HEAR 360°: GEOMETRY-AWARE PANORAMIC DEPTH ESTIMATION FROM MULTIPLE POSED ACOUSTIC OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the problem of audio-only panoramic depth estimation from multiple posed binaural observations. Existing audio-to-depth methods typically rely on a single forward-facing binaural pair and predict depth over a limited field of view, providing insufficient directional constraints for panoramic reconstruction. The challenge is further amplified by directional ambiguity and reverberation, which entangle geometric evidence from multiple propagation paths. We propose *HEAR 360°*, which formulates multichannel audio depth estimation as a geometry-aware association rather than a conventional channel-stacking approach. Our framework introduces: (i) pose-conditioned acoustic encoding that preserves observation-specific evidence while incorporating receiver heading and ear identity, and (ii) geometry-aware cross-attention that represents each panoramic location as a 3D ray and associates it with the most relevant posed acoustic observations. We evaluate *HEAR 360°* on panoramic depth benchmarks derived from Matterport3D and Replica. Experiments show that *HEAR 360°* outperforms existing echo-based methods and controlled channel-stacking baselines with substantially fewer parameters. Further analyses show consistent gains from additional observations, as well as robustness to irregular observation subsets and unseen receiver headings, supporting the effectiveness of pose-aware association.

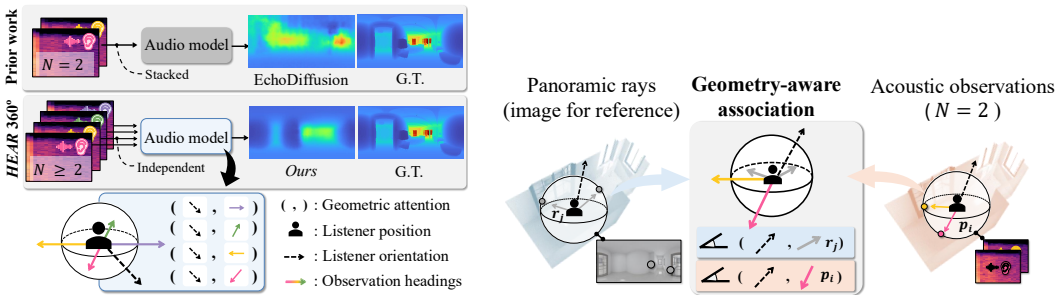
1 INTRODUCTION

Hearing is one of the fundamental senses that enables humans to perceive and interact with the world. By perceiving how sound reverberates within a space, humans can infer the approximate structure and scale of their surroundings even without vision (Kolarik et al., 2013; Dokmanić et al., 2013). For example, some blind individuals use self-generated sounds, such as mouth clicks, and interpret the returning echoes to detect nearby obstacles (Thaler et al., 2017). This is possible because reflections from surrounding walls and objects form reverberation that carries indirect but informative cues about the geometry of the environment.

Building on these observations, prior work show that acoustic observations support various spatial perception tasks, including indoor room geometry recovery (Dokmanić et al., 2013), localization and mapping (Kreković et al., 2016), and indoor structure reconstruction (Yeon et al., 2024). Among these tasks, we focus on depth estimation, which converts acoustic observations in the form of echo responses into a dense representation of the surrounding geometry.

Although echo-based spatial perception methods often collect observations from multiple microphone positions (Kreković et al., 2016) or use multichannel microphone arrays (Yeon et al., 2024) to obtain richer cues, recent echo-based depth estimation approaches primarily rely on binaural sensing, which uses two microphones to capture interaural time and level differences (Brunetto et al., 2023; Yun et al., 2023). This contrast raises a key question: Is a single forward-facing binaural observation sufficient for depth prediction? The question also extends to the output representation. Existing approaches typically predict front-facing perspective depth maps with a limited field of view. Since sound propagates and reflects throughout the surrounding space, and complementary microphone positions capture cues from different directions, panoramic depth may provide a more natural listener-centered representation of the environment.

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.



(i) Audio-only panoramic depth estimation **(ii)** Geometric association

Figure 1: **Motivation of *HEAR 360°*.** (i) Prior methods mainly rely on a single binaural observation or implicitly combine multiple signals through channel stacking, whereas *HEAR 360°* uses multiple posed acoustic observations. (ii) Each panoramic location is represented by a 3D ray, and its relationship with each posed acoustic observation provides an explicit cue for associating acoustic evidence with panoramic directions.

A single binaural observation is underconstrained for dense panoramic depth prediction due to two factors. *Directional ambiguity:* A binaural pair encodes sound direction through interaural time and level differences. However, distinct 3D points can produce similar binaural observations by yielding the same interaural time difference, as in the *cone of confusion* phenomenon (Blauert, 1996). Such ambiguity makes it difficult to associate acoustic evidence with the correct panoramic depth. *Reverberation entanglement:* Reverberation makes surrounding geometry acoustically observable, yet it is received as a superposition of multiple propagation paths. The direct path, reflections, and higher-order echoes can overlap. Thus, although reverberation contains geometric information, a single binaural observation provides insufficient spatial constraints to associate this mixed evidence with the correct panoramic directions.

If these ambiguities arise partly from insufficient directional coverage, a natural solution is to acquire additional acoustic observations at multiple poses, allowing ambiguous evidence from one heading to be examined from another. The remaining challenge is how to combine them. A straightforward approach is to stack their spectrograms along the input-channel dimension (Valverde et al., 2021; Jiang et al., 2022; Nguyen et al., 2022). However, channel stacking treats the observations as an ordered collection of signals without explicitly representing their orientations or geometric relationships, such as front–back correspondence and angular proximity. The network must therefore infer these relationships implicitly, which does not consistently improve depth prediction (see Tab. 1).

These observations suggest that multichannel audio depth estimation should not be treated solely through channel stacking. Instead, it can be formulated as a geometry-aware association problem, where the model determines which posed acoustic observations are most relevant to each panoramic direction, as in Fig. 1 (ii). Based on this idea, we propose *HEAR 360°* an echo-based framework for panoramic depth estimation from multiple posed acoustic observations. *HEAR 360°* extends single binaural sensing with observations from multiple known receiver headings. Each ear-specific observation is processed by a shared acoustic encoder and conditioned on the receiver orientation and ear identity, while each location in the equirectangular panorama is represented by a 3D viewing ray from the listener center. These geometric relationships guide attention, enabling each panoramic location to aggregate relevant acoustic evidence.

We evaluate *HEAR 360°* on panoramic depth benchmarks derived from Matterport3D (Chang et al., 2017) and Replica (Straub et al., 2019). *HEAR 360°* outperforms existing echo-based methods and channel-stacking baselines with substantially fewer parameters. Our analyses show that performance depends not only on angular coverage but also on associating panoramic directions with posed acoustic observations, and that pose-aware association remains robust to irregular observation subsets and unseen receiver headings.

Our contributions are summarized as follows:

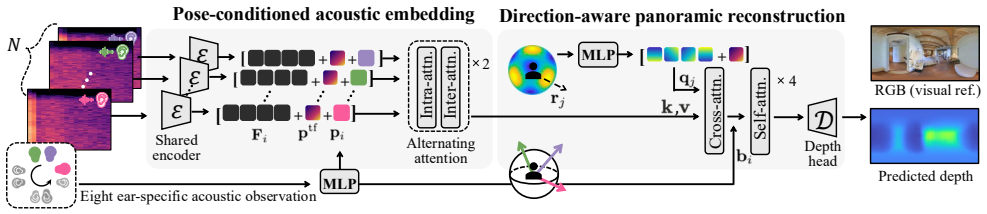


Figure 2: **Pipeline of *HEAR 360°*.** Given multiple posed acoustic observations, *HEAR 360°* extracts and fuses pose-conditioned acoustic features, associates them with panoramic directions through geometry-aware cross-attention, and decodes the resulting representation into a listener-centered panoramic depth map.

- We extend prior audio-to-depth estimation beyond single forward-facing binaural observations and limited-field-of-view predictions to panoramic depth estimation from multiple posed binaural observations.
- We formulate multichannel audio depth estimation as a geometry-aware association problem and propose *HEAR 360°*, a framework that identifies the acoustic observations most relevant to each panoramic 3D ray.
- We demonstrate that *HEAR 360°* outperforms existing echo-based methods and controlled channel-stacking baselines while using substantially fewer parameters, validating the effectiveness of its pose-aware design.

2 Related Work

Echo-based spatial perception. Echoes have been used for localization and mapping Kreković et al. (2016); Yang et al. (2022) and room geometry recovery Poschadel et al. (2021); Yeon et al. (2024). These tasks benefit from richer spatial acquisition. EchoSLAM Kreković et al. (2016) aggregates measurements along a trajectory, while room geometry methods use multichannel higher-order Ambisonics Poschadel et al. (2021) or spherical microphone arrays Yeon et al. (2024) to capture broader directional information.

Depth estimation represents another line of echo-based spatial perception. However, while broader spatial-perception methods have benefited from richer spatial acoustic acquisition, echo-based depth estimation has remained limited to a single binaural observation. VisualEchoes Gao et al. (2020) uses simulated echoes to learn spatial visual representations, while BatVision Christensen et al. (2019); Brunetto et al. (2023) predicts depth directly from binaural echoes. Subsequent methods improve geometric reasoning through angular-spectrum features Irie et al. (2022); Liu et al. (2025), diffusion-based prediction Zhang et al. (2025), and cross-modal distillation for omnidirectional scene understanding Yun et al. (2023). Nevertheless, existing approaches, including panoramic depth methods Yun et al. (2023); Liu et al. (2024), rely on a single binaural observation. In contrast, *HEAR 360°* estimates panoramic depth from multiple posed binaural observations and explicitly models the geometric relationships between receiver orientations and panoramic directions.

Direction-aware representation. Although the modalities differ, both posed echo observations and panoramic depth are associated with directions on the sphere. Thus, geometric representations developed for panoramic vision provide a natural basis for encoding their directional relationships. DA^2 Li et al. (2025) encodes polar and azimuth angles as spherical embeddings for cross-attention in panoramic depth estimation. PanoVGGT Guo et al. (2026) uses spherical positional embeddings for panoramic 3D reconstruction, while PanFusion Zhang et al. (2024) and SpheRoPE Hirschorn et al. (2026) employ spherical positional embeddings to model panoramic geometry and preserve spherical topology in generative models. Together, these approaches demonstrate the effectiveness of explicitly encoding spherical geometry for panoramic learning. Building on these approaches, we represent each panoramic pixel by its 3D ray and use the resulting ray embeddings to guide cross-attention with multichannel acoustic features, enabling direction-aware acoustic fusion.

3 METHOD

Our goal is audio-only panoramic depth estimation from multiple posed acoustic observations. To this end, we propose *HEAR 360°*, a framework that explicitly models the geometric relationships between acoustic observations and their panoramic directions rather than relying on implicit channel stacking. *HEAR 360°* consists of two main components: pose-conditioned acoustic encoding and direction-aware panoramic reconstruction. First, each acoustic channel is processed by a shared encoder and conditioned on its time–frequency position, receiver heading, and ear identity. The resulting features are fused through alternating intra- and inter-observations attention to produce a unified acoustic representation. Next, geometric queries over a coarse panoramic grid, represented by 3D viewing rays, attend to the acoustic features to encode the relationship between each ray and receiver pose. The resulting direction-aware panoramic features are refined and decoded into a dense depth map. An overview of *HEAR 360°* is shown in Fig. 2.

3.1 PROBLEM FORMULATION

We consider an acoustic sensing device at a fixed listener center. At one or more known receiver headings, the device emits a short sweep and records the corresponding binaural responses. Receiver heading and ear identity are known during acquisition. Since all observations share the same listener center, this information specifies their relative sensing geometry without requiring precise sensor calibration or full 6-DoF pose estimation. As in standard acoustic sensing, the excitation waveform and source level are fixed and known.

We represent each microphone channel recorded at a particular heading as a posed acoustic observation (X_i, ψ_i, e_i), where X_i denotes its magnitude spectrogram, ψ_i denotes the receiver heading represented by a yaw angle, and $e_i \in \{-1, +1\}$ indicates the left or right ear. The input set is

$$\mathcal{A} = (X_i, \psi_i, e_i)_{i=1}^{N}, \tag{1}$$

where N is the total number of heading–ear observations; thus, M binaural headings yield $N = 2M$ observations. Given $\mathcal{A}$, our goal is to predict a listener-centered panoramic depth map in panoramic coordinates:

$$\hat{\mathbf{D}} = f_{\mathbf{\Theta}}(\mathcal{A}) \in [0, d_{\max}]^{H \times W}, \tag{2}$$

where $\mathbf{\Theta}$ denotes the learnable model parameters, H and W are the height and width of the panoramic depth map, and $d_{\max}$ is the maximum sensing range. Each pixel of $\hat{\mathbf{D}}$ represents the distance from the listener center to the first surface intersected by its corresponding panoramic ray.

3.2 POSE-CONDITIONED ACOUSTIC EMBEDDING

Per-observation feature extraction. Each acoustic observation is acquired at a different receiver heading and captures complementary evidence. We process each observation independently using a shared acoustic encoder. Formally, the acoustic tokens for the i-th observation are computed as

$$\mathbf{F}_i = \mathcal{E}_{\text{audio}}(\mathbf{X}_i) \in \mathbb{R}^{K \times C}, \tag{3}$$

where K denotes the number of acoustic tokens, C their feature dimension, and $\mathcal{E}_{\text{audio}}$ the shared acoustic encoder. In a channel-stacked encoder, as used in prior works (Valverde et al., 2021; Jiang et al., 2022; Nguyen et al., 2022), standard convolutions mix information from all observations from the first layer by aggregating local time–frequency neighborhoods across input channels. This entangles the observations, requiring their geometric relationships to be learned implicitly. In contrast, our shared encoder maps all observations into a common feature space while retaining their observation-specific acoustic information distinct, allowing them to be fused later according to their receiver poses.

Time–frequency and pose conditioning. The convolutional encoder $\mathcal{E}_{\text{audio}}$ extracts local acoustic patterns and produces a time–frequency feature map. However, its output features do not explicitly encode their absolute time–frequency coordinates or the pose at which each observation was acquired. We therefore condition the acoustic tokens on both their time–frequency coordinates and observation. Let $\mathbf{p}^{\text{tf}} \in \mathbb{R}^{K \times C}$ denote the positional encoding of the normalized time–frequency

coordinates associated with the acoustic tokens. For the i-th observation, we encode its receiver heading and ear identity into a observation embedding $\mathbf{p}_i \in \mathbb{R}^C$ and update its acoustic tokens as

$$\begin{aligned} \mathbf{p}_i &= \mathrm{MLP}_{\mathrm{pose}}\left([\sin\psi_i,\ \cos\psi_i,\ e_i]\right), \\ \mathbf{F}_i &= \mathbf{F}_i + \mathbf{p}^{\mathrm{tf}} + \mathbf{p}_i. \end{aligned} \tag{4}$$

Alternating attention. To enrich the acoustic representation, we model local time–frequency relationships within each observation and complementary acoustic evidence across observations using an alternating-attention inspired by VGGT (Wang et al., 2025). *Intra-observation attention* independently refines the K tokens of each observation without exchanging information across observations. We condition the attention and feed-forward layers on the observation embedding $\mathbf{p}_i$ using Adaptive Layer Normalization (Peebles & Xie, 2023). The modulation layers are zero-initialized, such that the blocks initially perform pose-independent transformations and gradually learn pose-conditioned processing. *Inter-observation attention* exchanges information across the N observations, enabling each token to aggregate complementary evidence from other receiver headings and ears while accounting for their positional relationships through the observation embeddings. We alternate the intra- and inter-observation attention blocks twice.

3.3 Direction-aware Panoramic Reconstruction

We next associate the resulting acoustic features with panoramic directions to reconstruct the target depth map. The acoustic features are organized over time and frequency, whereas the depth map is defined over elevation and azimuth, resulting in no direct correspondence between their grid locations and creating a domain mismatch. To bridge this mismatch, we introduce a separate set of panoramic queries whose locations are represented by 3D rays. Each ray defines a viewing direction on the sphere and provides a geometric reference for identifying which posed acoustic observations contain relevant evidence for that direction.

Ray-conditioned panoramic queries. We define J learnable panoramic queries on a coarse panoramic grid. For the j-th query at elevation ϕ_j and azimuth θ_j, its unit ray $\mathbf{r}_j$ and feature $\mathbf{q}_j$ are given by

$$\mathbf{r}_j = \begin{bmatrix} \cos\phi_j \sin\theta_j \\ \sin\phi_j \\ \cos\phi_j \cos\theta_j \end{bmatrix}, \quad \mathbf{q}_j = \mathbf{q}_j^{\mathrm{learn}} + \mathrm{MLP}_{\mathrm{ray}}\left(\mathbf{r}_j\right), \tag{5}$$

where $\mathbf{q}_j^{\mathrm{learn}} \in \mathbb{R}^C$ is a learnable latent representation, and $\mathrm{MLP}_{\mathrm{ray}}$ projects the unit ray into the query feature space to explicitly encode its viewing direction. This geometric encoding allows the subsequent geometry-aware cross-attention to associate each panoramic query with acoustic evidence from the most relevant posed observations. We use a 16×32 panoramic grid, yielding $J = 512$ queries.

Geometry-aware cross-attention. We employ cross-attention to associate each panoramic query with the most relevant acoustic tokens. To make this association geometry-aware, we explicitly model the directional relationship between each panoramic query and the posed acoustic observations and introduce a geometry-dependent bias into the attention logits. Because the observations are acquired at different receiver yaw angles, the same panoramic direction has a different relative direction for each observation. For panoramic ray $\mathbf{r}_j$ and acoustic observation i, we therefore rotate the ray by the inverse receiver yaw:

$$\tilde{\mathbf{r}}_i = \mathrm{Rotate}(-\psi_i) \cdot \mathbf{r}_j, \tag{6}$$

where $\mathrm{Rotate}(-\psi_i)$ transforms the panoramic ray into the local coordinate frame of the i-th observation. The resulting local ray $\tilde{\mathbf{r}}_i$ specifies the queried panoramic direction relative to the corresponding receiver heading. We then combine the local ray with the ear identity to define the geometric relationship between the panoramic query and the observation:

$$\mathbf{b}_i = \mathrm{MLP}_{\mathrm{geo}}\left([\tilde{\mathbf{r}}_i,\ e_i]\right), \tag{7}$$

where e_i denotes the ear side of the i-th observation. The geometric MLP maps this relationship to an attention bias $\mathbf{b}_i$, allowing the subsequent cross-attention to favor observations geometrically aligned with the panoramic query. Using the initialized panoramic queries as queries and the fused acoustic features as keys and values, we obtain the panoramic features as:

$$\tilde{\mathbf{F}}_i = \mathrm{CrossAttention}\left(\mathbf{q}_j,\ \mathbf{F}_i,\ \mathbf{F}_i\right) + \mathbf{b}_i, \tag{8}$$

where $\mathbf{b}_i$ serves as a geometry-dependent bias added to the cross-attention output. Feature similarity captures the relevance of acoustic content, while the geometric bias accounts for its directional consistency with the panoramic query. Because the bias modulates the scores rather than imposing a fixed correspondence, each query can aggregate complementary acoustic evidence from multiple observations.

Feature refinement and depth decoding. The geometry-aware cross-attention produces one acoustic feature for each panoramic query. Four self-attention blocks further refine these features by modeling relationships across panoramic directions, promoting spatially coherent depth predictions. The refined features are reshaped into a 16×32 panoramic grid and progressively upsampled to the target resolution, where a depth prediction head regresses the panoramic depth map $\hat{\mathbf{D}}$.

3.4 Training Objective

Following BatVision (Brunetto et al., 2023), we train *HEAR 360°* using an L_1 depth loss. Ground-truth depths are normalized by the maximum sensing range of 10m, and the loss is evaluated only at valid pixels using a binary mask.

4 Experiment

4.1 Implementation Details

We convert each waveform into a magnitude short-time Fourier transform (STFT). Following prior work (Brunetto et al., 2023), the waveform is cropped according to the round-trip travel time for a maximum sensing range of 10m. We use an FFT size of 512, a window length of 400, and a hop length of 160, and resize the resulting spectrogram to 256×512. We train *HEAR 360°* for 40 epochs on a single NVIDIA RTX A6000 GPU using AdamW with a learning rate of 5×10^{-4}, a weight decay of 10^{-4}, and a batch size of 16. When more than four observations are available, we randomly mask a subset of observation tokens during training; otherwise, all tokens are retained. We use an exponential moving average of the model parameters with a decay factor of 0.999. Additional implementation and architectural details are in the suppl. material.

Dataset construction. At each listener center, M binaural headings provide left- and right-ear responses, yielding $N = 2M$ posed acoustic observations. Following prior work (Yun et al., 2023), we construct panoramic depth estimation benchmarks on Replica (Straub et al., 2019) and Matterport3D (Chang et al., 2017) using SoundSpaces 2.0 (Chen et al., 2022), which simulates acoustic responses from 3D scene meshes. At each listener position, we co-locate the emitter and binaural receiver, emit a 3ms sweep, and render binaural responses at headings of $0°$, $90°$, $180°$, and $270°$. This produces four binaural captures, or eight ear-specific acoustic observations, sampled at 48kHz. The corresponding listener-centered panoramic depth maps are rendered at a resolution of 256×512. For each scene, we uniformly sample 100 continuous listener positions over the navigable surface using the `PathFinder.get_random_navigable_point` interface of Habitat-Sim (Savva et al., 2019), yielding 400 binaural captures (800 acoustic observations) per scene. Additional dataset construction details are in the suppl. material.

Replica (Straub et al., 2019). Replica contains 18 indoor scenes. We adopt a scene-disjoint split with 12 scenes for training, three for validation, and three for testing. The validation scenes are *apartment_1*, *frl_apartment_4*, and *office_3*, while the test scenes are *apartment_2*, *frl_apartment_5*, and *office_4*. This split ensures evaluation on unseen environments and prevents geometrically and acoustically similar locations from appearing across training and evaluation sets.

Matterport3D (Chang et al., 2017). Matterport3D contains 90 indoor scenes. We use a scene-disjoint split with 72 scenes for training, nine for validation, and nine for testing.

4.2 Baselines

No existing method directly matches our setting in terms of input representation, number of observations, target task, and output resolution. We therefore adapt related methods to our setting and use them as baselines for a fair comparison.

Ch.	Method	#P	Matterport3D						Replica					
			RMSE↓	Abs Rel↓	$\log_{10}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE↓	Abs Rel↓	$\log_{10}$ ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
2	Pretrained ResNet*	30.0	1.453	0.611	0.191	0.420	0.636	0.772	0.709	0.283	0.100	0.681	0.844	0.914
	Pretrained ViT*	89.9	1.390	0.569	0.180	0.447	0.666	0.796	0.623	0.193	0.081	0.744	0.890	0.945
	EchoScan‡	41.3	1.455	0.627	0.194	0.410	0.625	0.765	0.666	0.195	0.080	0.732	0.897	0.952
	BatVision	54.5	1.398	0.544	0.179	0.442	0.664	0.796	0.574	0.179	0.075	0.761	0.905	0.955
	EchoDiffusion	156.1	**1.381**	0.526	**0.176**	**0.453**	**0.673**	**0.801**	**0.547**	0.166	0.072	0.776	0.906	0.956
	HEAR 360°	15.4	1.387	**0.522**	0.177	0.449	0.668	0.798	0.559	**0.157**	**0.067**	**0.797**	**0.918**	**0.961**
4	Pretrained ResNet*	30.0	1.402	0.557	0.178	0.446	0.666	0.796	0.687	0.281	0.099	0.687	0.847	0.916
	Pretrained ViT*	89.9	1.306	0.414	0.153	0.510	0.727	0.841	0.616	0.181	0.077	0.753	0.902	0.954
	EchoScan‡	41.3	1.437	0.583	0.188	0.420	0.639	0.777	0.658	0.194	0.079	0.742	0.901	0.954
	BatVision	54.5	1.310	0.461	0.157	0.494	0.716	0.835	0.545	0.171	0.072	0.772	0.911	0.959
	EchoDiffusion	156.1	1.289	0.399	0.150	0.526	0.738	0.847	0.523	0.157	0.067	0.791	0.915	0.960
	HEAR 360°	15.4	**1.276**	**0.376**	**0.144**	**0.539**	**0.749**	**0.855**	**0.514**	**0.140**	**0.061**	**0.826**	**0.932**	**0.967**
8	Pretrained ResNet*	30.0	1.427	0.507	0.180	0.431	0.655	0.793	0.663	0.263	0.096	0.689	0.852	0.920
	Pretrained ViT*	89.9	1.366	0.467	0.170	0.457	0.685	0.814	0.567	0.153	0.068	0.789	0.911	0.959
	EchoScan‡	41.3	1.460	0.681	0.197	0.404	0.624	0.763	0.637	0.187	0.076	0.754	0.906	0.956
	BatVision	54.5	1.321	0.446	0.157	0.500	0.721	0.837	0.516	0.164	0.067	0.792	0.920	0.963
	EchoDiffusion	156.1	1.236	0.386	0.140	0.544	0.757	0.862	0.513	0.150	0.065	0.803	0.920	0.963
	HEAR 360°	15.4	**1.225**	**0.346**	**0.137**	**0.553**	**0.768**	**0.870**	**0.481**	**0.127**	**0.055**	**0.847**	**0.938**	**0.972**

Table 1: **Panoramic depth estimation results.** Network parameters (#P) are reported in millions. The best and second-best results are shown in **bold** and underlined, respectively.

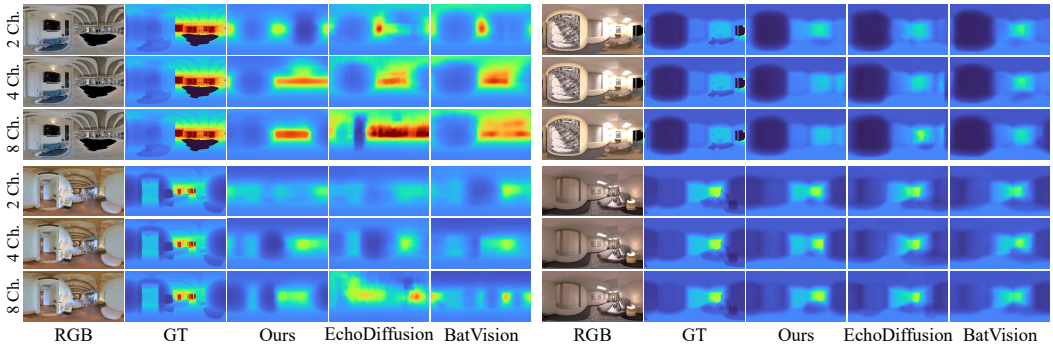


Figure 3: **Qualitative comparison on Matterport3D (Left) and Replica (Right)**.

Generic encoder-decoder baselines. We construct conventional encoder–decoder baselines using ImageNet-pretrained ResNet-50 (He et al., 2016) and ViT-B/16 (Dosovitskiy et al., 2021). The spectrograms from multiple observations are stacked along the channel dimension and projected to three channels using a learnable 1×1 convolution, making them compatible with the pretrained vision backbones. The resulting features are then passed to a depth decoder to predict the panoramic depth map. These baselines evaluate if a conventional encoder–decoder with simple channel adaptation can address the task. We denote them as ResNet* and ViT*.

4.3 Dataset

BatVision (Brunetto et al., 2023). BatVision (Brunetto et al., 2023) predicts pinhole depth from channel-stacked binaural spectrograms using a U-Net architecture. We retain its architecture and modify only the input channel count and output resolution to panoramic depth prediction. This baseline evaluates whether an existing audio-to-depth architecture can be directly extended to multiple posed observations without explicit geometric modeling.

EchoScan (Yeon et al., 2024). EchoScan predicts 3D room layouts, represented by a floor plan and room height, from six-channel waveforms using a 1D ResNet encoder. We preserve its waveform encoder, vary the input channel count, and replace its layout prediction heads with a depth decoder. This baseline evaluates whether a waveform-based room-geometry model can transfer to our depth estimation task. We denote the adapted model as EchoScan‡.

EchoDiffusion (Zhang et al., 2025). EchoDiffusion uses complementary waveforms and spectrograms for depth prediction. We retain both input pathways while adjusting their respective channel counts and the output resolution. Because its released implementation does not include the iterative diffusion procedure described in the paper, we adapt the publicly available deterministic single-pass

(i) Component ablation

Model	TF Emb.	Pose Emb.	Ray Emb.	Geo Bias	RMSE↓	Abs Rel↓	$\log_{10}$ ↓
(a)					0.987	0.421	0.149
(b)		✓	✓	✓	0.488	0.129	0.056
(c)	✓		✓	✓	0.542	0.148	0.065
(d)	✓	✓		✓	0.498	0.132	0.057
(e)	✓	✓	✓		0.502	0.136	0.059
(f)	✓	✓			0.560	0.161	0.067
Ours	✓	✓	✓	✓	**0.481**	**0.127**	**0.055**

(ii) Input-cue ablation

Model	Protocol	Pose Cue	Ear Cue	RMSE↓	Abs Rel↓	$\log_{10}$ ↓
(a)	inference		✓	1.122	0.502	0.179
(b)	inference	✓		0.963	0.436	0.142
(c)	inference		✓	1.015	0.473	0.155
(d)	retraining	✓		0.542	0.149	0.064
Ours	retraining	✓	✓	**0.481**	**0.127**	**0.055**

Table 2: **Ablation studies on Replica.** (i) Contribution of the proposed components. (ii) Effect of receiver-pose and ear-identity cues under inference-time removal and retraining.

architecture. This baseline evaluates whether combining waveform- and spectrogram-based representations can address our task without explicit geometric modeling.

4.4 Quantitative Evaluation

We evaluate panoramic depth estimation on the scene-disjoint Matterport3D and Replica benchmarks using different numbers of posed acoustic observations. Results are in Tab. 1. Key findings are as follows:

(1) *HEAR 360° gets the best overall performance with multiple observations.* With 2 observations, *HEAR 360°* remains competitive with EchoDiffusion, achieving the best Abs Rel on both datasets and leading five of six metrics on Replica. With 4 and 8 observations, *HEAR 360°*consistently achieves the best results across all metrics on both benchmarks.

(2) *Explicit geometric modeling enables effective use of additional observations.* The performance of *HEAR 360°* improves monotonically from 2 to 4 to 8 observations across every metric on both datasets. Although most baselines also benefit from additional observations, several show limited or inconsistent gains at 8 observations, particularly on Matterport3D, whereas *HEAR 360°* continues to improve. This demonstrates that explicitly modeling a geometric evidence-association problem is more effective than simply increasing the number of stacked input channels.

(3) *HEAR 360°provides strong performance with substantially fewer parameters.* At 8 observations, *HEAR 360°* reduces Abs Rel over EchoDiffusion by 10.4% on Matterport3D (0.346 vs. 0.386) and 15.3% on Replica (0.127 vs. 0.150). Notably, *HEAR 360°* achieves these gains with only 15.4M parameters, approximately one-tenth of EchoDiffusion's parameters.

4.5 Qualitative Evaluation

As shown in Fig. 3, *HEAR 360°* reconstructs the overall room structure while preserving local geometric details, including openings, nearby surfaces, and depth discontinuities. In contrast, EchoDiffusion and BatVision often produce smoother or blurred predictions, making individual structures less spatially distinct. Visualizations from the same listener position with different numbers of observations further show that additional posed observations improve spatial coherence in regions that are ambiguous from a single binaural pair. RGB panoramas are provided only for reference and are not used as input. More results are in the suppl. material.

4.6 Ablation Study

We conduct ablation studies to evaluate the contribution of each design choice. Unless otherwise specified, all experiments use 8 observations and are performed on Replica.

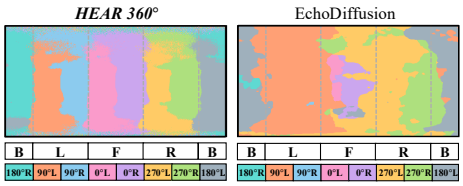
Effect of proposed components. Tab. 2 (i) evaluates each component by training every variant from scratch. Removing any individual component in variants (b) to (e) degrades performance relative to *HEAR 360°* while all variants still outperform the baseline model (a). Variant (f), which removes both the ray embedding and geometric bias, yields the largest degradation, highlighting the importance of explicitly associating panoramic directions with posed acoustic observations.

Effect of input cues in acoustic observations. We examine the impact of the receiver heading pose and ear identity cues defined in Eq. (1). For each cue, we evaluate both inference-time removal and

(i) Selection of acoustic observations

Setup	0°		90°		180°		270°		EchoDiff. (Abs Rel↓)	*HEAR 360°* (Abs Rel↓)
	L	R	L	R	L	R	L	R		
(a)	✓	✓			✓	✓			0.399	**0.376**
(b)			✓	✓			✓	✓	0.402	**0.385**
(c)	✓	✓		✓			✓		0.453	**0.378**

(ii) Acoustic observation attribution



(iii) Generalization to unseen heading orientations

Microphone Orientations	Model	RMSE↓	Abs Rel↓	$\log_{10}$ ↓
$\{45°, 135°, 225°, 315°\}$	BatVision	0.711	0.255	0.093
	EchoDiffusion	0.709	0.223	0.092
	Ours	**0.616**	**0.163**	**0.076**
$\{30°, 120°, 210°, 300°\}$	BatVision	0.731	0.264	0.095
	EchoDiffusion	0.726	0.222	0.092
	Ours	**0.626**	**0.163**	**0.073**
$\{-45°, 135°, 120°, 300°\}$	BatVision	0.920	0.410	0.138
	EchoDiffusion	0.871	0.338	0.129
	Ours	**0.685**	**0.199**	**0.093**

Figure 4: **Additional analysis of pose-aware acoustic aggregation.** (i) Performance under different subsets of four acoustic observations. (ii) Dominant acoustic observation for each panoramic direction under eight observations. (iii) Generalization to unseen receiver-heading configurations without retraining. F, L, R, and B denote front, left, right, and back, respectively.

retraining without that cue. As shown in Tab. 2 (ii), removing either cue degrades performance, with a substantially larger drop for pose cue. This indicates that receiver heading pose provides the primary geometric signal for associating posed acoustic observations with panoramic directions, while ear identity contributes complementary directional information.

Effect of observation selection. We compare three four-observation subsets with different binaural pairings and heading coverage in Fig. 4 (i). (a) and (b) retain two complete binaural pairs, under which both methods perform consistently. (c) combines one complete binaural pair with two single-ear observations from different headings. EchoDiffusion degrades substantially in this setup, whereas *HEAR 360°* remains stable, indicating that its pose-aware aggregation is robust to irregular observation subsets and does not require every observation to form a complete binaural pair.

4.7 Additional Analysis

Visualization of acoustic observation attribution. We visualize how *HEAR 360°* associates posed acoustic observations with panoramic directions. For each panoramic query, we aggregate its cross-attention weights over the tokens of each observation and assign it to the observation with the highest aggregate weight. As shown in Fig. 4 (ii), *HEAR 360°* produces contiguous regions broadly aligned with the corresponding receiver headings. EchoDiffusion exhibits a similar directional tendency but yields more fragmented patterns, suggesting a less structured association between acoustic observations and panoramic directions. These results illustrate that our pose-aware attention organizes observation attribution according to acquisition geometry.

Generalization to unseen receiver heading orientations. We evaluate *HEAR 360°* trained on fixed receiver headings $\{0°, 90°, 180°, 270°\}$, with three unseen configurations without retraining (Fig. 4 (iii)). The first two jointly rotate all headings by $45°$ and $30°$, respectively, while the third independently perturbs each heading. *HEAR 360°* achieves the best results across all configurations and degrades less than BatVision and EchoDiffusion relative to the original setup, indicating that our pose-aware design is more robust to unseen receiver headings, including irregular configurations.

5 Conclusion

In this paper, we introduce *HEAR 360°*, a framework for audio-only panoramic depth estimation that associates posed acoustic observations with listener-centered panoramic directions. By combining pose-conditioned acoustic encoding with geometry-aware cross-attention, our approach overcomes the limited directional constraints of single binaural sensing and the implicit geometric reasoning required by channel-stacking methods. Through extensive evaluation on Matterport3D and Replica, we demonstrate improved panoramic depth estimation with substantially fewer parameters, together with consistent gains from additional observations and robustness to irregular observation subsets and unseen receiver headings. Our results highlight the importance of explicitly associating acoustic evidence with panoramic geometry for dense spatial perception.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, manuscript drafting and language editing, and drafting and editing software code. All AI-assisted text and code were reviewed, revised, and, where applicable, tested by the authors. We take responsibility for the final content of this work, including text, claims, and code produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed implementation, training, evaluation, and dataset-construction protocols in the main paper and appendix. Upon publication, we will release the data-generation pipeline, scene splits, receiver poses and metadata, and training and inference code required to reproduce our benchmarks and experimental results.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

Jens Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 1996.

Amandine Brunetto, Sascha Hornauer, Stella X. Yu, and Fabien Moutarde. The audio-visual batvision dataset for research on sight and sound. In *IROS*, 2023.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *ICCV*, 2017.

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 *ECCV*, 2020.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS*, 2022.

Jesper Haahr Christensen, Sascha Hornauer, and Stella Yu. Batvision: Learning to see 3d spatial layout with two ears. *arXiv preprint arXiv:1912.07011*, 2019.

Ivan Dokmanić, Yue M. Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In *ECCV*, 2020.

Yijing Guo, Mengjun Chao, Luo Wang, Tianyang Zhao, Haizhao Dai, Yingliang Zhang, Jingyi Yu, and Yujiao Shi. Panovggt: Feed-forward 3d reconstruction from panoramic imagery. In *CVPR*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Or Hirschorn, Aaron Olender, Eli Alshan, Ianir Ideses, Lior Fritz, and Sagie Benaim. Spherope: Zero-shot optimization-free 360 panorama generation with spherical rope. *arXiv preprint arXiv:2606.32033*, 2026.

Go Irie, Takashi Shibata, and Akisato Kimura. Co-attention-guided bilinear model for echo-based depth estimation. In *ICASSP*, 2022.

Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. In *CVPR*, 2022.

Andrew J Kolarik, Shahina Pardhan, Silvia Cirstea, and Brian C J Moore. Using acoustic information to perceive room size: Effects of blindness, room reverberation time, and stimulus. *Perception*, 2013.

Miranda Kreković, Ivan Dokmanić, and Martin Vetterli. Echoslam: Simultaneous localization and mapping with acoustic echoes. In *ICASSP*, 2016.

Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. Da^2: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025.

Guangyao Liu, Weimeng Cui, Yuzhang Xi, Liu Yang, Peixuan Hu, He Kong, and Zhi Wang. Sagenet: Binaural echo-based 3d depth estimation with sparse angular queries and refined geometric cues. In *IROS*, 2025.

Xiaohu Liu, Amandine Brunetto, Sascha Hornauer, Fabien Moutarde, and Jialiang Lu. Pano-echo: Panoramic depth prediction enhancement with echo features. In *CAI*, 2024.

Thi Ngoc Tho Nguyen, Douglas L. Jones, Karn N. Watcharasupat, Huy Phan, and Woon-Seng Gan. SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays. In *ICASSP*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Nils Poschadel, Robert Hupke, Stephan Preihs, and Jürgen Peissig. Room Geometry Estimation from Higher-Order Ambisonics Signals using Convolutional Recurrent Neural Networks. In *AES Convention 150*. AES, Journal of the Audio Engineering Society, May 2021.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied ai research. In *ICCV*, 2019.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Lore Thaler, Galen M Reich, Xinyu Zhang, Dinghe Wang, Graeme E Smith, Zeng Tao, Raja Syamsul Azmir Bin Raja Abdullah, Mikhail Cherniakov, Christopher J Baker, Daniel Kish, et al. Mouth-clicks used by blind expert human echolocators–signal description and model based signal synthesis. *PLoS computational biology*, 2017.

Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *CVPR*, 2021.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.

Karren Yang, Michael Firman, Eric Brachmann, and Clément Godard. Camera pose estimation and localization with active audio sensing. In *ECCV*, 2022.

Inmo Yeon, Iljoo Jeong, Seungchul Lee, and Jung-Woo Choi. Echoscan: Scanning complex room geometries via acoustic echoes. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Heeseung Yun, Joonil Na, and Gunhee Kim. Dense 2d-3d indoor prediction with sound via aligned cross-modal distillation. In *ICCV*, 2023.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *CVPR*, 2024.

Wenjie Zhang, Jun Yin, Long Ma, Peng Yu, Xiaoheng Jiang, Zhen Tian, and Mingliang Xu. Echodiffusion: Waveform conditioned diffusion models for echo-based depth estimation. In *AAAI*, 2025.

A Appendix

Our supplementary material consists of two parts: (1) an interactive multimedia appendix (`.html`) containing trajectory-level qualitative results on Matterport3D, and (2) a technical appendix (*i.e.,* this document).

B Multimedia Appendix

In addition to the main paper and this technical appendix, we provide an interactive HTML appendix with trajectory-level qualitative results on the Matterport3D test scenes. We construct each sequence by sampling receiver locations along a continuous path. We use Matterport3D because its larger-scale environments allow relatively long paths. Due to the supplementary file-size limit, we include a representative subset of the constructed sequences.

At each sampled location, we acquire four posed binaural observations, yielding eight ear-specific acoustic observations. Panoramic depth is predicted independently at each location. The interactive viewer presents these samples in path order. It displays the RGB panorama, ground-truth depth, predicted depth, and receiver trajectory in bird's-eye view. Clicking a sample also plays the binaural chirp response acquired at the forward-facing ($0°$) heading. The RGB panorama is provided only for reference and is not used by the model.

These views make it possible to inspect how the estimated surrounding geometry changes across nearby locations along the same path. The results complement the qualitative examples at selected locations in the main paper. They also suggest the potential of audio-only panoramic depth estimation to provide useful spatial information across a traversed environment.

The remainder of this technical appendix provides additional evaluation and reproducibility details. Sec. B extends the analyses of observation selection and generalization to unseen receiver headings, while Sec. C presents additional qualitative results. Sec. D describes the implementation, training, and evaluation protocols. Sec. E details the dataset-construction procedure and simulation settings, and Sec. F discusses the limitations, licensing, and release plans.

C Additional Quantitative Results

We extend two analyses from the main paper that examine how *HEAR 360°* uses the geometry associated with multiple acoustic observations: observation selection (Tab. 4 in the main paper) and generalization to unseen receiver headings (Tab. 5 in the main paper). Unless otherwise stated, all experiments follow the same training and evaluation protocols as their corresponding main-paper experiments.

Generalization to unseen receiver headings. We extend the unseen-heading analysis in Tab. 5 of the main paper to a broader range of receiver configurations (Tab. S1). All models are trained using the canonical headings $\{0°, 90°, 180°, 270°\}$. All other configurations are evaluated zero-shot without retraining and include jointly rotated, unevenly spaced, and independently perturbed receiver headings. Unlike BatVision and EchoDiffusion, *HEAR 360°* receives the actual heading associated with each acoustic observation at test time. Across the unseen configurations, *HEAR 360°* consistently outperforms BatVision and EchoDiffusion. The advantage remains substantial

Headings (°)	Model	RMSE↓	AbsRel↓	$\log_{10}$ ↓
$\{45, 135, 225, 315\}$	BatVision	0.711	0.255	0.093
	EchoDiffusion	0.709	0.223	0.092
	Ours	**0.616**	**0.163**	**0.076**
$\{0, 120, 150, 315\}$	BatVision	0.733	0.278	0.099
	EchoDiffusion	0.657	0.223	0.091
	Ours	**0.570**	**0.161**	**0.073**
$\{0, 45, 210, 240\}$	BatVision	0.688	0.293	0.099
	EchoDiffusion	0.637	0.212	0.087
	Ours	**0.573**	**0.161**	**0.073**
$\{0, 135, 120, 300\}$	BatVision	0.802	0.314	0.111
	EchoDiffusion	0.705	0.250	0.100
	Ours	**0.567**	**0.162**	**0.073**
$\{0, 60, 225, 330\}$	BatVision	0.716	0.278	0.097
	EchoDiffusion	0.632	0.200	0.084
	Ours	**0.563**	**0.155**	**0.070**
$\{0, 150, 135, 210\}$	BatVision	0.763	0.309	0.106
	EchoDiffusion	0.654	0.229	0.092
	Ours	**0.579**	**0.159**	**0.073**
$\{0, 30, 240, 225\}$	BatVision	0.841	0.384	0.125
	EchoDiffusion	0.842	0.315	0.121
	Ours	**0.578**	**0.162**	**0.073**

Headings (°)	Model	RMSE↓	AbsRel↓	$\log_{10}$ ↓
$\{-30, 120, 150, 315\}$	BatVision	0.752	0.306	0.108
	EchoDiffusion	0.707	0.251	0.101
	Ours	**0.648**	**0.187**	**0.086**
$\{45, 45, 210, 240\}$	BatVision	0.824	0.389	0.122
	EchoDiffusion	0.726	0.272	0.106
	Ours	**0.669**	**0.194**	**0.090**
$\{-45, 135, 120, 300\}$	BatVision	0.920	0.410	0.138
	EchoDiffusion	0.871	0.338	0.129
	Ours	**0.685**	**0.199**	**0.093**
$\{60, 60, 225, 330\}$	BatVision	0.955	0.448	0.140
	EchoDiffusion	0.844	0.297	0.117
	Ours	**0.620**	**0.174**	**0.080**
$\{-60, 150, 135, 210\}$	BatVision	0.922	0.415	0.137
	EchoDiffusion	0.885	0.345	0.131
	Ours	**0.600**	**0.169**	**0.078**
$\{30, 30, 240, 225\}$	BatVision	0.955	0.472	0.146
	EchoDiffusion	0.952	0.384	0.141
	Ours	**0.661**	**0.192**	**0.088**

Table S1: **Generalization to unseen receiver headings.** Zero-shot evaluation of 8-observation models on diverse receiver configurations on Replica.

Setup	0° L	0° R	90° L	90° R	180° L	180° R	270° L	270° R	EchoDiffusion (Abs Rel↓)	*HEAR 360°* (Abs Rel↓)
(a)	✓	✓			✓	✓			0.399	**0.389**
(b)			✓	✓			✓	✓	0.402	**0.385**
(c)	✓	✓	✓	✓					0.433	**0.400**
(d)	✓	✓					✓	✓	0.439	**0.395**
(e)	✓	✓		✓			✓		0.453	**0.378**
(f)		✓		✓		✓		✓	0.435	**0.416**

Table S2: **Effect of observation selection on Replica.** Four of the eight available ear-specific acoustic observations are retained in each setup. Both methods use the same training and evaluation protocol for every subset. The better result in each row is shown in bold.

even under highly irregular layouts, including repeated or closely spaced receiver headings. These results further support the finding in the main paper that pose-aware association is not tied to the fixed receiver orientations used during training and can generalize to previously unseen sensing configurations.

Additional observation selections. The main paper compares three representative four-observation subsets with different binaural pairings and directional coverage. Tab. S2 extends this analysis to six subsets on Replica, including binaural pairs, mixed-ear observations spanning multiple headings, and observations distributed across the four available receiver directions. *HEAR 360°* outperforms EchoDiffusion for every evaluated subset. EchoDiffusion performs best for the two configurations containing diametrically opposed binaural pairs, *i.e.,* setups (a) and (b), and degrades for less symmetric or unpaired selections. In contrast, *HEAR 360°* achieves its best result with setup (e), which spans three receiver headings without requiring a complete binaural pair at each heading. This extends the observation-selection result in Tab. 4 of the main paper and shows that pose-aware aggregation can exploit directional coverage beyond fixed binaural pairings.

D Additional Qualitative Results

We complement the qualitative comparison in Fig. 4 of the main paper with additional Matterport3D examples in Fig. S2 and S3. For each listener position, we show predictions obtained using 2, 4, and 8 acoustic observations, allowing the effect of additional directional evidence to be compared at the same spatial location. RGB panoramas are included only for visual reference and are not used as input by any method. Across the additional scenes, *HEAR 360°* more consistently recovers the large-scale room structure while preserving local depth transitions. As more posed observations become available, the predicted panorama depth becomes more spatially coherent in regions that are

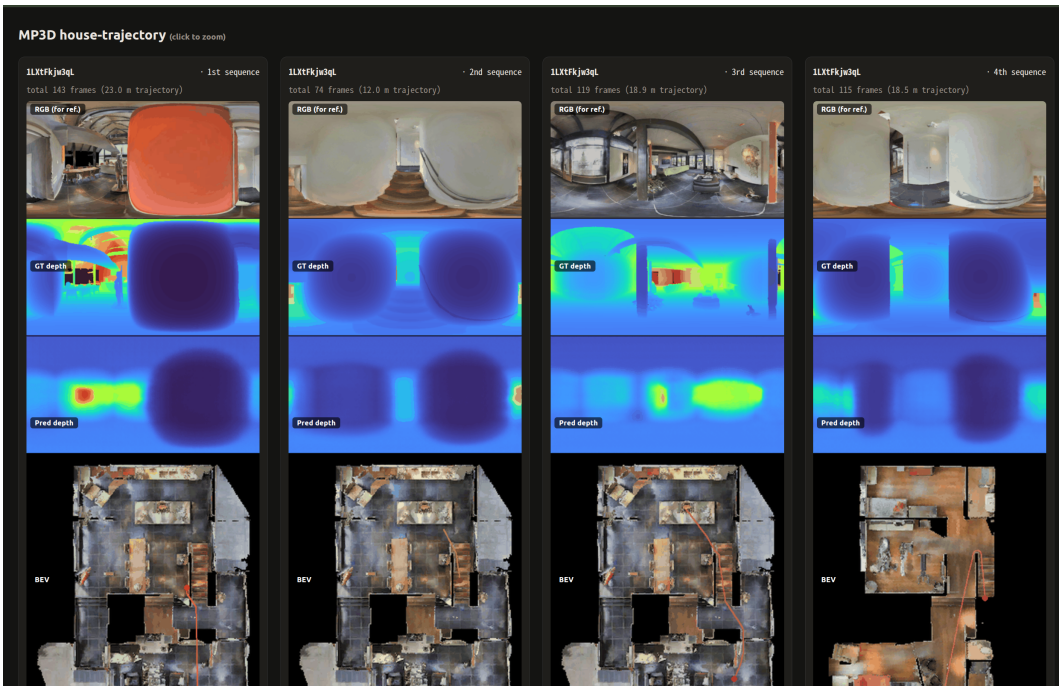


Figure S1: **Snapshot of the interactive multimedia appendix.** The HTML viewer displays the RGB panorama, ground-truth panoramic depth, predicted depth, and bird's-eye-view receiver trajectory for a sampled location along a Matterport3D path. The complete interactive viewer is included in the accompanying multimedia appendix.

ambiguous from a single binaural observation. These examples are consistent with the qualitative trends in the main paper.

Trajectory-level qualitative results. As shown in Fig. S1, the multimedia appendix (`.html`) visualizes panoramic depth predictions along continuous trajectories. Each sequence presents the rendered RGB panorama, ground-truth panoramic depth, the corresponding depth prediction, and a bird's-eye-view visualization of the receiver trajectory. Together, these views make it possible to inspect how the estimated surrounding geometry changes as the listener moves through the scene. Although each listener position is processed independently, the sequences provide a qualitative indication that echo-based panoramic depth estimation may support continuous spatial perception across changing listener positions. We encourage reviewers to view these sequences together with the static qualitative results in the main paper.

E Implementation Details

E.1 Input and Target Processing

Input processing. Each ear-specific acoustic observation is represented by the magnitude spectrogram of its raw impulse response. As described in Sec. 4.1 of the main paper, we crop each waveform to the round-trip propagation window for the maximum sensing range $d_{\max} = 10\,\mathrm{m}$. We use a Hann window and centered STFT frames, nearest-neighbor resizing is used instead of bilinear interpolation to avoid interpolation across the time–frequency bins containing echo structure. EchoScan (Yeon et al., 2024) and the waveform branch of EchoDiffusion (Zhang et al., 2025) use the same cropped waveforms without additional preprocessing.

Observation setup. At each listener position, we capture binaural observations at four headings separated by 90°. For each sample, one heading is defined as the reference direction, 0°, and the remaining headings are expressed relative to it. The panoramic depth map is aligned to the same

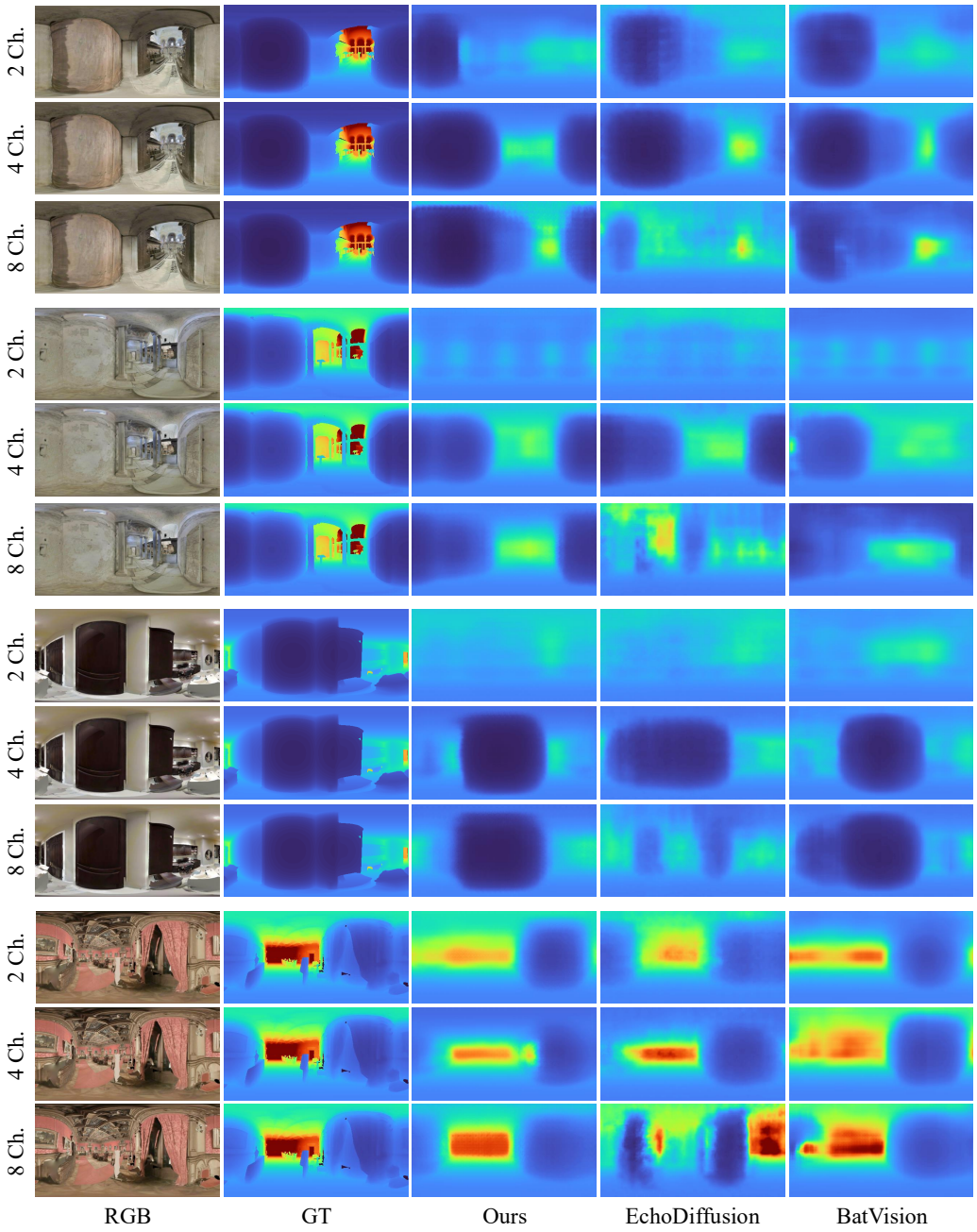


Figure S2: **Qualitative comparison of panoramic depth estimation on Matterport3D.**

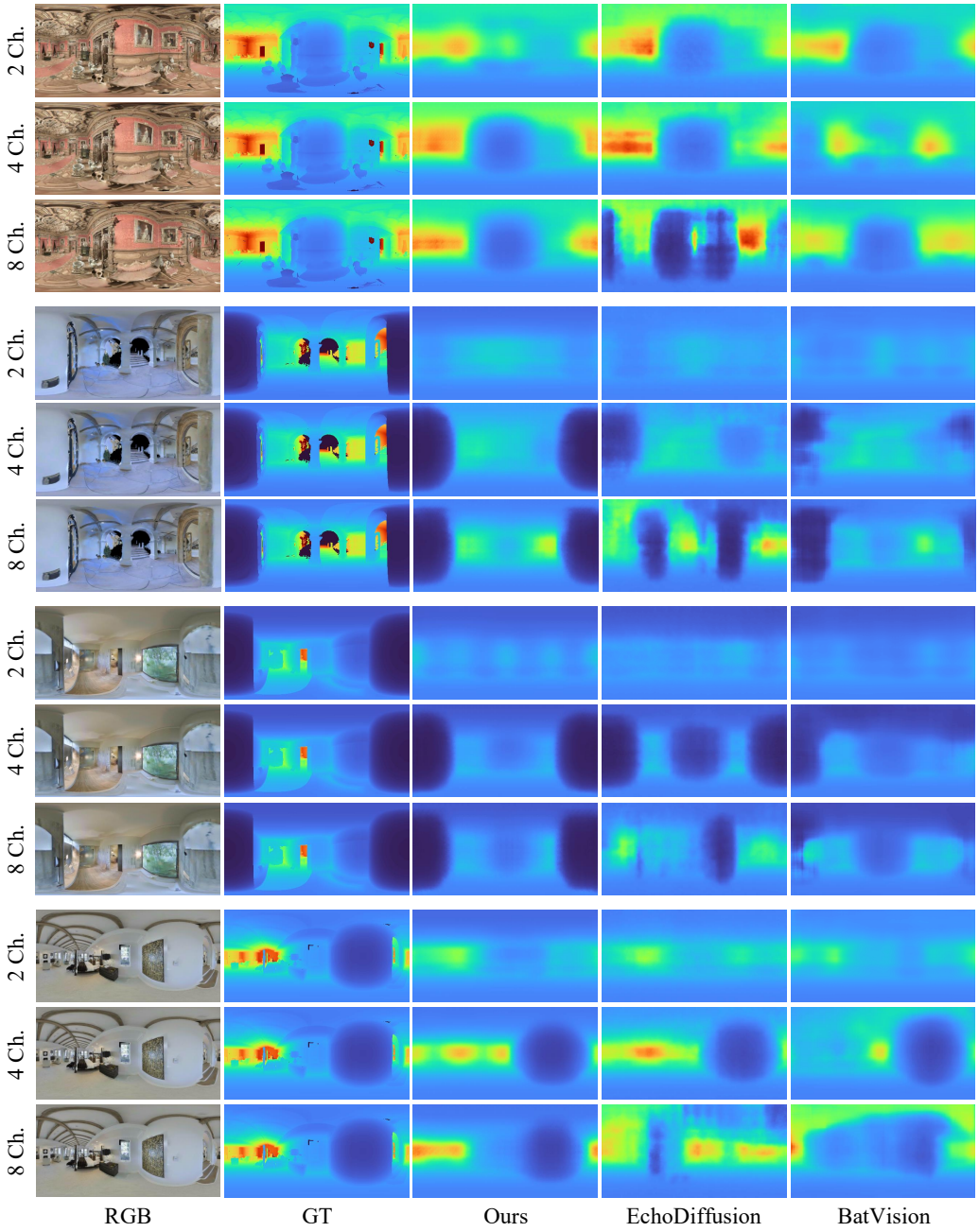


Figure S3: **Qualitative comparison of panoramic depth estimation on Matterport3D.**

listener-centered coordinate frame. The eight ear-specific observations are ordered as $[0°L,\ 0°R,\ 90°L,\ 90°R,\ 180°L,\ 180°R,\ 270°L,\ 270°R]$. The 2-, 4-, and 8-observation settings use $[0°L,\ 0°R]$, $[0°L,\ 0°R,\ 90°R,\ 270°L]$, and the full set, respectively. Each captured heading serves once as the reference direction, yielding four aligned samples per listener position.

E.2 Training Protocol

Depth supervision. We use listener-centered radial ERP depth, defined as the distance from the listener center to the first surface intersection along each panoramic ray. The native 512×1024 depth maps are downsampled to 256×512 using validity-masked average pooling over non-overlapping 2×2 regions. Each pooled value is computed using only finite, positive depth values; a pooled pixel is marked invalid when its corresponding region contains no valid values. The resulting depth is clipped at $d_{\max} = 10\,\text{m}$ and normalized as $\min(d, d_{\max})/d_{\max}$. All methods use the same depth targets and validity masks.

Training setup. All methods are optimized using AdamW with weight decay 10^{-4}, gradient-norm clipping at 1.0, linear warmup followed by cosine learning-rate decay, and a masked L_1 depth loss over valid pixels. Unless otherwise specified, models are trained for 40 epochs, and checkpoints are selected based on validation mean absolute error (MAE).

Per-method configurations. Tab. S3 shows architecture-specific learning rates, batch sizes, and warmup schedules. The unified setup are used on Replica and Matterport3D.

Method	lr	Batch (2/4/8 obs.)	Warmup
ResNet-50*	10^{-3}	64/64/48	2 ep
ViT-B/16*	10^{-3}	48/48/32	2 ep
EchoScan‡	10^{-3}	64/64/48	2 ep
BatVision	2×10^{-3}	64/64/48	1 ep
EchoDiffusion	10^{-4}	16/16/12/12	2 ep
HEAR 360°	5×10^{-4}	36/36/24/24	4 ep§

Table S3: **Per-method optimization settings.**

E.3 Evaluation Protocol

Scene-level splits. Receiver positions are sampled continuously within each scene. A position-level split could therefore place training and test receivers arbitrarily close to one another; we observed nearest-neighbor distances as small as 0.07 m. We instead use scene-disjoint splits to prevent geometrically and acoustically similar locations from appearing across training and evaluation sets.

Replica. We use 12 scenes for training, three for validation, and three for testing:

- **val (3):** `apartment_1`, `frl_apartment_4`, and `office_3`.
- **test (3):** `apartment_2`, `frl_apartment_5`, and `office_4`.

Validation and test contain scenes from the same scene families without sharing individual scenes.

Matterport3D. We use the following scene-disjoint validation and test sets:

- **val (9):** `1LXtFkjw3qL`, `1pXnuDYAj8r`, `EU6Fwq7SyZv`, `PuKPg4mmafe`, `Uxmj2M2itWa`, `ZMojNkEp431`, `fzynW3qQPVF`, `gxdoqLR6rwA`, and `r47D5H71a5s`.
- **test (9):** `8WUmhLawc2A`, `EDJbREhghzL`, `HxpKQynjfin`, `Z6MFQCViBuw`, `gTV8FGcVJC9`, `pLe4wQe7qrG`, `q9vSo1VnCiC`, `sT4fr6TAbpF`, and `uNb9QFRL6hY`.

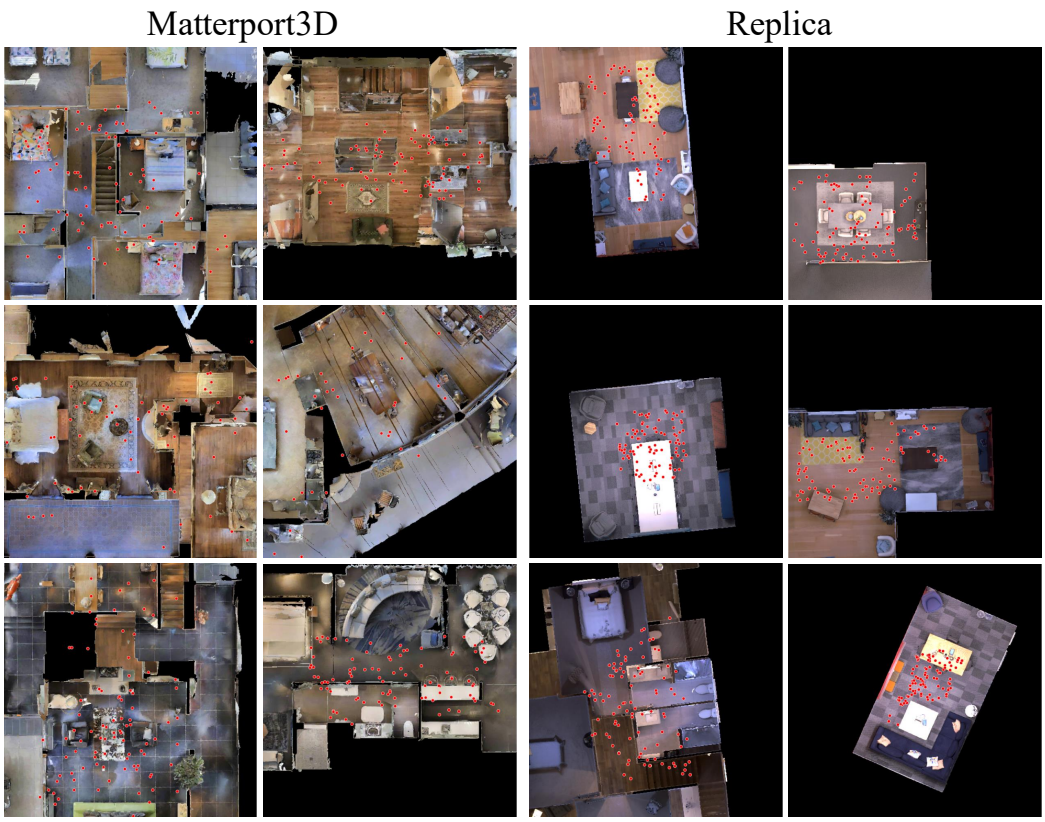


Figure S4: **Receiver placement in bird's-eye view.** Red dots indicate the receiver locations sampled in the simulator, showing that they are broadly distributed across the navigable area of each scene.

The remaining 72 scenes are used for training.

E.4 Details of Comparison Methods

Pretrained ResNet-50/ViT-B/16. For an acoustic input with C_{in} channels, where $C_{\text{in}} \in \{2, 4, 8\}$, a learnable 1×1 convolution projects the input into three pseudo-RGB channels. The projected input is processed by an ImageNet-pretrained backbone and a lightweight up-convolutional ERP decoder to produce a 256×512 panoramic depth map.

BatVision (Brunetto et al., 2023). The channel-stacked U-Net is kept unchanged; only the first convolution accepts C_{in} channels and the decoder output resolution is set to 256×512. We use the original depth branch (direct L_1 regression), the original higher CNN learning rate, and no architectural additions.

EchoScan (Yeon et al., 2024). The published 1D-convolutional waveform encoder is preserved and applied to the windowed waveforms. Its floor-plan and room-height heads are replaced by a small transposed-convolution decoder that maps the 4×8 encoder feature map to a 256×512 ERP depth map. The channel count of the first 1D convolution follows C_{in}.

EchoDiffusion (Zhang et al., 2025). The released implementation provides a deterministic single-pass network, while the iterative diffusion procedure described in the paper is not included in the public code. We therefore train the available deterministic path, which consists of an ASPP-ASFF spectrogram stem, an SD-shaped U-Net, and a CIDE branch. In the CIDE branch, a frozen wav2vec (Baevski et al., 2020) encoder processes the waveforms. The spectrogram stem accepts C_{in} input channels, and the output head predicts a 256×512 depth map through a sigmoid.

F Dataset Construction

Overview. Following prior work that constructs paired acoustic and geometric observations from scanned indoor environments (Gao et al., 2020; Yun et al., 2023), we generate our benchmarks using Habitat-Sim (Savva et al., 2019) and SoundSpaces 2.0 (Chen et al., 2022). This setup builds on the visual–acoustic simulation framework introduced by SoundSpaces (Chen et al., 2020), in which acoustic responses and visual observations are generated from the same underlying 3D scene. SoundSpaces 2.0 performs on-the-fly geometry-based acoustic rendering from a scene mesh and supports configurable source and receiver locations, acoustic materials, and microphone properties (Chen et al., 2022). Integrated with SoundSpaces 2.0, Habitat-Sim provides the corresponding geometric rendering and navigability information from the same scene representation and coordinate system (Savva et al., 2019; Chen et al., 2022).

For each sampled listener location, we render binaural room impulse responses at multiple known receiver headings together with the corresponding listener-centered panoramic depth. The acoustic observations, geometric supervision, and receiver-pose metadata are therefore registered within a shared scene coordinate frame. We sample listener locations over the navigable surface to cover diverse positions within each environment and use scene-disjoint training, validation, and test splits to evaluate generalization to unseen scenes. The following sections describe the receiver sampling, acoustic capture, depth-label construction, and dataset-specific acoustic configurations.

F.1 Simulation Protocol

Receiver placement. For each scene, we sample 100 listener positions uniformly over the navigable surface using Habitat-Sim's `get_random_navigable_point`. The sampled positions are continuous coordinates rather than points on a predefined grid. Figure S4 visualizes an example placement, showing the irregular spatial distribution of receiver positions over the navigable region.

Acoustic capture. Following the active echolocation setup of VisualEchoes (Gao et al., 2020), we co-locate the sound source and binaural receiver at each sampled listener position. As the emitted signal, we use the publicly released `3ms_sweep.wav` from the official VisualEchoes repository, which contains a 3-ms sweep. We render the corresponding binaural room impulse response at each receiver heading and convolve it with the sweep signal to obtain the echo observation.

Capture procedure. The released generation script consumes these pose files verbatim. Algorithm 1 summarizes one scene's capture loop.

Algorithm 1 Per-scene benchmark generation

```
Sample 100 navigable listener positions P
for p ∈ P, ψ ∈ Ψ do
    Place the listener and source at (p, ψ)
    Render binaural responses and panoramic data
    Convert D_pl to D_rad
    Save outputs and pose (p, ψ)
```

Radial depth conversion. Habitat renders panoramic depth from six posed pinhole views. Let $\mathbf{D}_{\mathrm{pl}}(u, v)$ denote the plane depth at pixel (u, v), measured along the optical axis of the corresponding pinhole camera. We convert it to listener-centered radial depth as

$$\mathbf{D}_{\mathrm{rad}}(u, v) = \frac{\mathbf{D}_{\mathrm{pl}}(u, v)}{\max\left(|r_x|, |r_y|, |r_z|\right)}, \tag{9}$$

where $\mathbf{r} = (r_x, r_y, r_z)$ is the unit viewing ray associated with pixel (u, v). Radial depth is better aligned with panoramic geometry because each panoramic pixel represents a viewing direction from the listener center, and its depth should measure the distance along that ray to the first surface. Invalid depth values remain 0, and we use $\mathbf{D}_{\mathrm{rad}}$ as the ground-truth depth for training.

Compute. Rendering used one workstation (32-core CPU, one RTX A6000, 4 parallel scene-workers). Average $\sim$60 s each ($\approx$120 h total for the base sets); the lean rotation-dense captures (binaural + depth) average 8.5–12.8 s, or $\approx$5.4 h wall-clock for all 4,800. The dominant cost is RLR ray tracing on CPU; GPU load is negligible.

Data organization and pose metadata. Each scene directory is organized into modality-specific subdirectories, including `audio_wav/`, `audio_npy/`, `erp_rgb/`, `erp_depth/`, and `bev/`, together with a `locations.json` file. Each entry in `locations.json` stores the listener position and receiver heading as follows:

```
{
"position":  [x, y, z],   # listener position
"rotation":  [w, x, y, z], # WXYZ quaternion
"pos_idx":  i,        # position index
"yaw_deg":  psi      # receiver heading
}
```

Quaternions follow WXYZ ordering, and $\psi = 0$ corresponds to Habitat's default heading ($-Z$). The stored `position` specifies the agent position, while the sensor and source are placed at a fixed height of 1.5 m during rendering. Depth maps are stored as float32 arrays in meters, with 0 indicating invalid pixels. Audio waveforms are stored as float32 PCM at 48 kHz, and the original unnormalized RIR is additionally stored as `*_original.npy`. File indices correspond one-to-one with entries in `locations.json`. Capture failures are recorded in `failures.json`, and each generation run stores its configuration in `summary.json`.

Matterport3D. Matterport3D contains 90 scenes, yielding 36,000 captures from 100 listener positions and four receiver headings per scene. We use the provided acoustic material configuration (`mp3d_material_config.json`) and load the corresponding semantic mesh for each scene.

Replica. Replica contains 18 scenes, yielding 7,200 captures under the same sampling procedure. We load the corresponding semantic mesh for each scene. Since Replica does not provide acoustic material annotations, all surfaces are assigned the simulator's default material. We use 12 scenes for training, three for validation, and three for testing.

Acoustic material configuration. The two datasets differ in their acoustic material modeling. Matterport3D uses scene-specific material annotations, whereas Replica uses the same default material for all surfaces. Thus, echoes in Replica remain informative about scene geometry but do not reflect surface-specific acoustic properties. Absolute performance values across the two datasets should therefore not be directly compared.

G Limitations and Licensing

Simulation-to-real gap. Our benchmarks are constructed in simulation using SoundSpaces 2.0 (Chen et al., 2020; 2022), on Replica (Straub et al., 2019) and Matterport3D (Chang et al., 2017). Thus, our evaluation does not include real-world acoustic recordings, and performance may be affected by differences between simulated and real acoustic responses.

Acoustic noise. Our experiments use simulated acoustic responses and do not explicitly model environmental or sensor noise. Such noise may obscure weak reflections and degrade depth estimation in real-world settings. Evaluating robustness under different noise conditions and real recordings remains an important direction for future work.

Licensing. Matterport3D meshes are used under the Matterport3D license and are not redistributed with our data. Instead, we release the generation pipeline, scene splits, and receiver poses needed to reproduce the benchmark from the original data. Replica, SoundSpaces 2.0, and Habitat-Sim are used under their respective licenses.

Privacy and ethics. Our benchmark is generated from existing indoor-scene datasets and contains no recorded speech or other audio captured from real environments. The acoustic observations used in our experiments are fully simulated.

Release. Upon publication, we will release the data-generation pipeline, benchmark metadata, and training and inference code required to reproduce our benchmarks and experimental results. For both Replica and Matterport3D, the benchmarks can be regenerated from the original scene assets using the released pipeline, subject to their respective licenses.