Acoustic Inverse Rendering for Room Reconstruction

Yiduo Hao[1†] **Yiwei Tang**[1†] **Jiayang Li**[1†] **Zitong Lan**[1†]
Haowen Lai[1] **Romit Roy Choudhury**[2] **Mingmin Zhao**[1]
[1]University of Pennsylvania
[2]University of Illinois Urbana-Champaign

Abstract

A room impulse response (RIR) encodes how sound reflects, scatters, and decays inside a room, and therefore carries information about what the room looks like and what it is made of. We formalize this observation as *Acoustic Inverse Rendering (AIR)*: recovering dense panoramic geometry, the depth image, and acoustic reflectance of a scene directly from its RIR. The task is challenging as the RIR renders all spatial propagation paths as a single temporal signal, and the resulting inverse mapping is severely ambiguous. AIR closes this gap by projecting the RIR onto the equirectangular grid and conditioning a diffusion model on the resulting directional features, yielding a distribution over plausible scenes rather than a single point estimate. The same spherical beamforming applies to first-order Ambisonics (FOA), binaural, and mono microphone setups. AIR produces coherent panoramic depth images from acoustic input alone, outperforming prior baselines on this inverse problem. Additional material are available at project website.

1 Introduction

Many animals perceive space through sound. Bats navigate dark caves by interpreting echoes of their ultrasonic calls [31], and dolphins localize underwater objects with acoustic clicks [5]. These biological systems reveal an intriguing fact: sound encodes not only acoustic events, but also the geometry of the scene itself. The same physics applies to everyday spaces we inhabit. Specifically, every room imprints a signature on the sound that propagates through it. This acoustic fingerprint is captured by the room impulse response (RIR) [34], which describes the acoustic transfer from a source to a receiver as sound reflects, scatters, and reverberates throughout the room [43, 41, 16].

This raises a natural question: *how much can we recover about a room from its acoustic fingerprint?* Beyond its intellectual appeal, the question also has broad practical value. In spatial audio for AR/VR and telepresence, knowing the listener's room lets synthesized sound match the surrounding space [54, 57, 65]. Audio processing tasks such as dereverberation, source localization, and speech enhancement also improve once the surrounding room is characterized [23, 45, 44]. In forensic and intelligence settings such as hostage recovery and human trafficking interdiction, inferring room geometry from audio recordings like voice notes or intercepted calls can facilitate their geolocation [3, 22, 49, 70].

A single RIR mixes acoustic arrivals from many surfaces and directions into one temporal signal, and distinct room layouts can produce similar reverberation patterns. Whether dense geometry can be recovered from such measurements is the central question of this paper.

Prior work on acoustics and scene geometry can be broadly divided into forward and inverse problems of sound propagation. In the forward setting, the goal is to predict the sound field given room geometry and material properties. This mapping can be accurately simulated using physical models such as ray tracing and the image-source method [33, 1, 8, 34], as well as learning-based approximations [51, 39, 47]. The inverse problem that infers scene structure from acoustic observations is

[†]core contributors

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

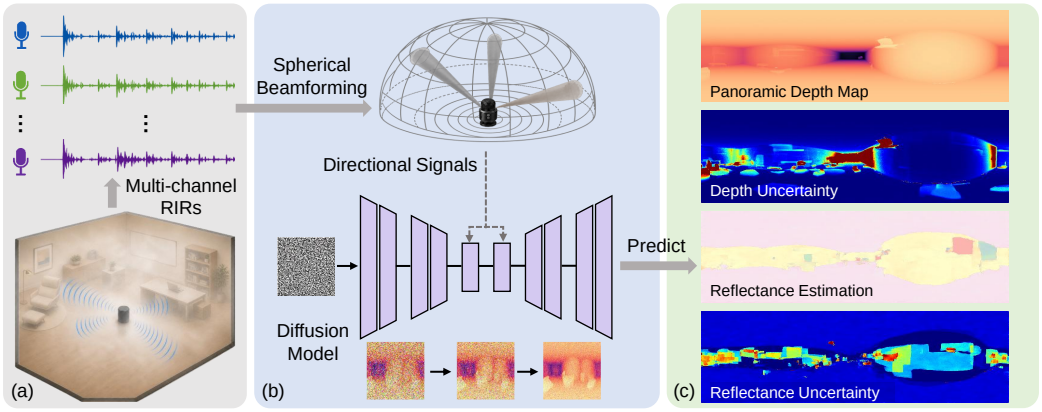

Figure 1: **Overview of Acoustic Inverse Rendering (AIR).** (a) A speaker emits a known probe signal, and a microphone array records room impulse responses (RIRs). (b) Spherical beamforming projects the RIRs into directional signals on an equirectangular grid, which condition the diffusion model throughout denoising. (c) From a single acoustic measurement, AIR reconstructs the panoramic scene geometry and acoustic reflectance of the environment, together with uncertainty estimates for both.

substantially more challenging, as it is ill-posed and admits multiple plausible solutions. In other sensing modalities, visual observations [63] and radio-frequency (RF) measurements [2, 50, 25] enable geometry reconstruction through frameworks inspired by neural radiance fields (NeRFs) [52]. In acoustics, classical inverse methods recover room geometry and acoustic boundary properties from sound measurements [19, 60, 18]. Some methods infer geometry by learning joint representations from paired audio and visual data [17, 24, 9], while echo-based methods estimate depth from binaural signals or raw waveforms [24, 56, 27, 9, 73]. EchoDiffusion [73] uses a diffusion model to estimate perspective depth and is closest to our setting. Building on these advances, we study joint panoramic depth and acoustic reflectance recovery, with the acoustic condition spatially aligned to the reconstruction.

In this paper, we introduce Acoustic Inverse Rendering (AIR), the acoustic counterpart of visual inverse rendering, and ask: *what can a single RIR tell us about the geometry and surface reflectance of the room?* As shown in Fig. 1(a), both targets are dictated by the physics of the forward process: an RIR is shaped by reflections off the full surrounding geometry and by the per-band reflection coefficients of every reflecting surface, motivating their joint recovery. Our method combines physics-aware acoustic signal spatialization with a learned scene prior. As shown in Fig. 1(b), for acoustic signal spatialization, we use spherical beamforming to project the multi-channel RIR into direction-dependent acoustic signals on the same equirectangular grid as the depth map, directly aligning the acoustic condition with the prediction target in a shared spatial domain. For the scene prior, we condition a diffusion model on these directional features, injecting acoustic condition into the UNet bottleneck layer so that the predicted scene remains anchored in the RIR measurement while the learned diffusion prior regularizes the solution space to geometrically coherent outputs. As shown in Fig. 1(c), the diffusion model also returns a distribution over plausible scenes rather than a single estimate, providing per-pixel uncertainty for estimation maps. The same spatialization applies uniformly to first-order Ambisonics (FOA), binaural, and mono RIRs, allowing the same model to operate across very different acoustic measurement setups.

We evaluate AIR using datasets containing 11600 simulated data pairs across 116 rooms. The full evaluation set contains simulated 240 samples, with 12 samples from each of 20 held-out rooms. In summary, our findings are as follows:

- Spherical beamforming is a strong performance lever: aligning the RIR to the panoramic ray grid before conditioning reduces root mean squared depth error (RMSE) from 0.607 to 0.502 meters and lifts threshold accuracy (Delta1) from 0.741 to 0.822 over a cross-attention conditioning variant.
- Predictive uncertainty from the diffusion model correlates with depth error (Pearson 0.580, Spearman 0.593 for FOA input), providing an empirical reliability signal for reconstructed depth.
- Different microphone arrays expose different amounts of the room information encoded in the RIR: on the same model, FOA achieves RMSE 0.502 meters, binaural degrades to 1.033 meters, and mono recordings (omnidirectional or cardioid) recover even less.

- Multi-bounce reflections help with recovery: including reflections beyond the first bounce improves depth prediction, indicating that these echoes carry useful information about room geometry.

2 Related Work

Signal-Based Room Reconstruction from Acoustics. Classical methods infer room geometry from echo arrival times and directions [22, 3, 49]. Dokmanić et al. [19] recover convex polyhedral rooms by associating multichannel echoes with walls, while Sprunck et al. [60] jointly estimate room parameters and boundary absorption under a shoebox image-source model. With known or approximate geometry, acoustic inverse methods estimate surface impedance [53, 4, 55] and frequency-dependent absorption [18], including formulations that jointly recover source location and boundary impedance [6]. Acoustic rendering methods also estimate surface reflection properties using supplied geometry or multi-view visual priors [62, 30]. AIR represents the reconstruction as a dense depth panorama centered at the receiver, rather than a simplified room model. From a single multichannel RIR, it estimates depth and reflectance without supplied room geometry or visual input.

Learning Geometry and Depth Estimation from Sound. Learning-based methods use scene priors to infer room geometry and dense depth from acoustic observations. RGI-Net [70] estimates room geometry from RIRs with hidden first-order reflections. Structure from Silence [17] learns geometric cues from ambient audio, while BatVision [24, 9] explore depth prediction from active binaural echoes with paired visual supervision. Other approaches combine echoes with images [56] or relate spectrogram and angular-spectrum features through co-attention [27]. EchoDiffusion [73] combines learned spectrogram features and waveform conditioning with diffusion for perspective depth estimation. AIR uses spherical beamforming to construct acoustic features indexed by the angular directions of the panorama. This gives the diffusion model an explicitly aligned spatial condition, rather than leaving audio-to-pixel correspondence to be learned entirely from data.

Audio-Visual and Acoustic Scene Understanding A large body of work leverages the coupling between sound and visual structure for spatial scene understanding. Platforms such as SoundSpaces [11, 12] provide simulation infrastructure and benchmarks for audio-visual navigation, demonstrating that acoustic signals carry rich spatial information about the environment. Neural acoustic field methods [51, 61, 46, 26] model continuous acoustic fields as implicit functions of position, enabling RIR synthesis at novel locations. Audio-visual neural fields [46, 7, 14] jointly represent visual and acoustic scene properties, while some work [58, 30] explicitly use multi-view RGB images to estimate RIRs. Broader work on acoustic scene reconstruction [15, 13, 62, 48] further confirms that acoustic signals encode fine-grained scene geometry. All of these approaches either require visual observations as input or focus on synthesizing audio from known geometry; we instead invert the relationship, recovering dense panoramic depth and reflectance directly from a measured impulse response without any visual input.

Monocular and Panoramic Depth Estimation Monocular depth estimation was established as a deep learning task [21, 35], and subsequent work has scaled this to large affine-invariant foundation models such as Depth Anything [68, 71] that generalize across diverse environments. The recent shift to diffusion-based depth, exemplified by Marigold [32], shows that repurposing image-domain generative priors for depth as a conditional generation problem yields strong zero-shot accuracy. Panoramic depth estimation [36, 37] extends this problem to equirectangular images, where the spherical projection introduces additional challenges related to spatial distortion and global layout consistency; dedicated architectures such as EGFormer [72] address these by incorporating equirectangular geometry biases. All such methods take visual images as input and are designed for camera-equipped systems. Our work establishes a fundamentally different input modality: we generate equirectangular depth maps conditioned solely on RIR signals, using a beamformed acoustic representation that is geometrically aligned to the panoramic output domain.

3 Method

We first define our task of room estimation from RIRs (§3.1) and discuss the challenges of solving it. We then present AIR, which consists of two key components: a spherical beamforming mechanism (§3.2) that transforms RIRs into spatially aligned acoustic representations, and a conditional diffusion model (§3.3) that leverages these representations to generate depth and per-band reflectance maps. An overview of the framework is shown in Fig. 2.

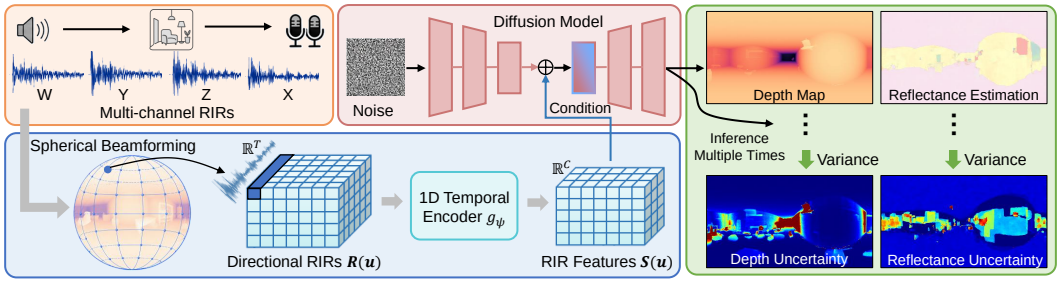

Figure 2: Overview of AIR. The model spatializes a multi-channel RIR with spherical beamforming, encodes each directional RIR with a shared temporal encoder, and conditions a diffusion model to jointly predict panoramic depth, per-band reflectance, and their uncertainties.

3.1 Task Definition and Challenges

We aim to recover the surrounding scene from sound alone. Given an RIR $\boldsymbol{u} \in \mathbb{R}^{N\times T}$, where N is the format-dependent signal dimension and T is the RIR length, recorded at a co-located source-receiver position, we predict equirectangular panoramas centered on the receiver: a depth map $\boldsymbol{D} \in \mathbb{R}^{H\times W}$ (H and W are height and width) describing the surrounding geometry, and a frequency-band reflectance map $\boldsymbol{P} \in \mathbb{R}^{H\times W\times K}$. For each panoramic ray, $\boldsymbol{D}$ stores the distance in meters from the receiver to the first intersected surface, and $\boldsymbol{P}$ stores that surface's frequency-dependent acoustic reflection coefficients. The equirectangular layout provides a common angular coordinate system for the target maps and acoustic conditioning. We predict $K = 5$ continuous coefficients at octave bands centered on 250, 500, 1000, 2000, and 4000 Hz. Our primary input is an FOA RIR; we additionally evaluate binaural and mono variants to study how the spatial richness of the input limits what can be recovered.

This inverse problem is hard for two reasons. First, the input and output live in different domains: an RIR is a temporal waveform that superposes reflections arriving from many directions and delays, whereas the targets are dense spatial fields indexed by viewing direction. A learned encoder must therefore re-discover *where* each acoustic cue belongs on the sphere before it can predict pixels. Second, the mapping is fundamentally ambiguous: rooms with different layouts but matched effective absorption can produce nearly identical reverberation, so a single deterministic prediction cannot faithfully represent what a measurement allows us to know.

3.2 Spherical Beamforming for Spatially Aligned Conditioning

A naive recipe to address this task is to feed the entire RIR to a global encoder and let the network learn which acoustic cue corresponds to which direction on the panorama purely from data. We argue this wastes a strong physical prior: the directional sensitivity of the acoustic measurement already records *where* each reflection came from, and we can exploit this structure analytically before any learning. Concretely, we algorithmically steer the same recording toward each panoramic ray and assemble a direction-indexed acoustic feature map aligned with the output panorama.

From RIRs to a spatialized RIR field. We discretize the sphere with the same equirectangular grid as the depth map, but at a coarser resolution $h \times w$ to keep the conditioning computationally tractable. For each pixel (i, j) on this grid, the corresponding viewing direction has azimuth $\theta_j = \pi\big(1 - \frac{2j+1}{w}\big)$ and elevation $\phi_i = \frac{\pi}{2}\big(1 - \frac{2i+1}{h}\big)$. For an FOA RIR $\boldsymbol{u}(t) = \big[\, u_W(t),\, u_Y(t),\, u_Z(t),\, u_X(t) \,\big]^\top$, the first-order spherical-harmonic steering vector for direction (θ, ϕ) is:

$$\boldsymbol{a}(\theta, \phi) = \big[\, 1,\ \sin\theta\cos\phi,\ \sin\phi,\ \cos\theta\cos\phi \,\big]^\top. \tag{1}$$

First-order Ambisonics (FOA) comprises one omnidirectional component and three directional components along the Cartesian axes. We apply fixed first-order spherical-harmonic beamforming to the FOA RIR, forming each directional signal through a direction-dependent weighted combination of its components. Evaluating this projection across the panoramic grid provides spatially aligned acoustic conditioning. The resulting directional RIR is:

$$r_{ij}(t) = \boldsymbol{a}(\theta_j, \phi_i)^\top \,\mathrm{diag}(1, \sqrt{3}, \sqrt{3}, \sqrt{3})\, \boldsymbol{u}(t). \tag{2}$$

Stacking $r_{ij}(t)$ over the grid yields a spatialized RIR field $\boldsymbol{R}(\boldsymbol{u}) \in \mathbb{R}^{h\times w\times T}$, in which every spatial location holds a one-dimensional time series that emphasizes acoustic reflections arriving from that

ray direction. Crucially, this projection is parameter-free and physically interpretable: the spatial index of each feature is tied to a known direction rather than learned from data. The same construction is microphone-agnostic, extending to binaural and mono inputs by substituting the FOA basis with the appropriate spatial decoder.

Shared per-ray temporal encoder. A 1-D encoder g_ψ, shared across all rays, maps each 1-D directional RIR waveform to a fixed C-dimensional feature vector $\boldsymbol{s}_{ij} = g_\psi(r_{ij})$. Sharing g_ψ across the grid forces it to learn direction-agnostic temporal primitives (arrival time, reflection density, decay rate), while the fixed beamforming grid carries direction. Stacking the encoded features gives the spatial acoustic condition $\boldsymbol{S}(\boldsymbol{u}) \in \mathbb{R}^{h\times w\times C}$ that is used by the diffusion model. We construct the acoustic grid at a reduced resolution relative to the output panorama to keep computation light. In our implementation, the output panorama is $H\times W = 256\times 512$ and the acoustic grid is $h \times w = 32\times 64$, an $8\times$ downsampling that matches the UNet's mid-resolution feature map.

3.3 Diffusion Prior and Uncertainty

The acoustic condition $\boldsymbol{S}(\boldsymbol{u})$ tells the model where reflections came from, but it does not pin down a unique scene: rooms with different geometries and matched effective absorption can produce nearly identical reverberation, and a deterministic regressor would be forced to commit to one of these solutions. We instead model the conditional distribution $p_\Theta(\boldsymbol{x}_0 \mid \boldsymbol{u})$ with a diffusion model parameterized by Θ, where $\boldsymbol{x}_0$ is the joint depth-and-reflectance panorama defined below. This serves two purposes simultaneously: it injects a learned prior over realistic indoor layouts and surface reflectance to regularize the under-determined inverse problem. Moreover, the distributional formulation naturally supports per-pixel uncertainty quantification, with predictive variance exposing how strongly the RIR constrains different spatial regions of the scene.

Spatial acoustic injection. Because $\boldsymbol{S}(\boldsymbol{u}) \in \mathbb{R}^{h\times w\times C}$ is already aligned with the UNet's mid-resolution stage, we inject it as an additional residual on the bottleneck spatial latent feature map rather than as a sequence of audio tokens. Let $\boldsymbol{F}_{\text{mid}}$ denote the UNet bottleneck feature; we directly add $\boldsymbol{S}(\boldsymbol{u})$ residually to this spatial feature map $\widetilde{\boldsymbol{F}}_{\text{mid}} = \boldsymbol{F}_{\text{mid}} + \boldsymbol{S}(\boldsymbol{u})$. This contrasts with the common alternative of compressing the RIR into a flat token sequence and consuming it via cross-attention, which leaves the audio-to-pixel correspondence for the network to discover implicitly. Our spatial injection enforces this correspondence by construction, pairing every acoustic feature with the panoramic ray it physically belongs to. We view this as a stronger inductive bias for geometry and reflectance recovery and quantify its benefit in our ablations.

Depth and reflectance target. We can either predict the panorama depth directly or stack depth and the K-band reflectance maps (K=5 in our setting) into a single panorama

$$\boldsymbol{x}_0 \;=\; \text{concat}\big(\boldsymbol{D},\, \boldsymbol{P}^{(1)},\, \ldots,\, \boldsymbol{P}^{(K)}\big), \tag{3}$$

normalized to $[-1, 1]$ before training (details in Appendix D). Joint denoising lets the network exploit the physical coupling between geometry and reflectance: surface positions determine reflection paths, while per-band reflection coefficients shape the frequency-dependent decay of those reflections.



(a) Simulated Rooms (b) Real-World Rooms

Figure 3: Examples from the simulated and real-world datasets. (a) Rendered simulated rooms from RIR-Room. (b) Photographs of indoor rooms used for real-world collection.

Training. We adopt a standard v-prediction objective [59]. At each training step we sample a timestep t, use cumulative noise schedule value $\bar{\alpha}_t$, and draw Gaussian noise $\boldsymbol{\epsilon}$. We then form the noisy panorama

$$\boldsymbol{x}_t \;=\; \sqrt{\bar{\alpha}_t}\,\boldsymbol{x}_0 \;+\; \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0},\, \boldsymbol{I}), \tag{4}$$

and train a UNet $\boldsymbol{v}_\Theta$ to predict the velocity target $\boldsymbol{v}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{x}_0$ by minimizing

$$\mathcal{L}_{\text{diff}} \;=\; \mathbb{E}_{(\boldsymbol{x}_0,\,\boldsymbol{u}),\,t,\,\boldsymbol{\epsilon}}\Big[\,\big\|\boldsymbol{v}_t - \boldsymbol{v}_\Theta\big(\boldsymbol{x}_t,\,t,\,\boldsymbol{S}(\boldsymbol{u})\big)\big\|_2^2\,\Big]. \tag{5}$$

Uncertainty. Since the model represents a distribution rather than a single estimate, we can quantify the residual ambiguity that the RIR alone cannot resolve. We define the per-pixel prediction uncertainty $\boldsymbol{U}$ as the posterior variance under the conditional model [37]:

$$\boldsymbol{U} \;=\; \mathrm{Var}_{\hat{\boldsymbol{x}}_0 \sim p_\Theta(\boldsymbol{x}_0 \mid \boldsymbol{u})}\big[\hat{\boldsymbol{x}}_0\big]. \tag{6}$$

In practice, we estimate this variance empirically from 8 independent diffusion samples per input RIR. The depth component of $\boldsymbol{U}$ gives a depth uncertainty map and the remaining K components give per-band reflectance uncertainty. Intuitively, regions weakly constrained by the RIR (acoustically shadowed surfaces, or geometries with ambiguous reflection paths) admit more diverse reconstructions and therefore higher sampling variance.

4 Experiment

4.1 Experimental Setup

Simulated Dataset. We evaluate on an RIR-Room dataset that we simulated with the AcoustiX [43] acoustic simulator. The dataset consists of simulated FOA RIRs with receiver-centered equirectangular panoramic depth maps at 256×512 resolution, as well as per-band reflectance maps encoding five frequency-band reflection coefficients for every pixel. The dataset is built from 116 indoor scenes spanning diverse room layouts, including compact rooms, hallways, and open living spaces, generated using a procedure similar to [74]. For each scene, we sample 100 speaker-microphone pairs. We use 96 of the scenes for training and hold out the remaining 20 rooms for evaluation. The full simulated evaluation set contains 240 samples across 20 held-out rooms, with 12 randomly selected locations per room. All RIR measurements are synthesized as FOA signals in WYZX convention. We also generate a binaural variant by linear combination of W and Y channels in the FOA format, as well as omnidirectional and cardioid mono variants from the same FOA signal, enabling direct comparison among different input formats. We visualize some of the generated scenes in Fig. 3(a). Details about the dataset generation process are provided in Appendix B.

Real-world Dataset. We also collect a small set of real-world RIR-depth pairs using a lightweight setup, where a smartphone records binaural acoustic responses and a loudspeaker plays the chirp signal. At each measurement location, the loudspeaker emits a one-second 50 Hz–20 kHz chirp, and the smartphone saves the received response as a binaural recording. The corresponding ground-truth panoramic depth map is captured using an Ouster OS0-64 LiDAR at the same location. The setup is illustrated in Fig. 4. We collect measurements in 5 indoor rooms with 180 total positions. We train on 3 rooms (120 recordings) and test on the remaining 2 unseen rooms (60 recordings). Example real-world scenes are shown in Fig. 3(b).

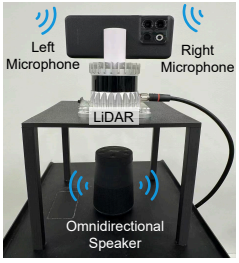

Figure 4: Real-world collection setup.

Baselines. We compare against EchoDiffusion [73], a waveform-conditioned diffusion model for echo-based depth estimation. EchoDiffusion conditions on a global embedding of the input waveform via cross-attention, without an explicit directional alignment between the audio and the output spatial field. We adapt it to our FOA setting by encoding the FOA RIR with its original waveform encoder and injecting the resulting tokens through cross-attention into our shared diffusion backbone. We also evaluate two no-audio baselines built from training depth panoramas. The *Mean Depth* baseline predicts the same mean panorama for every test input, while the *Random Depth* baseline samples a training panorama with a random yaw rotation.

AIR Implementation Details. Our model is trained with a null-condition probability $p = 0.1$. We evaluate the 50k-step checkpoints for CrossAttn and SphericalBF with a classifier-free guidance (CFG) scale of 1.5 unless otherwise noted. Simulated depth accuracy is evaluated using one diffusion prediction per input RIR. For real-world depth evaluation, we average eight independent diffusion

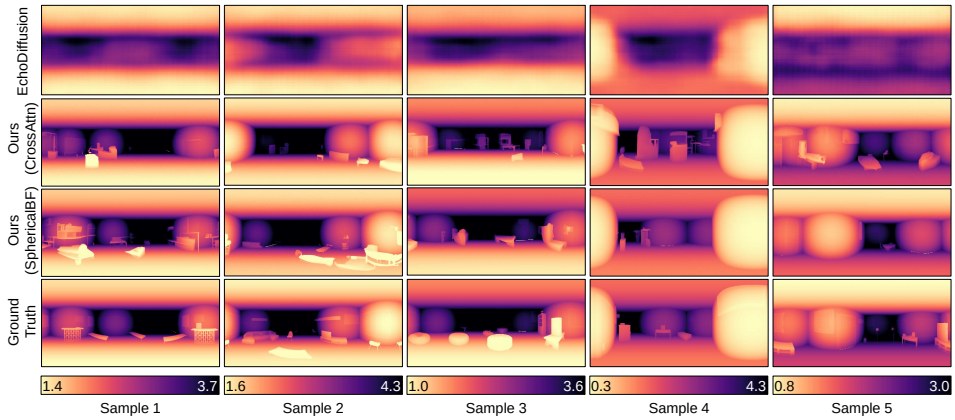

Figure 5: Qualitative panoramic depth predictions on the simulated test set.

predictions per input RIR. Uncertainty is estimated from eight independent diffusion samples per input RIR. We compare two conditioning strategies: *CrossAttn* is a cross-attention conditioning baseline that encodes RIR into multiple tokens and incorporates it into the diffusion UNet via cross-attention; *SphericalBF* is our proposed spatially aligned conditioning.

Evaluation Metrics. Let d, $\hat{d}$ denote the ground-truth and predicted depth. We evaluate panoramic depth quality using standard depth estimation metrics computed on valid (non-zero) depth pixels: root mean squared error (RMSE) in meters, log-scale RMSE (logRMSE), scale-invariant log error (SILog), and threshold accuracies Delta1, Delta2, and Delta3, defined as the fraction of valid pixels satisfying $\max(\hat{d}/d,\, d/\hat{d}) < 1.25^k$ for $k = 1, 2, 3$, respectively.

4.2 Results

Main Results. Tab. 1 compares our proposed spatially aligned conditioning against EchoDiffusion and the cross-attention conditioning on the simulated RIR-Room test set. All acoustic models use FOA input. SphericalBF consistently outperforms both EchoDiffusion and the CrossAttn variant across all metrics. The gap between CrossAttn and SphericalBF highlights the benefit of aligning the acoustic representation with the panoramic ray grid. Rather than leaving the model to discover spatial correspondence through cross-attention, the directional decomposition front-end imposes this alignment explicitly and reduces the geometric inference burden on the UNet. Fig. 5 shows representative qualitative examples, where SphericalBF better preserves the dominant panoramic room layout compared with the baselines.

Method	RMSE ↓	logRMSE ↓	SILog ↓	Delta1 ↑	Delta2 ↑	Delta3 ↑
Mean Depth	1.095	0.713	57.77	0.393	0.642	0.768
Random Depth	1.551	0.989	83.05	0.259	0.473	0.626
EchoDiffusion [73]	0.811	0.463	44.69	0.539	0.772	0.876
CrossAttn (ours)	0.607	0.310	29.31	0.741	0.887	0.943
SphericalBF (ours)	**0.502**	**0.282**	**26.34**	**0.822**	**0.911**	**0.951**

Table 1: Panoramic depth estimation on the simulated unseen-room test split. Acoustic methods use FOA input.

Input Format Comparison. Tab. 2 compares the influence of different input formats (FOA, binaural, and mono) for both conditioning strategies. The simulated binaural input is formed from sums and differences of the W and Y channels. Mono inputs are synthesized either with an omnidirectional response or with a fixed cardioid directivity pattern. FOA consistently outperforms binaural and mono input under both conditioning strategies. All mono variants outperform Random Depth but remain less accurate than Mean Depth (Tab. 1). With SphericalBF conditioning, both binaural and FOA inputs outperform these no-audio baselines, highlighting the value of directional acoustic information for panoramic reconstruction.

Uncertainty Analysis. Fig. 6 visualizes the predicted depth, estimation error, and model uncertainty for both binaural and FOA inputs. FOA produces more accurate and spatially consistent depth predictions compared to binaural input, particularly around challenging regions with large

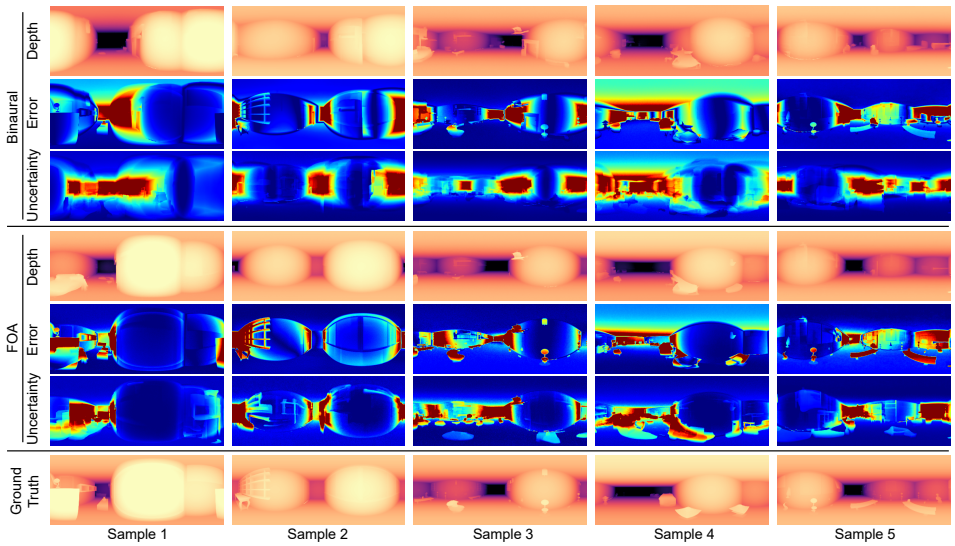

Figure 6: Depth prediction and uncertainty maps for binaural and FOA inputs. For each input format, rows show predicted depth, absolute error, and predictive uncertainty; FOA yields lower errors than binaural, especially near 0° and 180° azimuth.

Method	Input	RMSE ↓	logRMSE ↓	SILog ↓	Delta1 ↑	Delta2 ↑	Delta3 ↑
Random Depth	—	1.551	0.989	83.05	0.259	0.473	0.626
CrossAttn	Omnidirectional	1.264	0.806	78.11	0.357	0.600	0.742
	Cardioid	1.247	0.781	75.16	0.341	0.592	0.742
	Binaural	1.260	0.733	71.30	0.407	0.624	0.751
	FOA	0.607	0.310	29.31	0.741	0.887	0.943
SphericalBF	Omnidirectional	1.282	0.751	72.49	0.349	0.584	0.734
	Cardioid	1.320	0.756	71.24	0.332	0.573	0.723
	Binaural	1.033	0.579	55.61	0.467	0.729	0.833
	FOA	**0.502**	**0.282**	**26.34**	**0.822**	**0.911**	**0.951**

Table 2: Effect of acoustic input format on panoramic depth estimation for CrossAttn and SphericalBF on the simulated RIR-Room test split.

Method	Input	Correlation↑	Spearman↑	Kendall Tau↑	NLL_Laplace↓
CrossAttn	Binaural	0.546	0.418	0.298	2.311
	FOA	0.575	0.606	0.428	0.297
SphericalBF	Binaural	0.541	0.430	0.312	1.385
	FOA	0.580	0.593	0.424	-0.029

Table 3: Uncertainty quantification on the simulated test split.

geometric variation. In contrast, binaural predictions exhibit larger errors, especially around the 0° and 180° azimuth directions, where the binaural format provides weaker spatial cues. For uncertainty quantification, we report Pearson correlation, Spearman correlation, Kendall Tau, and Laplace negative log-likelihood (NLL_Laplace) between predicted uncertainty and depth error. The predicted uncertainty maps strongly correlate with the actual estimation errors. Regions with high reconstruction error consistently correspond to areas with higher uncertainty, showing that the model is able to capture its own confidence under ambiguous acoustic conditions. Compared to binaural input, FOA has generally lower uncertainty, further confirming the richer directional information provided by FOA. We provide uncertainty metrics comparison in Tab. 3 to quantify this. Example depth predictions for calculating the uncertainty are shown in Fig. 10 in Appendix A.

Reflectance Estimation. Fig. 7 shows representative qualitative results for per-band reflectance estimation across different frequency bands. The model successfully recovers dominant reflectance regions and large-scale room structures, with predictions that are spatially consistent with the ground truth. Large contiguous surfaces are generally estimated more accurately, likely because they contribute stronger and more stable acoustic reflections in the observed RIRs. Errors mainly

Method	CFG	RMSE↓	logRMSE↓	SILog↓	Delta1↑	Delta2↑	Delta3↑
CrossAttn	1.0	0.677	0.354	32.80	0.677	0.857	0.925
	1.5	0.607	**0.310**	**29.31**	0.741	0.887	0.943
	2.0	**0.584**	0.341	31.91	**0.755**	**0.897**	**0.946**
	3.0	0.604	0.379	34.79	0.748	0.892	0.940
	5.0	0.655	0.400	36.78	0.730	0.879	0.933
SphericalBF	1.0	0.521	0.286	**25.83**	0.793	0.901	0.949
	1.5	**0.502**	**0.282**	26.34	0.822	**0.911**	**0.951**
	2.0	0.507	0.314	29.04	**0.826**	0.910	0.950
	3.0	0.533	0.373	33.88	0.809	0.900	0.941
	5.0	0.579	0.435	39.25	0.792	0.893	0.934

Table 4: Effect of CFG scale on panoramic depth estimation.

occur around small isolated regions, and reflectance boundaries, where acoustic cues are weaker or more ambiguous. We also observe occasional over-smoothing between neighboring surfaces with similar reflection coefficients. For example, while the major room layout and wall regions are well reconstructed in both samples, some fine-grained local regions remain difficult to distinguish accurately. Additional qualitative examples and quantitative results are provided in Appendix A.

Noise Robustness. After noise-augmented fine-tuning, AIR maintains depth RMSE below 0.53 m at 20 and 30 dB SNR. RMSE increases to 0.678 m at 10 dB. We introduce environmental noise into simulated chirp recordings before RIR recovery; Appendix A.1 provides the full results and evaluation protocol.

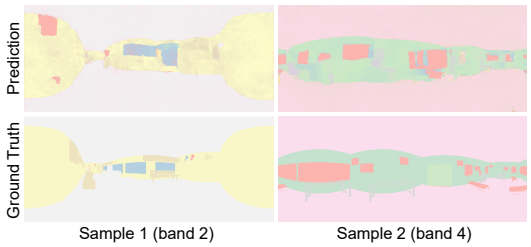

Figure 7: Reflectance predictions: 500 Hz (Sample 1) and 2000 Hz (Sample 2).

4.3 Ablation Study

We ablate two factors that directly affect acoustic inverse rendering: the strength of classifier-free guidance (CFG) during diffusion sampling and the maximum reflection order used to synthesize the RIR. The first ablation studies how strongly the sampler should follow the acoustic condition, while the second examines the contribution of multi-bounce reflections to depth estimation.

CFG guidance scale. Tab. 4 shows the effect of CFG scale on depth estimation. We observe that moderate CFG scales achieve the best performance, while no guidance (CFG $= 1.0$) or large guidance values degrade prediction quality. For both CrossAttn and SphericalBF conditioning, performance peaks around CFG values of 1.5–2.0. Excessively high CFG scales lead to increased artifacts and higher estimation error. Across all guidance scales, SphericalBF consistently outperforms CrossAttn.

Reflection Order	RMSE ↓	logRMSE ↓	SILog ↓	Delta1 ↑	Delta2 ↑	Delta3 ↑
1	0.557	0.311	28.89	0.801	0.900	0.951
2	0.483	0.304	28.33	0.840	0.928	0.955
4	0.488	0.259	24.72	0.837	0.924	0.958
8	0.502	0.282	26.34	0.822	0.911	0.951

Table 5: Effect of maximum simulated reflection order on panoramic depth estimation with FOA input and SphericalBF conditioning.

Reflection Order. We study how the maximum order of simulated reflections affects depth estimation accuracy in Tab. 5. We train and evaluate model instances on dataset variants with reflection orders of 1, 2, 4, and 8. Including multi-bounce reflections improves depth estimation compared with using only first-order reflections. The best maximum reflection order depends on the metric.

4.4 Real-World Experiments

Fig. 8 shows qualitative panoramic depth prediction results on unseen real-world environments after fine-tuning on the 120 recordings. Similar to the simulated dataset, the model recovers most of the large-scale room geometries and dominant structural layouts with spatial consistency.

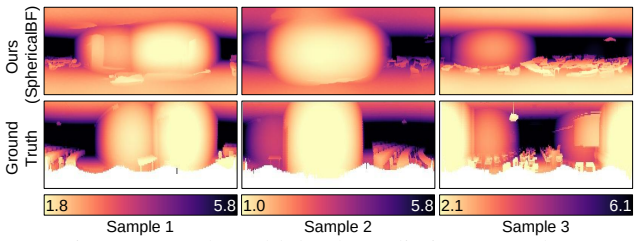

Figure 8: Real-world depth prediction examples.

Tab. 6 compares simulated and real-world evaluation under the binaural input setting. For each real-world input RIR, AIR averages eight independent depth predictions for stability. We do find that the performance degrades on real recordings due to the sim-to-real gap. Compared to simulation results, the real-world predictions exhibit noisier boundaries and local artifacts, particularly around fine-grained structures and distant regions. White regions indicate no LiDAR returns.

Setting	RMSE ↓	logRMSE ↓	SILog ↓	Delta1 ↑	Delta2 ↑	Delta3 ↑
Simulated	1.033	0.579	55.61	0.467	0.729	0.833
Real-world	1.826	0.584	52.8	0.353	0.596	0.759

Table 6: Simulated and real-world binaural evaluation with SphericalBF conditioning.

5 Discussion

Limitations and Future Work. Several limitations remain. First, real-world performance still lags behind simulation because commodity binaural recordings introduce hardware response, noise, and calibration differences that are not fully captured by the simulator. Second, the current experiments focus on indoor scenes with a co-located source-receiver setup. Acquisition assumes an approximately static scene during the one-second probe because motion changes the acoustic transfer function. Third, AIR recovers room-scale layout more reliably than small objects and fine boundaries. These local details remain weakly constrained by the available directional measurements and can vary across generated samples. Future work could reduce the sim-to-real gap through improved acoustic simulation and larger-scale real-world datasets, evaluate richer spatial measurements and outdoor or semi-enclosed scenes, and integrate acoustic inverse rendering with visual observations for richer multimodal scene understanding.

Conclusion. We presented AIR, a framework for recovering panoramic scene geometry and per-band acoustic reflectance directly from RIRs. By spatially aligning directional acoustic cues with the panoramic prediction domain through spherical beamforming, AIR enables diffusion models to infer plausible room structure from sound alone. Our experiments demonstrate the value of directional acoustic information and multi-bounce reflections for scene recovery, while uncertainty estimation further captures the ambiguity inherent in the inverse acoustic problem.

Acknowledgments and Disclosure of Funding

We thank the members of the WAVES Lab at the University of Pennsylvania for their valuable feedback. We are grateful to the anonymous reviewers for their insightful comments and suggestions.

References

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

[2] Chaitanya Amballa, Sattwik Basu, Yu-Lin Wei, Zhijian Yang, Mehmet Ergezer, and Romit Roy Choudhury. Can NeRFs "see" without cameras? In *Advances in Neural Information Processing Systems*, volume 38, pages 112023–112050. Curran Associates, Inc., 2025.

[3] Fabio Antonacci, Jason Filos, Mark Thomas, Emanuel AP Habets, Augusto Sarti, Patrick A Naylor, and Stefano Tubaro. Inference of room geometry from acoustic impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2683–2695, 2012.

[4] Niccolò Antonello, Toon van Waterschoot, Marc Moonen, and Patrick A. Naylor. Evaluation of a numerical method for identifying surface acoustic impedances in a reverberant room. In *Proceedings of the 10th European Congress and Exposition on Noise Control Engineering (EURONOISE)*, pages 185–190, 2015.

[5] Whitlow W. L. Au. *The Sonar of Dolphins*. Springer, 1993.

[6] Nancy Bertin, Srdan Kitić, and Rémi Gribonval. Joint estimation of sound source location and boundary impedance with physics-driven cosparse regularization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6340–6344. IEEE, 2016.

[7] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. *Advances in Neural Information Processing Systems*, 37:28920–28937, 2024.

[8] Jeffrey Borish. Extension of the image model to arbitrary polyhedra. *The Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[9] Amandine Brunetto, Sascha Hornauer, Stella X. Yu, and Fabien Moutarde. The audio-visual batvision dataset for research on sight and sound. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2023.

[10] Deli Cai, Shuyang Xu, Zhiyang Dou, Zitong Lan, Mingmin Zhao, Changxing Ding, and Taku Komura. MOMO: Human motion from moving sounds. arXiv preprint arXiv:2610.24522, 2026.

[11] 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*, pages 17–36. Springer, 2020.

[12] 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.

[13] Mingfei Chen, Israel D Gebru, Ishwarya Ananthabhotla, Christian Richardt, Dejan Markovic, Jake Sandakly, Steven Krenn, Todd Keebler, Eli Shlizerman, and Alexander Richard. Soundvista: Novel-view ambient sound synthesis via visual-acoustic binding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8331–8341, 2025.

[14] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. *Advances in Neural Information Processing Systems*, 37:141021–141044, 2024.

[15] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7853–7862, 2023.

[16] Ziyang Chen, Israel D. Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21886–21896, 2024.

[17] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 760–772, 2022.

[18] Stéphane Dilungana, Antoine Deleforge, Cédric Foy, and Sylvain Faisan. Geometry-informed estimation of surface absorption profiles from room impulse responses. In *Proceedings of the 30th European Signal Processing Conference (EUSIPCO)*, pages 867–871, 2022.

[19] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

[20] Zhenbang Du, Jun Luo, Zhiwei Zheng, Xiangchi Yuan, Kejing Xia, Dachuan Shi, Qirui Jin, Qijia He, Shaofeng Zou, Yingbin Liang, and Wenke Lee. PACT: Privileged trace co-training for multi-turn tool-use agents. *arXiv preprint arXiv:2606.16215*, 2026.

[21] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[22] J Filos, A Canclini, MRP Thomas, F Antonacci, A Sarti, and PA Naylor. Robust inference of room geometry from acoustic measurements using the Hough transform. In *Eur. Sig. Proc. Conf.(EUSIPCO), Barcelona, Spain*, 2011.

[23] Karim Guirguis, Christoph Schorn, Andre Guntoro, Sherif Abdulatif, and Bin Yang. Seld-tcn: Sound event localization & detection via temporal convolutional networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 16–20. IEEE, 2021.

[24] Jesper Haahr Christensen, Sascha Hornauer, and Stella Yu. Batvision with gcc-phat features for better sound to vision predictions. *arXiv preprint arXiv:2006.07995*, 2020.

[25] Yiduo Hao, Sohrab Madani, Junfeng Guan, Mohammed Alloulah, Saurabh Gupta, and Haitham Hassanieh. Bootstrapping autonomous driving radars with self-supervised learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15012–15023, 2024.

[26] Christopher Ick, Gordon Wichern, Yoshiki Masuyama, François Germain, and Jonathan Le Roux. Direction-aware neural acoustic fields for few-shot interpolation of ambisonic impulse responses. *CoRR*, abs/2505.13617, 2025.

[27] Go Irie, Takashi Shibata, and Akisato Kimura. Co-attention-guided bilinear model for echo-based depth estimation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4648–4652. IEEE, 2022.

[28] Derong Jin, Xiyi Chen, Ming C. Lin, and Ruohan Gao. SonoWorld: From one image to a 3D audio-visual scene. arXiv preprint arXiv:2603.28757, 2026.

[29] Derong Jin, Xiyi Chen, Ming C. Lin, and Ruohan Gao. SonoWorld4D: From one video to a 4D audio-visual scene via diffusion synchronization. arXiv preprint arXiv:2610.40640, 2026.

[30] Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 37–47, 2025.

[31] Gareth Jones and Marc Wilhelm Holderied. Bat echolocation calls: adaptation and convergent evolution. *Proceedings of the Royal Society B: Biological Sciences*, 274:905 – 912, 2007.

[32] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024.

[33] Asbjørn Krokstad, S. Strom, and Svein Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968.

[34] Heinrich Kuttruff. *Room acoustics*. Crc Press, 2016.

[35] Haowen Lai, Zitong Lan, and Mingmin Zhao. Non-line-of-sight 3D reconstruction with radar. *Advances in Neural Information Processing Systems*, 2025.

[36] Haowen Lai, Gaoxiang Luo, Yifei Liu, and Mingmin Zhao. Enabling visual recognition at radio frequency. *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, pages 388–403, 2024.

[37] Haowen Lai, Zhiwei Zheng, and Mingmin Zhao. Rf-based 3d slam rivaling vision approaches. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*, 2025.

[38] Zitong Lan, Yiduo Hao, and Mingmin Zhao. Guiding audio editing with audio language model. arXiv preprint arXiv:2509.21625, 2025.

[39] Zitong Lan, Yiduo Hao, and Mingmin Zhao. Resounding acoustic fields with reciprocity. *Advances in Neural Information Processing Systems*, 38:87218–87243, 2025.

[40] Zitong Lan, Yiduo Hao, and Mingmin Zhao. SmartDJ: Declarative audio editing with audio language model. In *The Fourteenth International Conference on Learning Representations*, 2026.

[41] Zitong Lan, Yiwei Tang, Yuhan Wang, Haowen Lai, Yiduo Hao, and Mingmin Zhao. Building audio-visual digital twins with smartphones. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services*, pages 825–838, 2026.

[42] Zitong Lan, Mutian Tong, Jiatao Gu, and Mingmin Zhao. Enabling immersive audio-visual experience from any video. arXiv preprint arXiv:2609.36295, 2026.

[43] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[44] Jean-Marie Lemercier, Eloi Moliner, Simon Welker, Vesa Välimäki, and Timo Gerkmann. Unsupervised blind joint dereverberation and room acoustics estimation with diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[45] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2724–2737, July 2023.

[46] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490, 2023.

[47] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023.

[48] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastià Vicenç Amengual Garí, Calvin Murdock, Ishwarya Ananthabhotla, Philip W. Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5732–5741, 2025.

[49] Michael Lovedee-Turner and Damian Murphy. Three-dimensional reflector localisation and room geometry estimation using a spherical microphone array. *The Journal of the Acoustical Society of America*, 146(5):3339–3352, 2019.

[50] Jiachen Lu, Hailan Shanbhag, and Haitham Al Hassanieh. GeRaF: Neural geometry reconstruction from radio frequency signals. In *Advances in Neural Information Processing Systems*, volume 38, pages 94200–94230. Curran Associates, Inc., 2025.

[51] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022.

[52] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pages 405–421. Springer, 2020.

[53] Gabriel Pablo Nava, Yosuke Yasuda, Yoichi Sato, and Shinichi Sakamoto. On the in situ estimation of surface acoustic impedance in interiors of arbitrary shape by acoustical inverse methods. *Acoustical Science and Technology*, 30(2):100–109, 2009.

[54] Annika Neidhardt, Christian Schneiderwind, and Florian Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms: Literature review and theoretical framework. *Trends in Hearing*, 26, 2022.

[55] Yukiko Okawa, Yasuaki Watanabe, Yusuke Ikeda, and Yasuhiro Oikawa. Estimation of acoustic impedances in a room using multiple sound intensities and FDTD method. In *Proceedings of the 27th International Congress on Sound and Vibration (ICSV27)*, 2021.

[56] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8268–8277, 2021.

[57] Boaz Rafaely, Vladimir Tourbabin, Emanuel Habets, Zamir Ben-Hur, Hyunkook Lee, Hannes Gamper, Lior Arbel, Lachlan Birnie, Thushara Abhayapala, and Prasanga Samarasinghe. Spatial audio signal processing for binaural reproduction of recorded acoustic scenes – review and challenges. *Acta Acustica*, 6:47, 2022.

[58] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27164–27175, 2024.

[59] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[60] Tom Sprunck, Antoine Deleforge, Yannick Privat, and Cédric Foy. Fully reversing the shoebox image source method: From impulse responses to room parameters. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1023–1033, 2025.

[61] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022.

[62] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11790–11799, 2024.

[63] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pages 27171–27183. Curran Associates, Inc., 2021.

[64] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending speech separation to noisy environments. In *Interspeech 2019*, pages 1368–1372, 2019.

[65] Justin Wilson, Nicholas Rewkowski, Ming C Lin, and Henry Fuchs. Echo-reconstruction: Audio-augmented 3d scene reconstruction. *arXiv preprint arXiv:2110.02405*, 2021.

[66] Shuyang Xu, Zhiyang Dou, Yiduo Hao, Zekun Li, Liang Pan, Jingbo Wang, Cheng Lin, Yuan Liu, Wenping Wang, Mingmin Zhao, and Taku Komura. MoSAT: Human motion generation from spatial audio and textual description. arXiv preprint arXiv:2609.23797, 2026.

[67] Shuyang Xu, Zhiyang Dou, Mingyi Shi, Liang Pan, Leo Ho, Jingbo Wang, Yuan Liu, Cheng Lin, Yuexin Ma, Wenping Wang, and Taku Komura. MOSPA: Human motion generation driven by spatial audio. In *Advances in Neural Information Processing Systems*, volume 38, pages 120436–120458, 2025.

[68] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *Advances in Neural Information Processing Systems*, 2024.

[69] Yue Yang, Fan-Yun Sun, Luca Weihs, Eli VanderBilt, Alvaro Herrasti, Winson Han, Jiajun Wu, Nick Haber, Ranjay Krishna, Lingjie Liu, Chris Callison-Burch, Mark Yatskar, Aniruddha Kembhavi, and Christopher Clark. Holodeck: Language guided generation of 3d embodied ai environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16227–16237, 2024.

[70] Inmo Yeon and Jung-Woo Choi. Rgi-net: 3d room geometry inference from room impulse responses with hidden first-order reflections. In *IWAENC*, 2024.

[71] Jinhao You, Shuo Lyu, Zhuohang Lyu, Tanxuan Li, Zibo Zhao, Jiaxiang Hu, Kai Tang, and Yichen Guo. Regimevggt: Layer-wise spatially preserving redundancy removal for visual geometry grounded transformer, 2026.

[72] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. EGformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6101–6112, 2023.

[73] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22578–22586, 2025.

[74] Zhiwei Zheng, Dongyin Hu, and Mingmin Zhao. Scalable rf simulation in generative 4d worlds. *arXiv preprint arXiv:2508.12176*, 2025.

[75] Zhiwei Zheng, Shibo Jin, Lingjie Liu, and Mingmin Zhao. Next-scale autoregressive models for text-to-motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16376–16386, June 2026.

A Additional Results

Reflectance Estimation. Fig. 9 provides additional qualitative examples of per-band acoustic reflectance estimation. Each example compares the predicted and ground-truth reflectance maps across the five frequency bands used in our dataset. Tab. 7 reports the corresponding quantitative reflectance metrics on the simulated test split. We use mean absolute error (MAE), RMSE, and threshold accuracies Acc@0.05 and Acc@0.10, where Acc@t is the fraction of valid pixels whose absolute reflectance error is below t. The model achieves an overall reflectance RMSE of 0.139, with 72.3% of valid pixel-band predictions within an absolute error of 0.10. Accuracy is highest at the lower-frequency bands and decreases for higher-frequency bands, consistent with the stronger ambiguity of frequency-dependent local acoustic reflectance. The predictions recover the dominant room-scale reflectance patterns and preserve the major frequency-dependent trends across bands. Large contiguous surfaces, such as walls and broad structural regions, are reconstructed more consistently than small isolated objects or thin boundary regions. The remaining errors are mostly over-smoothed local details and imprecise reflectance boundaries, which is expected because small surfaces contribute weaker and more ambiguous reflection cues to the observed RIR. These examples support the main-text observation that RIRs carry information about both geometry and surface reflectance, while fine-grained reflectance estimation remains harder than recovering large-scale layout.

Band	MAE ↓	RMSE ↓	Acc@0.05 ↑	Acc@0.10 ↑
Overall	0.093	0.139	0.506	0.723
250 Hz	**0.043**	**0.058**	**0.719**	**0.905**
500 Hz	0.060	0.081	0.503	0.820
1000 Hz	0.092	0.125	0.437	0.790
2000 Hz	0.125	0.173	0.435	0.580
4000 Hz	0.146	0.199	0.434	0.519

Table 7: Per-band reflectance estimation on the simulated test split with FOA input and SphericalBF conditioning. Overall metrics are averaged over all valid pixels and frequency bands.

Depth Prediction and Uncertainty. Fig. 10 visualizes stochastic depth predictions from the diffusion model, the corresponding ground-truth depth, the mean depth error, and the predicted uncertainty. Across samples from the same scene, the predicted depth maps preserve the dominant panoramic room structure while varying in locally ambiguous regions. This behavior is consistent with the distributional formulation of AIR, where multiple plausible reconstructions can be sampled from the same acoustic observation. The mean error maps show that mistakes concentrate around object boundaries, thin structures, cluttered regions, and far surfaces where acoustic evidence is weaker or harder to localize. The uncertainty maps assign higher uncertainty to many of these same regions, while large smooth walls and broad layout components are generally assigned lower uncertainty. This qualitative agreement between error and uncertainty complements the quantitative uncertainty analysis in Tab. 3, indicating that the model's predictive variance is a useful reliability signal rather than only a sampling artifact.

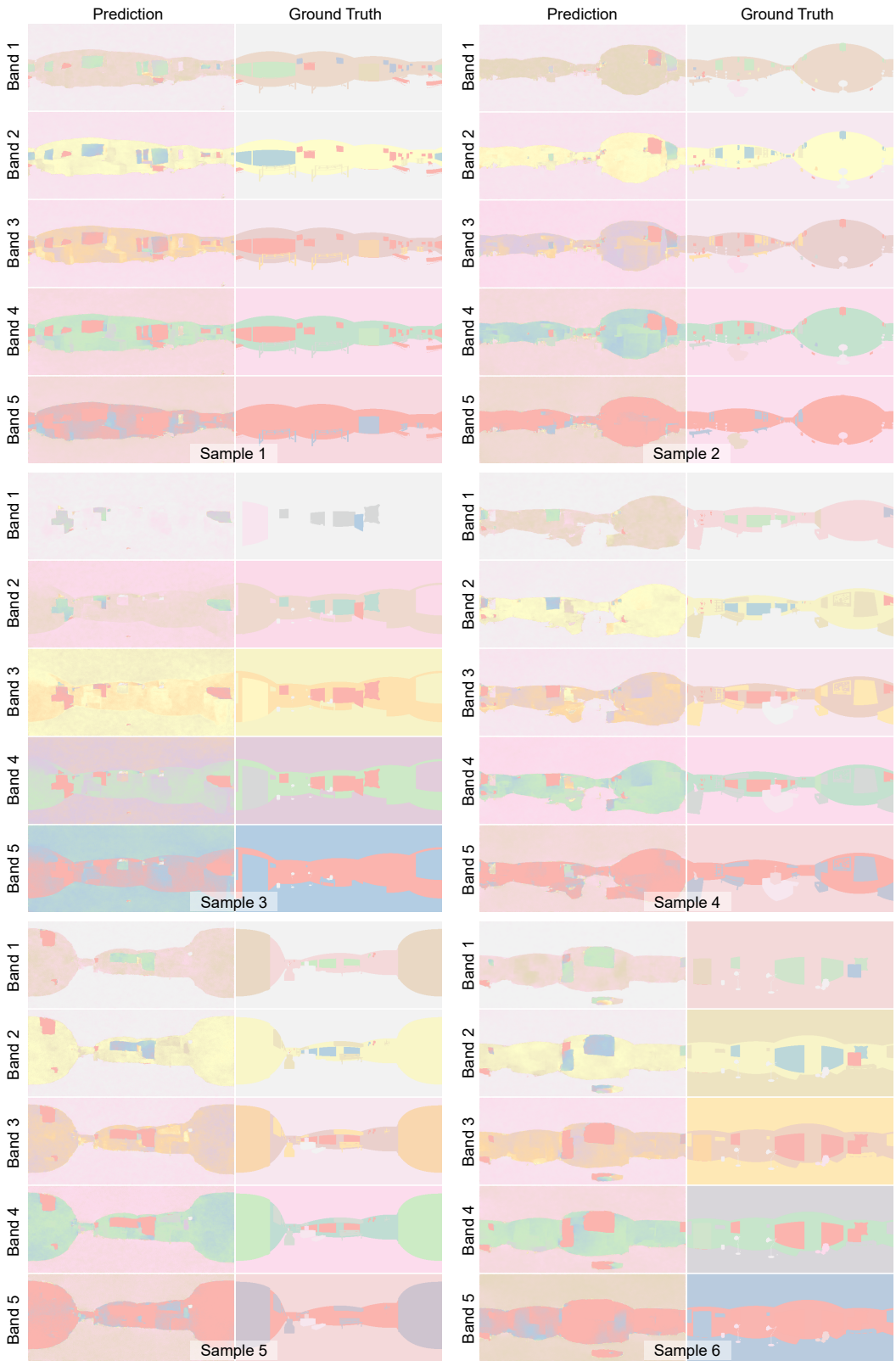

Figure 9: Additional qualitative per-band reflectance results. Each sample compares prediction and ground truth across the five frequency bands.

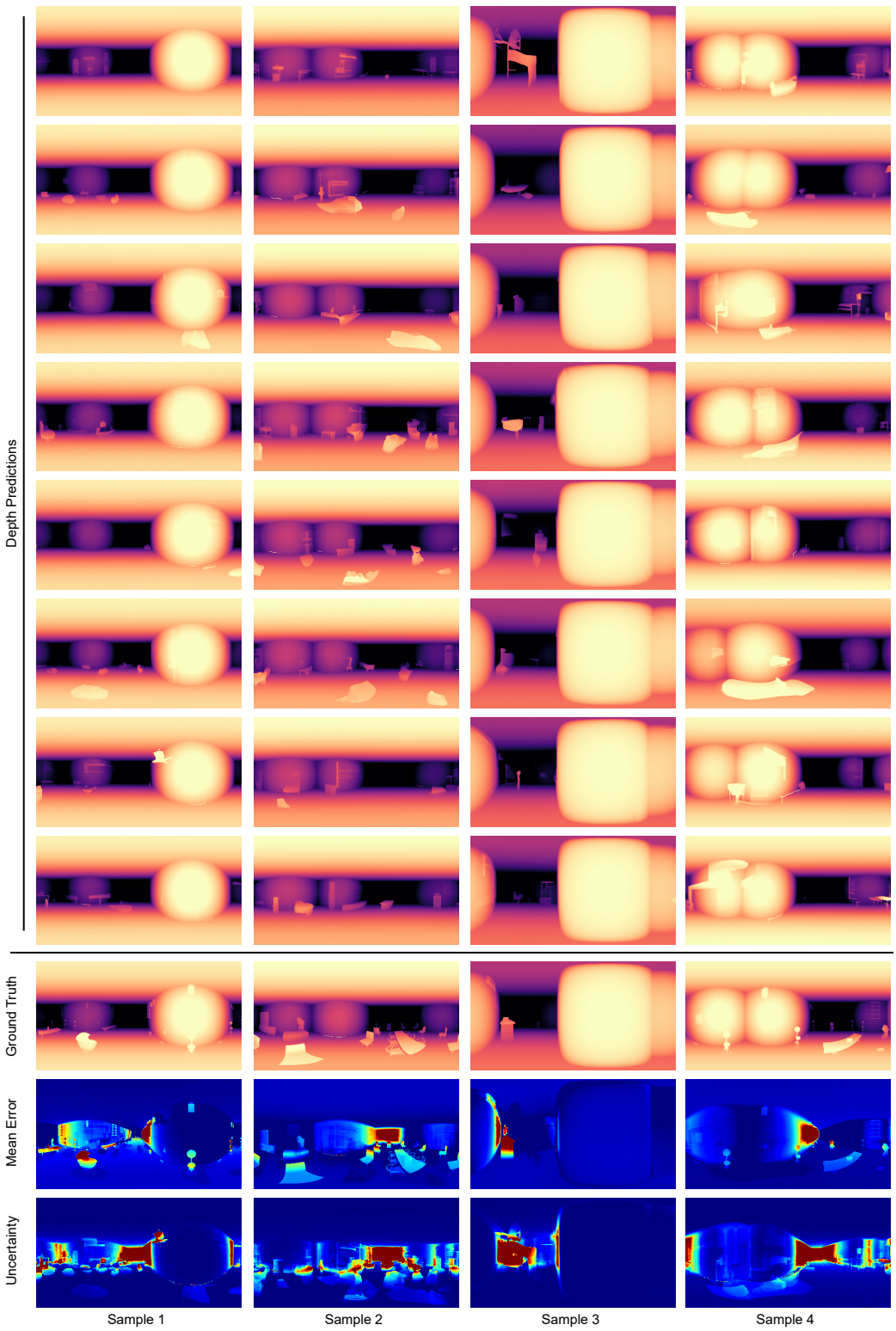

Figure 10: Additional stochastic depth prediction examples. Each column is one sample; rows show multiple generated depth samples, ground truth, mean error, and predictive uncertainty.

A.1 Noise Robustness

Measurement simulation. We evaluated noise robustness by introducing environmental noise into simulated acoustic measurements before RIR recovery. Each four-channel FOA RIR was convolved with a one-second chirp sampled at 48 kHz, sweeping from 50 Hz to 20 kHz. We mixed the resulting recordings with environmental noise from WHAM! [64] at SNRs of 10, 20, and 30 dB. We then applied the RIR-recovery pipeline used for our real-world measurements (Appendix C) to the noisy recordings.

Training and evaluation. We fine-tuned AIR with noise augmentation using WHAM! training clips and evaluated using WHAM! test clips. Each noisy condition was evaluated on the full simulated test set of 240 samples across 20 held-out rooms, with 12 randomly selected locations per room. Tab. 8 reports all six depth metrics. The noise-free reference is the FOA SphericalBF result from Tab. 1.

SNR	RMSE (m) ↓	logRMSE ↓	SILog ↓	Delta1 ↑	Delta2 ↑	Delta3 ↑
Noise-free	**0.5020**	**0.2820**	**26.34**	**0.8220**	**0.9110**	**0.9510**
30 dB	0.5165	0.2979	27.95	0.8183	0.9059	0.9466
20 dB	0.5273	0.3172	29.78	0.8116	0.9030	0.9443
10 dB	0.6783	0.4253	39.66	0.7336	0.8626	0.9152

Table 8: Depth RMSE remains below 0.53 m at 20 and 30 dB SNR. FOA depth estimation with environmental noise added before RIR recovery. Noisy conditions use noise-augmented fine-tuning; the noise-free reference is from Tab. 1.

B RIR-Room Dataset

Overview. We construct an RIR-Room dataset by pairing physically traced acoustic propagation with metric panoramic scene geometry. We generate indoor scenes using the procedure in [74], which builds on Holodeck [69]. We sample valid speaker-microphone pairs in each scene and render a microphone-centered equirectangular depth map $\boldsymbol{D}$ and per-band reflectance map $\boldsymbol{P}$ whose canonical forward direction is $+X$. For acoustic simulation, we build on AcoustiX [43] ray tracing simulation. In our pipeline, we use the AcoustiX backend to trace each speaker-microphone pair and store ray features, including propagation delay, amplitude, and arrival/departure directions. These ray features form an intermediate representation from which FOA RIRs can be synthesized and randomly azimuth-augmented at training time, while the depth/reflectance map is rotated by the same amount to preserve geometric alignment. For each sample, our dataset provides aligned ray features, synthesized FOA observations, panoramic depth/reflectance map, and surface-reflection properties.

Scene Generation. Recent work on multi-turn tool-use agents, such as PACT [20], studies how LLMs learn to coordinate external tools. Following a pipeline similar to WaveVerse [74], we create 116 indoor scenes, each exported as an XML file compatible with Mitsuba for depth queries and with AcoustiX for acoustic simulation. We generate 100 speaker-microphone pairs per room and use 96 rooms for training. The full simulated evaluation set contains 240 samples from the remaining 20 held-out rooms, with 12 randomly selected locations per room. We compute a top-down placement map for each room to obtain physically valid sample locations. The script estimates free floor regions, removes occupied cells, and inflates blocked regions so sources and receivers are not placed inside or too close to surfaces. Speaker-microphone pairs are then sampled from valid cells; in our current construction, they share the same horizontal grid cell, with the transmitter placed at a fixed vertical offset above the receiver. Fig. 11 shows examples of sampled points in the simulated environments.

Depth Map. For each sampled speaker-microphone pair, we generate two aligned outputs from the same scene: a receiver-centered panoramic depth map and a set of acoustic ray features. The depth map is rendered from the receiver position as a 256×512 equirectangular panorama. To render this, we first render six cubemap depth faces with a perspective camera, convert them to spherical coordinates, and store the resulting metric distance to the first visible surface for each panoramic ray. The saved panorama is aligned to the receiver coordinate convention used by the acoustic simulation, with the canonical forward direction corresponding to $+X$. The 2:1 panorama samples azimuth uniformly over 360° and elevation uniformly over 180°, following the angular convention in Section 3.2. Surfaces behind the first intersection are outside this representation, although they can contribute to the RIR through reflected paths. Consequently, scenes that differ only in such hidden geometry can share the same target panorama while producing different RIRs.

Per-band Reflectance Maps. We construct a lookup table that assigns each material in the scene library a vector of frequency-dependent acoustic reflection coefficients. The five entries correspond to octave bands centered on 250, 500, 1000, 2000, and 4000 Hz. For each dataset sample, we cast the same equirectangular ray grid used for the depth panorama into the scene. At every hit pixel, we read the material of the first intersected mesh and replace it with the corresponding five-dimensional reflection-coefficient vector, producing a per-band reflectance tensor $\boldsymbol{P} \in \mathbb{R}^{5\times H\times W}$. Materials in the finite library have fixed coefficient vectors, while model predictions are continuous. Pixels without a valid hit are assigned zero values and excluded by a binary validity mask.

Acoustic Scene Ray Tracing. For the acoustic ray tracing, we use AcoustiX [43] to model sound propagation. Each valid propagation path $[\tau,\ a,\ \phi_r,\ \theta_r,\ \phi_t,\ \theta_t]$ is stored as the following features: propagation delay τ, amplitude a, receiver-side arrival angles (ϕ_r, θ_r), and transmitter-side departure angles (ϕ_t, θ_t). Paths whose delays fall outside the target impulse-response window are discarded. We generate dataset variants with different maximum reflection depths, allowing the same geometric supervision to be paired with acoustic observations containing different orders of reflected paths. Since both the depth panorama and the ray features are generated from the same receiver pose and scene geometry, they share a consistent spherical reference frame.

Impulse Response Synthesis and Augmentation. We keep ray features as an intermediate representation rather than committing only to pre-rendered waveforms. We then synthesize a FOA RIR in the WYZX convention. For each valid acoustic path, the propagation delay determines its temporal shift, the path amplitude determines its contribution strength, and the receiver-side arrival direction determines the FOA encoding weights. The weighted contribution is accumulated into the FOA components, producing an FOA signal $\boldsymbol{x} \in \mathbb{R}^{4\times T}$ aligned with the model input representation. This ray-based representation also makes rotation-aware augmentation straightforward. During training, we sample a random azimuth angle, rotate the receiver-side ray azimuths before FOA synthesis, and

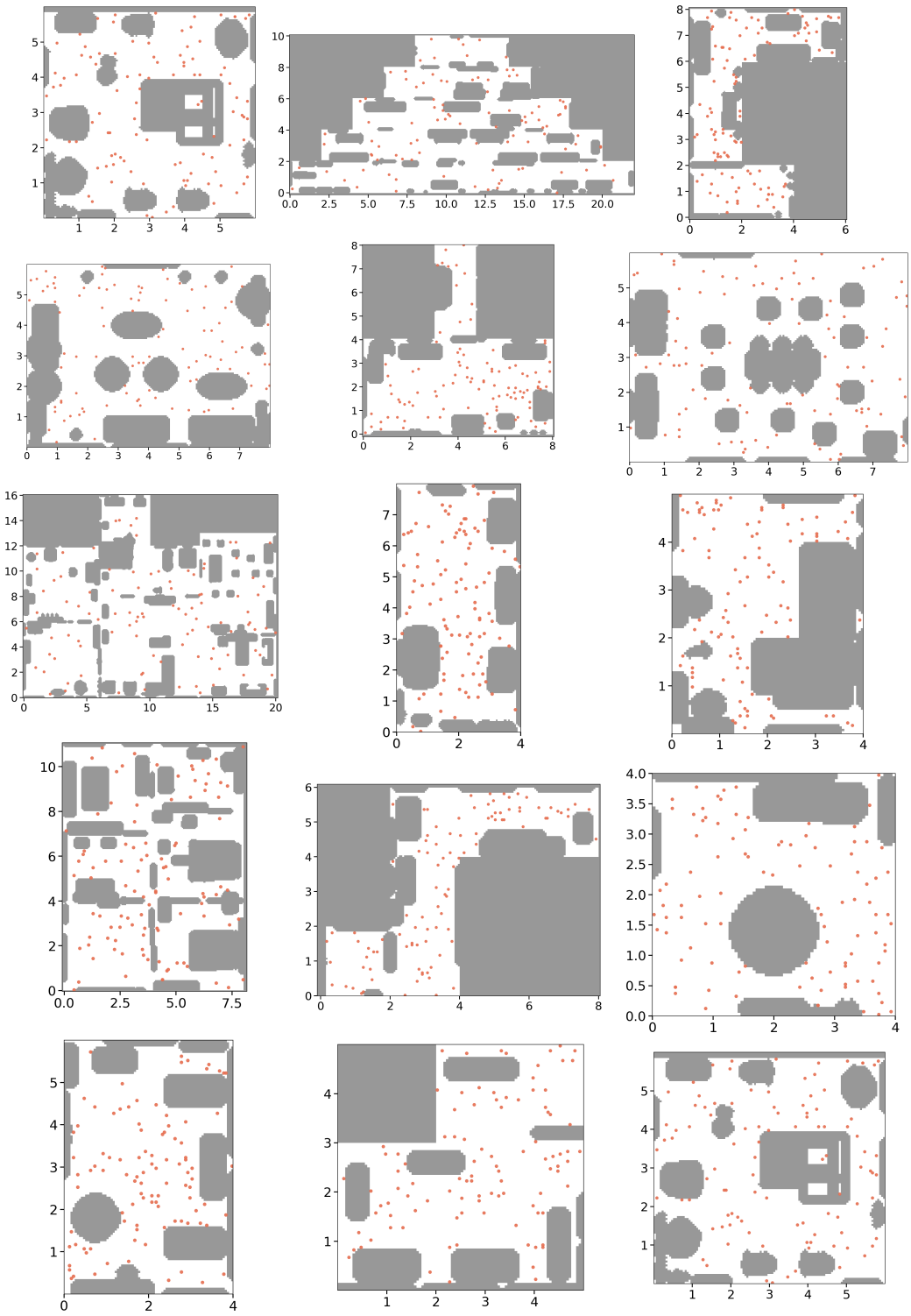

Figure 11: Top-down placement maps for simulated scenes. Red points indicate valid sampled co-located speaker-microphone positions, and gray regions indicate occupied or invalid areas.

roll the equirectangular depth map horizontally by the corresponding amount. For samples represented directly as FOA waveforms, we apply the equivalent azimuth rotation to the FOA components.

In both cases, the acoustic observation and target depth/reflectance map remain in the same spherical coordinate frame, encouraging the model to learn rotation-consistent acoustic-geometry cues.

C Real-World Dataset

To make RIR-based spatial perception practical and easy to deploy in real-world environments, we construct a real-world data collection pipeline using commodity sensing devices rather than specialized microphone arrays. As shown in Fig. 4, our acoustic sensing setup consists of a OnePlus 10T smartphone and a Bose SoundLink Revolve loudspeaker. The smartphone serves as the binaural recording device, while the loudspeaker serves as the acoustic source for signal playback. At each measurement location, the smartphone microphone and the loudspeaker are placed together and treated as a co-located acoustic receiver–source pair. This design avoids the need for calibrated microphone arrays or customized acoustic hardware, and better reflects a deployment scenario in which acoustic measurements are collected using widely available mobile devices.

For geometric supervision, we use an Ouster OS0-64 LiDAR to capture the ground-truth panoramic depth map at the corresponding measurement location. The LiDAR is used only during dataset construction to provide dense geometric labels, while the model input remains purely acoustic. This separation allows us to study whether room geometry can be inferred from realistic binaural acoustic measurements, without assuming access to depth sensors at inference time. The resulting setup therefore pairs mobile-device acoustic sensing with accurate geometric supervision, enabling controlled real-world evaluation of RIR-to-depth prediction.

At each measurement location, the co-located loudspeaker and smartphone microphone are positioned before acoustic recording. The Bose loudspeaker plays a one-second chirp signal sweeping from 50 Hz to 20 kHz, and the smartphone records the resulting binaural acoustic response through its built-in microphones. The recorded response contains the direct-path propagation as well as reflections induced by surrounding walls, furniture, and other scene structures. Since the transmitted chirp is known, we estimate the RIR using matched filtering between the recorded binaural signal and the chirp waveform. The matched-filtered response is then temporally cropped to a fixed one-second window, producing a one-second binaural RIR for each measurement.

Each training example is represented as a paired acoustic–geometric observation consisting of a binaural RIR and the corresponding panoramic depth map captured at the same measurement location. In total, we collect data from 5 real indoor rooms and in 180 paired binaural RIR–panoramic depth samples. We use 3 rooms (120 recordings) for training and reserve the remaining 2 rooms (60 recordings) as an unseen-room test split. This dataset allows us to evaluate RIR-to-depth prediction under realistic acoustic sensing conditions, while keeping the data collection procedure lightweight and reproducible. By relying on a smartphone as the recording microphone and a portable Bose speaker as the acoustic source, our collection protocol reduces the practical barrier to acquiring real-world RIR–depth data and supports the development of deployable acoustic spatial perception systems. On the real-world dataset, we average eight independent depth predictions per input RIR for stability.

D Implementation Details

Depth normalization. The model does not diffuse raw metric depth directly. Let $\boldsymbol{D}$ denote the raw metric depth map and D_{cap} denote the maximum valid depth in the training set. We first clip depth values to this global cap. We then apply logarithmic compression and map the result into the diffusion model's $[-1, 1]$ range:

$$\boldsymbol{D}_{\mathrm{clip}} = \min(\max(\boldsymbol{D}, 0),\ D_{\mathrm{cap}}), \tag{7}$$

$$\boldsymbol{D}_{\log} = \frac{\log(1 + \boldsymbol{D}_{\mathrm{clip}})}{\log(1 + D_{\mathrm{cap}})}, \tag{8}$$

$$\boldsymbol{D}_{\mathrm{model}} = 2\,\boldsymbol{D}_{\log} - 1. \tag{9}$$

This parameterization preserves metric ordering while compressing large depth values, which is especially important for panoramic scenes with long-range structure. During training the network

predicts the velocity target $\boldsymbol{v}_t$, with prediction $\hat{\boldsymbol{v}}_t$. The training loss is a masked mean squared error over valid-depth pixels indicated by binary mask $\boldsymbol{M}$, where $\odot$ denotes elementwise multiplication:

$$\mathcal{L} = \frac{\|\boldsymbol{M} \odot (\hat{\boldsymbol{v}}_t - \boldsymbol{v}_t)\|_2^2}{\max(1, \sum \boldsymbol{M})}. \tag{10}$$

We also apply conditioning dropout: with probability 0.1 the acoustic condition is replaced by null conditioning, enabling classifier-free guidance at inference time.

Acoustic normalization. We normalize the acoustic input per sample before encoding. Specifically, both the cross-attention encoder and the spatial conditioner divide the input FOA RIR by its peak absolute amplitude. This removes large scale variations across recordings while preserving the relative temporal and directional structure that the conditioner uses for geometry estimation.

We also apply rotation-aware training through random yaw augmentation. During training, the panoramic depth map is rolled horizontally, and the acoustic observation is rotated consistently: for native FOA RIRs we rotate the FOA components by the same yaw, while for ray-feature-derived inputs we rotate the ray azimuths before synthesizing the RIR. This encourages the model to learn rotation-consistent geometry cues on the sphere instead of overfitting to a fixed global yaw convention.

Model details. AIR's reported depth estimation results use a model trained and evaluated to predict $\boldsymbol{D}$ alone, while its reflectance estimation results use the same architecture trained and evaluated to predict $\boldsymbol{D}$ and $\boldsymbol{P}$ jointly. For our model, we leverage a piecewise temporal schedule that keeps the first 1200 RIR samples at full resolution and downsamples later samples more aggressively before the shared per-ray encoder. The spatial residual is injected at the UNet mid-resolution, and inference uses classifier-free guidance. We use the `DDIMScheduler` from Diffusers with 1,000 training diffusion timesteps, 50 inference steps, and a linear beta schedule from 10^{-4} to 0.02.

We train the model using Adam optimizer with a learning rate of 1×10^{-4}, a weight decay of 1×10^{-2}, and a batch size of 16 samples. We train for 50k steps with 1k learning rate warm up steps and evaluate the 50k-step checkpoints.

E Societal Impacts

This work studies whether indoor geometry and surface reflectance can be inferred from RIRs. A positive impact of this direction is that acoustic sensing could support spatial perception in settings where cameras are unreliable, unavailable, or undesirable, such as low-light navigation, assistive devices, and privacy-preserving indoor mapping. Because the proposed setup can use lightweight acoustic recordings, it may also lower the hardware barrier for collecting spatial information in ordinary indoor environments. Recovering room geometry from sound could also support immersive virtual environments, complementing advances in audio-visual scene generation [42, 29, 28], human motion generation [67, 66, 10, 75], and language-guided stereo audio editing [38, 40].

At the same time, acoustic scene reconstruction raises privacy and consent concerns. Although our method predicts coarse geometry and acoustic reflectance rather than human identity or speech content, room layout itself can reveal sensitive information about private spaces. Future deployment should therefore require clear user consent, visible or audible signaling during active measurement, and restrictions on collecting or storing acoustic data in spaces where participants have not agreed to sensing. Released datasets and models should also avoid exposing personally identifiable audio content and should document the collection protocol, intended use, and known limitations.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly state our contributions in the abstract and introduction and conduct extensive experiments on our dataset.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in the Discussion session.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: This work does not include theoretical proof.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We clearly detailed our methods and parameters for the training of the network.

   Guidelines:

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

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

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

Answer: [No]

Justification: Our code and dataset will be open-sourced later.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We have all the training and testing details, including data splits, hyperparameters, ablation studies on hyperparameters and the description of the used optimizer.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We exclude error bars because it would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We have all details in the experiment setup in the main text and the supplementary results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics and preserve anonymity

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We mention both potential societal impacts in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: Our model and data do not pose such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite the corresponding paper when needed.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We discuss the model architecture details, training details, dataset collection process in the paper.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: This work does not use LLMs.

Guidelines:

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