HEARVGO: BENCHMARKING SOUND-CONDITIONED TRAJECTORY ADAPTATION FOR URBAN WALKING

Anonymous authors
Paper under double-blind review

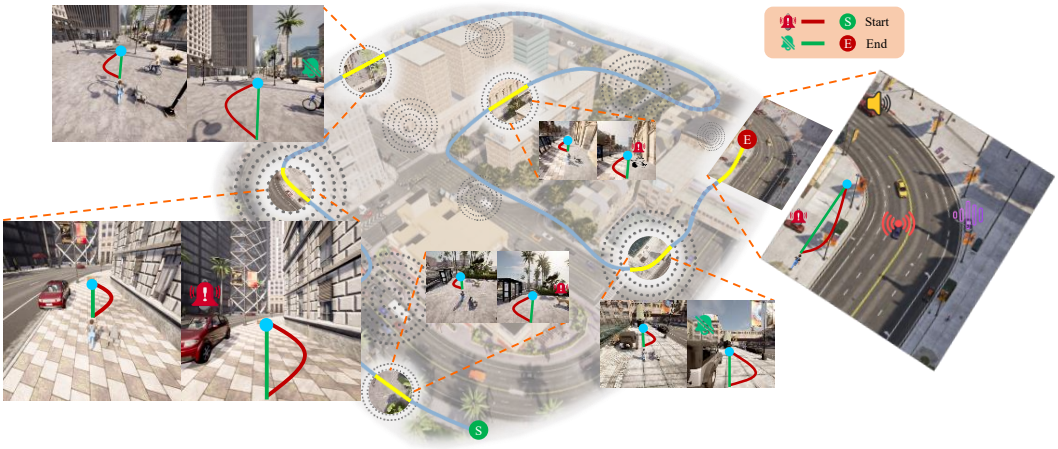


Figure 1: **Overview of HEARVGO.** An urban trajectory with diverse environment sounds, where highlighted local windows indicate segments potentially affected by sound. Given a goal ●, red paths (∼) indicate trajectory adaption *w.r.t.* sounds, while green paths (—) remain unchanged when sound has no effect.

ABSTRACT

Environmental sound can reveal nearby activity and potential hazards along a route. Yet outdoor navigation largely centers on vision, while audio-visual navigation often treats sound as a cue to the destination. How sound should instead guide route adjustments toward an independent goal remains underexplored. We introduce HEARVGO, a benchmark that tests this capability in omni models using brief multimodal observations. We augment CARLA with 1,288 environmental sound clips rendered as spatial audio and collect human walking demonstrations. The resulting 536 trajectory instances ask models to *Adapt* the path when sound warrants a change or *Preserve* the path otherwise. We evaluate 14 omni models under a common zero-shot protocol and complement the main comparison with controlled audio and vision ablations. These results show model-dependent gains from audio and few reversals of predicted avoidance direction when the left and right audio channels are exchanged. Some models retain substantial performance using only motion and the goal. Trajectory diagnostics show that goal progress can mask poorly formed waypoint sequences. Together, these results expose limits in the evaluated omni models' ability to determine when sound should alter a route and to express that decision as a coherent local trajectory. Code and data are publicly available: https://anonymous.4open.science/r/HearVGo.

1 INTRODUCTION

Safe outdoor navigation depends on scene geometry and the activity of objects along a route. Vision helps an agent locate objects and traversable space, and recent visual navigation models have made progress with these cues. Yet appearance may not reveal when an object becomes active or changes its behavior. Environmental sound can provide that missing evidence, but receives less attention in these models. When relevant, sound can prompt a path adjustment that creates clearance without changing the destination.

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.

Consider a pedestrian walking straight toward a fixed destination along a route that passes behind a stationary vehicle on the right. Passing behind it is reasonable and efficient while the vehicle remains still. If its reversing alarm sounds just before the pedestrian passes, the same route may no longer be appropriate. The pedestrian can move left to create more space, then continue toward the destination. The goal stays fixed, but the sound can change the local path.

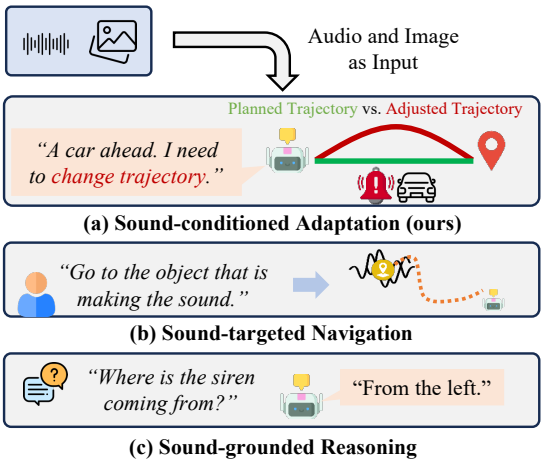


Figure 2: Comparison of how sound informs trajectory adaptation, navigation, and reasoning.

Existing work does not directly address this form of route adjustment. Outdoor navigation has largely relied on visual and geometric cues (Waghmare et al., 2025; Liu et al., 2025). Audio-visual navigation often seeks a sounding destination (Chen et al., 2020; 2021), while audio reasoning benchmarks ask for textual answers (Zheng et al., 2024; Sakshi et al., 2025). Figure 2 contrasts these tasks with ours, where sound may change the path toward an independent goal. A related study uses a jackhammer detector to adjust the safety margin of a control barrier function in a simulated construction site (Mootz & Akhavian, 2026). Its sound cue and control response are defined for that setting. Omni models can jointly process visual, acoustic, and textual inputs. Whether they can translate acoustic evidence into timely and coherent trajectory adaptation remains unclear.

Studying this question requires synchronized outdoor scenes, spatialized audio, and human walking trajectories. CARLA provides urban streets but lacks the diverse spatial sounds needed here (Dosovitskiy et al., 2017). We add 1,288 environmental audio clips from 32 classes in five groups, place their sources on semantically matched scene assets, and log each source's instance ID, position, and active state. A human operator controls a pedestrian toward destinations while viewing the scene and hearing the rendered audio. The recordings align RGB-D, binaural audio, actor poses, and source states. From them, we construct HEARVGO with 536 trajectory instances, including 379 `Adapt` cases requiring a sound-driven route change and 157 `Preserve` cases requiring no such change. Given recent multimodal observations and a goal independent of the sound source, an omni model predicts a local trajectory.

Our evaluation of 14 omni models on HEARVGO reveals two limitations in predicting a path toward an independent goal. **(1) Conditioning the future route on brief multimodal observations.** Some predicted paths progress toward the goal but show little lateral displacement toward the recorded human avoidance side. The effect of audio on Adapt and Preserve quality varies across models, and exchanging the left and right channels rarely reverses the predicted avoidance side in the five models tested. **(2) Forming a coherent local trajectory.** Several models progress toward the Local Goal but produce repeated or unevenly spaced waypoints or defer most movement to the final step. These results distinguish the ability to form a usable path from the ability to adjust that path when sound is relevant. Overall, this work makes three contributions.

- We extend CARLA with 1,288 environmental audio clips spanning 32 classes, semantically matched sound sources, and spatialized binaural rendering. This platform enables synchronized collection of visual observations, audio, actor poses, source states, and human walking demonstrations for studying sound informed outdoor navigation.
- We construct HEARVGO with 536 reviewed trajectory instances for local prediction toward a goal independent of the sound source. Its `Adapt` and `Preserve` cases test whether models modify the route when the acoustic context warrants a change and maintain it otherwise.
- We conduct a unified zero-shot evaluation of 14 omni models, complemented by controlled audio and vision interventions on selected models. The results expose inconsistent sound-conditioned route responses across models and trajectory degeneration in some models.

2 Related Work

Outdoor and audio-visual navigation. SANPO (Waghmare et al., 2025) provides egocentric outdoor videos with depth and scene annotations. CityWalker (Liu et al., 2025) learns urban policies from walking and driving videos, while Vid2Sim (Xie et al., 2025) builds interactive simulation from video. These efforts mainly use visual, geometric, and motion cues. Audio-visual navigation typically seeks a sounding source. SoundSpaces (Chen et al., 2020) introduced simulated binaural acoustics for this task, while SoundSpaces 2.0 (Chen et al., 2022) extended rendering to continuous locations. Later work considers intermittent sounds, continuous movement, and multiple sources (Chen et al., 2021; Shi et al., 2025; Zeng et al., 2026). In HearVGo, the destination is independent of the sound source, which may instead alter the route.

Audio understanding and evaluation. Audio benchmarks now extend beyond speech recognition. AIR-Bench (Yang et al., 2024) evaluates speech, natural sounds, and music. AudioBench (Wang et al., 2025) covers speech, acoustic scenes, and voice, while MMAU (Sakshi et al., 2025) emphasizes knowledge and reasoning across speech, environmental sounds, and music. BAT (Zheng et al., 2024) and STAR-Bench (Liu et al., 2026) probe spatial and temporal audio reasoning. These benchmarks use discrete or textual answers. HearVGo instead asks whether an omni model can use acoustic evidence to predict a metric trajectory toward an independent goal.

Audio informed motion and safety. Sound can also influence motion without defining the destination. Passive acoustic detection has been coupled with avoidance dives for underwater robots near motorized vessels (D'Este et al., 2015). ANAVI (Jain et al., 2024) predicts noise from a robot's own actions to plan quieter routes for nearby listeners. Closest to our setting, a jackhammer detector enlarges the safety margin of a control barrier function in a simulated construction site (Mootz & Akhavian, 2026). These systems use cues, objectives, or controllers tailored to their applications. HearVGo instead tests whether omni models can use multimodal observations to determine when and how diverse environmental sounds should change a local route.

3 HearVGo Platform

We integrate spatial audio into CARLA, record aligned observations and human walking demonstrations, and construct trajectory prediction instances, as summarized in Figure 3.

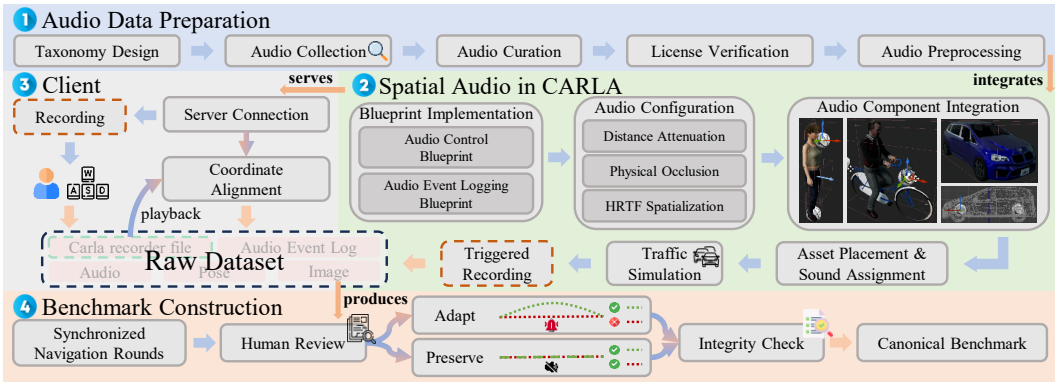
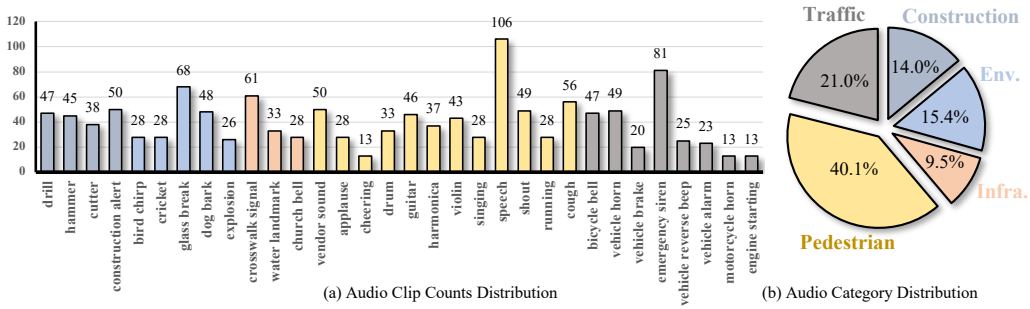


Figure 3: Construction pipeline of HearVGo. (1) Audio preparation collects, curates, and preprocesses environmental sounds. (2) The curated sounds are integrated with CARLA scene assets and spatial rendering. (3) The CARLA server serves a recording client that aligns audio, actor states, and replayed observations and produces the raw dataset. (4) Human review and integrity checks convert these records into Adapt and `Preserve` instances that form the canonical benchmark.

3.1 Audio Environment Construction

Environmental Audio Preparation. We collect environmental sounds from public audio resources and generate 50 construction announcements with `edge-tts`. We review the retrieved recordings

for duplicates and quality before preprocessing. Each clip is converted into a monaural source asset and standardized to a peak level of -6 dBFS before import into CARLA. Using monaural assets avoids carrying over left and right channel differences from the original recordings. Binaural cues are instead rendered from the source and listener geometry in simulation. The resulting collection contains 1,288 clips spanning 32 sound classes in five groups, as summarized in Figure 4. Appendix A summarizes the audio acquisition channels, while Appendix B reports the clip-duration distribution for each sound class.



(a) Audio Clip Counts Distribution (b) Audio Category Distribution

Figure 4: Distribution of the audio corpus. (a) Per-class number. (b) Distribution of five sound groups.

Spatial Audio Integration in CARLA. We import the curated audio assets into CARLA (Dosovitskiy et al., 2017) and implement a shared `SoundSource` Blueprint to manage audio selection and emission behavior. The Blueprint supports continuous, periodic, and random activation modes for different environmental events. We deploy its instances in two ways. Sources associated with native CARLA entities are attached to their corresponding blueprints, such as pedestrians, vehicles, bicycles, and traffic lights. Other fixed sources are represented by placing standalone `SoundSource` instances at semantically matched scene objects. This design keeps each source aligned with the object responsible for the sound while accommodating both native and manually placed scene assets.

All sources share distance attenuation, occlusion, and spatialization settings, while source gain can vary by sound class. Unreal Engine applies distance attenuation and geometry based occlusion, while Resonance Audio renders HRTF based binaural cues relative to the moving listener. Changes in source and listener geometry therefore alter the recorded audio as the scene evolves.

3.2 Benchmark Construction

Multimodal Data Collection. A human operator steers a pedestrian in CARLA while viewing camera feeds and listening to the binaural audio rendered by Unreal Engine. The interface initializes traffic and runs the simulation synchronously at 20 Hz. A robot dog and UAV follow the pedestrian to provide low and elevated views. CARLA Recorder stores actor and traffic states but not custom audio component states, so a separate Unreal Engine logger records each source's audio asset, semantic class, position, and active state at 20 Hz. When the operator issues a recording command, CARLA Recorder starts immediately. After a 2 s stabilization period, each tick captures 0.05 s of stereo audio and then saves the poses of the pedestrian, robot dog, and UAV. Each pose thus marks the end of its preceding audio block. The sound logger runs independently, and its records are aligned to poses by game time. The traveled distance and Recorder log are saved when recording ends.

After live navigation, we replay the Recorder log to capture images. Because the log starts 2 s before audio, pose, and source state capture, we advance 40 ticks before aligning the replay with the first saved pose. We then capture RGB, 16 bit metric depth, and semantic images every 20 ticks, or 1 Hz. These include forward and panoramic views from the pedestrian and robot dog and an elevated UAV view.

Some recordings contain brief zero dropouts and sample discontinuities. We repair these artifacts locally across both stereo channels. Appendix C reports the corresponding changes in channel levels, correlation, interaural level difference, and estimated interaural lag.

Trajectory Instance Construction. We review the aligned recordings to identify route adjustments associated with environmental sound. For each `Adapt` instance, we mark the start and end poses,

one primary sound source, and the left or right adjustment direction. The start pose, which we call the Query Pose, is the prediction origin, and the end pose is the Local Goal. The recorded poses between them form the reference trajectory. Multiple annotated adjustments within the same episode are treated as separate trajectory instances. For `Preserve` instances, we mark a Query Pose where no audio-driven route adjustment is needed. The episode's final pose is the Local Goal. The recorded poses from Query to this goal form the reference trajectory. Both conditions use the same trajectory prediction format.

After cross-modal consistency and integrity checks, HEARVGO contains 536 trajectory instances from 525 distinct collection rounds, including 379 `Adapt` and 157 `Preserve` cases. Table 1 compares HEARVGO with audio understanding and sound-targeted audio-visual navigation benchmarks.

Table 1: Comparison of dataset and benchmark in audio understanding and audio-visual navigation.

Work	Setting	Audio role	Goal	Output	Venue
Audio understanding					
AIR-Bench (Yang et al., 2024)	Dis-embodied	Evidence	None	Answer	ACL 2024
MMAU (Sakshi et al., 2025)	Dis-embodied	Evidence	None	Answer	ICLR 2025
STAR-Bench (Liu et al., 2026)	Dis-embodied	Evidence	None	Answer	ICLR 2026
Sound-goal audio-visual navigation					
SoundSpaces (Chen et al., 2020)	Indoor simulation	Goal cue	Sound	Actions	ECCV 2020
SoundSpaces 2.0 (Chen et al., 2022)	Indoor simulation	Goal cue	Sound	Actions	NeurIPS 2022
BeDAViN (Shi et al., 2025)	Indoor simulation	Goal cue	Sound	Actions	AAAI 2025
SAVN-CE (Zeng et al., 2026)	Indoor simulation	Goal cue	Sound	Actions	CVPR 2026
Sound-aware audio-visual navigation					
HearVGo (ours)	**Outdoor simulation**	**Route cue**	**Any point**	**Metric trajectory**	**2026**

3.3 TASK FORMULATION

From the Query Pose at time t_0, the model predicts a safe local walking path toward a given Local Goal. It should adjust the route when environmental sound is relevant to the current or upcoming path. Otherwise, it should continue toward the goal without an unnecessary detour.

The model input is

$$X = \left(\{V_t, \tilde{P}_t\}_{t \in \mathcal{H}}, A^L, A^R, g \right), \tag{1}$$

where $\mathcal{H}$ contains the available observations at $t_0 - 2$ s, $t_0 - 1$ s, and t_0. Missing early frames are omitted rather than repeated or padded. Each V_t has synchronized Walker and Drone RGB-D panel slots. Each panel pairs RGB with Depth, shown from the original 16 bit millimetre values on a fixed linear 8 bit grayscale scale with nearer regions lighter. The corresponding $\tilde{P}_t$ contains Walker position and heading and, when available, the Drone camera pose, all expressed relative to the Query Pose.

A^L and A^R are the left and right channels of up to 3 s of continuous audio ending at t_0. Their channel assignments are preserved in the shared 16 kHz signal derived from the original 48 kHz recording. The Local Goal $g \in \mathbb{R}^2$ uses the local frame centred at the Query Pose. Other than the Local Goal, the model receives no observations or recorded poses after Query, condition labels, annotated adjustment directions, or sound source metadata.

The output $\hat{Y} = (\hat{p}_1, \ldots, \hat{p}_{10}) \in \mathbb{R}^{10 \times 2}$ contains 10 ordered, approximately evenly spaced absolute future positions in metres, not stepwise displacements. The Query Pose is the origin, with $+x$ forward and $+y$ right, and is not included in the output. Waypoints have no timestamps. See Appendix D for coordinate conversion and reference sampling and Appendix E for the shared prompt.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate 14 omni models without task specific fine tuning. The 11 open-source models are **Phi 4 Multimodal** (Abouelenin et al., 2025), **Baichuan Omni 1.5** (Li et al., 2025), **Qwen2.5**

Omni (Xu et al., 2025a), **video SALMONN 2+** (Tang et al., 2025), **MiniCPM o 4.5** (Cui et al., 2026), **OmniVinci** (Ye et al., 2026), **TLive Omni** (Hu et al., 2026), **Gemma 4** (Team et al., 2026), **Qwen3 Omni Instruct and Thinking** (Xu et al., 2025b), and **Nemotron 3 Nano Omni** (Deshmukh et al., 2026). We also evaluate 3 closed-source models, **Gemini 3.8 Flash**, **Seed 2.0 Lite** (Seed, 2026), and **Qwen3.8 Omni Flash** (Team, 2026).

Evaluation protocol. Each model predicts 10 absolute future waypoints in the local frame centred at the Query Pose for the same frozen set of 536 trajectory instances, comprising 379 `Adapt` and 157 `Preserve` cases. All models share one task prompt and evaluator. Predictions are scored without geometric repair. Each intervention is paired with Full on the same trajectory instances. For paired ablation differences, we report marginal 95% confidence intervals from 2,000 bootstrap resamples of collection rounds. Input adaptation and statistical details appear in Appendix F.

4.2 Evaluation Metrics

Table 2 summarizes the evaluation measures. Adapt Quality (A-Q) and Keep Quality (K-Q) combine goal progress, the response required by each condition, and waypoint use within each trajectory instance. They are reported separately rather than averaged into one score. nDTW-GT measures agreement with the recorded human path. OVR applies only to the observed map after calibration against that path and is not a physical collision rate. The formulas, sampling rules, and parameter values are in Appendix G.

Table 2: Evaluation metric comparison. Arrows indicate the preferred direction. Coverage describes observation support rather than trajectory quality.

Metric	Set	Interpretation	Better
SoftRPE	All	Goal progress and path efficiency relative to the recorded path.	↑
WDS, ESU	All	Distribution and effective use of waypoint segments.	↑
BDS	Adapt	Lateral response on the annotated side, scaled by the recorded path.	↑
A-Q	Adapt	Joint goal progress, direction response, and waypoint use.	↑
SymPathDist	Preserve	Mean nearest path distance in both directions.	↓
KeepPath	Preserve	Bounded transform of SymPathDist.	↑
K-Q	Preserve	Joint goal progress, route preservation, and waypoint use.	↑
nDTW-GT	All	Ordered agreement with the recorded human path.	↑
OVR	All	Rate of sustained obstacle or road violations in observed space.	↓
Coverage	All	Fraction of predicted path supported by observed space.	–

4.3 Main Result

Overall performance. The main experiment provides Walker RGB-D, binaural audio, motion history, and the Local Goal, while leaving Drone observations blank. Table 3 reports `Adapt` and `Preserve` separately.

Performance varies substantially across models and conditions. Qwen2.5 Omni and MiniCPM o 4.5 lead `Adapt` A-Q. Seed 2.0 Lite and Gemini 3.8 Flash lead `Preserve` K-Q. The remaining metrics reveal two limitations.

Limited adaptation to sound under multimodal observations. In `Adapt` cases, BDS measures lateral response toward the annotated human avoidance side. nDTW_{GT} measures ordered agreement with the recorded path. Seed 2.0 Lite scores 2.27% on BDS and 36.27% on nDTW_{GT}. Qwen3 Omni Thinking scores 15.04% and 32.15%, respectively. Yet their SoftRPE scores reach 99.96% and 82.97%. Both models therefore make progress toward the goal without closely matching the human avoidance trajectory. On routes spanning several metres, these scores are consistent with predicted paths that follow the goal direction with only a slight lateral bend, rather than tracing the broader avoidance curves in the human reference. For the same models, nDTW_{GT} rises to 93.74% and 83.35% in `Preserve`. This contrast suggests that their predictions more closely match the reference when no detour is required, while missing the detours associated with sound in `Adapt`.

Table 3: Results on HearVGo. **Bold** and underline mark the highest and second highest scores.

Model	Adapt					Preserve					Observed Space	
	A-Q↑	SoftRPE↑	BDS↑	ESU↑	nDTW$_{GT}$↑	K-Q↑	SoftRPE↑	KeepPath↑	ESU↑	nDTW$_{GT}$↑	OVR↓	Coverage
Phi 4 Multimodal	7.93	30.24	17.92	25.55	10.63	4.07	30.46	35.94	9.56	25.22	16.23	50.34
Baichuan Omni 1.5	0.51	7.85	45.35	84.24	1.78	0.29	3.63	4.76	85.34	2.31	38.25	93.41
Qwen2.5 Omni	**41.15**	79.84	**72.02**	74.14	38.55	61.68	90.23	85.43	73.25	84.05	32.09	95.84
video SALMONN 2+	1.38	20.89	4.86	5.91	6.60	1.25	11.58	15.54	2.80	10.15	3.73	20.17
MiniCPM o 4.5	40.41	74.29	65.86	87.13	36.12	55.04	83.10	77.03	84.70	70.35	26.31	97.34
OmniVinci	26.57	56.34	60.51	76.48	24.18	38.68	66.83	63.25	75.27	53.90	31.90	91.51
TLive Omni	17.99	95.35	35.60	56.20	40.89	38.43	99.97	93.22	41.04	93.20	17.16	97.17
Gemma 4	18.79	98.26	24.56	83.98	42.31	86.63	99.22	91.35	95.52	91.13	11.01	97.29
Qwen3 Omni Instruct	23.23	81.96	41.43	81.29	34.36	71.08	90.67	79.46	94.19	77.32	20.90	97.08
Qwen3 Omni Thinking	6.60	82.97	15.04	95.24	32.15	82.52	91.93	84.32	98.20	83.35	14.37	96.02
Nemotron 3 Nano Omni	25.26	85.90	45.40	73.82	39.21	49.53	85.29	81.28	66.71	78.48	17.16	95.55
Gemini 3.8 Flash	37.30	99.71	37.71	**99.72**	**53.62**	91.29	99.91	91.36	**100.00**	91.34	7.84	97.11
Seed 2.0 Lite	2.15	**99.96**	2.27	98.18	36.27	**93.51**	**99.98**	**93.74**	99.76	**93.74**	7.65	97.32
Qwen3.8 Omni Flash	34.79	88.12	42.64	98.28	43.02	84.58	96.25	88.15	98.33	87.24	13.99	96.73

Uneven waypoint sequences limit trajectory quality. ESU measures how effectively the ten path segments share the total path length. Phi 4 Multimodal, video SALMONN 2+, and TLive Omni have the lowest ESU in both conditions. Their `Adapt` scores are 25.55%, 5.91%, and 56.20%, respectively. In `Preserve`, the scores fall to 9.56%, 2.80%, and 41.04%. ESU enters A-Q and K-Q for each instance, so poor waypoint use directly reduces these composite scores. TLive Omni illustrates the effect. It reaches 99.97% `Preserve` SoftRPE and 93.22% KeepPath, but its K-Q is only 38.43%. All three closed source models exceed 98% ESU in both conditions. Each exceeds every open source model in the corresponding condition. The next subsection examines the trajectory degeneration behind these differences.

OVR must be read alongside Coverage. video SALMONN 2+ has the lowest OVR at 3.73%, but its Coverage is only 20.17%. Seed 2.0 Lite has 7.65% OVR with 97.32% Coverage. Since violations along unobserved parts of a path cannot be assessed, the lower OVR for video SALMONN 2+ does not establish a safer route. OVR describes the reconstructed scene at Query time rather than collisions during future walking, as detailed in Appendix G.

Trajectory degeneration. Table 4 reports WDS, Terminal Step Share (Share for short) ℓ_{10}/L_P, and Exact Trajectory Reuse (Reuse for short) over all 536 predictions per model. Here ℓ_{10} is the final segment length, and Terminal Step Share is zero when $L_P = 0$. Exact Reuse is the fraction of queries whose ten waypoints, rounded to six decimal places, also occur for another query. The recorded trajectories receive the same diagnostics as a reference. None of these diagnostics removes predictions from the task scores.

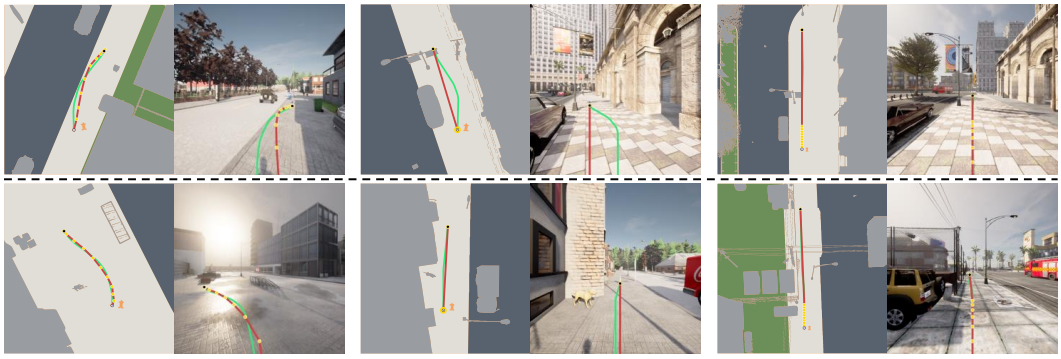
Table 4: Analysis of trajectory outputs in the main setting.

Model	WDS (%)	share (%)	reuse (%)
Human Reference	99.99	9.78	0.00
Phi 4 Multimodal	27.71	14.50	66.23
Baichuan Omni 1.5	92.20	13.07	87.87
Qwen2.5 Omni	87.86	29.76	24.44
video SALMONN 2+	5.90	10.77	82.28
MiniCPM o 4.5	94.17	8.23	5.60
OmniVinci	80.64	20.02	64.93
TLive Omni	73.37	51.58	0.75
Gemma 4	94.27	6.63	1.87
Qwen3 Omni Instruct	92.69	7.06	8.02
Qwen3 Omni Thinking	97.80	9.67	0.37
Nemotron 3 Nano Omni	83.72	31.91	14.93
Gemini 3.8 Flash	99.92	9.32	1.49
Seed 2.0 Lite	99.46	8.00	0.00
Qwen3.8 Omni Flash	99.30	9.02	3.17

Under the trajectory-degeneration diagnostics in Table 4, Gemini 3.8 Flash and Seed 2.0 Lite produce the most regular outputs, with WDS of 99.92% and 99.46% and exact reuse of 1.49% and 0.00%, respectively. In contrast, Phi 4 Multimodal and video SALMONN 2+ combine low WDS (27.71% and 5.90%) with high reuse (66.23% and 82.28%), while TLive Omni places 51.58% of its path length in the final step. Figure 5 illustrates representative degeneration patterns.

4.4 Ablation Studies

These paired interventions ask whether audio improves task quality, how much performance remains with only motion and the Local Goal, and whether visual context changes sensitivity to binaural



Correct avoidance | Repeated waypoints | Terminal jump

Figure 5: Visualization of adaptive trajectories. Green shows the human reference and red shows predictions. Yellow dots (● ●) mark waypoints, and black dots (●) mark the goals. Zoom in for better view.

direction. All Δ values are differences in percentage points. The corresponding 95% confidence intervals are in Appendix H.

Audio ablation. Silence replaces both audio channels with equal-duration zero waveforms while other inputs remain fixed. Table 5 reports Full minus Silence differences, with 95% intervals in Table 11. If sound supports only needed route adjustments, we expect higher `Adapt` A-Q and BDS with little change in `Preserve` K-Q. Phi 4 Multimodal gains 5.53 A-Q points and 11.24 BDS points, while Gemma 4 gains 7.20 and 9.45 points. Their A-Q and BDS intervals lie above zero, whereas their K-Q intervals include zero. Yet their Full A-Q scores are only 7.93% and 18.79% in Table 3. Qwen2.5 Omni and MiniCPM o 4.5 lead Full A-Q at 41.15% and 40.41%, but their A-Q and BDS difference intervals include zero. This does not prove that audio has no effect. It shows that high Full A-Q alone does not establish an audio benefit. K-Q also varies, increasing by 1.72 points for Qwen2.5 Omni and decreasing by 7.21 points for video SALMONN 2+, with intervals on opposite sides of zero. Overall, the measured audio benefit varies by model and does not track absolute `Adapt` quality. In practical terms, adding audio may alter a predicted trajectory without making the route better. This points to a limited ability to decide when sound warrants an adjustment and to turn that cue into a useful path.

Motion and goal baseline. Neither removes Walker RGB-D and audio but retains motion history and the Local Goal. Table 6 reports its scores and the Full minus Neither differences. The paired 95% intervals are in Table 12. MiniCPM o 4.5 still reaches 34.09% A-Q with Neither, only 6.32 points below Full. Its K-Q reaches 69.89%, which is 14.85 points above Full. Gemma 4 also reaches 91.69% K-Q with Neither, 5.06 points above Full. The intervals for these differences exclude zero. Qwen2.5 Omni shows a different pattern. Full raises its A-Q by 27.64 points and K-Q by 18.38 points, with positive intervals. Motion history and the goal can therefore support substantial trajectory quality, while added perceptual input affects the two conditions differently across models. The preceding Silence ablation does not resolve an A-Q benefit from audio for MiniCPM o 4.5 or Qwen2.5 Omni. Their high Full A-Q scores alone do not establish effective use of sound. Because Neither removes both visual and acoustic input, this comparison cannot assign any Full gain or loss to audio alone.

Visual context and binaural response. This test asks whether swapping the left and right audio channels makes a model avoid on the opposite side. It also asks whether that response emerges when Walker RGB-D is removed. Motion history and the Local Goal remain fixed. Table 7 reports L, Original, and Flip. L measures the size of the lateral waypoint change, but does not show whether the avoidance side reverses. Original counts `Adapt` paths on the annotated side before the swap. Flip counts those that move to the opposite side afterward. Both rates use all 379 `Adapt` instances. Appendix F gives the exact definitions, and Table 13 gives paired 95% intervals. With RGB-D, Qwen2.5 Omni has an Original rate of 82.32% but a Flip rate of 0.79%. Without RGB-D, its L rises from 1.67% to 4.50%, while Flip reaches only 1.85%. Across all five models and both visual conditions, Flip remains between 0.26% and 2.11%. Every interval for the change in Flip includes zero. Only 16 of the 46 observed flips pass basic duplicate and terminal jump checks. The channel swap can shift predicted waypoints without reliably reversing the avoidance side, even without vision.

Table 5: Effect of audio on trajectory scores. Values are score differences between the full input and silent audio, in percentage points. Positive values indicate higher scores with audio.

Model	ΔA-Q	ΔBDS	ΔK-Q
Phi 4 Multimodal	+5.53	+11.24	+0.94
Baichuan Omni 1.5	−0.08	−4.46	−0.38
Qwen2.5 Omni	+0.79	+1.16	+1.72
video SALMONN 2+	−1.00	−2.23	−7.21
MiniCPM o 4.5	+0.84	+1.76	−0.99
OmniVinci	−0.04	+0.65	−0.10
TLive Omni	+0.17	−0.46	+0.34
Gemma 4	+7.20	+9.45	+1.51
Qwen3 Omni Instruct	−0.19	−0.27	−2.39
Qwen3 Omni Thinking	+0.66	+0.92	−1.45
Nemotron 3 Nano Omni	+1.64	+1.49	−0.25
Seed 2.0 Lite	−0.10	−0.13	−0.14
Qwen3.8 Omni Flash	+0.67	+0.73	+1.31

Table 6: Analysis of the motion and goal baseline. Neither retains motion history and the Local Goal. Scores are percentages. The change is the Full score minus the Neither score in percentage points. Positive values mean higher Full scores.

Model	Metric	Neither	Δ
Qwen2.5 Omni	A-Q	13.51	+27.64
	K-Q	43.31	+18.38
MiniCPM o 4.5	A-Q	34.09	+6.32
	K-Q	69.89	−14.85
Gemma 4	A-Q	10.37	+8.42
	K-Q	91.69	−5.06
Qwen3 Omni Instruct	A-Q	25.30	−2.07
	K-Q	75.79	−4.71
Nemotron 3 Nano Omni	A-Q	24.09	+1.17
	K-Q	37.26	+12.27

This points to limited directional response to binaural audio under this test. The scene remains unchanged, so the swap tests sensitivity rather than the safety of a reversed route.

Table 7: Exchanging left and right audio channels with and without Walker RGB-D. All values are percentages. Δ is No RGB-D minus RGB-D. For Original / Flip rows, Δ applies to Flip. Metric definitions and calculations are in Appendix F.

Model	Metric	RGB-D	No RGB-D	Δ
Qwen2.5 Omni	L	1.67	4.50	+2.83
	Original / Flip	82.32 / 0.79	51.72 / 1.85	+1.06
MiniCPM o 4.5	L	2.88	2.92	+0.04
	Original / Flip	77.04 / 0.26	71.24 / 1.06	+0.79
Gemma 4	L	1.47	1.83	+0.36
	Original / Flip	31.40 / 1.58	27.70 / 1.58	0.00
Qwen3 Omni Instruct	L	2.09	0.92	−1.17
	Original / Flip	45.65 / 0.53	44.59 / 0.53	0.00
Nemotron 3 Nano Omni	L	1.46	2.18	+0.72
	Original / Flip	55.15 / 1.85	52.77 / 2.11	+0.26

5 Conclusion

HearVGo studies how environmental sound can adapt walking trajectories toward an independent outdoor goal. It provides synchronized observations and human demonstrations for 536 trajectory instances under `Adapt` and `Preserve` conditions. Across 14 omni models, goal progress can coexist with weak directional response or poorly formed waypoint sequences. Audio removal has model dependent effects, and exchanging the left and right channels rarely reverses the predicted avoidance side in the five models tested under that intervention. Together, these results show that the evaluated models do not consistently translate acoustic evidence into coherent local route adjustments.

Limitation and future work. HearVGo currently contains 536 trajectory instances, since the dataset is used only for model evaluation in this work. Its trajectories were collected manually by a human operator in CARLA, so the findings may not transfer directly to real streets. Each model produces one set of ten future waypoints from the available observations and Local Goal. This setting cannot measure how a model would revise its route over time. We plan to collect more trajectories across a wider range of sound events and walking scenes. A larger dataset could support model training and validation while retaining a separate test set. Future work will use real recordings and dynamic scenes to test how models adjust routes in response to sound.

AI USE STATEMENT

In this work, we have not used generative AI tools for tasks with required disclosure. Where relevant, these tasks were carried out by the authors without generative AI assistance, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist in drafting and editing parts of the manuscript to improve clarity, readability, and language quality. We have reviewed all AI assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, personally identifiable information, or sensitive personal data. The datasets and resources used in this study are publicly available and are used in accordance with their respective licenses and terms of use. We do not identify any additional ethical concerns associated with the experiments conducted in this work.

REPRODUCIBILITY STATEMENT

Details of the experimental setup, implementation, datasets, and evaluation protocols are provided in the main text and the appendix. Code and data are publicly available: https://anonymous.4open.science/r/HearVGo.

REFERENCES

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

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4218–4222, 2020.

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, 2020.

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.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 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.

Claire D'Este, Brett Seton, John McCulloch, Daniel Smith, and Chris Sharman. Avoiding marine vehicles with passive acoustics. *Journal of Field Robotics*, 32(1):152–166, 2015.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pp. 1–16. PMLR, 2017.

Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pp. 411–412, 2013.

Yibo Hu, Yu Qian, Mao Gu, Yingfan Tao, Yuhao Chen, Yongdong Luo, Zhuoqun Liu, Meiguang Jin, and Junfeng Ma. Tlive-omni: An omni-modal understanding model for e-commerce live streaming. *arXiv preprint arXiv:2608.20958*, 2026.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*, 2019.

Vidhi Jain, Rishi Veerapaneni, and Yonatan Bisk. Anavi: Audio noise awareness using visuals of indoor environments for navigation. *arXiv preprint arXiv:2410.18932*, 2024.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

Xinhao Liu, Jintong Li, Yicheng Jiang, Niranjan Sujay, Zhicheng Yang, Juexiao Zhang, John Abanes, Jing Zhang, and Chen Feng. Citywalker: Learning embodied urban navigation from web-scale videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6875–6885. IEEE, 2025.

Zihan Liu, Zhikang Niu, Qiuyang Xiao, Zhisheng Zheng, Ruoqi Yuan, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Jianze Liang, Xie Chen, et al. Star-bench: Probing deep spatio-temporal reasoning as audio 4d intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 134703–134731, 2026.

Johannes Mootz and Reza Akhavian. Control barrier functions with audio risk awareness for robot safe navigation on construction sites. *arXiv preprint arXiv:2602.12416*, 2026.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Sakshi Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 84929–84964, 2025.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 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.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3.8-omni: Towards native omni-modal agents, 2026. URL https://arxiv.org/abs/2609.25611.

Sagar M Waghmare, Kimberly Wilber, Dave Hawkey, Xuan Yang, Matthew Wilson, Stephanie Debats, Cattalyya Nuengsigkapian, Astuti Sharma, Lars Pandikow, Huisheng Wang, et al. Sanpo: A scene understanding accessibility and human navigation dataset. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 7855–7864, 2025.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4297–4316, 2025.

Ziyang Xie, Zhizheng Liu, Zhenghao Peng, Wayne Wu, and Bolei Zhou. Vid2sim: Realistic and interactive simulation from video for urban navigation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1581–1591. IEEE, 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.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 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.

Yichen Zeng, Hebaixu Wang, Meng Liu, Yu Zhou, Chen Gao, Kehan Chen, and Gongping Huang. Semantic audio-visual navigation in continuous environments. *arXiv preprint arXiv:2603.19660*, 2026.

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

A Source Audio Provenance

Table 8 groups the final clips by acquisition channel. Freesound (Font et al., 2013), FSD50K (Fonseca et al., 2021), Common Voice (Ardila et al., 2020), and LibriSpeech (Panayotov et al., 2015) are the named source collections. Openverse, YouTube, and Wikimedia Commons are additional retrieval channels. The 50 construction announcements were synthesized with `edge-tts`[1].

We first defined 32 sound classes for urban walking scenes and searched these sources for matching recordings. We listened to the candidates, kept those with relevant sounds, trimmed unnecessary sections when needed, and converted the selected clips to a consistent format for simulation.

Table 8: Acquisition channels for the 1,288 final audio clips. Durations are measured after trimming and processing.

Source	Clips	Minutes
Freesound direct	552	54.6
FSD50K via Freesound	511	51.4
Openverse	43	4.8
Common Voice via VoxBox	83	6.0
LibriSpeech via OpenSLR	10	1.5
YouTube	37	4.2
Wikimedia Commons	2	0.2
`edge-tts`	50	2.2
Total	**1,288**	**124.9**

B Source Audio Duration by Class

The 1,288 clips contain 7,493.8 s of audio, with a pooled median duration of 4.27 s. Table 9 reports the clip count, median, interquartile range, and observed range for each of the 32 sound classes.

C Binaural Audio Repair Audit

Table 10 compares the original and repaired 48 kHz recordings at the episode level. The 540 eligible recordings used for this audio quality audit are distinct from the 536 trajectory instances in the benchmark. Before and After are medians across recordings. The change columns summarize absolute paired differences, with the 95th percentile computed by the nearest-rank rule. RMS levels are $20 \log_{10}(\mathrm{RMS})$ in dBFS. For ILD and interaural lag, the Before and After columns use absolute values so that opposite source directions do not cancel, while paired changes are computed from the signed values. Interaural lag is a recording-level estimate in samples at 48 kHz, not a source-localization measurement. The largest estimated lag change is 13 samples, so this audit does not establish exact preservation of every spatial cue.

D Coordinate Conversion and Reference Sampling

For a query pose at (x_q, y_q) with yaw θ_q in radians, a world position (x, y) is expressed locally as

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta_q & \sin\theta_q \\ -\sin\theta_q & \cos\theta_q \end{bmatrix} \begin{bmatrix} x - x_q \\ y - y_q \end{bmatrix}. \tag{2}$$

The axes point forward and right relative to the query. CARLA yaw is converted from degrees to radians for the matrix. Historical heading differences remain in degrees and are wrapped to $[-180°, 180°)$. The original Pose path from Query to the selected end has arc length $H_q = L_R$, which supplies the SoftRPE reference length. This path is sampled at 0.25 m intervals to store R and at $kH_q/10$ for reference waypoint $k = 1, \ldots, 10$, excluding the origin and including the goal. For direction response, SymPathDist, and nDTW-GT, the evaluator resamples the stored R at intervals no greater than 0.25 m to obtain $\widetilde{R}$.

[1] https://github.com/rany2/edge-tts

Table 9: Duration statistics for the final 1,288 source audio clips. All durations are in seconds.

Sound class	Clips	Median	Q1–Q3	Min–Max
Construction (180 clips)				
drill_sound	47	9.00	6.23–9.50	2.56–24.00
hammer_sound	45	5.97	3.37–9.50	0.82–28.43
cutter_sound	38	9.37	5.73–9.99	1.49–15.36
construction_voice_announcement	50	2.58	2.35–3.07	1.55–3.99
Environment (198 clips)				
bird_chirp	28	2.30	1.47–3.24	0.80–9.66
cricket	28	10.78	6.63–14.41	1.89–26.06
glass_break	68	0.97	0.67–1.19	0.26–4.84
dog_bark	48	1.42	0.87–1.92	0.21–6.37
explosion	26	1.86	1.50–2.71	0.86–9.84
Infrastructure (122 clips)				
crosswalk_signal	61	7.26	4.81–11.35	1.48–20.36
water_landmark	33	9.50	9.50–21.45	4.38–54.50
church_bell	28	6.35	5.03–7.25	2.38–28.84
Pedestrian (517 clips)				
vendor_sound	50	6.93	3.46–9.50	1.57–12.03
street_applause	28	4.28	3.50–5.72	2.18–9.96
street_cheering	13	3.78	3.26–7.60	1.22–15.53
street_music_drum	33	4.57	2.78–5.65	0.49–10.97
street_music_guitar	46	8.02	2.59–11.15	0.90–28.00
street_music_harmonica	37	6.55	4.93–13.50	1.05–27.59
street_music_violin	43	8.64	7.57–10.49	3.09–26.64
street_singing	28	8.66	5.87–17.68	1.96–26.40
pedestrian_speech	106	4.33	3.36–5.83	2.11–25.79
pedestrian_shout	49	1.95	1.38–2.89	0.66–5.94
footsteps_running	28	5.69	2.57–8.95	1.74–13.93
pedestrian_cough	56	1.14	0.86–1.56	0.53–6.32
Traffic (271 clips)				
bicycle_bell	47	1.16	0.95–1.55	0.34–7.86
vehicle_horn	49	1.40	0.89–2.51	0.40–7.17
vehicle_brake	20	1.48	1.17–2.39	0.48–6.10
emergency_siren	81	6.20	4.88–8.97	0.65–16.74
vehicle_reverse_beep	25	9.90	5.63–12.25	2.28–24.46
vehicle_alarm	23	8.83	5.76–11.26	2.36–20.16
motorcycle_horn	13	1.05	0.82–1.18	0.38–2.56
engine_starting	13	2.84	2.05–3.30	1.21–4.49
All classes	**1,288**	**4.27**	**1.89–8.31**	**0.21–54.50**

Table 10: Binaural recording properties before and after local stereo repair across 540 eligible episode recordings.

Property	Before	After	Median $\lvert\Delta\rvert$	95th $\lvert\Delta\rvert$	Max $\lvert\Delta\rvert$
Left RMS (dBFS)	-44.655	-44.661	.0031	.0264	.2228
Right RMS (dBFS)	-44.315	-44.324	.0035	.0291	.1284
Left/right correlation	.162909	.162877	.000101	.001230	.006681
Absolute ILD (dB)	6.0620	6.0600	.0011	.0145	.1692
Absolute lag (samples)	2	2	0	1	13

E EVALUATION PROMPT

The prompt below is shared by all models. At inference, `$sample_data` is replaced by the sample specific panel count, motion history, audio duration, and Local Goal. The corresponding RGB-D panels and binaural audio are supplied through each model's media interface.

```
Task

Plan one safe and walkable local path for an outdoor walking agent from the current Query
Pose to the Local Goal.

Input Interface

The input interface has a fixed structure across all samples. Use only evidence contained
in the supplied observations. A uniformly blank visual slot, a silent waveform, a null
walker motion entry, or a null drone camera pose contains no evidence. Do not infer the
scene condition or desired route from the absence of a signal.

Visual observations are synchronized panel pairs ordered from oldest to newest. At each
time step, the WALKER panel is followed by the DRONE panel. Both panels in a pair share
the same panel index and timestamp. The WALKER panel is captured from the walker
perspective. The synchronized DRONE panel is captured from an external drone perspective
and shows the same agent and surrounding scene. Every panel places RGB on the left and
synchronized Depth on the right. WALKER Depth and DRONE Depth use the same fixed linear
metric scale across all frames and samples. Depth is encoded as 8 bit grayscale. Pixel
value 255, shown as white, corresponds to 0 m. Pixel value 0, shown as black, corresponds
to 65.535 m or any farther depth clipped to this range. Intermediate pixel values vary
linearly with metric distance. Each Depth image measures distance from the camera that
produced its panel. A Depth pixel is not a position in the coordinate frame centered at
Query. The Sample Data explicitly associates every visual time step with one WALKER panel,
 one DRONE panel, one walker motion row, and one synchronized drone camera pose. Every
drone camera pose gives panel_index, time_from_query_s, x_forward_m, y_right_m, z_up_m,
pitch_from_query_deg, yaw_from_query_deg, and roll_from_query_deg. Drone positions are
relative to the Walker Query Pose. Positive z is upward. Drone orientations are relative
to the Query frame. Positive yaw turns toward the right. Positive pitch tilts upward.
Positive roll lowers the right side of the camera and raises the left side. The newest
WALKER and DRONE panels are the Query observations at time zero.

The walker motion history is aligned with the visual panel pairs. Each walker motion row
gives its panel index, time relative to Query, forward and rightward position in metres,
and heading relative to the Query heading in degrees. Positive heading turns toward the
right.

The binaural observation consists of synchronized LEFT EAR and RIGHT EAR signals from one
continuous recording. Preserve their temporal alignment and ear identity. Sound content
and timing may provide semantic evidence. Differences between the ears may provide spatial
 evidence.

The Local Goal gives the desired future position as x_forward_m and y_right_m in the
coordinate frame centered at Query.

Sample Data

$sample_data

Planning Rules

Use the visual, depth, motion, binaural audio, and goal evidence jointly when that
evidence is present. Infer visible or reasonably supported walkable space and construct
one continuous route toward the Local Goal.

The audio may contain several environmental sounds or no relevant sound. Determine whether
 the acoustic evidence indicates a potential hazard that can affect the current or
upcoming path. Consider sound semantics, timing, and binaural spatial evidence. Do not
assume that every audible sound requires avoidance.

When the combined evidence supports a relevant potential hazard, adjust the route to pass
it with reasonable clearance and then continue toward the Local Goal. When no acoustic
evidence is relevant, avoid an unnecessary detour caused only by audio and follow a
natural route toward the goal using the remaining evidence.

The route must make progress toward the Local Goal and remain smooth and continuous. Avoid
 collisions, unnecessary detours, backtracking, abrupt jumps, repeated positions, and
zigzag motion.

Coordinate Rules
```

```
All positions use an ego coordinate frame centered at Query. Query is (0, 0). Positive x
is forward. Positive y is right. Negative y is left. The Local Goal, walker motion
positions, and drone camera positions already use this frame. Every predicted waypoint is
an absolute future position in this frame and is not a displacement from the preceding
waypoint.

Output Rules

The evaluator adds Query point (0, 0). Do not include Query in the waypoint list.

Predict exactly 10 distinct future waypoints in travel order. Space them approximately
uniformly by travelled distance along the planned route. Waypoint 1 must be a new future
position. Waypoint 10 should be at or near the Local Goal. Every consecutive pair must
have nonzero separation.

Return only one valid JSON object following the exact form below. Use finite numbers.
Include no explanation, Markdown, comments, extra fields, or alternative paths.

{"waypoints":[[x1,y1],[x2,y2],[x3,y3],[x4,y4],[x5,y5],[x6,y6],[x7,y7],[x8,y8],[x9,y9],[x10,
y10]]}
```

F Evaluation Protocol Details

Each query provides up to 3 synchronized Walker RGB-D panels and motion rows, including the Query observation, and up to 3 s of continuous binaural audio ending at Query. The Local Goal remains available in every condition. In the main experiment, Drone panels contain blank images and their camera poses are null. Common preprocessing fixes the media selection and Query centred coordinates before model adapters format the inputs for their respective interfaces. The task prompt is unchanged across models and interventions. The shared validator accepts exactly 10 finite two dimensional waypoints and does not repair invalid responses or predicted geometry.

The recorded audio is 48 kHz stereo. All Full results in the main table use the same derived 16 kHz left and right ear signals.

For each intervention, predictions are paired with Full on the same queries, with all inputs outside the intervention held fixed. We use 2,000 bootstrap resamples clustered by complete collection round for paired effect intervals. Left and Right scores use the same resampled rounds. The reported 95% intervals reflect variation across collection rounds, not repeated generation from closed-source models, and are not adjusted for multiple comparisons.

Channel exchange diagnostics. For each query and visual condition, L averages the absolute change between original and swapped signed lateral offsets at the first nine waypoints, then divides by the distance from Query to the Goal and converts the result to a percentage. The final waypoint is excluded from this lateral comparison. Table 7 averages L over all 536 instances. To calculate Original and Flip, we resample the path from Query through the ninth waypoint at intervals no greater than 0.25 m and take its mean signed offset from the straight line connecting Query and the Goal. Original counts `Adapt` instances whose unswapped path lies more than 0.25 m on the annotated side. Flip additionally requires the swapped path to lie more than 0.25 m on the opposite side. Both rates use all 379 `Adapt` instances as their denominator. The threshold is below the smallest recorded `Adapt` mean offset of 0.37 m. Passing the duplicate and terminal-jump checks does not establish collision safety.

G Metric Computation Details

Paths and sampling. For each trajectory instance, $P = (p_0, \ldots, p_K)$ joins the Query Pose $p_0 = (0, 0)$ to the $K = 10$ predicted waypoints. The stored reference polyline $R = (r_0, \ldots, r_M)$ is sampled from the recorded pose path and runs from $r_0 = p_0$ to the Local Goal $r_M = g$. The number M varies by instance. Let L_P be the arc length of P and L_R the arc length of the original recorded pose path. For direction response, SymPathDist, and nDTW-GT, $\widetilde{P}$ and $\widetilde{R}$ denote paths resampled at intervals no greater than 0.25 m, with endpoints retained.

Goal progress and path efficiency. Let $d_g = \|p_K - g\|_2$ be the distance from the predicted endpoint to the Local Goal. SoftRPE gives partial credit for goal progress and penalizes predicted paths longer than the recorded path

$$\mathrm{SoftRPE}(P, R, g) = \max\left(0, 1 - \frac{d_g}{L_R}\right) \frac{L_R}{\max(L_R, L_P)}. \tag{3}$$

The score lies in $[0, 1]$, with higher values indicating greater goal progress and path efficiency relative to the recorded path.

Waypoint utilization. For the $K = 10$ segments from Query to the predicted waypoints, let $\ell_k = \|p_k - p_{k-1}\|_2$. When $L_P > 0$, let $\rho_k = \ell_k / L_P$ and compute

$$\mathrm{WDS}(P) = -\frac{\sum_{k=1}^{K} \rho_k \log \rho_k}{\log K}, \qquad \mathrm{ESU}(P) = \frac{K^{\mathrm{WDS}(P)} - 1}{K - 1}. \tag{4}$$

We use $0 \log 0 = 0$. When $L_P = 0$, ρ_k is not evaluated and $\mathrm{WDS} = 0$ by definition. This gives $\mathrm{ESU} = 0$ for a stationary path or a path with all displacement in one segment.

Direction response and `Adapt` quality. For `Adapt` instance i, the nonzero Local Goal g_i defines the unit goal direction $u_i = g_i / \|g_i\|_2$ and its rightward normal $n_i = (-u_{i,2}, u_{i,1})$. For $X \in \{P, R\}$, let $\widetilde{X}_i$ be its resampled path and $|\widetilde{X}_i|$ its point count. The mean signed lateral offset is $\bar{d}_i^X = |\widetilde{X}_i|^{-1} \sum_{x \in \widetilde{X}_i} x^\mathsf{T} n_i$. Set $a_i = 1$ for an annotated Right response and $a_i = -1$ for Left. The audited reference satisfies $D_i^R = a_i \bar{d}_i^R > 0$. The predicted direction score is $s_i = \mathrm{clip}(a_i \bar{d}_i^P / D_i^R, 0, 1)$, where clip limits its value to $[0, 1]$.

Let $\mathcal{C} = \{\mathrm{Left}, \mathrm{Right}\}$ be the direction labels. For each $c \in \mathcal{C}$, let $\mathcal{A}_c$ be the set of `Adapt` instances annotated with c and $|\mathcal{A}_c|$ its size. For any per-instance values $f = (f_i)$, define their mean in this set as $m_c(f) = |\mathcal{A}_c|^{-1} \sum_{i \in \mathcal{A}_c} f_i$. Thus $m_c(s)$ averages s_i, while $m_c(\mathrm{SoftRPE}\, s\, \mathrm{ESU})$ averages $\mathrm{SoftRPE}_i s_i \mathrm{ESU}_i$. Balanced Direction Score (BDS) and `Adapt` Quality (A-Q) are

$$\mathrm{BDS} = \frac{1}{2} \sum_{c \in \mathcal{C}} m_c(s), \qquad \text{A-Q} = \frac{1}{2} \sum_{c \in \mathcal{C}} m_c(\mathrm{SoftRPE}\, s\, \mathrm{ESU}). \tag{5}$$

Both scores lie in $[0, 1]$ and weight Left and Right equally. BDS measures lateral response on the annotated side relative to the recorded path. Neither score establishes the safety of an alternative path or proves that the model used audio.

Symmetric path distance. Both the predicted path and the recorded path are resampled at intervals no greater than 0.25 m, retaining their endpoints. With the resulting point sets $\widetilde{P}$ and $\widetilde{R}$, the `Preserve` path distance is

$$\mathrm{SymPathDist}(P, R) = \frac{1}{2} \left(\frac{1}{|\widetilde{P}|} \sum_{p \in \widetilde{P}} \min_{r \in \widetilde{R}} \|p - r\|_2 + \frac{1}{|\widetilde{R}|} \sum_{r \in \widetilde{R}} \min_{p \in \widetilde{P}} \|r - p\|_2 \right). \tag{6}$$

The 2 directions prevent a short predicted path from receiving a low distance solely because it lies on part of the recorded route. Each minimum is over resampled points, not over continuous line segments.

`Preserve` quality. For `Preserve` instance i, let $\delta_i = \mathrm{SymPathDist}(P_i, R_i)$ in metres. Define $\mathrm{KeepPath}_i = \exp(-\delta_i / d_{\mathrm{th}})$, where $d_{\mathrm{th}} = 1.0\,\mathrm{m}$ is the Goal tolerance. For the set $\mathcal{I}_{\mathrm{preserve}}$ of `Preserve` instances, Keep Quality (K-Q) is

$$\text{K-Q} = \frac{1}{|\mathcal{I}_{\mathrm{preserve}}|} \sum_{i \in \mathcal{I}_{\mathrm{preserve}}} \mathrm{SoftRPE}_i \mathrm{KeepPath}_i \mathrm{ESU}_i. \tag{7}$$

The product is computed for each instance before averaging. K-Q lies in $[0, 1]$ and combines goal progress, agreement with the recorded `Preserve` route, and effective waypoint use.

Ordered reference agreement. Normalized dynamic time warping (Ilharco et al., 2019) is computed as

$$\mathrm{nDTW}_{\mathrm{GT}}(P, R) = \exp\left[-\frac{\mathrm{DTW}(\widetilde{P}, \widetilde{R})}{|\widetilde{R}| d_{\mathrm{th}}}\right]. \tag{8}$$

Here DTW is the accumulated Euclidean alignment cost, $|\widetilde{R}|$ is the number of sampled reference points, and d_{th} is the same Goal tolerance used for KeepPath. The GT suffix denotes the recorded human reference, not the only valid route.

Observed space reconstruction. For all 536 trajectory instances, we align the replayed overhead semantic and depth observations to the Query Pose and project them to a 0.05 m ground plane grid. The overhead reconstruction provides the primary occupancy evidence. Where the walker is occluded from above, synchronized Walker observations fill only cells without overhead ground or obstacle evidence. They do not replace classified overhead cells. Supported obstacle evidence is expanded by the 0.25 m walker footprint. Road cells represent motor vehicle road intrusion. Cells outside reconstructed evidence remain unknown rather than being treated as free space.

Before scoring any model, we remove obstacle and road evidence within a shared 0.35 m corridor around the recorded path. The same calibrated map is used for every prediction of that query. This calibration makes the recorded path free of measured violations by construction, so the observed space metric does not independently certify the safety of human demonstrations. A predicted path is sampled at intervals no greater than 0.05 m. Its violation indicator is 1 when an uninterrupted run of at least 0.20 m lies in the union of retained obstacle and road masks. If Query initially overlaps this union, only the continuous occupied prefix is ignored. For an episode end goal, a connected terminal goal object with semantic ID 20, 21, or 22 is exempted if detected within 0.8 m of the goal. This map-search tolerance differs from the 1.0 m trajectory Goal tolerance d_{th}. No such exemption applies to other goals. These tests describe the reconstructed Query time scene, not future obstacle motion or closed loop collisions.

Observed space violation and coverage. For path i, let $b_i \in \{0, 1\}$ mark a sustained obstacle or motor vehicle road violation, J_i be its number of sampled points, and $o_{ij} \in \{0, 1\}$ indicate whether point j lies in observed space. Across N paths, Observed Space Violation Rate (OVR) and Coverage are

$$\mathrm{OVR} = \frac{1}{N}\sum_{i=1}^{N} b_i, \qquad \mathrm{Coverage} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{J_i}\sum_{j=1}^{J_i} o_{ij}. \tag{9}$$

Coverage must accompany OVR because violations in unknown space cannot be assessed. OVR is not a physical collision rate.

H Ablation Confidence Intervals

Tables 11, 12, and 13 give the 95% confidence intervals for the paired differences reported in the corresponding main-text ablation tables. All interval limits are in percentage points.

Table 11: Confidence intervals for the Full minus Silence differences in Table 5.

Model	ΔA-Q	ΔBDS	ΔK-Q
Phi 4 Multimodal	$[3.76, 7.23]$	$[7.86, 14.60]$	$[-1.04, 3.21]$
Baichuan Omni 1.5	$[-0.44, 0.25]$	$[-7.82, -1.33]$	$[-1.20, 0.18]$
Qwen2.5 Omni	$[-0.37, 1.97]$	$[-0.56, 3.07]$	$[0.03, 3.68]$
video SALMONN 2+	$[-2.01, 0.01]$	$[-5.15, 0.55]$	$[-10.21, -4.36]$
MiniCPM o 4.5	$[-1.15, 2.73]$	$[-0.64, 4.11]$	$[-2.70, 0.66]$
OmniVinci	$[-0.83, 0.73]$	$[-1.05, 2.31]$	$[-0.43, 0.08]$
TLive Omni	$[-0.35, 0.67]$	$[-1.25, 0.24]$	$[-0.45, 1.31]$
Gemma 4	$[5.12, 9.27]$	$[6.80, 12.29]$	$[-0.01, 3.32]$
Qwen3 Omni Instruct	$[-1.79, 1.47]$	$[-2.45, 1.97]$	$[-5.89, 0.78]$
Qwen3 Omni Thinking	$[-1.65, 2.95]$	$[-3.08, 5.01]$	$[-6.56, 3.72]$
Nemotron 3 Nano Omni	$[-0.19, 3.49]$	$[-1.07, 4.00]$	$[-4.81, 3.94]$
Seed 2.0 Lite	$[-1.10, 0.85]$	$[-1.29, 0.97]$	$[-0.54, 0.21]$
Qwen3.8 Omni Flash	$[-3.47, 4.48]$	$[-4.20, 5.67]$	$[-1.72, 4.63]$

Table 12: Confidence intervals for the Full minus Neither differences in Table 6.

Model	Metric	Δ interval
Qwen2.5 Omni	A-Q	$[24.88, 30.43]$
	K-Q	$[13.65, 23.60]$
MiniCPM o 4.5	A-Q	$[3.20, 9.43]$
	K-Q	$[-18.15, -11.45]$
Gemma 4	A-Q	$[5.69, 11.07]$
	K-Q	$[-6.51, -3.58]$
Qwen3 Omni Instruct	A-Q	$[-4.36, 0.32]$
	K-Q	$[-9.22, -0.39]$
Nemotron 3 Nano Omni	A-Q	$[-1.83, 3.92]$
	K-Q	$[5.32, 18.50]$

Table 13: Confidence intervals for No RGB-D minus RGB-D differences in Table 7. The second interval column corresponds to Flip.

Model	ΔL interval	ΔFlip interval
Qwen2.5 Omni	$[1.10, 4.68]$	$[-0.53, 2.65]$
MiniCPM o 4.5	$[-0.75, 0.79]$	$[-0.26, 2.10]$
Gemma 4	$[-0.09, 0.83]$	$[-1.85, 1.82]$
Qwen3 Omni Instruct	$[-1.77, -0.61]$	$[-1.06, 1.05]$
Nemotron 3 Nano Omni	$[0.17, 1.32]$	$[-1.58, 2.11]$