Hearing Anything Anywhere

Mason Long Wang[1*] Ryosuke Sawata[1,2*] Samuel Clarke[1]
Ruohan Gao[1,3] Shangzhe Wu[1] Jiajun Wu[1]
[1]Stanford University [2]Sony AI [3]University of Maryland, College Park
masonlwang.com/hearinganythinganywhere

Abstract

Recent years have seen immense progress in 3D computer vision and computer graphics, with emerging tools that can virtualize real-world 3D environments for numerous Mixed Reality (XR) applications. However, alongside immersive visual experiences, immersive auditory experiences are equally vital to our holistic perception of an environment. In this paper, we aim to reconstruct the spatial acoustic characteristics of an arbitrary environment given only a sparse set of (roughly 12) room impulse response (RIR) recordings and a planar reconstruction of the scene, a setup that is easily achievable by ordinary users. To this end, we introduce DIFFRIR*, a differentiable RIR rendering framework with interpretable parametric models of salient acoustic features of the scene, including sound source directivity and surface reflectivity. This allows us to synthesize novel auditory experiences through the space with any source audio. To evaluate our method, we collect a dataset of RIR recordings and music in four diverse, real environments. We show that our model outperforms state-of-the-art baselines on rendering monaural and binaural RIRs and music at unseen locations, and learns physically interpretable parameters characterizing acoustic properties of the sound source and surfaces in the scene.*

1. Introduction

Much of the impetus to realize immersive virtual reality (VR) stems from the desire to recreate and share *real* scenes and experiences. Motivated by this goal, recent progress in 3D computer vision and computer graphics has led to tools that can virtualize real-world 3D environments using simple consumer devices (e.g., cellphone cameras) for numerous Mixed Reality (XR) applications. Alongside immersive visual experiences, immersive auditory experiences are equally vital to our holistic perception of an environment. For instance, while the interior of Carnegie Hall in New York City is visually beautiful, one cannot fully appreciate the majesty of its design without experiencing a musical performance in-person and hearing its unique acoustics.

In this paper, our goal is to capture the acoustic intrinsics of a real-world scene using a sparse set of measurements, in order to render arbitrary source audio at any location, hence the name, "Hearing Anything Anywhere". This is analogous to the task of sparse-view novel view synthesis (NVS) in computer vision and graphics [6, 48, 68].

However, there are two key differences between light and sound that make common approaches to visual NVS inapplicable to audio. First, light is typically emitted from continuous sources and travels steadily and almost instantly through space, resulting in a largely stationary visual scene. In contrast, sound signals are usually time-varying and travel through space at a much slower pace, resulting in a constantly changing 4D acoustic field with both numerous early reflections and late reverberations. Second, a single camera captures *millions* of pixels in a split second, each recording a distinct light ray from a *particular* direction. In contrast, a typical microphone only records an amalgamation of sound waves arriving to a *single* location from *all* directions, with different times-of-arrival. Therefore, while it is possible to capture the appearance of a 3D scene by simply walking through it with a camera, the same approach falls short to record the entire 4D acoustic field.

Thus, capturing a fully immersive acoustic field often necessitates setting up hundreds of microphones densely across the space [43, 54, 56, 63], which is impractical for many consumer use cases. In this work, we attempt to capture real-world acoustic spaces with a *basic* hardware setup, e.g., 12 microphones, which can be easily scaled to arbitrary environments.

To capture the acoustic properties of the scene, we measure a room impulse response (RIR) between the sound source and each microphone location. An RIR is a time-series signal that estimates how a perfect impulse emitted from the source, traveling and bouncing in the room, would be perceived at the listener location. RIRs effectively capture a room's intrinsic acoustic properties between source

*Equal contribution.

and listener points, and are thus widely used in acoustic simulation [4]. In order to simulate the sound of an arbitrary source for a particular listener location in a room, the RIR associated with the source-listener pair is simply convolved with the source audio [41].

We thus formulate our *Hearing Anything Anywhere* task as inferring RIRs and music at novel listener locations from a sparse set of RIRs measured between a single source and a small set of microphone locations spatially distributed within the scene. Towards this goal, we introduce a fully differentiable impulse response rendering framework DIFFRIR that reasons about the individual contributions of each acoustic reflection path between the source and the receiver, including the time delay and magnitude of the sound on each path, as well as the influence of reflections from each surface in the scene.

By explicitly modeling the sound source location, the directivity map of the source, and the reflection properties of the surfaces in the scene in a fully differentiable audio rendering framework, we can characterize the parameters of each model through an analysis-by-synthesis paradigm by optimizing the output of DIFFRIR against the known subset of measured RIRs. After optimizing the interpretable parameters of our model, we can estimate the RIR from any unseen location in the scene.

To validate our method, we collect a dataset that contains RIR measurements from four real-world environments that represent a diverse range of room materials, shape, and complexity. Through experiments comparing our framework with current state-of-the-art methods, DIFFRIR shows greater robustness in real, data-limited scenarios. Moreover, with the explicit and interpretable models of source and surface reflection properties, we can easily synthesize novel auditory experiences with different speaker orientations and locations, which can be useful in applications such as virtual reality and acoustics-aware interior design. In addition, the differentiable and interpretable models of our framework allow us to estimate acoustic parameters of the sound source and surfaces in the room, which can be useful in applications like robotics and architectural design for acoustics.

Our contributions are threefold. First, we contribute DIFFRIR, a differentiable acoustic inverse rendering framework that can recover the fully immersive acoustic field of a room from a set of 12 sparsely located RIR measurements. Second, we contribute a new dataset of real-world RIRs measured from hundreds of locations in four different real environments. Third, we compare our method to existing methods across various settings, demonstrating that our method is more effective than existing methods on real data in our data-limited scenarios, predicting more accurate RIRs and music at unseen locations. Code and data are available at the project website.

2. Related Work

Learning-Based Room Acoustics Prediction. While many acoustical learning frameworks model room acoustics implicitly, others explicitly interpolate and predict RIRs at novel points. Frameworks that predict RIRs at novel points in a room vary not only in their underlying techniques, but also in their inputs. Some methods do not use vision or geometry to make their estimates, but instead learn to directly approximate a function mapping spatial coordinates to RIRs [54, 56]. These methods can require large training set sizes on the order of 1,000 RIRs from a room to effectively interpolate RIRs to novel points within the same room. Alternatively, some methods use geometric features of the scene [43], such as [63], which learns a diffuse reflection model from a small subset of points in the mesh of the environment, to achieve a performance improvement over pure audio-based methods. Our method uses environment geometry to explicitly model specular reflections on each surface. To validate our approach, we compare against three baselines, including one audio-only method [56] and two methods that use scene geometry [43, 63].

Audio-Visual (AV) Room Acoustics Prediction. Other methods learn relationships between visual inputs and room acoustics to perform tasks such as predicting the dereverberated signal from an audio recording and a panoramic image of the recording environment [17], or predicting how an input audio signal would sound in a target space based on an image of the space [14]. Many works use visual inputs to explicitly perform the novel view acoustic synthesis (NVAS) task. For instance, Chen et al. [16] proposed the Visually-Guided Acoustic Synthesis (ViGAS) network, which outputs the spatial audio of the speech of a human in corresponding visual frames. Furthermore, by using audio-visual features as well as geometric ones, Ahn et al. [1] show that the important sub-tasks of NVAS, e.g., sound source localization, separation, and dereverberation, can be jointly solved. AV-NeRF [42] improved the performance of both NVS and NVAS tasks via multi-task training by using an audio-based Neural Radiance Field (NeRF). Their audio NeRF estimates variations in the magnitudes of audio perceived from varying locations, whereas we explicitly estimate the RIR, a much more holistic characterization of the environment acoustic properties.

Similar to our binaural prediction task, Garg et al. [27] predict binaural audio from an AV scene's monaural audio and visual features extracted from the scene's video frames. Although AV approaches can sometimes outperform uni-modal audio-only models at estimating environment acoustics, collecting large enough datasets of synchronized audio-visual pairs for these models can be laborious. Perhaps for this reason, many such models, even one boasting few-shot generalization [45], present results from eval-

uating exclusively on simulated data.

Geometry-Based RIR Simulation. Many of the aforementioned works use datasets of simulated RIRs generated by the SoundSpaces framework [15], a fast acoustic simulator based on geometric acoustic methods. They simulate the acoustics of virtualized versions of real rooms from datasets of meshes reconstructed from RGBD scans of real rooms in home and workplace environments, such as the Matterport3D dataset [12] or the Replica dataset [62]. The Geometric-Wave Acoustic (GWA) dataset uses a hybrid propagation algorithm combining wave-based methods [31] with geometric acoustic methods, intending to model low-frequency wave effects more accurately, albeit at the cost of longer run-time. The input meshes are from a dataset of professionally designed virtual home layouts [26]. The Mesh2IR framework uses the GWA dataset to learn a conditional generative adversarial network (cGAN) to more quickly predict RIRs from meshes of rooms [55]. The authors do not show how their cGAN's estimates of RIRs compare to measured RIRs from real rooms.

Differentiable Acoustics. The previously mentioned simulators are not differentiable, which precludes gradient-based optimization techniques which can be used in solving inverse problems. Differentiable audio rendering techniques have been used to solve such inverse problems estimating acoustic properties of musical instruments [25] and everyday objects [20], as well as the reverberation properties of the environments they are in. The authors of [19] implemented a differentiable acoustic ray tracer for inverse tasks in underwater acoustics, such as estimating the absorption of the seabed on simulated 2D data. We use similar principles for estimating absorption parameters of surfaces in 3D environments from our real, airborne sound data.

3. Method

We first lay out the definition of our task, and then introduce our proposed DIFFRIR framework to approach it.

3.1. Task Formulation

To achieve our goal of virtualizing real acoustic spaces, our method should require information about the room that is as easy as possible to obtain. With this objective in mind, we show that our method produces accurate results, while only requiring the following:

1. A small set of omnidirectional RIR recordings captured at sparse locations (e.g., 12), with the xyz coordinates at which they were captured.
2. The room's rough geometry, expressed as a small number of planes.

RIRs can be easily captured by playing a sine sweep from the source location and recording it from a microphone at the listener location. In our setup, we assume a stationary audio source whose orientation and position are unknown. With this information, our goals are to simulate monoaural and binaural RIRs and music at arbitrary listener locations and orientations in the room.

3.2. The DIFFRIR Framework

To achieve this task, we design a differentiable RIR rendering framework, dubbed DIFFRIR. As an overview of the DIFFRIR framework, we use the sound source and microphone location, along with the planar decomposition of the environment, to trace all specular reflection paths between the source and a listener location, up to a certain number of reflections. We estimate the sound arriving to the listener from each path using a series of parametric models for the sound source directivity and impulse response, as well as the acoustic reflection of each surface. Each model is fully differentiable, with interpretable parameters. We compute each RIR as the sum of contributions of the sound arriving from each path, combined with a learned residual. We use these models in a differentiable audio renderer to optimize parameters according to a loss function comparing our estimates to the known subset of ground-truth RIRs. We describe each model in detail below.

3.2.1 Characterizing the Sound Source

Source Localization. We first estimate the location of the sound source for all subsequent steps. Based on the known subset of RIRs we use their locations and the timing of the first peak to localize the source using a traditional time-of-arrival method. More details are provided in Appendix E.

Source Directivity. Most real sound sources do not radiate sound uniformly in all directions. For instance, a loudspeaker will usually be much louder from the front, and human speakers also have distinct directivity patterns [52]. The source's *directivity* describes the way in which the source radiates sound differently in different directions and is generally frequency dependent. For example, a loudspeaker will overall sound much louder from the front, with the higher-frequency components radiating in especially narrow beams and lower-frequency components more omnidirectionally. The sound source's directivity has a significant impact on the acoustic field of the room and is therefore important to model.

We model the filtering effect of exiting the sound source in any particular direction with the *directivity response*. Let $\vec{d_p}$ be the absolute direction (given as a unit vector) in which the sound path exits the speaker. Our goal is to fit $D(\vec{d_p})$, a function mapping $\vec{d_p}$ to a magnitude frequency response that accounts for the effect of exiting the speaker in the direction of $\vec{d_p}$. When a sound exits the speaker in the direction of $\vec{d_p}$, the frequency content of the sound wave is *multiplied* by $D(\vec{d_p})$.

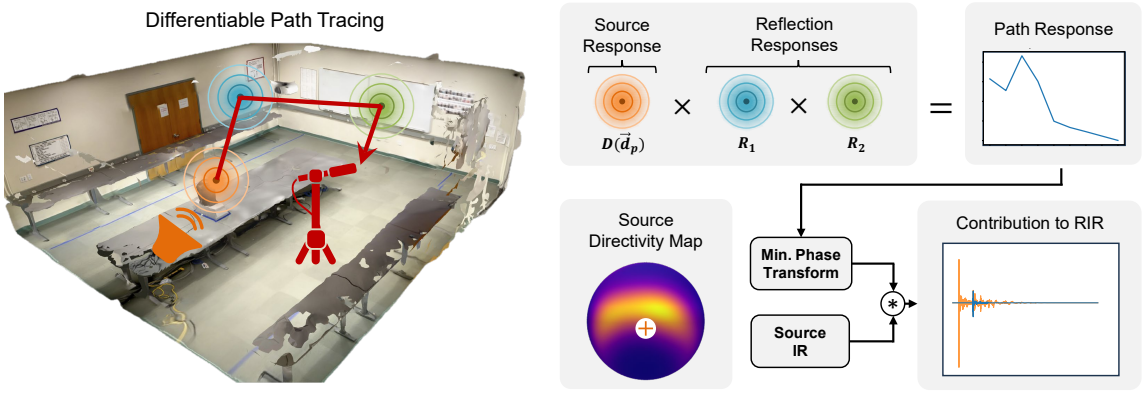


Figure 1. **Differentiable Room Impulse Response Rendering Framework (DIFFRIR).** Our model renders the contribution to the RIR of a single traced reflection path. After computing a reflection path, we characterize it by the direction at which it exits the speaker, its length, and the surfaces on which it reflects. The sound source has a learned frequency response that depends on the outgoing direction, and each surface has a different learned frequency response. We multiply each of these responses to estimate the overall path response. To determine the reflection path's time-domain contribution to the final RIR, we apply a minimum-phase inverse-Fourier transform to the path response, convolve it with the source impulse response, and then shift the result in time based on the path length and the speed of sound.

To model the direction-dependent frequency response, we fit F different heatmaps on unit spheres centered on the speaker, one heatmap for each of F octave-spaced center frequencies comprising vector $\mathbf{f}$. To do this, we distribute 128 points evenly along the surface of the unit sphere, using a Fibonacci lattice [32]. We denote this set of points L. Let $A_{\vec{x},f_o}$ be the log-amplitude gain for sound traveling out of the speaker in the direction of $\vec{x}$ at frequency f_o. To determine the log-amplitude gain at f_o in direction $\vec{d_p}$, we interpolate between the points on the heatmap using a spherical Gaussian weighting function, inspired by [67]:

$$A_{\vec{d_p},f_o} = \frac{\sum_{x\in L} A_{x,f_o} e^{-\lambda(1-\vec{d_p}\cdot x)}}{\sum_{x\in L} e^{-\lambda(1-\vec{d_p}\cdot x)}}, \tag{1}$$

where λ is a fixed sharpness value shared across all heatmaps. In order to obtain the full frequency response for the direction d, we linearly interpolate between the log-amplitude gains as in [33], and then exponentiate them to convert them to linear amplitude values:

$$D(\vec{d_p}, f_o) = \mathrm{e}^{\ell(\mathbf{A_d},\mathbf{f},\mathrm{f_o})}, \tag{2}$$

where ℓ represents linear interpolation on the vector of decibel values $\mathbf{A_d}$ indexed by center frequencies $\mathbf{f}$, based on query frequency f_o.

Source Impulse Response. Since the room impulse response relates the source signal fed to the speaker to the sound heard in the room, we must also account for the way that the source modifies the source signal being fed to it. For instance, if the source is a loudspeaker, it may attenuate or boost certain frequencies. We model these effects by learning a source impulse response IR_s in the time domain, thus approximating the source's response as a linear system [8] and convolving it with our RIR.

3.2.2 Modeling and Characterizing Reflections

We trace each specular reflection path and model the acoustic effects of each reflection along the path, with unique reflection parameters for each surface in the environment.

Reflectivity. When a sound wave encounters a surface, a fraction of the sound wave's energy will be specularly reflected, while the remaining energy will be absorbed, transmitted, diffusely reflected, or diffracted. These effects vary by frequency, depending on the texture and material properties of each surface.

For each surface s, we fit a vector $\mathbf{V_s}$ of F different values representing the magnitude of sound specularly reflected by the surface at each of F octave-spaced centered frequencies in vector $\mathbf{f}$. We apply the sigmoid function to these values to determine the *energy* reflection coefficients (the proportion of specularly reflected sound energy) at each frequency. Next, we determine the *amplitude* reflection coefficients (the amount that the surface attenuates the incoming sound at each frequency in terms of linear amplitude gain) by taking the square root of the energy reflection coefficients [39]. Using the amplitude reflection coefficients at the F center frequencies, we obtain the amplitude gains for arbitrary frequencies through linear interpolation. This gives us the *reflection response* R_s, a magnitude frequency response representing the surface's effect on incoming audio of different frequencies. Thus, the formula for R_s is:

$$R_s(f_r) = \ell\left(\sqrt{\sigma(\mathbf{V_s})}, \mathbf{f}, f_r\right). \tag{3}$$

Here, σ denotes the sigmoid function, and ℓ is a linear interpolation from the coefficients $\mathbf{V_s}$ based on the relation of the query frequency f_r to the center frequencies $\mathbf{f}$.

Reflection Paths. Given the estimated source location S_{xyz}, a listener location L_{xyz}, and a planar representation of the room's geometry, we use the image-source method [3] to efficiently compute all of the specular reflection paths between the source and listener in the room, up to a particular order N (e.g., 5). The method considers all permutations from 1 to N of these surfaces with repetition and, for each permutation, determines if there is a valid reflection path that travels from the source to the listener after reflecting specularly off of each of the surfaces in order. For each valid reflection path p from source to listener, we track the length of the reflection path l_p, the ordered list S_p of reflection surfaces along the path, and the direction from which the path exits the source $\vec{d_p}$.

Rooms often contain parallel surfaces, which lead to prominent higher-order reflections. These reflections result in "axial modes," which are powerful room resonances with especially long reverberation times [57]. Thus, in addition to computing all N^{th}-order reflection paths for all possible orderings of surfaces, our image-source algorithm also computes all valid reflection paths for pairs of parallel walls, up to a much higher order, e.g., 50. This modification, which we call *axial boosting*, improves the model's performance (see Appendix D.4) in adversarial cases like the Hallway, with a computational overhead that scales linearly rather than exponentially with reflection order. We discuss additional surface interactions, such as diffuse reflection, in Section 3.2.3.

3.2.3 Combining Models

We combine these reflection and sound source models to estimate the contribution of each reflection path. We then sum the contributions across all paths and add a residual to estimate the RIR for a given source and listener location.

Contribution of a Single Reflection Path. In summary, for each individual reflection path p, the outgoing direction $\vec{d_p}$ from the source, the ordered list S_p of reflected surfaces, and the total path length l_p each have distinct effects on rendering the path's contribution. $D(\vec{d_p})$ characterizes the frequency response of the source from the path's outgoing direction. The reflection of each surface $s \in S_p$ attenuates the amplitude of the sound in a frequency-dependent fashion parameterized by R_s. The total reflection-based attenuation is the product of the frequency response across all $s \in S_p$. Finally, we use the path length l_p to compute the time of arrival t_p by dividing the path length l_p by the speed of sound. We also use l_p to estimate the attenuation of the amplitude due to spherical propagation, where the amplitude is inversely proportional to l_p, as well as air absorption, which we characterize by air absorption coefficient α [60].

Thus, the function K that computes the time-domain contribution of each individual path is:

$$K(d_p, S_p, t_p) = \frac{\alpha^{t_p}}{\rho}\tau \left[\mathcal{M}\left(D(d_p) \odot \prod_{s \in S_p} R_s \right), t_p \right], \tag{4}$$

where $\odot$ is the element-wise product, ρ is the length of the reflection path in meters, and τ_t is the time-shift operator, which delays its input signal by t_p seconds. $\mathcal{M}$ is a minimum-phase inverse Fourier transform, which computes a time-domain filter from a magnitude frequency response, assuming minimum phase. The minimum phase assumption can be used to approximate the phase of an acoustic reflection given a desired magnitude frequency response [46]. More details are in Appendix E.

Modeling Residual Effects. For the purposes of gradient-based optimization, we require a model that is fast, simple, and differentiable. Consequently, we do not explicitly model many physical phenomena, including diffuse reflection, diffraction, transmission, refraction, and higher-order specular reflections. Modeling all of these effects would increase our model's computational footprint, impeding the iterative process of fitting to a real scene. Instead, we approximate these effects as spatially uniform, with some theoretical justifications. As the reflection order increases, the number of reflection paths grows exponentially, making individual reflections less distinguishable. This comprises a sound field that, in real rooms, is approximately uniform and isotropic [37, 50, 51]. Diffuse reflections in particular can contribute to the uniformity of the sound field [64]. We approximate the total effect of high-order specular reflections, diffuse reflections and other effects as uniform, modeling them with a spatially-invariant residual signal r.

Overall Formula. Given respective source and listener locations S_{xyz} and L_{xyz}, we render the early-stage RIR by summing the contributions from all reflection paths, then convolve the result with the source's impulse response IR_s.

$$\text{RIR}(S_{xyz}, L_{xyz}) = \gamma \left[IR_s \circledast \sum_{p \in P} K(d_p, S_p, t_p) \right] + (1-\gamma) r \tag{5}$$

In this formula, $\circledast$ denotes convolution, and P is the set of all paths between the source and listener locations. As r is intended to capture higher-order reflections, its effects are likely to become more dominant later in the impulse response, whereas the traced paths are intended to characterize the early-stage reflections. For this reason, we fit 16 points on a temporal spline γ that interpolates a relative weighting between the contributions of the late-stage residual and those of explicitly computed reflection paths.

(a) Classroom (b) Dampened Room (c) Hallway (d) Complex Room

Figure 2. Photos of each room used for the DIFFRIR Dataset, each shown in its base configuration.

3.2.4 Fitting and Inference

We estimate the parameters of each acoustic model in the environment in an iterative analysis-by-synthesis process. Inspired by [20] and [25], we optimize according to a multi-scale log-spectral loss comparing rendered RIR $\hat{W}$ with the ground-truth RIR W measured at the same location. The specific loss formulation is in Eq. 6 in Appendix E.

For inference, we simply compute Equation 5 for a point at a novel location, computing all the specular paths below the maximum order between the source and the novel location, etc., and using the parameters we determined from the analysis-by-synthesis process.

Binauralization. We train our model on single-channel RIRs recorded using omnidirectional microphones. However, immersive spatial audio requires binauralization - the process of converting single-channel audio into left and right channels, in a way that mimics human perception. The shape of the head, the acoustic shadow it casts, and the differences in time-of-arrival between the left and right ears all result in distinct perceptual cues that help place the listener in the scene [28, 66]. These effects are typically modeled by head-related impulse responses (HRIRs). There is a different HRIR for each incoming audio direction. To render binaural audio, the incoming audio from each reflection path is convolved with an HRIR sampled from the SADIE II dataset [5] corresponding to its incoming direction. This allows our model to approximate perceptually accurate binaural audio, which captures the effects of the human head, with merely monaural supervision.

4. The DIFFRIR Dataset

To evaluate methods of rendering and interpolating RIRs, we collect a novel dataset of real monoaural and binaural RIRs and music data in four different rooms, as illustrated in Figure 2. Table 1 further summarizes the dimensions and reverberation time measurements of each room. In particular, we choose the following rooms to represent a wide range of room layouts, sizes, geometric complexities, and reverberation effects:

1. **Classroom.** A standard classroom with 13 rectangular tables combined into three groups, a chalkboard, two whiteboards, drywall walls, a carpeted floor, office tile ceiling, and three doors. There is ventilation noise.
2. **Dampened Room.** A semi-anechoic chamber with a carpeted floor, all four walls covered with jagged acoustic foam wedges, and specialty acoustic tile ceiling.
3. **Hallway.** A narrow, highly reverberant hallway, with two wooden doors, a tile floor, and drywall ceiling and walls.
4. **Complex Room.** A room with an irregular shape that resembles a pentagonal prism. Portions of the side wall and ceiling are covered with acoustic panels. There are three pillars in the middle of the room, one slanted diagonally. A portion of the rear wall is glass which is internally covered with paper posters. There are 7 tables, one of which is in a figure-eight shape. There are exposed air ducts, six hanging lights, water pipes, monitors, and chairs, as well as various large objects, such as a shelf. There is significant ventilation noise.

To collect audio recordings, we place a QSC K8.2 Loudspeaker in a particular location and orientation in the room and play sine sweeps to measure real RIRs in several hundred precisely-measured listener locations using a custom-built microphone array. In addition, we play and record several 10-second music clips selected from the Free Music Archive dataset [24] from the same listener and speaker locations. The music and RIRs are recorded using multiple time-synchronized Dayton Audio EMM6 omnidirectional microphones, as well as a 3Dio FS XLR microphone, which features ear-shaped silicone microphones to model human hearing and captures binaural audio.

Additional Configurations. We also collect additional subdatasets in some rooms where we slightly modify each room configuration. In each such subdataset, we vary the location and/or orientation of the speaker, or the presence and location of standalone whiteboard panels in the room. We use these additional configurations to evaluate zero-shot virtual speaker rotation and translation, and panel insertion and relocation. We include these evaluations and details on these configurations in Appendix C. While previous RIR datasets include varying room configurations [29, 47, 65]

Room	Size (m)	RT60 (s)	# of Points
Classroom	$7.1 \times 7.9 \times 2.7$	0.69	630
Dampened	$4.9 \times 5.2 \times 2.7$	0.14	768
Hallway	$1.5 \times 18.1 \times 2.8$	1.41	936
Complex	$8.4 \times 13.0 \times 6.1$	0.78	672

Table 1. Characteristics of each room and corresponding subdataset. The last column is the number of distinct microphone-speaker location pairs for which both RIRs and music are recorded, across all configurations. RT60 reverberation times are each room's average across frequencies and sub-configurations. For the Complex room, the size of its bounding box is reported.

the DIFFRIR Dataset is the first to our knowledge that also includes monoaural and binaural music recordings.

5. Experiments

For each room in our collected dataset, we evaluate our performance on the tasks of rendering both omnidirectional RIRs and music at unseen listener locations. In each room configuration, we select 12 omnidirectional RIRs to train our model. We then use our model to render RIRs at unseen locations in the test set, and compare our rendered RIRs to the ground-truth RIRs using metrics we detail in Section 5.1. To simulate music playing in the room, we convolve our rendered RIRs with five different source music files, and compare the result to real recordings of the same music files being played in the room, across the same metrics.

Baselines. We compare our method with nearest neighbor (NN) and linear interpolation baselines, which are widely used to interpolate RIRs [16, 43, 56]. We also compare with Deep Impulse Response (DeepIR) [56] and Neural Acoustic Fields (NAF) [43], which are both deep-neural-network-based (DNN-based) frameworks. DeepIR predicts the monaural RIR at novel locations based only on the location's coordinates, while NAF uses the location combined with local geometric features to estimate the RIR. In addition, NAF was originally designed for binaural rendering. Thus, we modify NAF to output monaural audio for the monaural RIR estimation task. We also compare our method with Implicit Neural Representation for Audio Scenes (INRAS) [63], which uses a combination of DNNs to more explicitly model specular and diffuse reflections at a subset of points in a scene's 3D mesh.

Additional details on baselines and any necessary adjustments we made to them are included in Appendix F.

5.1. Results

Metrics. We compare rendered audio to ground-truth audio using two metrics:

1. **Multiscale Log-Spectral L1 (Mag)**. A comparison of rendered and GT waveforms in time-frequency domain at multiple temporal and frequency resolutions [20, 25].
2. **Envelope Distance (ENV)**. The L1 distance between the log-energy envelopes of the ground-truth and rendered waveforms. Energy decay envelopes are used to extract the decay curve of the RIR, which characterizes the room's reverberant qualities [23]. We compute the signal's energy envelope by taking the envelope of the squared signal [9]. Satoh et al. [58] directly use this log-energy (squared) envelope of an RIR to measure the room's RT60 reverberation time, which is a common way of characterizing the room's acoustics [40].

Analysis. Our results for the base monaural prediction task are shown in Table 2. For the monaural prediction task, our model significantly outperforms all baselines on our metrics, across all rooms. Results for the binaural prediction task are shown in Appendix D.1.

5.2. Interpretability

We show the physically interpretable parameters our model learns for the source's directivity and reflection coefficients.

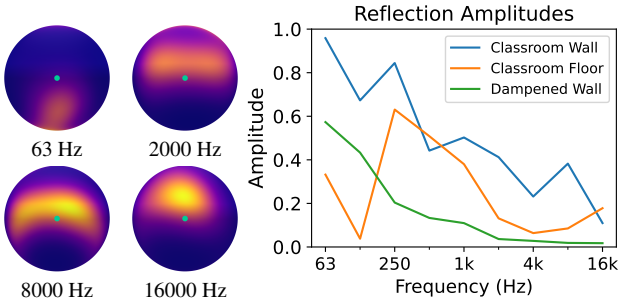


Figure 3. Visualization of our model's learned parameters. The left images show sample spherical heatmaps that our model fits to the speaker's directivity pattern when trained on 12 points from the Classroom subdataset. The green dot indicates the direction the speaker is facing, and the yellow regions indicate higher volume. The right image shows reflection amplitude responses that our model learns for various surfaces.

Directivity Maps. The left side of Figure 3 shows the source directivity heatmaps at various frequencies, learned from 12 training points in the Classroom subdataset. The area near the front of the speaker emits the loudest sound across most frequencies, as expected. The figures also confirm that higher frequencies are more directionally emitted than lower ones, evident in the narrowing yellow directivity "beam" with increasing frequency. Additionally, the fact that higher frequencies are typically emitted by the loudspeaker's tweeter at the top front of the speaker, is reflected in our heatmaps, where the yellow regions appear above the speaker's center for higher frequencies.

Reflection Amplitude Responses. The right side of Figure 3 shows the specular reflection amplitude responses that

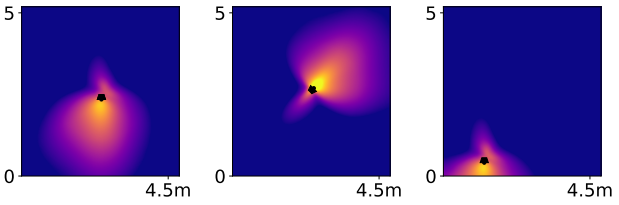
	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
NN	5.99	1.10	2.95	1.42	1.36	0.61	1.99	1.36	10.14	3.04	2.62	1.32	5.52	0.99	2.39	1.42
Linear	6.44	1.52	3.34	1.82	1.55	0.652	2.43	1.66	11.63	4.49	3.11	1.75	6.03	1.43	2.74	1.74
DeepIR	9.23	2.81	3.15	1.65	3.09	3.41	3.39	2.22	15.71	10.34	2.97	1.47	8.08	2.80	2.62	1.65
NAF	6.36	1.38	3.32	1.75	2.00	0.73	3.38	1.54	12.26	3.82	3.13	1.46	6.10	1.31	2.87	1.71
INRAS	9.99	4.52	4.45	1.75	4.20	2.48	6.22	5.35	14.52	9.19	3.70	1.58	9.02	2.58	3.61	1.66
DIFFRIR (ours)	**5.22**	**0.94**	**2.71**	**1.36**	**1.21**	**0.56**	**1.59**	**1.19**	**9.13**	**2.95**	**2.59**	**1.25**	**4.86**	**0.92**	**2.25**	**1.41**

Table 2. Experimental results on the task of predicting monaural RIRs and music at an unseen point. Lower is better for all metrics. Errors for RIRs are multiplied by 10.

	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
DIFFRIR	**5.22**	**0.94**	**2.71**	**1.36**	**1.21**	**0.56**	**1.59**	**1.19**	**9.13**	**2.95**	**2.59**	**1.25**	**4.86**	**0.92**	**2.25**	**1.41**
w/o Directivity Pattern	5.47	0.97	3.02	1.49	1.64	0.63	3.02	1.54	9.98	3.09	2.98	1.34	5.13	0.94	2.45	1.46
w/o Source IR	5.39	0.99	2.79	1.48	1.36	0.63	1.73	1.45	9.38	3.04	2.76	1.38	5.07	0.96	2.38	1.49
w/o Residual Component	6.90	1.37	3.07	1.40	1.37	0.61	1.77	1.38	15.49	4.80	2.81	1.27	6.24	1.30	2.46	1.47

Table 3. Ablation results. In each row, the ablated parameter is frozen to its initial value during training, i.e., the Source IR is assumed to be an ideal impulse, the Directivity Pattern is assumed to be uniform at all frequencies, and the Residual Component is assumed to be zero.

our model fits to some surfaces in the Classroom and Dampened Room. Our model correctly infers that the carpeted floor seems to be more absorptive than the wall, which consists of more rigid and smooth materials. The wall in the Dampened Room is even more absorptive, as our model predicts nearly no reflection above 2 kHz.



(a) Base Model (b) Virtual Rotation (c) Virtual Translation

Figure 4. RIR loudness heatmaps generated from DIFFRIR trained on 12 points in the Dampened Room's base subdataset.

Virtual Rotation and Translation. Since our model learns physically interpretable parameters, we can simulate changes to the room layout that are unseen in the training data. In Figure 4, we train our model on the Dampened subdataset, and use it to simulate virtual speaker rotation and translation. We visualize these changes by plotting RIR loudness heatmaps. Since the DIFFRIR Dataset also includes real data where the speaker is rotated or translated, we include quantitative evaluations on virtual speaker rotation and translation in the Appendix C.3, as well as evaluations on virtual panel insertion and relocation.

5.3. Ablation Study

We ablate three major components of our model (the residual, modeling the source's directivity, and modeling the source's impulse response) to determine their individual contributions. Table 3 shows our results. The results suggest that these components are all necessary for effectively rendering accurate RIRs at novel locations. More ablations experiments are in Appendix D.4.

5.4. Additional Experiments and Visualizations.

Along with additional RIR loudness maps, Appendix B.2 shows that our model can reconstruct the modal structure of the soundfield at a low frequency. In Appendix D.2, we show that our model trained on 6 points outperforms all baselines trained on 100 points. Appendix D shows that our model is robust to geometric distortions and experiments with modeling the effects of transmission.

6. Conclusions

We presented DIFFRIR, a differentiable RIR renderer capable of accurately rendering the room's acoustic impulse response at new locations, given a small set of microphone recordings and the room geometry. Future work could focus on modeling a room's acoustics implicitly by recording natural audio, thus obviating the need to measure RIRs.

Acknowledgments. The work is in part supported by NSF CCRI #2120095, RI #2211258, RI #2338203, ONR MURI N00014-22-1-2740, Adobe, Amazon, and Sony.

References

[1] Byeongjoo Ahn, Karren Yang, Brian Hamilton, Jonathan Sheaffer, Anurag Ranjan, Miguel Sarabia, Oncel Tuzel, and Jen-Hao Rick Chang. Novel-view acoustic synthesis from 3d reconstructed rooms, 2023. 2

[2] Thibaut Ajdler, Luciano Sbaiz, and Martin Vetterli. The plenacoustic function and its sampling. *IEEE TIP*, 54(10): 3790–3804, 2006. 13

[3] 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. 5

[4] Adil Alpkocak and Kemal Sis. Computing impulse response of room acoustics using the ray-tracing method in time domain. *Archives of Acoustics*, 35, 2010. 2

[5] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11):2029, 2018. 6, 17

[6] Shai Avidan and Amnon Shashua. Novel view synthesis in tensor space. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1034–1040. IEEE, 1997. 1

[7] Per Bak, Chao Tang, and Kurt Wiesenfeld. An explanation of 1/f noise. *Physical Review Letters*, 59:381–384, 1987. 20

[8] Alexis Benichoux, Laurent Simon, Emmanuel Vincent, and Remi Gribonval. Convex regularizations for the simultaneous recording of room impulse responses. *Signal Processing, IEEE Transactions on*, 62:1976–1986, 2014. 4

[9] Boualem Boashash. *Time-frequency signal analysis and processing: a comprehensive reference*. Academic press, 2015. 7

[10] S. Butterworth. On the Theory of Filter Amplifiers. *Experimental Wireless & the Wireless Engineer*, 7:536–541, 1930. 14

[11] Diego Caviedes-Nozal, Nicolai A.B. Riis, Franz M. Heuchel, Jonas Brunskog, Peter Gerstoft, and Efren Fernandez-Grande. Gaussian processes for sound field reconstruction. *Journal of the Acoustical Society of America*, 149(2):1107–1119, 2021. Funding Information: The authors would like to thank Manuel Hahmann for the fruitful discussions. This work is part of the MONICA project and has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 732350. It is partly supported by the VILLUM foundation (Grant No. 19179, "Large scale acoustic holography"). Publisher Copyright: © 2021 Acoustical Society of America. 13

[12] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 3

[13] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020. 22

[14] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 2

[15] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022 Datasets and Benchmarks Track*, 2022. 3

[16] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *CVPR*, pages 6409–6419, 2023. 2, 7, 17, 18

[17] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[18] Ziyang Chen, David F Fouhey, and Andrew Owens. Sound localization by self-supervised time delay estimation. *European Conference on Computer Vision (ECCV)*, 2022. 20

[19] Mandar Chitre. Differentiable ocean acoustic propagation modeling. In *OCEANS 2023-Limerick*, pages 1–8. IEEE, 2023. 3

[20] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *Conference on Robot Learning*, pages 662–673. PMLR, 2022. 3, 6, 7, 20

[21] Orchisama Das, Paul Calamia, and Sebastia V. Amengual Gari. Room impulse response interpolation from a sparse set of measurements using a modal architecture. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, 2021. 13

[22] Gary Davis and Ralph Jones. *The Sound Reinforcement Handbook*. Hal Leonard, 1987. 20

[23] Simona De Cesaris, Dario D'Orazio, Federica Morandi, and Massimo Garai. Extraction of the envelope from impulse responses using pre-processed energy detection for early decay estimation. *The Journal of the Acoustical Society of America*, 138(4):2513–2523, 2015. 7

[24] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016. 6, 21

[25] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020. 3, 6, 7, 20

[26] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 3

[27] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. *International Journal of Computer Vision*, pages 1–15, 2023. 2

[28] Michele Geronazzo, Erik Sikström, Jari Kleimola, Federico Avanzini, Amalia De Götzen, and Stefania Serafin. The impact of an accurate vertical localization with hrtfs on short explorations of immersive virtual reality scenarios. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 90–97. IEEE, 2018. 6

[29] Georg Götz, Sebastian J Schlecht, and Ville Pulkki. A dataset of higher-order ambisonic room impulse responses and 3d models measured in a room with varying furniture. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–8. IEEE, 2021. 6

[30] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. 21

[31] Brian Hamilton. Pffdtd software, 2021. https://github.com/bsxfun/pffdtd. 3

[32] D. P. Hardin, T. J. Michaels, and E. B. Saff. A comparison of popular point configurations on $\mathbb{S}^2$, 2016. 4

[33] Laszlo Hars. Frequency response compensation with dsp. *Signal Processing Magazine, IEEE*, 20:91– 95, 2003. 4

[34] Sahar Hashemgeloogerdi and Mark Bocko. Invertibility of acoustic systems: An intuitive physics-based model of minimum phase behavior. page 055002, 2015. 20

[35] Jun-Hyeok Heo, Deok-Ki Kim, and Byoung-Duk Lim. Application of minimum phase condition to the acoustic reflection coefficient measurement. *Transactions of the Korean Society for Noise and Vibration Engineering*, 15, 2005. 20

[36] M. S. Howe. *Introduction*, page 1–24. Cambridge University Press, 2002. 24

[37] Cheol-Ho Jeong. Diffuse sound field: challenges and misconceptions. *Proceedings of 45th International Congress and Exposition on Noise Control Engineering*, pages 1015–1021, 2016. 5

[38] Ole Kirkeby and Philip A. Nelson. Reproduction of plane wave sound fields. *The Journal of the Acoustical Society of America*, 94(5):2992–3000, 1993. 13

[39] V.D. Landon. A study of the characteristics of noise. *Proceedings of the Institute of Radio Engineers*, 24(11):1514–1521, 1936. 4

[40] Eric A. Lehmann and Anders M. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *The Journal of the Acoustical Society of America*, 124(1):269–277, 2008. 7

[41] Yan Li, Peter F. Driessen, George Tzanetakis, and Steve Bellamy. Spatial sound rendering using measured room impulse responses. In *2006 IEEE International Symposium on Signal Processing and Information Technology*, pages 432–437, 2006. 2

[42] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis, 2023. 2

[43] 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. 1, 2, 7, 12, 21

[44] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, page 3. Atlanta, GA, 2013. 21

[45] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 35:2522–2536, 2022. 2

[46] J. Gregory McDaniel and Cory L. Clarke. Interpretation and identification of minimum phase reflection coefficients. *The Journal of the Acoustical Society of America*, 110(6):3003–3010, 2001. 5

[47] Thomas McKenzie, Leo McCormack, and Christoph Hold. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis. *arXiv preprint arXiv:2111.11882*, 2021. 6

[48] 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 *ECCV*, 2020. 1, 21, 23

[49] Henrik Møller and Christian Sejer Pedersen. Hearing at low and infrasonic frequencies. *Noise & health*, 6(23):37–57, 2004. 23

[50] Mélanie Nolan, Marco Berzborn, and Efren Fernandez-Grande. Isotropy in decaying reverberant sound fieldsa). *The Journal of the Acoustical Society of America*, 148(2):1077–1088, 2020. 5

[51] Beth Paxton. Room acoustics, sixth ed., heinrich kuttruff. crc press (2017). isbn: 978-1-4822-6043-4. *Applied Acoustics*, 126:90–91, 2017. 5

[52] Christoph Pörschmann and Johannes M Arend. Analyzing the directivity patterns of human speakers. *Proceedings of the 46th DAGA*, pages 16–19, 2020. 3

[53] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 21

[54] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. IR-GAN: Room Impulse Response Generator for Far-Field Speech Recognition. In *Proc. Interspeech 2021*, pages 286–290, 2021. 1, 2

[55] Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 924–933, 2022. 3

[56] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *ICASSP*, pages 3209–3213. IEEE, 2022. 1, 2, 7, 21

[57] Jens Holger Rindel. Modal energy analysis of nearly rectangular rooms at low frequencies. *Acta Acustica united with Acustica*, 101(6):1211–1221, 2015. 5

[58] Furniaki Satoh, Yoshito Hidaka, and Hideki Tachibana. Reverberation time directly obtained from squared impulse response envelope. In *Proc. Int. Congr. Acoust*, pages 2755–2756, 1998. 7

[59] Robin Scheibler, Eric Bezzam, and Ivan Dokmanic. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018. 19

[60] Julius O. Smith. *Physical Audio Signal Processing*. https://ccrma.stanford.edu/~jos/pasp/, accessed 2023. online book, 2010 edition. 5

[61] Julius O. Smith. *Spectral Audio Signal Processing*. https://ccrma.stanford.edu/~jos/sasp/, accessed ¡date¿. online book, 2011 edition. 20

[62] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 3

[63] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 7, 12, 22

[64] Chiara Visentin, Matteo Pellegatti, and Nicola Prodi. Effect of a single lateral diffuse reflection on spatial percepts and speech intelligibility. *The Journal of the Acoustical Society of America*, 148(1):122–140, 2020. 5

[65] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. In *Advances in Neural Informaion Processing Systems*, 2023. 6

[66] Shu-Nung Yao. Headphone-based immersive audio for virtual reality headsets. *IEEE Transactions on Consumer Electronics*, 63(3):300–308, 2017. 6

[67] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[68] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 286–301. Springer, 2016. 1

Hearing Anything Anywhere

– Supplementary Material –

Website: masonlwang.com/hearinganythinganywhere

Code: github.com/maswang32/hearinganythinganywhere

Dataset: zenodo.org/records/11195833

Contents

A. Qualitative Results and Video

Please see the supplementary video on the website for an in-depth qualitative analysis and comparative evaluation against baseline models. This video showcases a simulation of a song played in two distinct environments: the Dampened Room and the Hallway. The purpose is to demonstrate the immersive quality and perceptual accuracy of the audio rendered by our model, reflecting the true characteristics of the real scenes. To achieve this, we rendered 100 room impulse responses at various locations, convolved them with the chosen source audio, and smoothly interpolated between these convolved signals. For an optimal experience of these qualitative results, we recommend using earbuds or headphones while viewing the video.

Furthermore, the video features a side-by-side comparison of our binaural audio results with those from baseline models, highlighting the enhanced realism and compelling nature of the audio generated by our model. This comparison underscores the significant qualitative improvements our model offers in creating an immersive auditory experience. In addition, the video provides visualizations explaining our method, and the task setup.

B. RIR Heatmap Visualizations

B.1. Broadband RIR Heatmaps

After our model is trained, we can use it to visualize how the loudness of the rendered acoustic field varies spatially. To do this, we use the model trained on each of the base subdatasets to render RIRs on a dense 2D-grid of listener locations. We visualize of the root mean square (RMS) volume level of the RIRs in Figure 5, on a decibel (logarithmic) color scale. The visualizations shown are similar to those in [43, 63].

We observe several differences in the heatmaps for the different rooms. In the Dampened Room, the surfaces are less reflective, and thus, much of the soundfield's loudness is concentrated in the region in front of the speaker. This effect is reduced in the Classroom, where the soundfield is more spread out. In the Hallway, which is the most reflective room, the soundfield's volume is even more spread out, and the region behind the speaker is significantly louder than it is in any of the other rooms.

B.2. Soundfield Reconstruction

When observed at a single frequency, the spatial variations in sound pressure for a given sound field often exhibit modal patterns. Reconstructing the pressure levels of a sound field from a sparse set of observations is a problem of longstand-

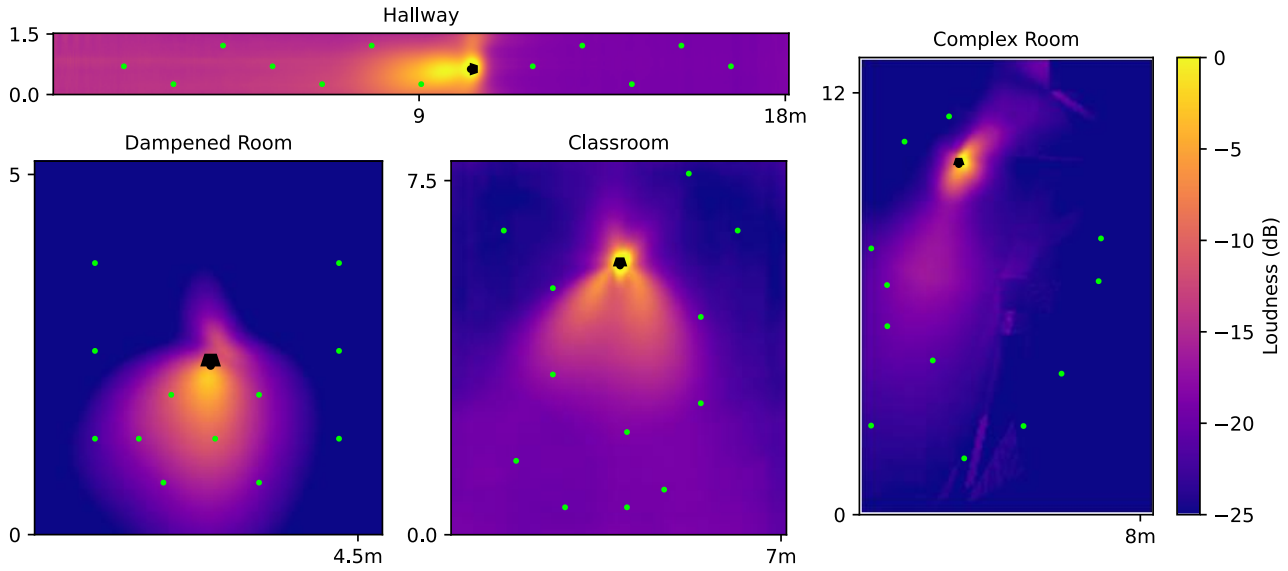


Figure 5. Visualization of RIR loudness maps generated from our model trained in each of the four base subdatasets. We measure loudness by rendering an RIR at a given listener location and measuring its RMS volume level. For each RIR rendered, we fix the height of the listener location to be 1 meter above the floor. The resolution of each xy-grid is approximately 5 centimeters in both the x and y directions. We fix the location and orientation of the speaker (indicated by the black icon) to where it was during RIR measurement. The color scale is in decibels and is consistent between rooms. The green dots indicate the xy locations of the 12 training points, which are projected onto the $z = 1$ plane.

ing theoretical and practical interest [2, 11, 21, 38]. Using the RIRs measured in the Classroom subdataset, we calculate the sound pressure level at 70 Hz at all locations in our subdataset, plotted in Figure 6a. We also use the predicted RIRs from each method to predict the sound pressure level at 70 Hz at every spatial location. We find that our model learns to predict the modal structure of the RIR sound field without explicitly modeling it, while other baselines fail to do this. Note that our model approximately predicts the locations of the sound field's nodes and anti-nodes (regions of high and low intensity), even without observing training data in those locations.

C. Results on Additional Room Configurations

C.1. Description of Additional Subdatasets

In addition to the base subdatasets collected in each of the four rooms, we collect additional data in different room configurations, where we vary the location of the speaker, the orientation of the speaker, or the presence and number of rectangular whiteboard panels. We collect this additional data for two reasons:

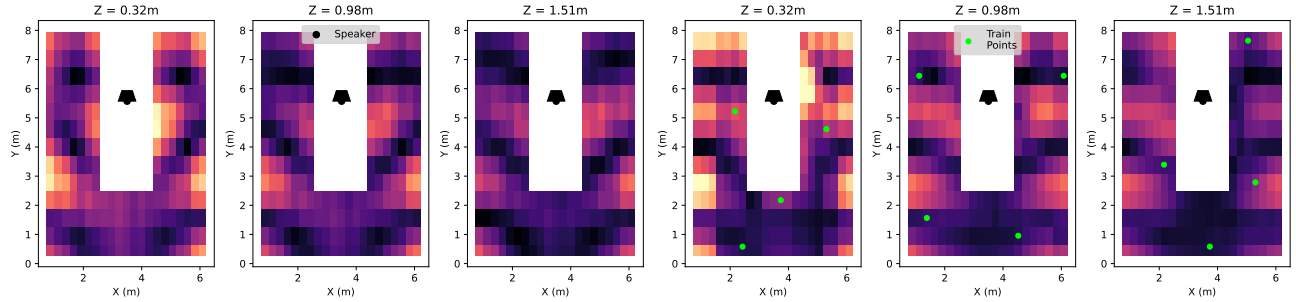
- To test our method's effectiveness on various room layouts, including those where the speaker is occluded.
- To evaluate acoustic interpolation methods on the task of zero-shot generalization to changes in room layouts, by virtually simulating speaker rotation and translation, and panel relocation and insertion.

The locations and orientations of the speakers as well as the positions of the panel(s), are provided as part of the DIFFRIR Dataset. Photographs of each additional configuration are shown in Figure 7.

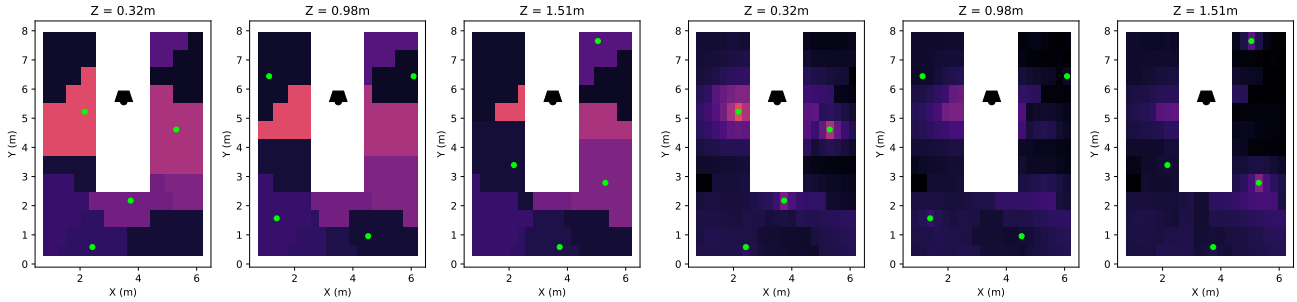
Rotation Subdatasets. In the Dampened Room, Hallway, and Complex Room, we collected 120, 72, and 132 additional datapoints where the speaker was rotated by 225°, 90°, and 90°clockwise, respectively. The location of the speaker and all surfaces otherwise remain the same.

Translation Subdatasets. In the Dampened Room, Hallway, and Complex Room, we collected 120, 72, and 132 additional datapoints where the speaker was translated to another part of the room, but the orientation of the speaker is was kept the same. In the Dampened Room, we move the speaker such that it is near one corner of the room and facing a wall. In the Hallway, we move the speaker to the far end of the Hallway, such that the speaker faces the entire length of the Hallway. The Complex Room is roughly divided into two halves by the table and pillars in the middle of the room. In the Complex Room, we collect additional datapoints where the speaker is translated from the one half to the other.

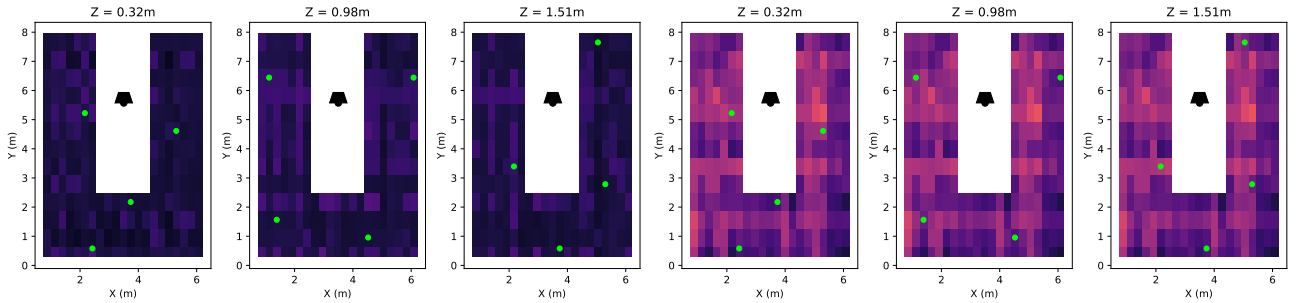
Panel Subdatasets. In the Dampened Room and Hallway, we place 1-2 whiteboard panels in the room. In the Dampened Panel subdataset, we place the panel directly in front



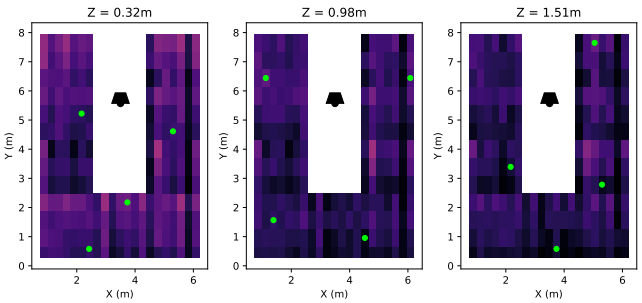
(a) GT (b) Ours



(c) Nearest Neighbors (d) Linear



(e) NAF (f) INRAS



(g) DeepIR

Figure 6. Visualization of RIR loudness at 70 Hz in the Classroom subdataset. The sound field intensity at a given location is measured by filtering the ground-truth or predicted RIR around 70 Hz using a 2nd order Butterworth filter [10] and measuring the RMS volume level of the filtered signal. Subfigure a) shows the intensity of the 70hz sound field at all locations in the subdataset. Subfigure b) shows predicted intensities at these same locations using our model trained on 12 points. We indicate the spatial locations of these 12 training points with green dots, and the speaker's location and orientation with a black icon. Subfigures c) through g) show the sound field intensity as predicted by each of our baseline models. Note that in subfigure d), the Linear baseline underestimates the soundfield intensity at locations far away from the training locations, since the linear interpolation at these locations is a weighted average of roughly uncorrelated signals whose mean is roughly zero.

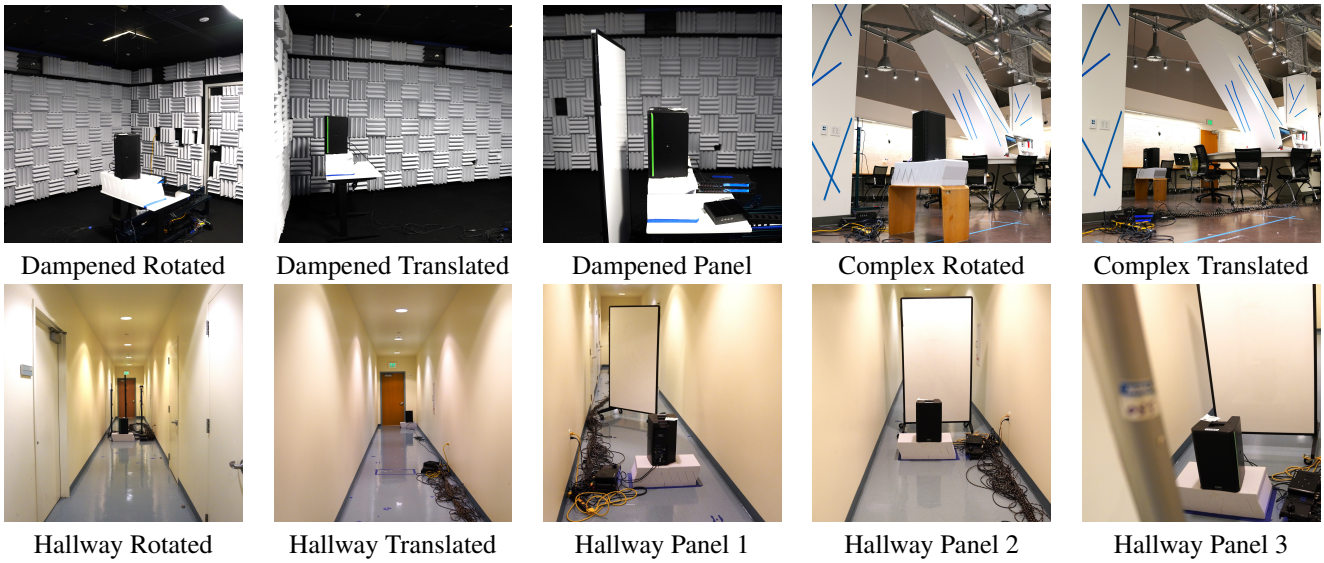


Figure 7. Photographs of all additional configurations in the DIFFRIR Dataset. Note that the Hallway Panel 1 photo is taken from behind the speaker.

of the speaker. In the Hallway subdataset, there are three panel configurations. In Hallway Panel 1, we place one whiteboard panel in front of the speaker at a slanted angle. In Hallway Panel 2, we place one whiteboard panel directly behind the speaker. In Hallway Panel 3, we place whiteboard panels both in front of and behind the speaker.

C.2. Evaluations on Configurations

We evaluate our model on each of these configurations independently in Table 4, training and testing on the same subdataset. For each configuration, we select 12 training points from each of the subdatasets, and evaluate our rendered RIRs on a test set of held-out data.

C.3. Quantitative Results on Virtual Room Layout Modifications

Since our model learns physically interpretable parameters for the speaker's directivity, we expect to be able to virtually simulate rotations or translations of the speaker that are unobserved in the training data. We simulate rotating the speaker by rotating the speaker's learned directivity map, and translation by moving the speaker's estimated location during path-tracing.

These predicted changes in the speaker's location or orientation can be evaluated against real data, since the DIFFRIR Dataset contains additional configurations that modify the base subdataset in each room by moving or rotating the speaker.

The quantitative results in Tab. 5, 6, 7, and 8 show the usefulness of the DIFFRIR Dataset in benchmarking the performance of methods of virtual room layout modification. Future work can use the DIFFRIR Dataset to improve the performance of these tasks.

	RIR		Music	
Room/Configuration	Mag	ENV	Mag	ENV
Dampened	1.21	0.56	1.59	1.19
w/ Rotated Speaker	1.14	0.44	1.49	1.36
w/ Translated Speaker	0.68	0.39	0.91	1.18
w/ Panel	1.23	0.60	1.62	1.47
Hallway	9.13	2.95	2.59	1.25
w/ Rotated Speaker	8.40	2.86	2.58	1.27
w/ Translated Speaker	8.91	3.02	2.84	1.25
Panel Config. 1	8.47	2.99	2.58	1.32
Panel Config. 2	8.52	3.61	2.63	1.36
Panel Config. 3	8.39	2.94	2.67	1.35
Complex	4.86	0.92	2.25	1.41
w/ Rotated Speaker	4.33	0.83	2.13	1.41
w/ Translated Speaker	4.38	1.19	2.22	1.44

Table 4. DIFFRIR's performance on additional configurations in the DIFFRIR Dataset, on the task of predicting monaural RIRs and music at an unseen point. Lower is better for all metrics. Errors for RIRs are multiplied by 10. Each DIFFRIR model is trained on 12 points.

Virtual Speaker Rotation. As an experiment, we take the DIFFRIR model trained on each base subdataset with a corresponding rotated subdataset, virtually rotate the speaker by rotating the learned directivity heatmap, and predict RIRs and music at locations in each of the corresponding rotated subdatasets. We evaluate these predictions

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Dampened w/ Rotation				
Trained on Rot. Data	**1.14**	**0.44**	**1.49**	**1.36**
Trained on Base w/ Virt. Rot.	1.39	0.51	1.88	1.48
Hallway w/ Rotation				
Trained on Rot. Data	**8.40**	**2.86**	**2.58**	**1.27**
Trained on Base w/ Virt. Rot.	9.83	3.22	2.88	2.50
Complex w/ Rotation				
Trained on Rot. Data	**4.33**	**0.83**	**2.13**	**1.41**
Trained on Base w/ Virt. Rot.	4.84	0.89	2.27	1.59

Table 5. Results on Virtual Speaker Rotation. Evaluations are done on the test set of the rotated subdataset.

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Damp. → Hall. 1.				
Hall. 1 Model	**8.47**	**2.99**	**2.58**	**1.32**
Virtual Insertion	9.32	2.96	2.69	1.33
Damp. → Hall. 2.				
Hall. 2 Model	**8.52**	**3.61**	**2.63**	**1.36**
Virtual Insertion	9.31	3.45	2.62	1.38
Hall. 1.→ Damp.				
Damp. Panel Model	**1.23**	**0.600**	**1.62**	**1.47**
Virtual Insertion	1.84	0.660	3.70	1.56
Hall. 2.→ Damp.				
Damp. Panel Model	**1.23**	**0.600**	**1.62**	**1.47**
Virtual Insertion	1.84	0.660	3.70	1.56

Table 6. Results on Virtual Panel Insertion. 'Damp.→Hall 1.' means that take the DIFFRIR model from the Hallway Base subdataset (no panel). Then, we virtually insert a panel to simulate the Hallway Panel 1 subdataset, by borrowing the reflection coefficients of the panel from the DIFFRIR model trained on the Dampened w/ Panel subdataset. We then evaluate the virtual insertion on the recordings from the Hallway Panel 1 subdataset. As a baseline, we compare to a model that is trained on the same panel subdataset that it is tested on.

against ground-truth RIRs and music recordings from the rotated subdatasets. In addition, we compare our virtual rotation with the performance of the DIFFRIR model both trained and tested on the rotated subdatasets. The results are shown in Table 5. Although the model both trained and tested on the rotated subdatasets outperforms our virtually-rotated model, the results are quite close in the Dampened and Complex Rooms. The results in the Hallway are worse, perhaps because the Hallway's narrow nature means that the set of direct paths from the speaker to the training locations cover a narrow range of outgoing angles.

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Hall. 1.→ Hall. 2.				
Baseline	**8.52**	**3.61**	**2.63**	**1.36**
Virtual Panel Relocation	8.91	3.59	2.71	1.39
Hall. 2.→ Hall 1.				
Baseline	**8.47**	**2.99**	**2.58**	**1.32**
Virtual Panel Relocation	8.89	3.13	2.72	1.39

Table 7. Results on Virtual Panel Relocation. 'Hall 1.→ Hall 2.' means that take the DIFFRIR model from the Hallway Panel 1 subdataset (no panel). Then, we virtually move this panel to its location in the Hallway Panel 2 subdataset. We then evaluate on the recordings from the Hallway Panel 2 subdataset.

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Dampened w/ Translation				
Trained on Trans. Data	**0.68**	**0.39**	**0.91**	**1.18**
Trained on Base w/ Virt. Trans.	1.22	0.53	1.26	1.61
Hallway w/ Translation				
Trained on Trans. Data	**8.91**	**3.02**	**2.84**	**1.25**
Trained on Base w/ Virt. Trans.	9.28	3.05	**2.84**	1.28
Complex w/ Translation				
Trained on Trans. Data	**4.38**	**1.19**	**2.22**	**1.44**
Trained on Base w/ Virt. Trans.	4.79	**1.19**	2.24	1.54

Table 8. Results on Virtual Speaker Translation. Evaluations are done on the test set of the translated subdataset.

Virtual Speaker Translation. We perform a similar experiment with virtual speaker translation, evaluating against ground-truth recordings from the corresponding subdataset. The results are shown in Table 8.

Virtual Panel Relocation. We would like to see if we can learn the reflective characteristics of a surface in one room, then 'virtually move' the surface to another location in the same room. In the Hallway, we collect two subdatasets (Hallway Panel 1 and Hallway Panel 2 in Figure 7), where the room layouts are identical except for the location and orientation of a single whiteboard panel. In our experiments, we train on the first panel configuration, then move the location of the whiteboard panel to that of the second configuration before performing inference. We then evaluate our predicted audio against ground-truth audio from the second configuration. Results are shown in Table 7. The baseline shown is one where we train on the same subdataset that we evaluate on.

Virtual Panel Insertion. We would like to see if we can learn the reflective characteristics of a surface in one room,

then 'virtually insert' the surface into another room. Three of our base subdatasets also include a version with a single inserted whiteboard panel. In each of our four experiments, we take the base subdataset (e.g., the Dampened Base subdataset), and the coefficients learned for the whiteboard panel from another room (e.g., the Hallway Panel Config. 1 subdataset). We then virtually insert the whiteboard panel into the base subdataset, and evaluate the virtual insertion against the version of the base dataset with a panel in it (e.g., the Dampened Panel subdataset). Results are shown in Table 6. The baseline shown is one where we train on the same subdataset that we evaluate on.

D. Additional Experiments and Ablations

D.1. Results on Binaural Rendering

We evaluate our method on the task of rendering a binaural RIR at an unseen location. We collect binaural RIRs at several locations in all rooms using our 3Dio binaural microphone, and compare these to predicted RIRs that we binauralize from single-channel audio as described in the Methods section.

We compare our binauralized audio with the ground-truth audio using the left-right energy ratio error between the ground-truth and predicted recordings, which is used in [16]. To compute the left-right energy ratio, we compute compute the ratio of total energy between the left and right channels of the RIR or music recordings. We then compute the mean squared error between the left-right energy ratio of the predicted and ground-truth RIRs or music. Results are shown in Table 9.

Since the baselines do not have a way of generating binaural RIRs from monaural ones, we binauralize these baselines by rendering two monaural RIRs at the locations of the left and right ears of the 3Dio microphone, and combining them into left and right channels.

Our method outperforms our baselines across most metrics. Note that it is difficult to compare a binaural RIR recorded from our binaural microphone with binauralized audio originally recorded from a different microphone. Our rendered binaural audio will have characteristics of the monaural microphone and the microphones used in the SADIE dataset [5] used to record the HRIRs that we convolve our monaural recordings with. The binaural recordings in our dataset will h ave different characteristics, since they are recorded using a different microphone with different spectral characteristics and directionality. Because of this, we include qualitative binauralization results in the supplementary video.

D.2. Performance vs Number of Training Points

We conduct an ablation study with varying numbers of training points N on each subdataset and compare against the baselines. As shown in Figure 8, the performance increases with N, and our model consistently outperforms the baselines when $N \geq 2$. Note that in all rooms, our model trained on only 6 locations outperforms all baselines trained on 100.

Note that our model's hyperparameters are optimized for performance in data-limited scenarios. When the number of training points is higher, it is possible that increasing the number of parameters (for instance, increasing the resolution of the heatmap or the number of reflection coefficients) leads to even better performance.

D.3. Robustness to Inaccurate Geometry.

Our method requires measuring the room's geometry. In our dataset, we do this using a tape measure or laser distance measure, which both provide sufficiently accurate measurements. In order to explore the effect of inaccurate geometric measurements, we conduct an additional experiment to measure the performance after adding random artificial distortions to the surfaces. In the Classroom, we select 8 random directions to move each of the 11 vertices defining the walls, ceiling, floors and the corners of the tables that are exposed. We move each vertex by 0-2 meters in its corresponding random direction. Results are shown in Figure 9. Observe that unless we distort *all* vertices in the room by over 1.5 meters, our model outperforms the best baseline (Nearest Neighbors). We conclude that our method is robust to geometric distortion.

Geometric distortion can affect our model's rendering in one of three ways: It can change the distance of reflection paths, which affects its time-of-arrival and amplitude; it can eliminate reflection paths, or it can add new reflection paths. Since our model is optimized against a frequency-domain loss whose smallest window size is 256 samples (or 1.8 meters at the speed of sound), our model is robust to perturbations in times-of-arrival.

D.4. More Ablations

In the Methods section, Section E.4, and Section E.3, we discuss several minor components of our model (axial boosting, time-of-arrival perturbation, hop size 1 loss, etc.) that provide a boost to our model's performance and/or robustness. Results with each of these components ablated are in Table 10. Our model performs the best on a plurality of evaluations, proving that these performance boosts are good on balance. However, we should also observe that even in evaluations where our model does not perform the best, it is never worse than the best performing ablation by a significant margin. We cannot say the same for the Interpolation Spline ablation, which also performs the best in the same number of evaluations (six), but significantly underperforms our model in several settings.

	Classroom		Dampened Room		Hallway		Complex Room	
	RIR	Music	RIR	Music	RIR	Music	RIR	Music
NN	1.27	0.516	5.64	2.57	0.062	0.034	0.345	0.166
Linear	1.29	0.531	5.48	2.09	**0.045**	**0.008**	0.335	**0.157**
DeepIR	1.10	0.529	6.20	5.90	0.048	0.036	0.350	0.397
NAF	1.93	0.743	5.93	2.37	0.108	0.012	0.320	0.176
INRAS	1.25	0.383	5.86	4.35	1.60	4.41	0.332	0.183
DIFFRIR (ours)	**0.43**	**0.091**	**2.94**	**0.316**	0.097	0.012	**0.287**	0.288

Table 9. Experimental results from the task of predicting binaural RIRs and music at an unseen point from a model trained on 12 monoaural RIRs. We use the left-right energy ratio error metric [16]. Lower is better. All errors are multiplied by 10.

	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
DIFFRIR	5.22	**0.942**	2.71	**1.36**	**1.21**	**0.555**	1.59	1.19	**9.13**	2.95	2.59	1.25	**4.86**	0.917	2.25	**1.41**
w/o Time-of-Arrival Perturbation	**5.19**	0.962	**2.70**	1.43	1.23	0.582	1.61	1.36	**9.13**	2.93	2.60	1.27	**4.86**	**0.913**	2.23	1.42
w/o Axial Boosting	**5.19**	0.969	2.71	1.43	1.22	**0.555**	1.59	1.20	9.14	2.95	2.59	1.30	**4.86**	0.934	2.25	1.44
w/o Hop Size 1 Loss	5.26	0.988	2.74	1.41	1.25	0.559	1.67	**1.16**	9.22	2.98	2.60	**1.24**	4.90	0.962	2.27	1.42
w/o Interpolation Spline	5.60	0.973	2.72	1.41	1.63	0.565	**1.53**	**1.16**	9.47	**2.92**	**2.56**	**1.24**	5.24	0.920	**2.21**	1.42

Table 10. Ablation results from the task of predicting monaural RIRs and music at an unseen point. In the Interpolation Spline ablation, the *Residual Component* and the contributions from explicitly computed reflection paths are simply added together, instead of being blended using the learned temporal spline γ. Lower is better for all metrics. Errors for RIRs are multiplied by 10.

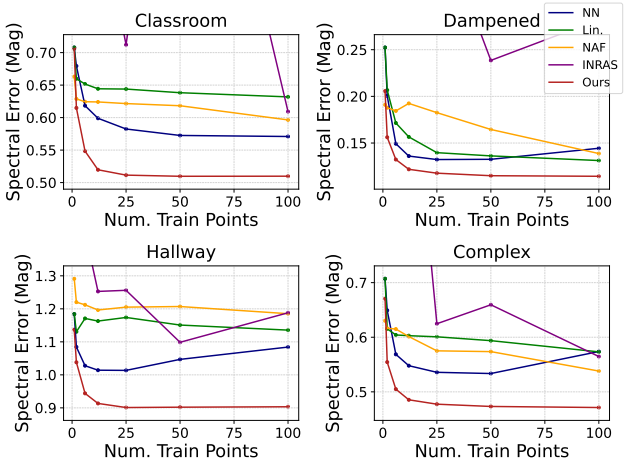


Figure 8. Evaluations of our method and baselines with different numbers of training points. We use the Multiscale Log-Spectral L1 Loss (Mag), and train with $N \in \{1, 2, 6, 12, 25, 100\}$. All training locations are selected as nested subsets of one another, and we evaluate on a fixed test set. Note that the DeepIR baseline's error was too large to fit into the range of the plot.

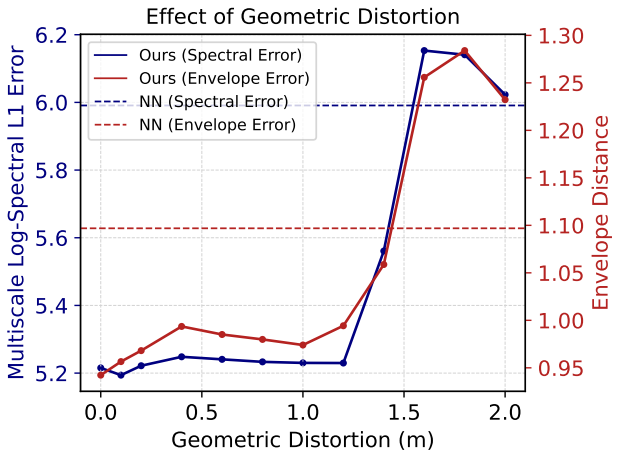


Figure 9. Effect of geometric distortion on RIR prediction performance in the Classroom subdataset The blue line shows our model's performance according to the Multiscale Log-Spectral L1 metric, and the red line shows our model's perfomance according to the envelope distance metric. The red and blue dashed lines indicate the performance of the nearest-neighbors baseline according to the multiscale log-spectral L1 metric and the envelope distance metric, respectively.

D.5. Modeling the Effects of Transmissions

Our model assumes that sound energy encountering a surface is either reflected or absorbed by the surface. This is for the sake of simplicity. We also conduct an experiment in which we consider surface transmission as well. This means that we modify our tracing algorithm to consider reflection paths that pass through surfaces, and assume that a proportion of the sound energy at each frequency can be *transmitted* through these surfaces in a frequency-dependent manner. Our modified training procedure then fits *surface transmission coefficients* in a manner identical to the way it fits

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Classroom				
DIFFRIR (ours)	**5.22**	**0.942**	**2.71**	**1.36**
w/ Transmission	5.23	0.951	2.72	**1.36**
Dampened Room w/ Panel				
DIFFRIR (ours)	**1.23**	**0.604**	**1.62**	1.47
w/ Transmission	**1.23**	**0.604**	**1.62**	**1.45**
Hallway w/ Panels				
DIFFRIR (ours)	8.39	2.94	2.67	1.35
w/ Transmission	**8.38**	**2.92**	**2.64**	**1.34**
Complex Room				
DIFFRIR (ours)	**4.86**	0.917	2.25	1.41
w/ Transmission	**4.86**	**0.915**	**2.24**	**1.38**

Table 11. Evaluations of DIFFRIR vs DIFFRIR with Transmission modeling. Lower is better for all metrics, and RIR errors are multiplied by 10.

surface reflection coefficients. Table 11 contains quantitative results from a model that models both transmission and reflection, and shows that in our settings, modeling transmission is not necessary. However, in other rooms with surfaces of different materials, modeling transmission may be important.

D.6. Comparison to Traditional Acoustic Simulations

We compare our method to a widely-used image-source audio simulator, Pyroomacoustics [59]. For each room in our dataset, we simulate RIRs by providing the dimensions and the speaker location, and selecting the closest material coefficients for each surface from its pre-defined database (e.g., drywall, ceiling tiles, carpet). Table 12 reports the accuracy of the simulated RIRs compared to the ground truth.

Room/Configuration	RIR		Music	
	Mag	ENV	Mag	ENV
Classroom				
DIFFRIR (ours)	**5.22**	**0.942**	**2.71**	**1.36**
Pyroomacoustics	18.64	3.67	3.26	1.68
Dampened Room				
DIFFRIR	**1.21**	**0.555**	**1.59**	**1.19**
Pyroomacoustics	2.14	0.798	2.17	1.96
Hallway				
DIFFRIR	**9.13**	**2.95**	**2.59**	**1.25**
Pyroomacoustics	32.01	4.03	3.39	1.70

Table 12. Comparison of our model against Pyroomacoustics. Lower is better for all metrics, and RIR errors are multiplied by 10.

E. Method Details

E.1. Details on Source Localization

Our method does not require a ground-truth source location measurement. Instead, we use a simple time-of-arrival technique to estimate the sound source's location to a degree of accuracy sufficient for the subsequent steps of the method. For each Room Impulse Response (RIR) in the training set, we determine its first peak, which is proportional to the distance of the direct path between the microphone and source locations. We locate the first peak of the RIR by measuring when the RIR first exceeds a quarter of its absolute maximum. We then determine the distance from the source to the target microphone by multiplying by the speed of sound, assumed to be 343 m/s.

We use a gradient descent optimization method to fit the optimal source location. We initialize the source location to the origin, which is at a corner of the room. We perform an optimization process that updates the optimal source location's position at each step. At each iteration of the optimization process, we compute the estimated times-of-arrival for each of the microphone locations, based on the current estimate for the source location. We then calculate the L1 loss between the estimated times-of-arrival and the times-of-arrival as measured by locating the first peak of the ground-truth RIR as described in the previous paragraph. We perform a gradient update on the estimated source location to minimize this L1 loss. We optimize for 1000 steps, and use the final estimate for the source location as our estimated source location.

In all of the base configurations, our method is able to predict a source location that is inside the location of our QSC loudspeaker. We used the estimated location in all configurations except for the Complex Rotation and Complex Translation configurations, where our localization method failed.

E.2. Minimum-Phase Transform

Our model learns the frequency-domain response curve for each of the surfaces in the room and for each outgoing direction from the source, allowing us to determine how the frequency profile of sound traveling along that reflection path is altered. However, this frequency-domain response is not enough to determine the reflection path's time-domain contribution to the RIR, because it contains magnitude information, but no phase information. In order to invert our reflection path's frequency profile into a time-domain signal, we need to provide phase values at each frequency, so we can perform the inverse-Fourier transform.

In our analysis, we adopt the minimum-phase assumption to calculate phase values for acoustic reflections, a

method widely recognized and justified within acoustic research [34, 35]. This assumption posits that for each frequency, the phase delay introduced by the reflection is minimal, implying that the time delay contributed by the path of reflection at any given frequency is as short as possible. From a physical standpoint, this is akin to assuming that sound is reflected off surfaces with negligible delay, thereby behaving as if the reflections are 'instantaneous' while still preserving the unique frequency-dependent characteristics of the reflection. We compute the phase values using the method described in [61].

E.3. Specific Loss Formulation

Loss Formulation and Equations. We define the loss for a given short-time Fourier transform (STFT) window size s_w and hop size h in Equation (6). This is the sum of the L1 distance between the magnitude-spectrograms of the ground-truth and synthesized RIRs and the log-magnitude spectrograms of the ground-truth and synthesized RIRs.

In the formula, W and $\hat{W}$ indicate the ground-truth and predicted RIRs, respectively. h indicates the hop length, s_w indicates the STFT window size, and S is the short-time Fourier transform, or spectrogram, whose arguments are the time-domain signal to transform, the window size, and the hop length, respectively. H indicates the hop ratio, or the hop length divided by the window size. We set $H = 0.25$.

Equation (7) provides the total loss, which sums across multiple window sizes, and adds a loss term that uses a hop size of 1.

Hop Size 1 Loss. We use a spectral loss term with hop size 1 to ensure that the early part of the RIR has accurate time-domain characteristics, since the hop length of 1 allows for high-resolution in the time domain. We take inspiration from [18] for this term, and discover it improves performance, as seen in Table 10.

Modifications from Related Work. Our multi-scale spectral plus log-spectral loss is identical to those used in [20] and [25], with two exceptions: First, is the introduction of the loss term with hop size one. Second, the minimum window size in our loss is 256, instead of 32 or 64. This is because there will be error in the time-of-arrival of certain reflection paths, due to geometric measurement error (which increases with reflection order) or errors in the speed of sound approximation. This means that the placement in time of a reflection path's contribution to our synthesized RIR may be off from its placement in the ground-truth RIR by some amount. Using larger window sizes compensates for this error, since larger windows are more likely to contain both the reflection path's contribution to our synthesized RIR and its contribution to the ground-truth RIR.

E.4. Small Efficiency and Performance Boosts

Efficiency Boosts. Since each rendered RIR combines hundreds of reflection paths, we compute all the reflection path contributions in parallel to minimize runtime. In addition, all reflection paths for the training points are precomputed before training starts.

Time-of-Arrival Perturbation. Since our measurements of each room are not necessarily precise, to make our model more robust, especially with an extremely limited number of measurements, we would like to perturb the surfaces during training. However, reflection paths for all training locations are precomputed before the training process begins. Perturbing each surface would require retracing at each iteration, which is computationally inefficient. As a proxy to this, we perturb the time of arrival of all paths by adding Gaussian noise to it, with a standard deviation of 7 samples. We found that this improved the interpretability of the estimated parameters and led to minor perfomance boosts, as shown in Table 10.

Regularization via Convolution with Pink Noise. Since RIRs are often used as a means to simulate sounds in an acoustic environment, we would like to not only ensure that our rendered RIRs are accurate, but also that the sounds we simulate via convolution with the RIR are accurate. Minimizing the spectral loss between ground-truth and predicted RIRs does not always accomplish this, since convolving the RIRs with other waveforms results in significant changes in the spectrograms.

Pink Noise is a special type of noise whose power spectral density is inversely proportional to frequency. It is ubiquitious in nature [7], and is often used as a test signal to calibrate sound systems and loudspeakers, since its frequency profile is similar to that of music [22] and other sounds the speaker might play.

To encourage our model to maintain accuracy post-convolution, we implement a regularization strategy using pink noise. For the latter half of training iterations, we convolve both our predicted and the ground-truth RIRs with five seconds of randomly generated pink noise, compute the loss between them, and add it to the loss computed between RIRs at each iteration. Convolving RIRs with pink noise simulates the speaker playing of a pink noise test signal. It is equivalent to reshaping the RIR's spectrum according to the profile of pink noise, and applying a random phase shift at each frequency.

Table 13 shows that this form of regularization results in improvements in both RIR prediction and music prediction. Such forms of regularization should be the study of future work and theoretical study.

With the goal of improving rendered music in mind, we also tried a similar form of regularization, where we convolve both our ground-truth and predicted RIRs with

$$L_{s_w,h}(W,\hat{W}) = |S(W,s_w,h) - S(\hat{W},s_w,h)| + |\log S(W,s_w,h) - \log S(\hat{W},s_w,h)| \tag{6}$$

$$L(W,\hat{W}) = \left[\sum_{s_w \in (512,\ldots 4096)} L_{s_w,Hs_w}(W,\hat{W})\right] + L_{256,1}(W,\hat{W}) \tag{7}$$

five seconds of music randomly sampled from the FMA dataset [24] at each iteration after training is halfway done. Convolution with the music files simulates the speaker playing them. Results for this form of regularization are also shown in Table 13, although we prefer the performance and simplicity of pink noise regularization.

E.5. Computational Cost

Training and Path-Tracing Time. In all of our experiments, we trained our model for 1000 epochs. In Table 14, we report the amount of time it took for our model to train on each of the base room configurations. Note that since the Complex Room is only traced up to order 4, there are substantially fewer valid reflection paths, and thus training is faster. In all other rooms, we trace up to order 5. Tracing is slower in rooms with more surfaces.

Main Contributions to Training Time. We also measured the different steps in the training process to see which ones took the longest. Each training location is associated with hundreds of reflection paths that must be added together to form the the RIR. While rendering these contributions is done in parallel, compiling them requires placing them in at the right locations in time and is done sequentially. In practice, 37.7% of the time during the 1000 epochs is spent on this compilation, 61.9% on the backwards passes, and 0.4% on everything else.

F. Baseline Implementation Details

Linear. The Linear baseline computes a RIR at a given test location by taking a linear combination of the four nearset points in the training data. The weights on each of these four training points are inversely proportional to the distance to the test location. We also experimented with taking a weighted combination of all the training data, where the weights are inversely proportional to distance. This alternative linear baseline performs quite poorly, with error increasing with the number of training points. This is because the training RIRs are roughly uncorrelated with mean zero, so the average of N RIRs tends towards zero as N increases.

Neural Acoustic Fields (NAF) [43]. To compare our method to NAF, we utilized NAF's official code,[1] as open-sourced by authors. However, in order to apply NAF to our dataset and experimental settings, we modified this code in some minor ways. Specifically, the original NAF was designed to estimate arbitrary stereo RIRs constrained to lie on a 2D horizontal plane within a 3D room, i.e., it did not consider a z-axis and thus does not output RIRs at arbitrary heights. Therefore, we added the height on the z-axis as an input, embedding it by using the same positional encoding [48, 53] as the authors' code. The corresponding elements of the network architecture, e.g., the number of units in the input layer, were also modified. The architecture we used for NAF in our experiments consisted of 8 linear layers with Leaky-ReLU activations [44]. Note that we only changed the number of the number of units in the input layer, from 126 to 168, due to the aforementioned addition of z-axis features. In addition, the NAF we used in our experiments was designed to output only magnitude-spectrograms, i.e., without any phase information, because the official code also does not have the phase-related loss and corresponding phase output. We utilized the Griffin-Lim algorithm [30] to estimate the phase of each magnitude spectrogram and render the time-domain RIRs. For training, we followed the same process in their official code and used the model's weights after the final training epoch for inference and evaluation. Finally, we used a 48000 Hz sample rate rather than the original 22050 Hz. All other settings, such as the optimizer, number of epochs, learning rate, etc., are the same as their official implementation.

Deep Impulse Responses (DeepIR) [56]. Unlike NAF, the authors of DeepIR have not open-sourced an official codebase. Therefore, we implemented DeepIR ourselves, based on the details in their paper. Specifically, we built a simple multi-layer perceptron (MLP) consisting of 6 linear layers, each followed by leaky-ReLU activations. The input feature vector consists of (x, y, z, t), which are the desired spatial coordinates and the time index, respectively. Similar to NAF, we applied positional encoding to all inputs before passing them into the MLP. Hence, the number of units in the input layer is d_{emb}, whereas all other layers have 512 units. DeepIR directly outputs the t^{th} time sample of the RIR to produce an estimate $\hat{\text{IR}}$ of the full RIR. We then convolve this with the arbitrary dry source audio x, to produce an estimate $\hat{y}$ of the sound of the arbitrary audio being recorded from the specified source and listener location in the room, i.e., $\hat{y} = x * \hat{\text{IR}}$. We optimized $\hat{y}$ according to

[1] https://github.com/aluo-x/Learning_Neural_Acoustic_Fields

	Classroom				Dampened Room				Hallway				Complex Room			
	RIR		Music		RIR		Music		RIR		Music		RIR		Music	
	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV	Mag	ENV
Pink Reg	5.22	**0.942**	**2.71**	**1.36**	**1.21**	**0.555**	**1.59**	**1.19**	**9.13**	**2.95**	**2.59**	**1.25**	4.86	0.917	**2.25**	**1.41**
No Reg.	5.22	0.973	2.76	1.47	1.23	0.579	1.62	1.33	9.17	2.99	2.71	1.35	**4.84**	0.908	2.26	1.45
Music Reg.	**5.20**	0.952	2.72	1.40	1.22	0.569	1.62	1.27	9.14	2.96	2.65	1.31	4.84	**0.903**	2.25	1.42

Table 13. Comparison of our model trained with no regularization, regularizing by convolving with pink noise, and regularization by convolving with music, on the tasks of monoaural RIR and music prediction. Lower is better for all metrics, and RIR errors are multiplied by 10.

Room	Training Time (Hours)	Inference Time (s)	N. Surfaces	Tracing Time (s)	Avg N. Reflection Paths
Classroom	9.61	0.90	9	4.3	874
Dampened	5.75	0.56	6	0.83	675
Hallway	8.97	0.90	6	1.5	853
Complex	2.82	0.37	33	47	439

Table 14. In all of our experiments, we train our model for 1000 epochs and report the training time that this takes in each room, in the base configuration. In addition, we report the inference time, or the time it takes our model to render a single RIR. Before training begins, we precompute the reflection paths that go between the source and listener locations, up to a certain maximum reflection order, so we also report this tracing time to trace reflection paths, per listener location of each room and its corresponding subdataset. We also report the number of valid reflection paths found by the tracing algorithm, as an average across all points in the subdataset.

an L2 loss comparing the log-magnitude spectrogram with that of the corresponding ground-truth audio y_{gt}. We omitted the noise model, since our dataset did not include artificially added noise, and the noise in our recordings was minimal. We set other hyperparameters for DeepIR such as the optimizer, learning rate, the number of epochs, etc., to the same values as NAF.

Implicit Neural Representation for Audio Scenes (INRAS) [63]. The authors of the INRAS baseline released their code in the Supplementary Materials of their submission.[2] We use their code with some minor modifications. The framework is originally trained and tested on data from the SoundSpaces dataset [13], which provides simulated binaural recordings within virtual environments. The architecture is built around consuming this data, where each simulated recording represents a stereo, binaural recording with the head positioned at one of the four cardinal angles. Our training sets use exclusively monaural recordings from omnidirectional microphones. Thus, in order to make our changes to the network as minimal as possible, we duplicated our mono-channel recordings to stereo-channel recordings and assumed them to all be at the $0°$ angle. We then took only the left channel of the stereo output as the framework's estimate of the monaural RIR. Since INRAS consumes environment meshes, we provide it with a 3D scan of each room. Otherwise, we used mostly the same hyperparameters as the original, with the exception that we increased the sample rate from 22050 to 48000 Hz. Since our training set of 12 recordings per subdataset was approximately four orders of magnitude smaller than the datasets on which the authors had trained, we increased the initial learning rate from to 0.001 instead of 0.0005, slowed the learning rate's exponential decay schedule to decay rate $\gamma = 0.1$ over 3000 epochs rather than 50, and trained for 5000 epochs rather than 100. We evaluated the model against a validation set every 100 epochs. For our test evaluations of INRAS, we used the weights and consequent outputs of the model with the best performance across all such validation evaluations.

[2] https://openreview.net/forum?id=7KBzV5IL7W

G. Data Collection Procedure Details

We use a custom-built microphone frame designed to accommodate 12 Dayton Audio EMM6 measurement microphones, as well as one 3Dio FS XLR binaural microphone, all of which were rigidly mounted at precisely measured positions on the frame. Figure 10 shows a photo of the microphone frame used to collect the data. We set the origin of each room such that there is one wall representing $x = 0$ and one wall representing $y = 0$. Before each recording, we positioned the frame within the room and measured the distance from the edge of the frame to each of the origin walls using a tape measure or a Bosch GLM20 laser distance measure, which have 1 and 3 millimeter measurement resolutions, respectively. We use the measured position of the frame's corner as well as the pre-measured offset of each microphone from the frame's corner in order to annotate

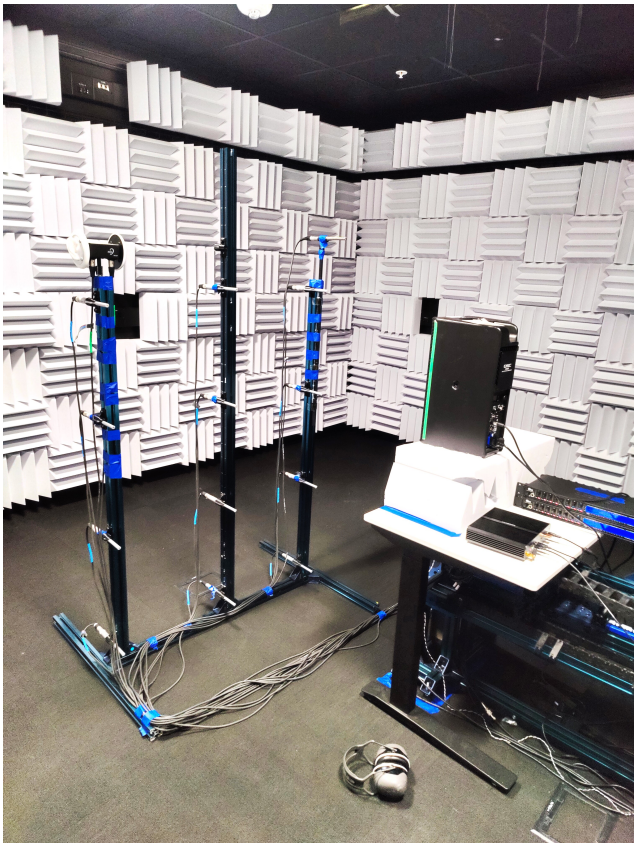


Figure 10. A photo of the data collection procedure in the Dampened Room. The custom microphone frame holds 12 microphones, as well as a 3Dio FS XLR binaural microphone.

each microphone's position in the room to sub-centimeter precision for our dataset.

G.1. Estimating the Room Impulse Response (RIR)

In order to measure each RIR, we played a logarithmic sine sweep through the speaker. The sweep spanned from 20 Hz to 24 kHz for 10 seconds, followed by 4 seconds of silence. This sine sweep was recorded from each of the microphones simultaneously at each gantry position. While sending the sine sweep signal from the audio interface to the speaker, we also recorded loopback signal by wiring the audio interface's output to one of its inputs. We used this loopback signal to estimate and correct for the latency in the system.

To compute the RIR $r[t]$, we take

$$r[t] = IFFT\left(\frac{FFT(a[t])}{FFT(l[t])}\right),$$

where FFT and $IFFT$ are the Fast-Fourier Transform and its inverse respectively, $a[t]$ is the digital recording of the sine sweep, and $l[t]$ is the digital loopback signal. Note that we deconvolve the loopback signal from the recording, instead of deconvolving the source signal sent to the speaker from the recording. We assume that the loopback signal is the same as the source signal, but delayed in time by the latency of the system. Deconvolving from a delayed copy of the source signal instead of directly from the source signal thus corrects for the delay in the system. We remove the last 0.1 seconds of the 14-second RIR to eliminate anti-causal artifacts.

In addition, to account for differences in microphone sensitivity, we adjust the volume of each sweep recording according to the sensitivity of the microphone used to record it. Specifically, we look up each EMM6's microphone's response at 1000 Hz in dB from its calibration sheet, and reduce the overall volume of its recordings by the same amount.

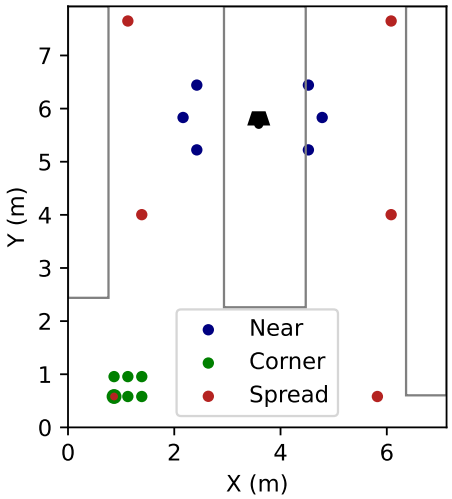


Figure 11. The distributions of three different sets of training points in the Classroom subdataset. The grey lines indicate the locations of tables in the subdataset.

G.2. Room Geometry Estimation

As the wavelengths of audible sound typically range from 2 cm - 17 m [49], the prominent sound waves are likely to bypass or diffract around smaller surfaces. Hence, we only focus on modeling salient surfaces (e.g., walls, pillars, table tops), which are often characterized by planes, and simply trace the reflection paths using image source methods. For the rooms we captured in our dataset, we also measured the walls and surfaces and manually created planar mesh-based reconstructions of them. With the recent progress in visual 3D scene reconstructions [48], our geometric estimation can also easily be replaced by automatic algorithms or even mature customer tools such as Polycam.

	RIR		Music	
Training Point Configuration	Mag	ENV	Mag	ENV
Near	5.89	1.14	3.25	1.61
Spread	**5.39**	**0.976**	**2.80**	**1.36**
Corner	5.88	1.07	3.12	1.41

Table 15. Evaluations of DIFFRIR on different datasets of size 6, with varying spatial distributions. All microphone locations are selected from $Z = 0.98$, and all locations used for testing and evaluation are also selected from $Z = 0.98$. Lower is better for all metrics, and RIR errors are multiplied by 10.

H. Guidelines for Microphone Placement.

To maximize efficiency, we found it empirically beneficial to spread our RIR locations in all three dimensions. This allows us to 1) cover a variety of angles around the speaker, which likely leads to better speaker directivity estimates, 2) disentangle the effects of individual reflections, and 3) better estimate the diffuse sound field, which is approximated in our model as spatially uniform.

To study this effect, we conducted an experiment in the Classroom subdataset. We select three different sets of training locations (shown in Figure 11), each of which contain 6 RIR recordings from 6 different locations. For simplicity, these training locations were selected in the 2D plane defined by $Z = 0.98$. We evaluated DIFFRIR trained on each of these sets of training locations on a test set comprised of other points selected in the $Z = 0.98$ plane.

Our best performance across all metrics is achieved in the 'Spread' configuration of training points, confirming our intuition. Interestingly enough, the 'Near' Configuration performed the worst. We believe this could be due to the model overfitting to the near-field of the speaker [36], which can be substantially different than the sound field at other locations in the room.