Danial Mohseni Taheri, Ilia Karmanov, Guilin Liu, Jarno Seppanen, Arushi Goel, et al. Nemotron 3 nano omni: efficient and open multimodal intelligence. *arXiv preprint arXiv:2604.24954*, 2026.

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

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

Junyu Gao, Xuan Yao, and Changsheng Xu. Fast-slow test-time adaptation for online vision-and-language navigation. *arXiv preprint arXiv:2311.13209*, 2023.

Keji He, Ya Jing, Yan Huang, Zhihe Lu, Dong An, and Liang Wang. Memory-adaptive vision-and-language navigation. *Pattern Recognition*, 153:110511, 2024.

Haodong Hong, Sen Wang, Zi Huang, Qi Wu, and Jiajun Liu. Navigating beyond instructions: Vision-and-language navigation in obstructed environments. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7639–7648, 2024.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

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

Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Audio visual language maps for robot navigation. In *International Symposium on Experimental Robotics*, pp. 105–117. Springer, 2023.

Intern Robotics. InternVLA-N1: An open dual-system vision-language navigation foundation model with learned latent plans. Technical report, Intern Robotics and Shanghai AI Laboratory, 2025. URL `https://internrobotics.github.io/internvla-n1.github.io/static/pdfs/InternVLA_N1.pdf`.

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

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 3765–3773, 2024.

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. *arXiv preprint arXiv:2602.18527*, 2026.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

Bingchen Miao, Rong Wei, Zhiqi Ge, Shiqi Gao, Jingzhe Zhu, Renhan Wang, Siliang Tang, Jun Xiao, Rui Tang, Juncheng Li, et al. Towards physically executable 3d gaussian for embodied navigation. In *International Conference on Learning Representations*, volume 2026, pp. 22945–22965, 2026.

Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. Ambix-a suggested ambisonics format. In *Ambisonics Symposium*, volume 2011, 2011.

Bowen Pan, Rameswar Panda, SouYoung Jin, Rogerio Feris, Aude Oliva, Phillip Isola, and Yoon Kim. Langnav: Language as a perceptual representation for navigation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 950–974, 2024.

Sudipta Paul, Amit Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *Advances in Neural Information Processing Systems*, 35:6236–6249, 2022.

Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Zhanbo Shi, Lin Zhang, Linfei Li, and Ying Shen. Towards audio-visual navigation in noisy environments: A large-scale benchmark dataset and an architecture considering multiple sound-sources. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 14673–14680, 2025.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13139–13150. IEEE, 2024a.

Liuyi Wang, Xinyuan Xia, Hui Zhao, Hanqing Wang, Tai Wang, Yilun Chen, Chengju Liu, Qijun Chen, and Jiangmiao Pang. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9455–9465. IEEE, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Shuo Wang, Yucheng Wang, Guoxin Lian, Yongcai Wang, Maiyue Chen, Kaihui Wang, Bo Zhang, Zhizhong Su, Yutian Zhou, Wanting Li, et al. Progress-think: Semantic progress reasoning for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4076–4086, 2026.

Zehao Wang, Minye Wu, Yixin Cao, Yubo Ma, Meiqi Chen, and Tinne Tuytelaars. Navigating the nuances: A fine-grained evaluation of vision-language navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4681–4704, 2024c.

Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986. IEEE, 2023.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

Meng Wei, Chenyang Wan, Peng Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 12380–12396, 2026.

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

Jie Wu, Chunlei Wu, Xiuxuan Shen, Fengjiang Wu, and Leiquan Wang. Recursive bidirectional cross-modal reasoning network for vision-and-language navigation. *Expert Systems with Applications*, 270:126442, 2025.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

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

Zeyuan Yang, Jiageng Liu, Peihao Chen, Anoop Cherian, Tim K Marks, Jonathan Le Roux, and Chuang Gan. Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16251–16261, 2024.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

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

Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 8(2):928–935, 2023.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot, 2024. URL https://arxiv.org/abs/2412.02612.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

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

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7641–7649, 2024.

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

Franz Zotter and Matthias Frank. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

A APPENDIX

A.1 DATA GENERATION DETAILS

A.1.1 SPATIAL AUDIOVISUAL SIMULATION FOR DATA SYNTHESIS.

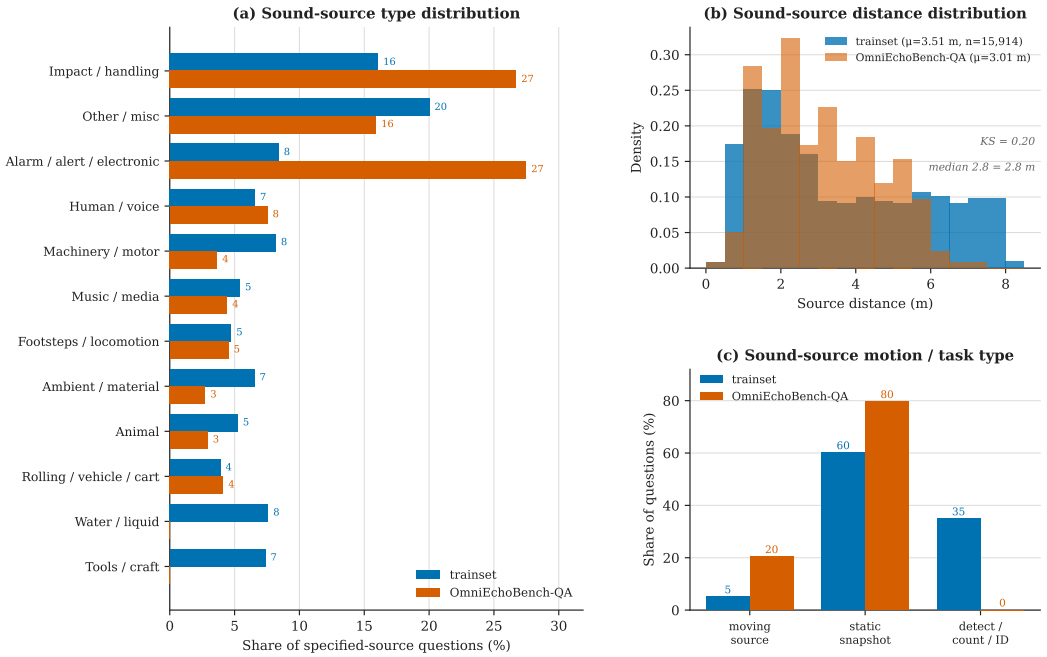


Figure 7: Distributional alignment between our synthetic training corpus (**trainset**, 145 103 samples; blue) and the real-recorded **OmniEchoBench-QA** benchmark (orange), across three axes. **(a)** *Sound-source type.* Normalised over specified-source questions, the two corpora now share a comparable semantic mix (JSD $= 0.14$ bits); the residual gap is the benchmark's heavier emphasis on alarm/electronic tones, against the trainset's broader coverage of machinery, tools, and water sounds. **(b)** *Source distance.* The two distance distributions now largely overlap, with an identical median of $2.8\,\mathrm{m}$ and a Kolmogorov–Smirnov statistic of $D = 0.20$. **(c)** *Motion / task type.* We further augment the training set with non-spatial detect/count/identify tasks, aiming to improve the model's overall capability in spatial-audio perception.

We first construct spatial audiovisual QA data from generated videos. A large language model (LLM) acts as a *director*: given previously generated scene descriptions to encourage diversity, it produces a structured JSON scene containing N foreground sources and one ambient source. For each source, the scene specifies a semantic label, target loudness (LUFS), onset and offset times, and a full trajectory $\{(t_k, \theta_k, \phi_k, r_k)\}_k$. The same call also emits a full-scene video prompt describing all sounding entities and camera behavior, together with one isolated *close-up* prompt per source that depicts only that sounding body in an otherwise silent environment. The director is driven by the system prompt shown in the box below (translated from the original and abridged; diversity, visibility-dynamics, and fine-grained grid/angle rules are elided with `[...]`). A set of auxiliary LLM agents subsequently refines the proposed scene—adjusting per-source onset/offset timing and target loudness, and resolving textual–geometric direction/distance contradictions—before any media are generated.

Scene director — system prompt (translated, abridged)

```
You are a Spatial Audio-Video Scene Director. Freely design one creative spatial-audio
scene and write a matching video-generation script (prompt).

The scene contains N_FOREGROUND foreground sources (e.g. a dripping tap, a cat, footsteps),
role="foreground", and exactly 1 ambient/background source (e.g. rain, wind, traffic),
```

```
role="ambient".

COORDINATES AND SPATIAL LAYOUT
- Right-handed frame; the listener faces +y.
  azimuth in [-180,180], 0 = straight ahead, +90 = left, -90 = right (+left / -right
convention).
  elevation in [-90,90], + = up.  distance in [0.5m, 15m].
- The listener sits at the origin, position=[0,0,0].
- Each source is described by a trajectory (>=2 timestamped points); speed between adjacent
  points < 10 m/s.  Initial azimuths of different sources should be >= 30 deg apart.

CAMERA / LISTENER ORIENTATION
- The camera is bound to the listener; horizontal FOV ~= 90 deg (half-FOV 45 deg).
- listener.orientation.yaw is the camera azimuth and may vary over time via a trajectory.
- Visibility: camera_relative_azimuth(t) = object_azimuth(t) - listener_yaw(t), normalized
to
  [-180,180]; the source is visible when |...| < 45 deg. Each source outputs a
  visibility_window {video_enter_t, video_exit_t}.

PHYSICAL-CONSISTENCY RULE FOR CAMERA ROTATION (critical)
- yaw increasing to positive => "camera rotates left"; decreasing to negative => "rotates
right".
- When the camera turns toward X, the centered object must slide out toward not-X while a
new
  object on the X side enters from the X edge to the center; violating this makes the
  video model produce reversed or frozen motion.  [... do/don't examples omitted ...]

OUTPUT FORMAT (strict) -- output ONE JSON object only, no markdown, no explanation:
{
  "scene": {
    "scene_id", "duration",
    "environment": {"type":"indoor|outdoor", "room":{"dimensions":[w,l,h],"rt60":0.3-0.8},
                    "description"},
    "listener": {"position":[0,0,0],
                 "orientation":{"yaw","pitch","roll","trajectory":[{"t","yaw"}, ...]}},
    "objects": [ {"id","type","label","role","loudness_lufs","snr_db","onset_t","offset_t",
                  "trajectory":[{"t","azimuth","elevation","distance"}, ...],
                  "visibility_window":{"video_enter_t","video_exit_t"}, "wav_path"} ]
  },
  "sub_prompts": [ "close-up video prompt per source (silent env, only that source sounding
)", ...],
  "video_prompt": "full-scene prompt (3 parts: scene description; camera motion; METADATA
appendix)"
}

LOUDNESS: near (<2m) foreground -14..-18 LUFS; mid (2-5m) -20..-26; far/ambient -30..-40;
  ~6 dB per distance doubling. snr_db is relative to the quietest (ambient=0 dB) source.
TIMING: at least one foreground onset_t=0; continuous/ambient sources span the whole clip.
SUB-PROMPTS: one close-up per source showing only that sounding body in an otherwise silent
 but
  reverberant environment; on-screen direction must match the source azimuth sign; static
sources
  must be told to stay still.
DIVERSITY: vary environment (indoor/outdoor, city/nature/home/industrial, day/night, quiet/
noisy)
  and source types across scenes; do not repeat similar scenes.  [... extensive diversity,
  dynamic-visibility, and grid/angle-annotation rules omitted ...]
```

Before media generation, we validate the proposed scene for source counts, timing, spatial ranges, velocities, and visibility intervals. Invalid scenes are regenerated. We further compare directional and distance language in the prompts against the numerical trajectories, and a second LLM revises inconsistent descriptions. This checking–revision loop is repeated for up to three rounds. Additional planning adjusts source activity schedules and target loudness so that ambient sounds span the scene, transient events occur at plausible times, and foreground sounds remain audible over the background.

We then submit the prompts to Seedance (Seedance et al., 2026). The full-scene prompt yields the rendered video V, while each close-up prompt yields an isolated clip whose audio track is extracted as that source's monophonic stem s_i. Stems are trimmed or loop-extended to the intended duration, loudness-normalized to their target LUFS, and gated or faded before rendering.

Because the director's trajectories are only approximately consistent with the generated pixels, we recover source geometry directly from the rendered video. Under a static-camera assumption verified by dense optical flow, objects are localized by open-vocabulary text-prompted segmentation and tracked across frames, and each sounding source is matched to a detected visual instance by label.

For a detected center pixel (c_x, c_y) in a frame of size $W \times H$, we recover the viewing direction under a pinhole camera model. Let

$$f = \frac{W/2}{\tan(\mathrm{FOV}/2)}$$

denote the horizontal focal length and f_v the vertical focal length. We first compute normalized image-plane coordinates

$$u = \frac{W/2 - c_x}{f}, \qquad v = \frac{H/2 - c_y}{f_v}, \tag{2}$$

and then recover azimuth and elevation as

$$\theta = \mathrm{atan2}(u, 1), \qquad \phi = \mathrm{atan2}\big(v, \sqrt{1 + u^2}\big). \tag{3}$$

Monocular depth on key frames provides relative range estimates, which are anchored to the director's scene scale to obtain metric distance. The resulting trajectories are smoothed and resampled at a fixed rate. Sources never observed on screen retain their scripted trajectory but are forced outside the FOV; dynamically exiting sources are linearly extrapolated; and disagreements between detected instance counts and the script trigger instance splitting or masking. This yields a corrected scene in which every source has a per-timestep (θ, ϕ, r) trajectory together with an explicit visibility state.

Starting from the corrected scene, we augment the sample with additional off-screen events and camera motion. On-screen sources are kept unchanged, while 2–4 off-screen sources are appended at azimuths guaranteed to remain outside the possibly rotating FOV. We also account for sources entering or leaving the image and simulate their off-screen continuation. Each appended source follows a world-frame trajectory chosen from a static location, a straight walk, or a polyline with a single turn of at least 90°, and is voiced by a borrowed stem from a cross-sample library after loudness normalization. A camera yaw trajectory $\psi(t)$ is written into the scene so that the rendered audio rotates consistently with the cropped or panned image.

Given the final augmented scene, all stems are rendered into FOA using Eq. equation 1, with optional indoor reflections, and the resulting multichannel audio is muxed into the cropped or panned video without re-encoding the visual stream. We then generate QA pairs directly from the final trajectories rather than from an LLM response. At each queried time, we compute a source's world-frame direction, its direction relative to the current camera, its distance, and whether it lies inside the camera frustum; over intervals, we compute changes from the same trajectories. From these quantities we instantiate four categories of questions—direction (coarse 8-way and fine-grained degrees), motion (world heading and change), 3D localization, and camera rotation—covering eight subtypes across camera, camera-body, and world coordinate frames. Each video contributes 15–20 questions, with source and time selection balanced to cover both binary outcomes and the full set of direction labels. This process yields 26,585 training questions from 1,484 synthesized videos and the data distribution is shown in Figure 7.

Audio-only Spatial Data We also synthesize spatial-audio questions without visual input. Each scene contains one to three independently generated sounds. We assign every source a time interval and either a static, constant-speed, accelerating, or decelerating 3D trajectory, then render the resulting mixture as FOA. Because the trajectories are known, we can directly ask about direction, elevation, distance, source count, motion, speed change, approach or recession, closest approach, and relations between sources. An LLM rewrites the questions and expresses the trajectory-based reasoning in natural language, while the answers remain fixed by the rendered scene. We obtain 18,518 questions from 2,000 moving-source scenes.

The same single-source recordings are reused to construct 100,000 simpler static scenes for spatial-audio pretraining. Each clip contains one to four sound categories placed at independently sampled azimuths, elevations, and distances. These clips support the semantic-alignment and query-conditioned localization stages of the FOA encoder described in Section 4.

A.1.2 Sound-Guided Navigation Data Synthesis

We convert four existing VLN corpora—R2R, RxR, VLN-PE, and ScaleVLN—into sound-guided navigation supervision through a six-stage pipeline: (i) destination sound-event proposal, (ii) audio synthesis, (iii) FOA rendering and sample construction, (iv) acoustic augmentation, (v) omni-modal SFT formatting, and (vi) chain-of-thought (CoT) annotation.

Destination sound-event proposal. For each episode, a vision–language model (Qwen3.7-flash) observes the last five first-person frames captured at the destination (front camera, $1.25\,\mathrm{m}$ height) and proposes a plausible sound event, under a hard constraint that neither the sound content nor the derived task may reveal the destination location, room name, or nameable furniture, so that the agent must localize the target acoustically. Proposals are balanced roughly $50/50$ between two categories: (1) a short *spoken utterance* (a person calling the robot, with diverse intents and an accompanying voice-style description of gender, age, timbre, emotion, and pace), and (2) a *non-speech ambient event* (e.g., knocking, running water, dragging furniture, an animal, a phone, or an appliance). The model returns a structured record containing the sound content, a natural-language task description, the voice/acoustic style, and its scene reasoning. The system prompt is reproduced verbatim below.

Destination sound-event proposal — system prompt

```
You are a scenario designer for an indoor "sound-source navigation" task. A robot starts
somewhere in a house and must reach a specific target object/location ONLY by following a
sound
emitted from that target (plus a short text instruction). You are shown the LAST few first-
person
camera frames of a navigation episode; these frames show the destination area (the target).
 Your
job: invent a plausible sound event that would emit from this destination, so the robot can
localize it acoustically.

Design two categories of scripts; pick ONE per episode (respect the requested category when
given, and keep the two categories roughly balanced across the dataset):
  Category 1 - "spoken utterance that calls the robot over": the sound itself is a short
spoken
    English utterance from a person at the destination. This is NOT limited to emergencies
or
    calls for help - ANY natural reason a person would summon the robot is valid, and you
should
    keep them DIVERSE across episodes, e.g.: asking for help ("Robot, I fell down, please
come and
    find me."), requesting an item be fetched or carried, inviting it to come look at
something,
    assigning a task, a casual/curious call, a playful summon, an impatient shout, etc.
Vary the
    speaker, wording, tone and intent widely. The text_instruction can briefly restate the
intent
    (e.g. "Go to the person who is calling you.").
    For Category 1 you MUST also fill "voice_style": one English sentence describing the
SPEAKER'S
    VOICE for downstream TTS - gender, approximate age, timbre, emotion/tone, speaking pace
, and
    optionally accent. Make these rich and varied across episodes.
  Category 2 - "ambient sound source + text command": the sound is a common non-speech
sound
    believable in this room (knocking, running water / flushing, dragging furniture, a pet
- dog
    barking / cat meowing, human chatter, music, TV, kettle whistle, phone ringing, etc.).
The
    text_instruction is a short English command telling the robot what to home in on, e.g.
"Find
    where the knocking sound is coming from." For Category 2, set "voice_style" to a short
English
    description of the sound's acoustic character or "" if not applicable.

Use the visible scene ONLY to make the sound realistic (bathroom -> water/flush; kitchen ->
clattering dishes / kettle; living room -> TV / music; bedroom -> alarm clock / phone).

HARD CONSTRAINTS:
- NEVER reveal the destination's location, room name, or nameable furniture in either
  sound_content or text_instruction. The robot must find it by sound, not by being told
where it
  is. (Do NOT say things like "in the bathroom" / "near the toilet".)
- sound_content, text_instruction and voice_style must be in ENGLISH.
- Keep each field to one short natural sentence.

Return ONLY a JSON object, no markdown fences, with exactly these keys:
{
  "category": 1 or 2,
  "sound_content": "<the audio content: cat.1 = the spoken line; cat.2 = a short sound
description>",
  "text_instruction": "<short English instruction given to the robot>",
```

```
  "voice_style": "<cat.1 = the speaker's voice; cat.2 = the sound's acoustic character or
empty>",
  "scene_reasoning": "<one short sentence: why this sound fits what you see>"
}
```

Table 5: Audio synthesis models used for sound-guided navigation data.

Sound category	Synthesizer	Key settings
Spoken utterance	Custom-voice TTS (Qwen3-TTS) (Hu et al., 2026)	7 speakers (4 F / 3 M), style-conditioned, ≤ 2048 tokens
Non-speech event	Seedance (Seedance et al., 2026)	clip 10 s

Table 6: Rendering and augmentation configuration for sound-guided navigation data.

Component	Setting
Source placement	episode endpoint (center of final destination view)
Distance attenuation	inverse distance $\propto 1/r$, floor $r_{\min} = 0.3$ m, no peak normalization
FOA format	first-order ambisonics, ACN/SN3D, 4 channels, order $[W, Y, Z, X]$
Listener / camera height	1.25 m
History frames	up to 8 (uniformly sampled)
Waypoints	$K = 8$, $[x_{\text{fwd}}, y_{\text{left}}, \psi]$ relative to current pose
Room reverberation	R2R / RxR: image-source early reflections + cross-room attenuation (MP3D)
	ScaleVLN / VLN-PE: direct-path propagation only
Directional noise (optional)	1–3 plane-wave interferers (pink/white); el. $\pm 45°$; $\geq 30°$ from source
Noise SNR	sampled from $\{20, 15, 10, 5, 0\}$ dB w.r.t. the W channel

Audio synthesis. Spoken utterances are synthesized with a custom-voice text-to-speech model conditioned on the voice-style description, and non-speech events with Seedance, as summarized in Table 5.

FOA rendering and sample construction. The proposed source is placed at the episode endpoint (the center of the final destination view). At every decision frame, we transform the endpoint into the listener's current camera frame to obtain its forward/left/up direction components and distance r; the mono stem is inverse-distance attenuated ($\propto 1/r$, floored at $r_{\min} = 0.3$ m, with no per-clip peak normalization so that loudness preserves a distance cue) and encoded into 4-channel FOA via Eq. equation 1. Trajectory targets are the next $K = 8$ waypoints $[x_{\text{fwd}}, y_{\text{left}}, \psi]$ (forward metres, left metres, yaw radians) expressed relative to the current pose. Key constants are listed in Table 6.

Acoustic augmentation. For corpora with room geometry (R2R and RxR on Matterport3D), the shared renderer adds first-order image-source early reflections and attenuates direct sound across rooms (Allen & Berkley, 1979), and RxR further adds low-level directional background noise; geometry-free corpora (ScaleVLN on HM3D, VLN-PE) use direct-path propagation with distance attenuation only. On top of this, we provide an optional directional-noise augmentation that injects 1–3 plane-wave interferers (pink or white noise) at random azimuths (elevation within $\pm 45°$, at least $30°$ from the true source direction), mixed at a signal-to-noise ratio sampled per clip from $\{20, 15, 10, 5, 0\}$ dB referenced to the omnidirectional (W) channel (Table 6).

Omni-modal SFT format and CoT. Each SFT example interleaves up to eight uniformly sampled history frames, the current frame, and the co-located 4-channel FOA clip, followed by the task text; the target is the $K = 8$ waypoint list. A vision–language model then writes a first-person CoT rationale conditioned on the ground-truth source bearing (bucketed into eight directions with $22.5°/67.5°/112.5°/157.5°$ edges), distance (near < 1.5 m, far > 4 m), and two trajectory-derived meta-actions, without altering the trajectory targets. We additionally construct in-place turning examples (sources to the side or behind the agent). Active-stop supervision converts stationary end-of-episode waypoints into a fixed eight-slot trajectory carrying a majority of `<stop>` tokens (≥ 5 of 8). The eight-way direction bucket, the near/far distance bucket, and the two trajectory-derived meta-actions are computed deterministically from geometry and injected into the CoT prompt as ground truth; the prompt template is shown below.

Navigation chain-of-thought — system and task prompt

```
[SYSTEM]
You are writing first-person chain-of-thought (CoT) reasoning traces AS a sound-source
navigation
robot. You (the robot) hear a spatial-audio cue and must move toward the sound source while
avoiding obstacles. You are given YOUR OWN recent egocentric RGB frames (oldest to newest,
the
last one is your current view), the task the sound is asking you to do, the GEOMETRICALLY
COMPUTED
direction and distance of the sound source relative to you, the ground-truth meta-action
sequence,
and the ground-truth future trajectory. Write a single coherent, physically plausible
reasoning
trace, STRICTLY IN THE FIRST PERSON (use 'I', 'me', 'my', 'ahead of me'; NEVER say 'the
robot').
The computed direction/distance and the meta-action sequence are the SOURCE OF TRUTH: never
contradict them. Describe the visual scene ONLY from what you actually see in the frames.

[TASK]
Write a first-person chain-of-thought reasoning trace for my next navigation step.

=== Task the sound is asking me to do ===
{instruction}

=== Computed sound-source cue (ground truth -- do NOT contradict) ===
- Direction of the sound source relative to me: {direction}  (azimuth {az:+.0f} deg, where
+ is to
  my left, 0 is straight ahead)
- Distance: {dist:.1f} m ({dist_bucket})
- Elevation: {el:+.0f} deg (roughly at my ear level unless large)

=== Ground-truth motion (derived from the future trajectory -- do NOT contradict) ===
- My two main meta-actions (first half of my path, then second half): {meta_seq}
- Net motion over this horizon: {net_fwd:+.2f} m forward, {net_left:+.2f} m to my left; my
heading
  changes by {net_yaw:+.0f} deg.
- Overall travel direction of my path: {path_dir}
- {deviation_note}

=== What to write ===
Produce ONE first-person reasoning trace (a few sentences, natural English, no bullet
headers, no
markdown) that flows through these stages IN ORDER:
1. SCENE: briefly describe what I see in front of me / around me in my current view.
2. INSTRUCTION: restate, in one clause, what the sound is asking me to do.
3. SOUND DIRECTION: state where I hear the sound source ({direction}, {dist_bucket}) and
tie it to
   what I see.
4. PLAN: give a short high-level plan that reconciles what I see (obstacles/openings) with
where I
   hear the sound. If my path direction differs from the straight line to the source,
explain it as
   going around an obstacle or leaving my current room; if the source is far, consider it
may be in
   another room.
5. DECISION: conclude in ONE short first-person sentence that commits to my two sequential
moves --
   what I do over the first half of my path, then the second half -- consistent with: {
meta_seq}.
   Do NOT write a "Meta actions:" label, do NOT enumerate sub-moves, and do NOT output
coordinates.

Write everything in the FIRST PERSON. Do NOT refer to "the robot" in the third person. Do
NOT
output the numeric trajectory or any coordinates. Stop at the meta-action level.

Output STRICT JSON and nothing else, exactly:
{"cot": "<the first-person reasoning trace as a single string>"}
```

A.2 Benchmark Details

A.2.1 OmniEchoBench-QA: Real-World Capture and Annotation

Capture rig and formats. Each script is performed and recorded live, producing three synchronized streams in a single take: a first-person main-view video (1920 px wide, 30 fps), a 360° panoramic

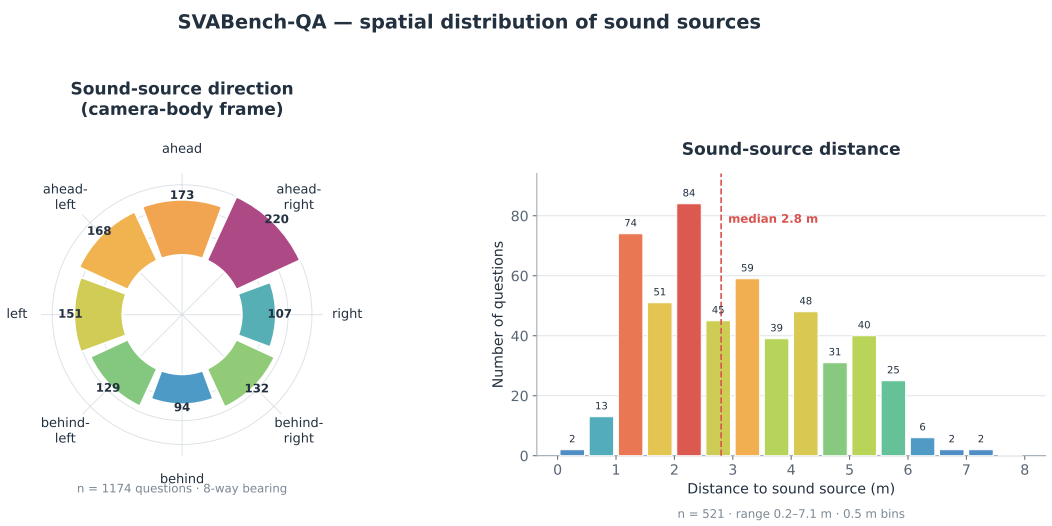


Figure 8: Direction and Distance distribution of OmniEchoBench-QA.

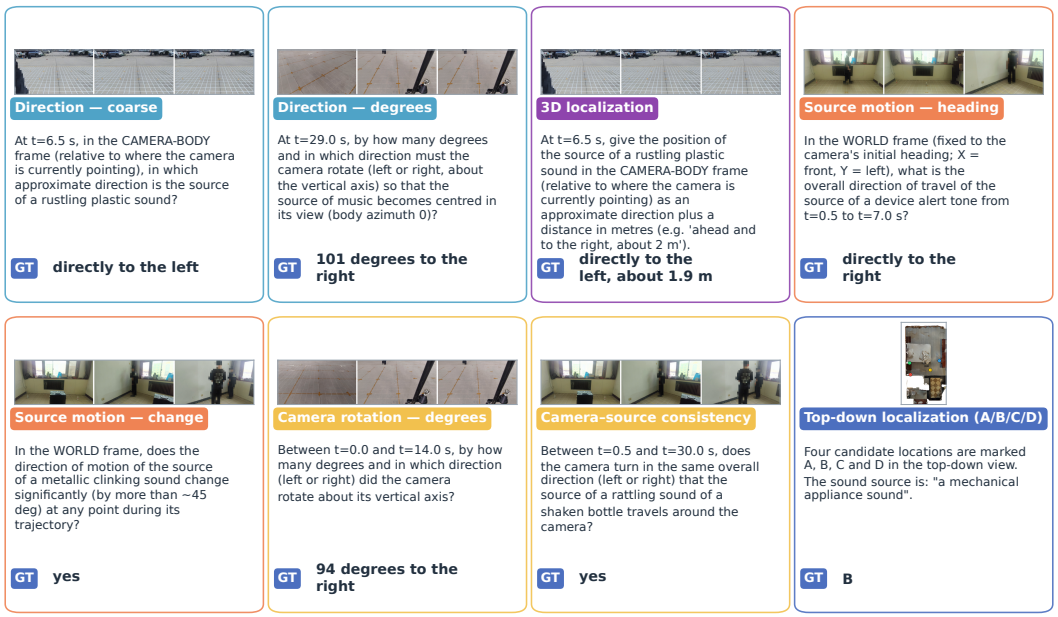


Figure 9: Representative examples from OmniEchoBench-QA, covering eight spatial reasoning question types.

video, and a four-channel FOA audio track recorded at 48 kHz. The FOA audio is stored with the omnidirectional channel first, and its horizontal axes are aligned so that the front direction corresponds to the $+X$ axis and the listener's left to the $+Y$ axis. The 197 scripts are distributed across small indoor (50), medium indoor (50), large indoor (77), and open outdoor (20) settings.

Annotation. Using the panoramic recording as ground-truth reference, annotators label the real-time position of every active sound source at 2 Hz (every 0.5 s), together with per-timestep flags for whether the source is visible in the main view, visible in the panorama, and audible. The panorama supplies the "present but off-screen" evidence that the main view cannot, and is used only during annotation. Two capture regimes are used: a *camera-fixed* regime (the original FOA is kept) and a *camera-moving* regime, in which the camera yaw is recovered by optical flow and calibrated against annotated rotation anchors, and the FOA is rotated into the current camera frame so that audio and

Table 7: OmniEchoBench-QA temporal-reasoning categories, subtypes, coordinate frames, and counts (from `spatial_qa_test`). The four geometric categories (2,372 QA) are reported in the main paper; Visibility is additionally annotated.

Category	Subtype(s)	Frame	Count
Source Direction	approx. 8-way / degrees	camera-body	1,069
3D Localization	bearing + distance	camera-body	521
Source Motion	world heading / change	world	473
Camera Rotation	degrees / consistency	camera / mixed	309

image remain aligned. At test time, the model receives the main-view video (or a single frame) muxed with the processed four-channel FOA.

Question construction. All ground-truth answers are computed deterministically from the annotated trajectories rather than from a language model: at a queried time we compute the source's world and camera-body bearing, its distance, and its visibility; over an interval we compute changes from the same trajectories. Answers are short phrases drawn from closed buckets (eight-way direction words, *N degrees to the left/right*, yes/no, or a direction plus a metric distance), and per-(category, subtype) answer-bucket balancing is applied. The resulting temporal-reasoning categories and counts are listed in Table 7. The main paper reports the four geometric categories (Direction, 3D Localization, Motion, Camera Rotation; 2,372 QA); a Visibility category is additionally annotated and available. The distribution of direction and distance is shown in Figure 8. The Cognitive Map subset (600 QA) is built on the navigation scenes below. Example QAs from OmniEchoBench-QA are shown in Figure 9.

Coordinate conventions. The world frame is fixed to the camera's initial heading, with $+X$ ahead and $+Y$ to the left; the camera-body frame at time t rotates with the recovered camera yaw $\theta(t)$, and body azimuth equals world azimuth minus $\theta(t)$. Azimuth is measured counter-clockwise (0° ahead, 90° left, 180° behind, 270° right).

A.2.2 OmniEchoBench-Nav: Real-World Scene and Acoustic-Field Collection

Scenes. The benchmark contains 30 real-scanned indoor environments, split into 10 single-room and 20 multi-room (cross-room) layouts spanning two to four connected rooms. Each scene is provided as a textured mesh (`.glb`) together with a top-down render and a pixel-to-mesh affine calibration; walkability, collision checks, and shortest-path planning are provided by an occupancy-grid navmesh derived from the mesh.

Acoustic field. Each scene contains 30 navigation samples (900 in total). In every sample a single mono source is placed at a known 3D position, and its signal is captured at a dense grid of 12–17 receiver ("microphone") positions on an approximately 1 m grid at $\sim$ 1 m height, each stored as a world-aligned four-channel FOA recording (ACN/SN3D, order $[W, Y, Z, X]$, 48 kHz), forming a spatially dense acoustic field of the environment. Every receiver's coordinates are stored explicitly. Sources span non-speech environmental events (505 samples) and spoken commands (395 samples), across a set of playback devices; each sample is annotated with the source device, source room, semantic description, spoken-command text (when applicable), speaker description, and a natural-language navigation-target description. Table 8 summarizes the composition.

Test-time perception. At each step, the agent is served the FOA recording from the receiver nearest to its current position. Although the recordings are stored in a world-aligned coordinate frame, their horizontal directional components are rotated into the agent's current egocentric frame according to the camera yaw before being passed to the model. The visual observation is rendered from the same pose and orientation ($+x$ forward, 90° horizontal FOV, 15° downward pitch), ensuring that the FOA forward direction remains aligned with the camera view as the agent turns. The Cognitive Map QA subset (600 four-option A/B/C/D questions) is built on these scenes: it pairs the egocentrically aligned nearest-receiver FOA with a top-down map on which the true source and three distractor floor points (each $\geq 40^\circ$ apart in bearing) are marked, and asks which marked location is the source. In the ego-up variant the map is rotated so image-up is the robot's forward direction and the robot is centered.

Table 8: OmniEchoBench-Nav composition. 30 scenes $\times$ 30 samples $=$ 900 navigation samples; each sample provides a dense FOA receiver grid.

Property	Value
Scenes	30 (10 single-room, 20 multi-room / 2–4 rooms)
Navigation samples	900 (30 per scene)
Receivers per sample	12–17 (dense $\sim$1 m grid, $\sim$1 m height)
Source types	non-speech 505, spoken command 395
FOA format	4-channel, ACN/SN3D $[W, Y, Z, X]$, 48 kHz, world-aligned
Scene assets	textured `.glb` mesh + top-down render + occupancy navmesh
Cognitive-Map QA	600 four-option (A/B/C/D) questions over all 30 scenes

Table 9: FOA encoder architecture ($\sim$ 11.9 M parameters). The encoder is shared across all three training stages and frozen at Stage 3.

Module	Configuration
Input feature map	$5 \times 128 \times T$: log-mel of W + DirAC intensity (i_x, i_y, i_z) + diffuseness
STFT front-end	16 kHz, FFT 400, hop 160, 128 mel bins
Conv. frontend	3$\times$ Conv2d (stride 2, GELU), 8$\times$ time downsampling ($\sim$12.5 tokens/s)
Transformer	5 layers, $d_{\text{model}} = 384$, 6 heads, FFN 1536, pre-norm, sinusoidal PE
Semantic head (Stage 1)	learnable-query attention pooling $\rightarrow$ 512-d clip embedding
Text encoder (frozen)	CLIP ViT-B/16 (Radford et al., 2021), 512-d joint space, $C = 644$ sound labels
Localization head (Stage 2)	query-conditioned cross-attention $\rightarrow$ (az, el, dist)

A.2.3 Data Privacy and De-identification

Because OmniEchoBench contains audiovisual recordings captured in real-world environments, we process the data to minimize the disclosure of personally identifiable information before any sample is included in the benchmark or prepared for release. In all visual streams, every visible face and human figure, including incidental bystanders, is mosaic-blurred. Only these de-identified visual streams and privacy-preserving audio recordings are included in the benchmark and prepared for release. The released data are intended for research on spatial audio–visual perception and navigation, not for identity recognition, person re-identification, speaker identification, or other biometric analysis.

A.3 Training Details

OmniEcho is trained in three stages: two stages pretrain the standalone FOA encoder (semantic alignment, then query-conditioned localization), and a third stage integrates the frozen encoder into the Qwen3-Omni backbone via a trainable projector. The FOA encoder is a compact ($\sim$ 11.9 M-parameter) convolution–Transformer network; its architecture is summarized in Table 9.

Stage 1: Semantic alignment of the FOA encoder. We first pretrain a lightweight FOA audio encoder f_θ that maps an FOA audio clip to a temporal sequence of d-dimensional tokens ($d = 384$). Each clip is converted into a 5-channel input map $\mathbf{X} \in \mathbb{R}^{5\times128\times T}$ comprising the log-mel spectrogram of the omnidirectional (W) channel together with the three active-intensity components (i_x, i_y, i_z) and the diffuseness δ, all in the ACN/SN3D convention. A convolutional frontend downsamples time by $8\times$ (yielding $\sim$12.5 tokens/s), and a 5-layer Transformer produces token embeddings. An attention-pooling semantic head aggregates these tokens into a clip embedding $\mathbf{a} \in \mathbb{R}^{512}$.

To inherit an open-vocabulary sound-semantic space, we align the clip embedding with frozen CLIP text embeddings using the balanced sigmoid objective defined below. This balancing is important because positive audio–label pairs constitute only $\sim$0.3% of all clip–label pairs in our training setup. Stage 1 is trained on a 100k-clip corpus of synthetic static FOA scenes (1–4 sources, 2–8 s), rendered with an order-1 spherical-harmonic encoder from mono stems at known azimuth, elevation, and distance.

Balanced semantic alignment objective. Let $\mathbf{a}_i \in \mathbb{R}^{512}$ denote the audio embedding of clip i, and let $\mathbf{t}_j \in \mathbb{R}^{512}$ be the frozen CLIP text embedding of label j, where $j \in \{1, \dots, C\}$ and $C = 644$. Before similarity computation, both embeddings are ℓ_2-normalized:

$$\tilde{\mathbf{a}}_i = \frac{\mathbf{a}_i}{\|\mathbf{a}_i\|_2}, \qquad \tilde{\mathbf{t}}_j = \frac{\mathbf{t}_j}{\|\mathbf{t}_j\|_2}. \tag{4}$$

We then form sigmoid logits as

$$s_{ij} = \alpha\, \tilde{\mathbf{a}}_i^\top \tilde{\mathbf{t}}_j + b, \qquad \alpha = \exp(\gamma), \tag{5}$$

where α is a learnable logit scale and b is a learnable bias. For clip i, let $\mathcal{P}_i \subseteq \{1, \dots, C\}$ be the set of present labels and let $\mathcal{N}_i = \{1, \dots, C\} \setminus \mathcal{P}_i$ be the set of absent labels. We use a per-clip class-balanced sigmoid loss,

$$\mathcal{L}_{\text{sem}}^{(i)} = \frac{1}{2}\left(\frac{1}{|\mathcal{P}_i|} \sum_{j \in \mathcal{P}_i} \text{BCE}(s_{ij}, 1) + \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \text{BCE}(s_{ij}, 0) \right), \tag{6}$$

and average it over the mini-batch:

$$\mathcal{L}_{\text{sem}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{sem}}^{(i)}. \tag{7}$$

Here

$$\text{BCE}(s, y) = -\big[y \log \sigma(s) + (1 - y) \log(1 - \sigma(s)) \big]. \tag{8}$$

The same $\mathcal{L}_{\text{sem}}$ is used in both Stage 1 and Stage 2.

Stage 2: Query-conditioned spatial localization. Initialized from Stage 1, the encoder is jointly fine-tuned for sound-source localization. Given a text query embedding $\mathbf{q}$ describing one source present in the clip, a localization head attends over the encoder tokens via query-conditioned cross-attention and regresses the source direction and range. Rather than predicting azimuth a and elevation e directly, the head emits their $(\sin, \cos)$ encodings together with a log-distance, i.e., a raw 5-vector $(\hat{s}_a, \hat{c}_a, \hat{s}_e, \hat{c}_e, \hat{\ell})$. The two angular pairs are ℓ_2-normalized before the loss is applied, which avoids the $\pm 180^\circ$ wrap-around discontinuity.

Stage 2 minimizes a localization regression loss while retaining the same semantic objective $\mathcal{L}_{\text{sem}}$ as a regularizer to prevent semantic drift:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{ang}} + \lambda_d\, \mathcal{L}_{\text{dist}} + \lambda_s\, \mathcal{L}_{\text{sem}}, \qquad \lambda_d = \lambda_s = 0.5. \tag{9}$$

The angular term is the mean-squared error between the normalized predicted $(\sin, \cos)$ pairs and the ground-truth pose (a, e):

$$\mathcal{L}_{\text{ang}} = \Big\| \widehat{\mathbf{u}}_a - (\sin a, \cos a) \Big\|_2^2 + \Big\| \widehat{\mathbf{u}}_e - (\sin e, \cos e) \Big\|_2^2, \tag{10}$$

with

$$\widehat{\mathbf{u}}_a = \frac{(\hat{s}_a, \hat{c}_a)}{\|(\hat{s}_a, \hat{c}_a)\|_2}, \qquad \widehat{\mathbf{u}}_e = \frac{(\hat{s}_e, \hat{c}_e)}{\|(\hat{s}_e, \hat{c}_e)\|_2}. \tag{11}$$

The range is supervised in the log domain with an ℓ_1 term against the ground-truth distance r:

$$\mathcal{L}_{\text{dist}} = \big|\, \hat{\ell} - \log r \,\big|. \tag{12}$$

Both encoder stages are trained on the same 100k-clip synthetic FOA corpus; the optimization settings are listed in Table 10.

Figure 10 shows the encoder pretraining dynamics. In Stage 1, the semantic loss drops to ~ 0.23 while open-vocabulary retrieval Top-1 rises to $\sim 37\%$ over 644 labels; in Stage 2, the mean azimuth error falls to $\sim 16^\circ$, elevation error to $\sim 7^\circ$, and distance error to ~ 1.1 m, confirming that the compact encoder becomes both semantically grounded and spatially discriminative before it is frozen.

Table 10: FOA encoder pretraining hyperparameters (Stages 1 and 2).

Setting	**Stage 1 (semantic)**	**Stage 2 (localization)**
Training corpus	100k synthetic FOA clips, 1–4 sources, 2–8 s	
Objective	SigLIP (balanced) (Zhai et al., 2023)	$\mathcal{L}_{\text{ang}} + 0.5\mathcal{L}_{\text{dist}} + 0.5\mathcal{L}_{\text{sem}}$
Initialization	from scratch	warm-start from Stage 1
Optimizer	AdamW ($\beta = 0.9, 0.95$), weight decay 0.05, grad clip 1.0	
Peak / min LR	3e-4 / 1e-5	1e-4 / 1e-6
Schedule / warmup	cosine, warmup fraction 0.05	
Batch size (per GPU)	32	32
Epochs / GPUs	10 / 2	10 / 2
Precision	AMP	

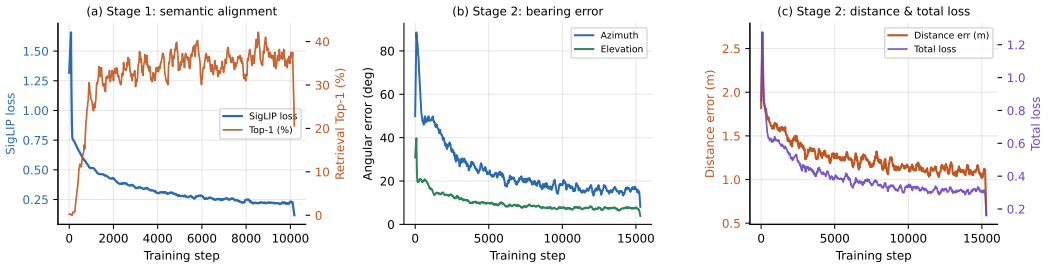


Figure 10: FOA encoder pretraining curves (smoothed). (a) Stage 1 semantic alignment: SigLIP loss and open-vocabulary retrieval Top-1 accuracy. (b) Stage 2 query-conditioned localization: mean azimuth and elevation error. (c) Stage 2 distance error and total loss.

Stage 3: Integration into the Omni backbone. We graft the frozen FOA encoder into Qwen3-Omni-30B-A3B (a Mixture-of-Experts model with $\sim$ 3B active parameters). Each FOA clip is processed along two paths: its W channel is down-mixed to mono and fed to the original frozen audio tower to preserve the native semantic audio tokens, while the full four-channel clip is passed through the frozen encoder to produce spatial tokens. A trainable projector (LayerNorm, then $384 \to 2048 \to 2048$ with GELU) maps the encoder tokens into the language embedding space; the spatial tokens are time-resampled by linear interpolation to match the co-located audio span's token count, and are written into reserved `<foa_pad>` placeholder rows bracketed by learnable `<foa_bos>`/`<foa_eos>` tokens with no vocabulary expansion. Only the projector and the language-model parameters are updated; the FOA encoder, audio tower, and visual tower remain frozen. The final training composition is summarized in Table 12.

A.4 Evaluation Details

A.4.1 Spatial Audio-Visual QA Evaluation

Because model outputs are short free-form phrases, we evaluate each prediction with a two-stage protocol that combines lightweight LLM-based extraction with deterministic rule-based judging. In the first stage, the model output is first stripped of any `<think>` reasoning, and an extractor LLM normalizes the remaining free text into a small set of structured fields: `yesno` (`yes`/`no`), `dir_set` (a subset of {`left`, `right`, `front`, `back`}), `degrees` (float), `distance_m` (float, in meters), `letter` (`A`/`B`/`C`/`D`), and a free-text `answer`. Extraction is performed with a text LLM, for which we use `qwen3.7-flash`. The extractor is queried with a fixed system prompt together with a user message containing only the question and the model answer. The LLM is used solely for answer normalization, not for judging correctness; final correctness is determined entirely in the subsequent rule-based stage, which keeps scoring fully reproducible.

In the second stage, correctness is decided by fixed rules through comparison with the gold reference whenever a task-specific rule applies, as shown in Table 13. Multiple-choice questions (*Cognitive Map*) require exact letter match. Directional questions use combinations of the four axis tokens

Table 11: Stage 3 integration hyperparameters.

Setting	Value
Base model	Qwen3-Omni-30B-A3B (Xu et al., 2025b) (MoE, $\sim$3B active)
Trainable / frozen	FOA projector + LLM / FOA encoder, audio tower, visual tower
Hardware	32$\times$ A100 (4 nodes $\times$ 8), $\sim$4 days
Parallelism	FSDP2, Ulysses seq.-parallel $= 4$, expert-parallel $= 8$
Precision	bf16 mixed precision, gradient checkpointing
Global / micro batch size	128 / 4
Max sequence length	32,768
Optimizer	AdamW, weight decay 0.1, max grad norm 1.0
Peak / min LR	7e-6 / 7e-7
Schedule / warmup	cosine, warmup ratio 0.033
Epochs	2

Table 12: Training data composition. The navigation examples are sampled from much larger corpus pools with probability $\sim$0.3 after action rebalancing.

Source	Full pool	Used (final mix)
Spatial audio-visual QA	—	26,585
Spatial audio QA	-	18,518
ScaleVLN (Wang et al., 2023)	403,215	139,142
RxR (Ku et al., 2020)	324,342	92,418
R2R (Anderson et al., 2018)	112,448	31,577
VLN-PE (Wang et al., 2025)	78,006	28,969
In-place rotation data (from ScaleVLN)	—	25,984
Total	—	363,193

`front`/`back`/`left`/`right`, giving eight-way resolution; a prediction is correct only if its direction set exactly matches the ground-truth set, with the opposite direction absent. Angular questions require the reported magnitude to fall within a tolerance of $\pm 45^\circ$ and the turn direction to be correct. 3D localization questions require the reported distance to fall within ± 2 m and the bearing to be correct. Motion questions are scored using the corresponding direction or degree rule depending on subtype. Overall accuracy is computed as the plain average of correctness over all evaluated samples (micro-average), i.e., the total number of correct predictions divided by the total number of samples.

Extractor system prompt

```
You normalize a model's free-text answer for a spatial-audio video QA benchmark. You are
given the QUESTION and the MODEL_ANSWER (which may contain reasoning; the real answer is
usually its last line). Extract, FROM THE MODEL_ANSWER ONLY, [yesno / dir_set / degrees /
distance_m / letter / answer]. Map ahead/forward$\rightarrow$front, behind/rear/backward$\
rightarrow$back. For multiple choice, pick the model's FINAL choice, not options it merely
discusses. Return ONLY a JSON object.
```

Table 13: QA scoring rules and tolerances by category.

Category	Correctness rule
Cognitive Map (A/B/C/D)	exact letter match
Visibility	yes/no matches ground-truth bucket
Source Direction (8-way)	axis set equals GT set, opposite absent
Direction / Camera Rotation (degrees)	$\lvert\Delta\theta\rvert \leq 45^\circ$ and correct turn direction
3D Localization	$\lvert\Delta\text{dist}\rvert \leq 2$ m and correct bearing
Source Motion	direction / degrees rule per subtype

A.4.2 SOUND-GUIDED NAVIGATION EVALUATION

Closed-loop setup. Navigation is evaluated in the Habitat simulator using the same episode set (identical start and goal) as the text-instruction baselines. In the sound-guided setting, the agent receives FOA observations together with a short task instruction that identifies the sound event to follow but does not reveal the destination or provide route-level spatial guidance. At each decision step, the agent's live pose selects the nearest receiver in the acoustic field, and the recording's horizontal directional components are remapped into the agent's egocentric frame before being passed (truncated to 10 s) to the model server. The model is queried once per decision step and returns $K = 8$ relative waypoints $[x_{\text{fwd}}, y_{\text{left}}, \psi]$; these are converted to discrete actions and executed before the next replan. Evaluation is parallelized by sharding scenes across GPUs, each running an independent model server and worker, and per-shard metrics are merged by an episode-weighted average.

Action space and caps. The discrete action space is {`stop`, `move_forward`, `turn_left`, `turn_right`} with a forward step of 0.25 m and a turn angle of 15°. A waypoint segment is converted to round(Δpos/0.25) forward steps when its displacement dominates, otherwise to round($|\Delta\psi|/15°$) turn actions. Episodes are capped at 50 model decisions (500 executed steps), and a stuck detector forces termination if net displacement over a 16-step window stays below 0.25 m.

Stopping condition. All reported navigation experiments use agent self-stop. The model must terminate on its own: a `stop` is triggered when it emits an empty trajectory or when the maximum waypoint displacement is below 0.3 m with negligible yaw (pure in-place turns are executed, not treated as arrival).

Metrics. We report the metrics with a 1 m success radius in OmniEchoBench-Nav and 3 m success radius in VLN-CE R2R val-unseen, defined in Table 14. NE and TL are averaged with `nanmean` to exclude errored episodes.

Table 14: Navigation metrics The success radius r_s is 1 m for OmniEchoBench-Nav and 3 m for VLN-CE R2R val-unseen, the same as in prior work.

Metric	Definition
SR ↑	success rate: agent stops within r_s of the source
SPL ↑	success weighted by (shortest path)/(max(path, shortest path))
NE ↓	navigation error: final distance to goal (m)
OS ↑	oracle success: ever within r_s along the trajectory
TL	trajectory length: total path length (m)

A.5 FURTHER EXPERIMENTS

A.5.1 SIM2REAL ANALYSIS

Table 15: **Sim-to-real transfer results.** OmniEchoBench is partitioned into two scenario-disjoint subsets, with 1,000 samples allocated for training and the remaining samples reserved for testing. Fine-tuning on the real-data subset improves performance on the held-out subset across the evaluated task categories, suggesting that real-data adaptation can help address limitations of training solely on synthetic data.

Training with real-world data	Camera Rotation	Source textbfDirection	3D Localization	Source textbfMotion	Cognitive Map	Overall
✗	41.7	22.6	14.2	24.7	47.5	28.5
✓	44.3	31.0	21.7	35.6	48.0	34.9

Table 15 presents the sim-to-real transfer results on OmniEchoBench, where the benchmark is split into training and test subsets. Fine-tuning with real-world data yields a clear improvement in overall performance, raising the average score from 28.5 to 34.9. Performance improves across all task categories, including Camera Rotation (+2.6), Direction (+8.4), 3D Localization (+7.5), Motion (+10.9), and Cognitive Map (+0.5). These results suggest that even a limited amount of real-world

data can substantially enhance the model's spatial-audio reasoning and generalization ability in real environments. More importantly, these results are consistent with a noticeable sim-to-real gap in OmniEchoBench for certain question types, while yielding smaller adaptation gains on other tasks (e.g., Camera Rotation and Cognitive Map).

A.5.2 VLN Failure Modes

Table 16: OmniEcho on OmniEchoBench-Nav.

Aspect	Metric	Value
Trajectory	Success Rate (SR)	16.22%
	SPL	11.54%
	Oracle Success (OS, once $\leq$1 m)	52.89%
	Navigation Error (NE)	4.09 m
	Trajectory Length (TL)	7.02 m
Direction (coarse)	Step-wise closer rate	60.2%
	Net-approach episode rate	64.7%
	Mean net approach	+2.07 m
Distance	MAE $\lvert\text{pred} - \text{true}\rvert$	2.20 m
	Median AE	1.61 m
	Bias (pred$-$true)	+0.70 m
	Pearson r	0.27

Table 17: Episode outcome breakdown on OmniEchoBench-Nav (900 episodes).

Outcome	Rate	Note
Successful termination within the goal region ($\leq$1 m & stopped)	16.22%	—
Entered the goal region without successful termination	36.67%	OS$-$SR; active-stop failure
Never entered the goal region ($>$1 m all steps)	47.11%	100$-$OS; navigation failure

Failure-mode analysis. On OmniEchoBench-Nav, the agent moves closer to the source on 60.2% of decision steps and achieves a net reduction in source distance in 64.7% of episodes. These statistics indicate a tendency to approach the target, although they do not directly measure sound-direction accuracy. As shown in Table 17, 52.89% of episodes enter the 1 m goal region at least once, whereas only 16.22% terminate successfully. Specifically, 36.67% enter the region without successful termination, and 47.11% never enter it. Among episodes that enter the goal region, approximately 30.7% ultimately succeed. This reveals a substantial gap between visiting the target vicinity and completing the task.

The outcome breakdown highlights limitations in both target approach and successful termination. Table 16 additionally reports a distance-estimation MAE of 2.20 m and a Pearson correlation of 0.27, indicating limited reliability in estimating target proximity. Under our execution protocol, multiple actions are executed between model decisions, so an agent may enter and leave the goal region before its next opportunity to select a stop. Consequently, the OS–SR gap cannot be attributed exclusively to incorrect stopping decisions. More reliable proximity estimation, adaptive replanning near the target, and a history of recent acoustic observations are promising directions for improving task completion; their individual effects remain to be evaluated.

A.5.3 VLN with GT direction

To examine how ground-truth directional guidance affects closed-loop navigation, we evaluate a GT-direction variant of OmniEcho that receives the target's ground-truth egocentric direction at each decision step. The action-execution and agent self-stopping protocols remain unchanged. As shown in Table 18, this variant increases SR from 16.2% to 23.7% and SPL from 11.5% to 22.1%. However, OS decreases from 52.89% to 34.33%, while NE remains similar (4.09 m versus 4.14 m). These results are compatible because OS measures whether a trajectory ever enters the goal

Table 18: **Performance of OmniEcho with GT direction on our OmniEchoBench-Nav.** OmniEcho† receives the GT egocentric source direction at each decision step under the same action-execution and agent self-stopping protocol. The GT-direction variant achieves higher SR and SPL but lower OS, showing different effects on successful task completion and goal-region visitation.

Model	Base Model	Guidance	Val-Unseen				
			SR↑	SPL↑	NE↓	OS↑	TL
OmniEcho	Qwen3-Omni-30B-A3B	FOA audio	16.2	11.5	4.09	52.89	7.02
OmniEcho†	Qwen3-Omni-30B-A3B	GT direction	23.7	22.1	4.14	34.33	3.47

region, whereas SR additionally requires successful termination there. Among episodes that enter the region, the fraction ending successfully increases from approximately 30.7% to 69.0%. This is a descriptive comparison over the trajectories generated by each variant; it does not isolate stopping performance on a common set of reached episodes. The reduction in mean TL from 7.02 m to 3.47 m may reflect more direct successful trajectories, earlier termination of unsuccessful episodes, or both. Thus, the benefits of ground-truth directional guidance are primarily reflected in successful termination and path efficiency, rather than an increased likelihood of entering the goal region.

A.6 Visualization Examples

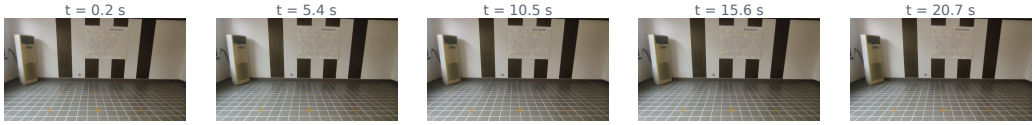
Figures 11–14 visualize representative predictions on OmniEchoBench-QA. The successful examples cover sound-source motion, 3D localization, and camera rotation, illustrating the model's ability to recover spatial states from FOA audio and visual observations. Figure 15 shows two failure examples. In the 3D localization example, OmniEcho captures the coarse direction but predicts the source as directly to the right rather than ahead-right and underestimates its distance. In the camera-rotation example, it identifies the correct rotation direction but substantially overestimates the angle. Together, these visualizations highlight the remaining gap between coarse spatial understanding and fine-grained geometric estimation.

Figures 16–19 visualize successful closed-loop episodes from four OmniEchoBench-Nav scenes. Each example pairs the global trajectory with intermediate egocentric observations, estimated sound directions, and planned actions, showing how spatial-audio cues are converted into scene-conditioned movement. Figure 20 provides a further visualization on a physical robot. Across successive steps, the robot updates the estimated direction and distance of an initially out-of-view alarm, plans around intervening obstacles using spatial audio and visual observations, and progressively approaches the source.

Sound-source motion 360° panorama (equirectangular) · t = 10.5 s



t = 0.2 s | t = 5.4 s | t = 10.5 s | t = 15.6 s | t = 20.7 s



Q: In the WORLD frame (fixed to the camera's initial heading; X = front, Y = left), what is the overall direction of travel of the source of a toy-car motor sound from t=0.5 to t=13.5 s?

GT		**directly to the right**
OmniEcho (Ours)	✓	**directly to the right**
Qwen3-Omni-30B	✗	directly ahead
Qwen2.5-Omni-7B	✗	directly to the left
SO-7B	✗	ahead

Figure 11: A successful sound-source motion example from OmniEchoBench-QA.

Sound-source motion 360° panorama (equirectangular) · t = 18.7 s



t = 0.2 s | t = 9.5 s | t = 18.7 s | t = 27.9 s | t = 37.1 s



Q: In the WORLD frame (fixed to the camera's initial heading; X = front, Y = left), what is the overall direction of travel of the source of a rustling plastic sound from t=12.0 to t=21.0 s?

GT		**directly to the left**
OmniEcho (Ours)	✓	**directly to the left**
Qwen3-Omni-30B	✗	directly behind
Qwen2.5-Omni-7B	✗	directly to the right
SO-7B	✗	behind

Figure 12: A successful sound-source motion example from OmniEchoBench-QA.



Figure 13: A successful 3D localization example from OmniEchoBench-QA.



Figure 14: A successful camera-rotation example from OmniEchoBench-QA.

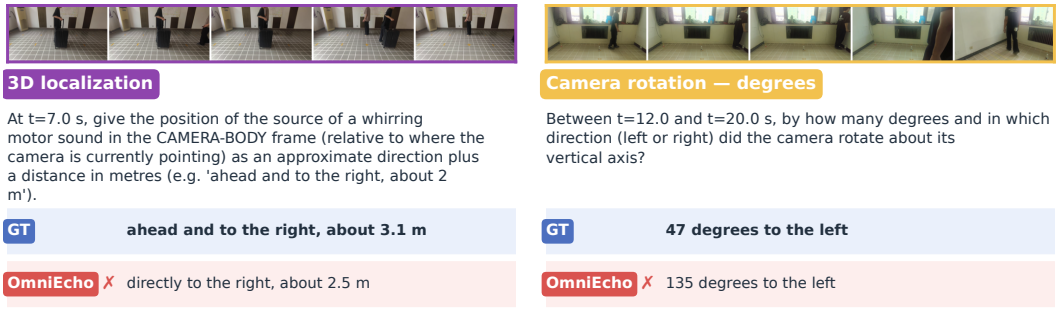


Figure 15: Selected failure cases on OmniEchoBench-QA. OmniEcho preserves part of the correct coarse orientation but misses finer geometric details: it predicts an ahead-right source as directly to the right and estimates a 135° leftward camera rotation instead of 47°.

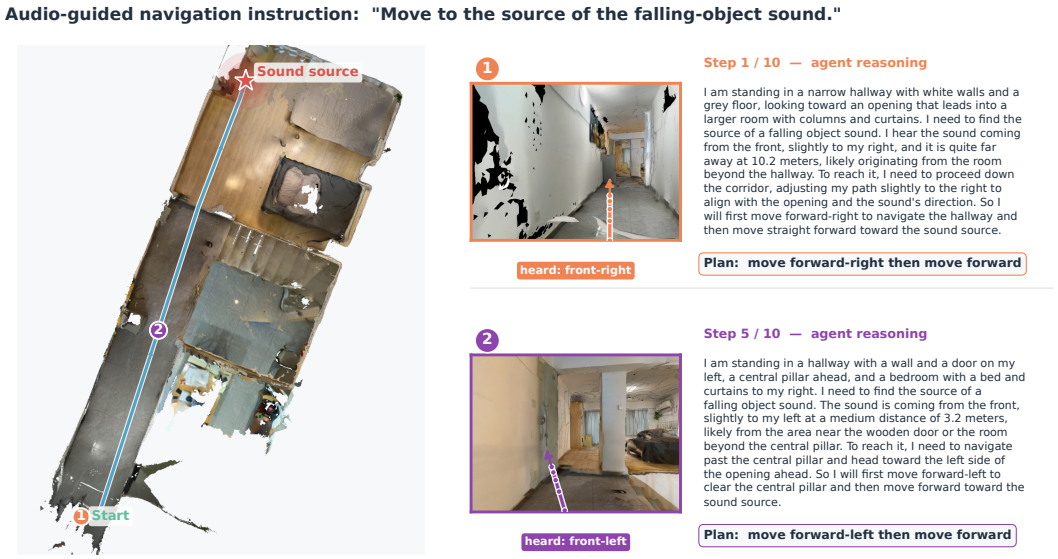


Figure 16: Selected successful navigation trajectories on OmniEchoBench-Nav.

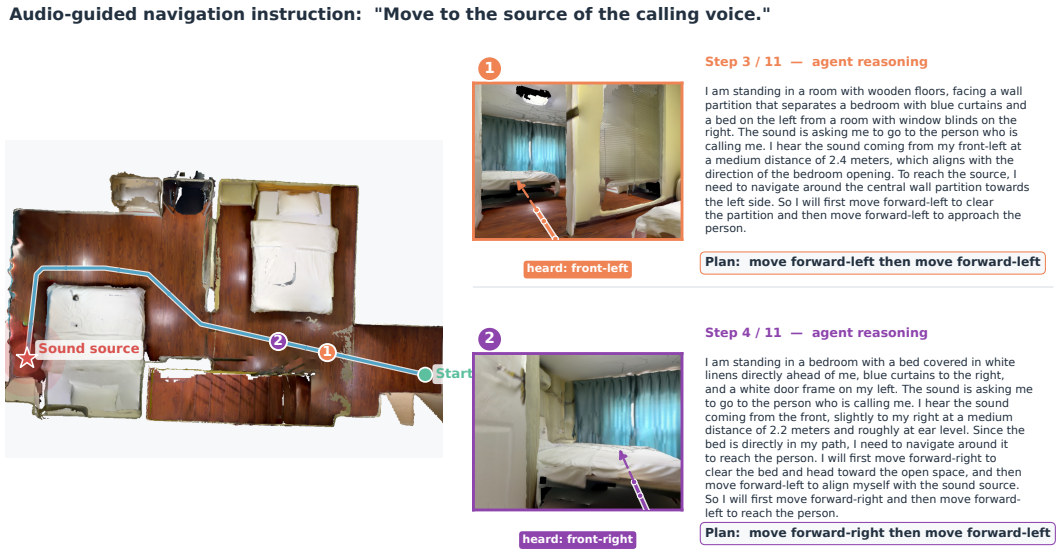


Figure 17: Selected successful navigation trajectories on OmniEchoBench-Nav.

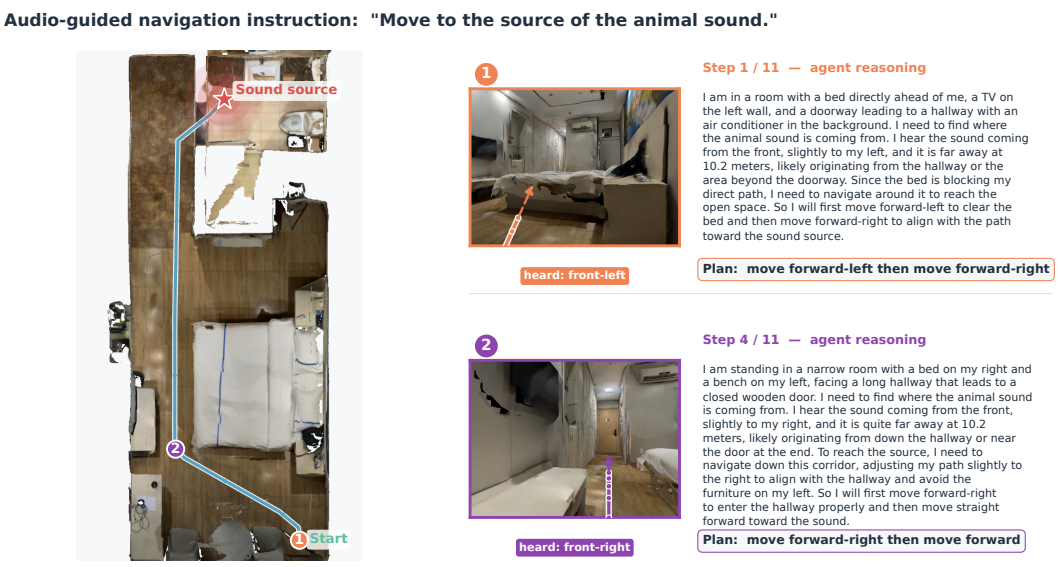


Figure 18: Selected successful navigation trajectories on OmniEchoBench-Nav.

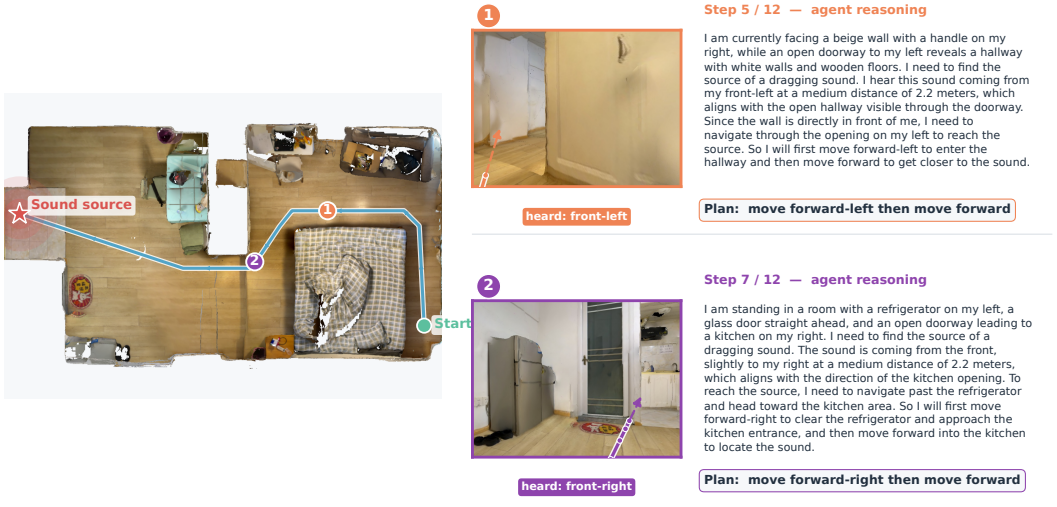


Figure 19: Selected successful navigation trajectories on OmniEchoBench-Nav.



Figure 20: Qualitative example of OmniEcho deployed on a physical robot for sound-guided navigation. The robot progressively localizes and approaches an initially out-of-view alarm by jointly reasoning over egocentric spatial audio and visual observations while navigating around obstacles.