Differentiable Room Acoustic Rendering with Multi-View Vision Priors

Derong Jin
University of Maryland, College Park
djin77@umd.edu

Ruohan Gao
University of Maryland, College Park
rhgao@umd.edu

Abstract

*An immersive acoustic experience enabled by spatial audio is just as crucial as the visual aspect in creating realistic virtual environments. However, existing methods for room impulse response estimation rely either on data-demanding learning-based models or computationally expensive physics-based modeling. In this work, we introduce Audio-Visual Differentiable Room Acoustic Rendering (*AV-DAR*), a framework that leverages visual cues extracted from multi-view images and acoustic beam tracing for physics-based room acoustic rendering. Experiments across six real-world environments from two datasets demonstrate that our multimodal, physics-based approach is efficient, interpretable, and accurate, significantly outperforming a series of prior methods. Notably, on the Real Acoustic Field dataset,* AV-DAR *achieves comparable performance to models trained on 10 times more data while delivering relative gains ranging from 16.6% to 50.9% when trained at the same scale. Project Page: https://humathe.github.io/avdar/.*

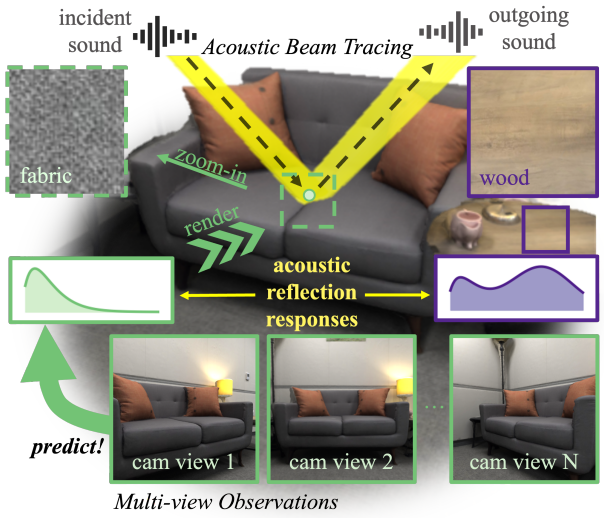

Figure 1. Our differentiable room acoustic rendering framework combines multi-view visual observations and acoustic beam tracing for efficient and accurate room impulse response (RIR) prediction. By analyzing the visual cues of surfaces (*e.g.*, fabric vs. wood), it infers acoustic reflection responses for accurately rendering RIRs through physics-based, end-to-end optimization.

1. Introduction

Spatial audio is a fundamental component of immersive multimedia experiences and is often regarded as "half the experience" in virtual and augmented reality (VR/AR) applications. For example, spatialized sound can help users navigate real environments [58], and precise spatial positioning of audio sources enhances natural remote communication and engagement [48]. Accurately modeling spatial audio unlocks a wide range of immersive applications, from entertainment and gaming to education and telepresence.

Recreating the spatial acoustic experience is analogous to novel-view synthesis [42, 44] in vision, where the goal is to learn from sparse images to synthesize photorealistic images from arbitrary viewpoints. Similarly, novel-view acoustic synthesis [33, 38, 40, 59, 65] aims to render the received sound at any listener location within a scene. A widely used representation for this task is the *Room Impulse Response* (RIR) [6, 55], which describes how an emitted sound propagates through space—including reflections and diffractions—before reaching a listener. Accurately estimating RIR for every possible source-listener pair would, in principle, enable realistic spatial acoustic rendering.

Existing methods for estimating RIRs generally fall into two broad categories: *learning-based* and *physics-based*. Learning-based approaches [5, 37, 39, 40, 49, 59] treat RIR estimation as a regression task trained on densely measured ground-truth RIRs. Although effective for fitting simulated scenes and offering efficient inference, they can be difficult to deploy in real-world settings due to high data requirements and the lack of physically grounded guarantees. In contrast, physics-based approaches [33, 65] rely on explicit acoustic models such as the image-source method [65] or volume rendering [33]. However, both are impractical for large scenes with sparse training samples: the image-source method requires significant computation to enumerate reflection paths and becomes prohibitively expensive in com-

plex environments, while volume rendering demands extensive 3D samples for RIR reconstruction and considerable training data to learn a neural acoustic field.

Our key insight is that while sound fundamentally differs from light—traveling more slowly and exhibiting time-of-arrival variations—both are influenced by the same geometry and surface material properties within a given scene. As shown in Fig. 1, a surface region's acoustic property often strongly correlates with visual appearance due to the same underlying materials. For example, smooth, hard surfaces like glass tend to reflect high-frequency sound, whereas rough, deformable materials such as carpets primarily absorb high-frequency components and reflect lower-frequency waves. Leveraging visual observations to estimate surface acoustic properties in a physics-based differentiable framework could potentially enable more accurate and efficient learning of room acoustic parameters.

To realize this intuition, we introduce an *Audio-Visual Differentiable Room Acoustic Rendering* (AV-DAR) framework, which leverages visual cues to guide the differentiable rendering of RIRs in an end-to-end manner. By aligning multi-view image features from camera space to 3D scene space via a cross-attention mechanism, our approach decouples view-dependent visual information, forming a unified, material-aware representation in the scene space. This representation serves as a robust foundation for learning reflection properties and enables accurate RIR estimation at unobserved locations. Additionally, we employ beam tracing [21, 29] to search for specular reflection paths and model the acoustic field, which requires fewer training samples than volume rendering and significantly reduces computation time compared to the image-source method.

Experiments on six real-world environments [13, 65] show that our approach significantly outperforms both learning-based and physics-based baselines. For example, on the Real Acoustic Field dataset [13], our model achieves comparable performance to existing methods trained on roughly 10$\times$ RIR measurements while delivering 16.6% to 50.9% improvement when trained at the same scale. Moreover, qualitative analysis confirms that our method learns material-specific acoustic reflection responses that closely align with the visual characteristics of the scene.

Our main contributions are threefold: First, we propose a *physics-based differentiable* room acoustic rendering pipeline that not only learn from sparse, real-world RIR measurements but is also *efficient*, *interpretable*, and *accurate*. Second, we are the first to integrate acoustic beam tracing within an end-to-end differentiable framework, enabling efficient computation of reflection responses. Third, our approach leverages multi-view images to capture material-aware visual cues that correlate with acoustic reflection properties, achieving significantly more accurate RIR rendering than prior methods.

2. Related Work

Learning-Based RIR Prediction. A growing body of work leverages machine learning techniques to approximate room impulse responses (RIRs) directly. Early methods often rely on large sets of measured or simulated RIRs to train neural networks capable of predicting RIRs at new positions. For instance, Ratnarajah *et al*. [51] use a generative adversarial network (GAN) to synthesize RIRs, while later efforts incorporate scene meshes [50] or visual signals [39, 49] to condition RIR generation. Continuous implicit neural representations have also been explored to model high-fidelity acoustic fields within individual scenes [5, 37, 38, 40, 52, 59]. However, unlike our approach, which can learn from sparse RIR measurments, these methods often rely on dense measurements and/or simulated data, requiring up to 1,000 measured RIRs in order to accurately interpolate to new positions within the same environment [52].

Physics-Based Room Acoustics Modeling. Classical acoustic modeling generally follows one of two paradigms: *wave-based* or *geometric* approaches. Wave-based methods directly solve the wave equation to model how sound propagates through air and interacts with surfaces [2, 28, 46, 57, 60]. These methods accurately capture inference and diffraction but can become computationally prohibitive for large domains or higher frequencies. In contrast, geometric acoustics approximates sound propagation using rays or beams. Techniques such as the *image-source method* [3] account for specular reflections by spawning virtual sources, but the number of virtual sources grows exponentially with reflection order [10]. Other geometric approaches, such as *ray tracing* [32, 54] and *beam tracing* [21, 29, 34, 62], use stochastic sampling to achieve more efficient rendering and more accurate late reverberation modeling compared to image-source methods [53]. Our method is also physics-based and uses beam tracing, but we integrate it into our differentiable room acoustic rendering pipeline for end-to-end optimization.

Audio-Visual Room Acoustics Learning. Recent inspiring work integrates visual and audio signals on an array of interesting tasks related to room acoustics, including predicting how a given audio signal would sound in a scene visually depicted in images and videos [7, 11, 17, 36, 56], converting monaural audio to spatial audio based on visual spatial cues from the environment [24, 26, 27, 35, 45], learning image representations, scene structures, and human locations through echolocation and ambient sound [14, 18, 23, 64], navigating in simulated and real room environments based on audio-visual observations [8, 22, 25], and synthesizing binaural sound from novel viewpoints [1, 9, 12, 15, 37, 41]. Different from all of them, we incorporate multi-view visual cues into a differentiable room acoustic rendering pipeline for accurate RIR prediction.

Differentiable Acoustic Rendering. Differentiable rendering has become a powerful tool in graphics and vision, enabling image-based training for a wide range of reconstruction tasks [31, 61, 67]. A similar approach can be applied to acoustic rendering, allowing inverse problems to be solved by optimizing acoustic-related physical properties using gradient-based optimization [16, 19, 20, 30]. In room acoustics, DiffRIR [65] introduces a differentiable image-source renderer that jointly learns source and scene properties, and AVR [33] applies a differentiable volume renderer to train a neural acoustic field. Differently, we leverage beam tracing [21, 29] to efficiently search for specular reflection paths, achieving higher accuracy than ray tracing and significantly reduced computation time compared to volume rendering or image-source methods [53].

3. Approach

3.1. Preliminaries

Our goal is to learn a time-domain room impulse response function from sparse training data:

$$\widehat{\mathrm{RIR}}(\mathbf{x}_a, \mathbf{x}_b, \mathbf{p}_a, t). \tag{1}$$

Here, $\mathbf{x}_a$, $\mathbf{x}_b$, $\mathbf{p}_a$ denote the speaker location, the listener position, and the source orientation, respectively.

During training, besides using a sparse set of ground-truth RIR measurements, we leverage multi-view images to capture the scene's visual information. Formally, we assume a set of N_c images with known camera intrinsics π and extrinsics $P^{(i)}$:

$$\left\{ \{I^{(i)}, \pi, P^{(i)}\} \,\middle|\, i = 1, \cdots, N_c \right\}. \tag{2}$$

Our motivation of using multi-view images arises from the fact that, while RIR measurements encode the final results of acoustic wave propagation—including direct sound, early reflections, and late reverberations—they make it difficult to directly infer the underlying surface reflection properties. Multi-view images, on the other hand, capture visual material and geometric cues that are essential for predicting plausible reflection responses, effectively complementing the acoustic modality.

In practical applications, once $\widehat{\mathrm{RIR}}$ is learned, convolving it with an arbitrary input signal $h(t)$ yields the predicted audio at $\mathbf{x}_b$:

$$h_b(t) = h(t) \ * \ \widehat{\mathrm{RIR}}(\mathbf{x}_a, \mathbf{x}_b, \mathbf{p}_a, t), \tag{3}$$

thus enabling realistic spatial acoustic rendering.

3.2. Overview of the AV-DAR Framework

We now introduce our framework, AV-DAR, which estimates room impulse responses (RIRs) at novel source-listener pair locations from sparse RIR measurements, multi-view images, and rough geometry of the room (*e.g.*, expressed as a small number of planes). Following the decomposition in [65], we model the RIR as:

$$\widehat{\mathrm{RIR}}(t) = \sum_{\tau} s(\tau; \Theta_1) \cdot R(t - \tau; \Theta_2) + r(t; \Theta_3), \tag{4}$$

where:

- $s(t; \Theta_1)$ is a learnable time-indexed vector representing the speaker's source impulse response.
- $R(t; \Theta_2)$ is the integrated reflection response. We compute this term using beam tracing and differentiable rendering (Section 3.3), which is efficient compared to image-source methods and more accurate than ray tracing. To adapt to the beam tracing approach, we develop a multi-scale reflection response (Section 3.4). In addition, our multi-view vision module (Section 3.5) extracts visual material cues to condition the reflection response prediction explicitly on surface properties.
- $r(t; \Theta_3)$ is the position-dependent residual impulse response that models high-order reflections, diffraction, and late reverberation (Section 3.6).

Overall, AV-DAR integrates all components into an end-to-end differentiable pipeline, enabling gradient-based optimization for accurate RIR rendering.

3.3. Acoustic Beam Tracing

Here we aim to develop a differentiable acoustic rendering model that supports gradient-based optimization of reflection responses. This requires an efficient method to compute the integrated reflection response while maintaining differentiability for backpropagation.

Traditional image-source methods [3] yield accurate specular reflections but are computationally prohibitive in large scenes due to their exponential complexity. Although ray tracing [32, 53, 54] offers faster, stochastic sampling of paths, it often fails to capture specular reflections when the source and listener are modeled as infinitesimal points. To overcome these limitations, we adopt beam tracing [21, 29, 62] as our acoustic primitive. In contrast to ray tracing—which treats sound as infinitesimally thin lines—beam tracing represents sound as volumetric, cone-shaped beams (see Supp. for an illustration); a listener is considered "hit" if it lies within a beam, ensuring robust detection of specular paths without relying on artificial volumetric approximations.

We uniformly sample N_d beams from the source using a Fibonacci lattice to ensure even angular coverage and select a small apex angle to prevent beam overlap while maximizing spatial efficiency. Assuming narrow beams, we neglect beam splitting. The beam tracing function defines the set of valid reflection paths between source $\mathbf{x}_a$ and listener $\mathbf{x}_b$ as:

$$\mathcal{P}(\mathbf{x}_a, \mathbf{x}_b) = \{\tilde{\mathbf{x}}_k\}_{k=1}^{N}. \tag{5}$$

For each path $\tilde{\mathbf{x}}$, we compute the frequency response by combining the frequency-dependent reflection response at

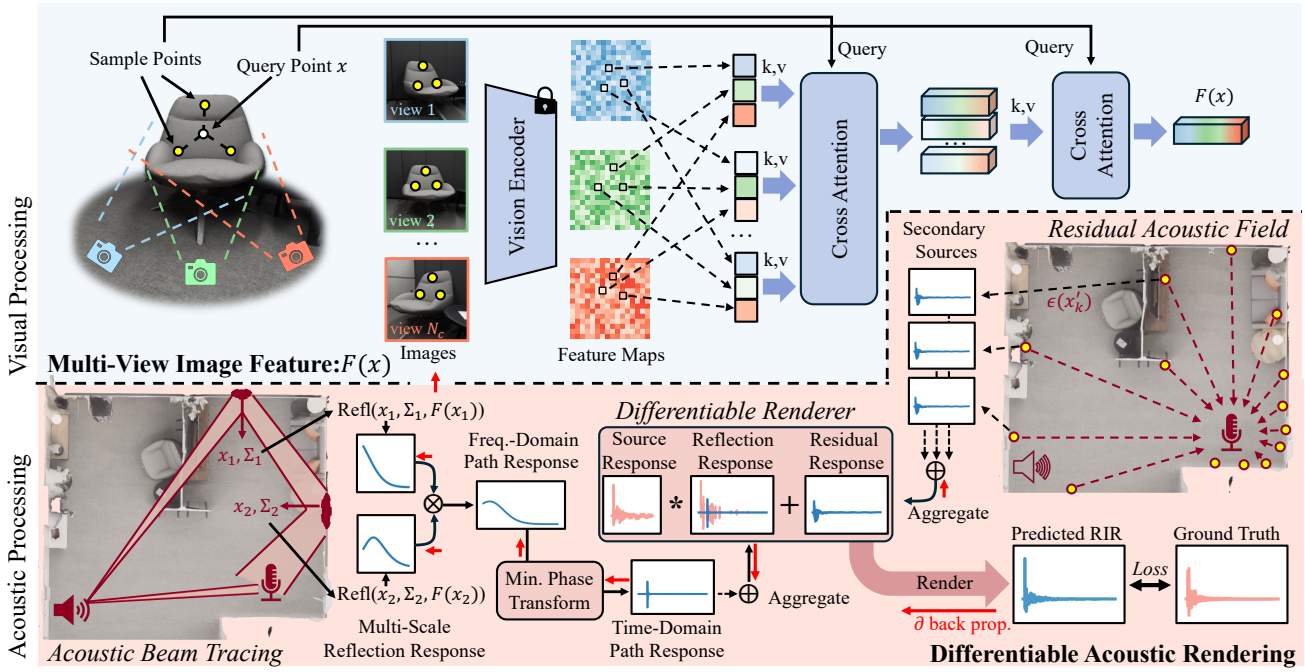

Figure 2. **Method Overview.** Our framework contains two main components for rendering the room impulse response (RIR): **(1) Visual Processing (top)**: Multi-view images of the scene are passed through a pre-trained vision encoder to extract pixel-aligned features at sampled points on the room surface. We then apply cross-attention both across views for each sampled point and across sampled points of query $\mathbf{x}$ to obtain a unified, material-aware visual feature $\mathrm{F}(\mathbf{x})$ (detailed in Section 3.5). **(2) Acoustic Processing (bottom):** On the left, we illustrate our acoustic beam tracing procedure (Section 3.3), where we sample specular paths and compute the path reflection response, conditioned on both the positional encoding (Section 3.4) and the visual feature $\mathrm{F}(\mathbf{x})$. On the right, we show how we model the residual acoustic field (Section 3.6) by treating every point on the surface as a secondary sound source and integrating its contribution via Monte-Carlo integration. The entire pipeline is fully differentiable, enabling end-to-end optimization of both acoustic and visual parameters.

each hit point, $\mathrm{Refl}(\mathbf{x})[f]$, with the source's directional response, $\mathbf{D}_{\tilde{x}}$. The total frequency-domain attenuation is:

$$\prod_{\mathbf{x}_j \in \tilde{\mathbf{x}}} \mathrm{Refl}(\mathbf{x}_j)[f] \cdot \mathbf{D}_{\tilde{x}}[f]. \tag{6}$$

This product is transformed into the time domain via a minimum phase transform [43]:

$$\kappa(\tilde{\mathbf{x}}, t) = \mathrm{MinPhase}\left\{ \mathbf{D}_{\tilde{x}} \circ \prod_{\mathbf{x}_j \in \tilde{\mathbf{x}}} \mathrm{Refl}(\mathbf{x}_j) \right\}(t). \tag{7}$$

In addition, we account for attenuation and delay due to air absorption and propagation loss using a propagation operator:

$$\mathcal{S}_\tau\{h\}(t) = \frac{\exp(-a_0 \tau)}{v_{\text{sound}}\, \tau}\, h(t-\tau), \tag{8}$$

where $\exp(-a_0\tau)$ models air absorption, $(v_{\text{sound}}\,\tau)^{-1}$ represents propagation loss, and $h(t-\tau)$ applies the appropriate delay. For a path with total travel time $\tilde{t}$, the final contribution to the reflection response is $\mathcal{S}_{\tilde{t}}\{\kappa(\tilde{\mathbf{x}}, t)\}$.

Finally, the integrated reflection response is obtained by summing the contributions of all valid paths:

$$R(t) = \sum_{\tilde{\mathbf{x}}_k \in \mathcal{P}} \mathcal{S}_{\tilde{t}_k}\{\kappa(\tilde{\mathbf{x}}_k, t)\}\,.$$

Our differentiable formulation, combined with beam tracing, enables efficient optimization of reflection response learning and visual feature extraction, as discussed next.

3.4. Multi-Scale Reflection Response

In a volumetric beam framework, the *contact region* between a beam and a surface grows as the beam propagates farther from the source, causing the reflected energy to span a larger patch of the surface. This phenomenon naturally introduces a *multi-scale* effect: near the source, reflections come from a smaller region; farther away, reflections integrate over a larger area. Due to small apex angle, the physical contact region is considered as an ellipse in most cases. However, our beam tracing algorithms only returns a single hit-point $\mathbf{x}$ for each reflection. To model the overall reflection response from the area, we approximate it by a small Gaussian distribution around the sampled hit point $\mathbf{x}$. Concretely, we let $\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \Sigma)$, where Σ encodes the elliptical patch size and orientation. Intuitively, Σ grows with the travel distance l and depends on the reflection angle θ and

half-apex angle φ. For the detailed calculation of Σ, please refer to Supp.

Next, we introduce our modeling of the reflection response. Recall from Equation 6 that $\mathrm{Refl}(\mathbf{x})$ is a frequency-dependent surface function. We discretize the frequency axis into F key frequencies and learn reflection values at those key frequencies, linearly interpolating to produce a continuous frequency response. We thus seek:

$$\mathrm{Refl}(\gamma(\mathbf{x}, \Sigma); \mathbf{\Theta}_2) \in \mathbb{R}^F, \tag{9}$$

where $\gamma(\mathbf{x}, \Sigma)$ is the integrated positional encoding (IPE) of surface location $\mathbf{x}$. We adopt the formulation of IPE from Mip-NeRF [4] to handle the spread in the contact region, meaning that we replace the point-based Fourier positional encoding $\gamma(\mathbf{x})$ with an "averaged" encoding:

$$\gamma(\mathbf{x}, \Sigma) = \mathbb{E}_{\mathbf{x}' \sim \mathcal{N}(\mathbf{x}, \Sigma)}[\gamma(\mathbf{x}')] \tag{10}$$

$$= \gamma(\mathbf{x}) \circ \exp\left(-\tfrac{1}{2}\mathrm{diag}(\Sigma_\gamma)\right) \tag{11}$$

Though not explicitly integrating over the elliptical region, this encoding incorporates the local variance of the beam contact, allowing reflection predictions at the same surface point to vary when the beam's elliptical region differs.

3.5. Multi-View Vision Feature Encoder

We incorporate *multi-view images* to guide reflection response prediction. Specifically, we learn a vision feature encoder $\mathrm{F}(\mathbf{x}, \Sigma; \mathbf{\Theta}_2')$ that captures the appearance information around the surface $\mathbf{x}$. The reflection response is then:

$$\mathrm{Refl}\Big(\gamma(\mathbf{x}, \Sigma), \mathrm{F}(\mathbf{x}, \Sigma); \mathbf{\Theta}_2\Big) \in \mathbb{R}^F \tag{12}$$

Assume we have a set of N_s basis sample points on geometry $\mathcal{M}$, denoted by $\{\mathbf{z}_j\}_{j=1}^{N_s}$, and N_c captured images, denoted by $\{I^{(i)}\}_{i=1}^{N_c}$. We construct F in three stages:

1. *Per-view feature extraction.* This stage encodes each image into a feature map, then sample per-sample features from each image.
2. *Multi-view feature aggregation.* This stage aggregates multi-view features of each sample point $\mathbf{z}_j$ into a single feature vector $\mathbf{v}_j$.
3. *Sample-level neighborhood fusion.* This stage fuses the features of the k-nearest sample points around a query $\mathbf{x}$ into the unified vision feature vector $\mathrm{F}(\mathbf{x}, \Sigma)$.

Per-View Feature Extraction. Let $I^{(i)}$ be the i-th image with known camera intrinsics π and extrinsics $P^{(i)}$. We first extract a pixel-aligned feature map $W^{(i)} = \mathcal{E}(I^{(i)})$ using a pre-trained vision encoder $\mathcal{E}$ (*e.g.*, DINO-v2[47]). Then, for each 3D sampled point $\mathbf{z}_j$, we project it into image i with $P^{(i)}$ and π. We then bi-linearly sample from $W^{(i)}$ to obtain the vision feature $\mathbf{v}_j^{(i)}$:

$$\mathbf{v}_j^{(i)} = \begin{cases} W^{(i)}\Big(\pi\left(P^{(i)}\mathbf{z}_j\right)\Big) & \text{if } \mathbf{z}_j \text{ is visible in } I^{(i)}, \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{13}$$

Multi-View Feature Aggregation. Given the per-view feature $\mathbf{v}_j^{(i)}$, we next aggregate them across images to form a single vision feature $\mathbf{v}_j$ for each sample $\mathbf{z}_j$. We adopt cross-attention mechanism [63]:

$$\mathbf{v}_j^{\mathrm{raw}} = \sum_{i=1}^{N_c} \rho\Big(\phi(\gamma(\mathbf{z}_j))^\top \psi(\mathbf{v}_j^{(i)} \oplus P^{(i)}) + m_j^{(i)}\Big) \cdot \alpha(\mathbf{v}_j^{(i)}) \tag{14}$$

where $\mathbf{v}_j^{\mathrm{raw}}$ is the intermediate aggregated feature. The functions ψ, ϕ, and α are linear projections to produce keys, quires, and values, respectively. The mask $m_j^{(i)}$ is 0 if $\mathbf{z}_j$ is visible in $I^{(i)}$ and $-\infty$ if otherwise. $\gamma(\cdot)$ is a Fourier positional encoding, and ρ denotes the softmax function. Finally, a 2-layer feed-forward network (FFN) refines this raw feature:

$$\mathbf{v}_j = \mathrm{FFN}(\mathbf{v}_j^{\mathrm{raw}}). \tag{15}$$

Sample-Level Neighborhood Fusion. With $\mathbf{v}_j$ computed for each sampled point $\mathbf{z}_j$, we then derive a vision feature for an arbitrary query point $\mathbf{x}$. Let

$$N(\mathbf{x}) = \{\mathbf{z}_j^*\}_{j=1}^k \tag{16}$$

be the set of k-nearest samples to $\mathbf{x}$, with corresponding vision features $\{\mathbf{v}_j^*\}_{j=1}^k$. We then fuse these neighborhood features using a point-transformer [68]:

$$\mathrm{F}(\mathbf{x}, \Sigma) = \sum_{j=1}^{k} \rho\Big(g\big(\phi'(\gamma(\mathbf{x}, \Sigma)) - \psi'(\mathbf{v}_j^*) + \delta_j\big)\Big)\big(\alpha'(\mathbf{v}_j^*) + \delta_j\big), \tag{17}$$

where g is a MLP producing the non-normalized attention coefficients from the key-query difference, and ϕ', ψ', and α' are key, query, value projections. δ_j is a positional encoding produced by another MLP ϑ':

$$\delta_j = \vartheta'(\mathbf{x} - \mathbf{x}_j^*). \tag{18}$$

In this way, our two-level fusion across both views and sampled points yields a robust vision feature that captures local geometry, visibility, and appearance. These features then drive more accurate reflection response predictions in the subsequent acoustic modules.

3.6. Position-Dependent Residual Component

Finally, we introducethe residual component $\mathrm{r}(t; \mathbf{\Theta}_3)$, which is responsible for modeling high-order reflections, diffuse reflections, diffraction, and late reverbarations. We treat each point on the geometry $\mathbf{x} \in \mathcal{M}$ as a secondary sound source and model the residual IR as the integral of these secondary sources from the listener position $\mathbf{x}_b$.

We use a 4-layer MLP ϵ, to predict the differential time-domain response $h(t)$ per solid angle ω at any location $\mathbf{x}$:

$$\mathrm{h}(t) = \epsilon(\mathbf{x}, \omega, t, \mathbf{x}_a, \mathbf{p}_b; \mathbf{\Theta}_3). \tag{19}$$

The residual IR is then calculated by the integral:

Method	Scale	RAF-Empty				RAF-Furnished			
		Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓	Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓
NAF++ [13, 40]	1%	6.05	2.10	94.5	23.9	6.61	2.10	74.9	23.0
INRAS++ [13, 59]	1%	3.69	2.59	100.3	23.5	2.96	2.61	92.6	25.0
AV-NeRF [37]	1%	3.16	2.52	96.4	21.8	2.92	2.64	96.7	24.5
AVR [33]	1%	3.00	2.19	87.3	24.1	2.97	2.33	72.3	17.9
Ours	0.1%	3.14	1.81	86.6	16.9	2.45	1.98	80.1	15.2
Ours	1%	2.50	1.42	56.2	10.7	1.68	1.29	47.4	9.61

Table 1. Results on the Real Acoustic Field dataset [13] (0.32 s, 16 kHz). Cells highlighted in green denote the best performance, and yellow indicates the second best. Note that our model trained on only 0.1% of the data already achieves lower C50 and T60 errors than baseline methods, and significantly outperforms all baselines when using the same amount of training data.

$$\mathrm{r}(t;\boldsymbol{\Theta}_3) = \int_{\mathbb{S}^2} \mathcal{S}_\tau\{\epsilon\}(\mathbf{x}'(\omega), -\omega, t; \boldsymbol{\Theta}_3) p_{\mathrm{u}}(\omega) d\omega \quad (20)$$

$$\approx \sum_{k=1}^{N_r} S_{\tau_k}\{\epsilon\}(\mathbf{x}'_k, -\omega_k, t; \boldsymbol{\Theta}_3). \quad (21)$$

In Equation 20, $\mathbf{x}'(\omega)$ is the intersection of a ray in direction ω from $\mathbf{x}$ with room geometry $\mathcal{M}$, and p_{u} is the uniform distribution over $\mathbb{S}^2$. Equation 21 approximates Equation 20 via Monte Carlo integration with N_r sampled directions ω_k from distribution p_{u}.

4. Experiments

4.1. Experiment Settings

Datasets. We evaluate our method on two real-world datasets: the *Real Acoustic Field* (RAF) [13] dataset and the *Hearing Anything Anywhere* (HAA) [65] dataset, which are the only available real-world RIR datasets with accompanying visual capture. The RAF dataset contains densely measured monaural RIRs recorded in two office settings (*Empty* and *Furnished*) using tens of thousands of source–listener pairs. We use 0.32 s RIR segments resampled at 16 kHz, following prior work [13, 33]. Each room also has a 3D reconstruction and a dense set of images [66].

The HAA dataset comprises four rooms with diverse structures and acoustic characteristics. Each room has manually crafted planar geometry. Following [65], we train on 12 listener locations per room and test on half of the remaining unseen locations. For HAA, RIR segments of 2.0 s are used, and they are resampled at 16 kHz.

Evaluation Metrics. Following [13, 33, 59], we evaluate perception-related energy decay patterns using *Clarity* (C50), *Early Decay Time* (EDT), and *Reverberation Time* (T60). To account for differences in overall RIR magnitude, we also adopt a loudness metric defined as:

$$\text{Loudness Error} = \left| 10 \log_{10}\left(\frac{E_{\text{pred}}}{E_{\text{gt}}}\right) \right|, \quad (22)$$

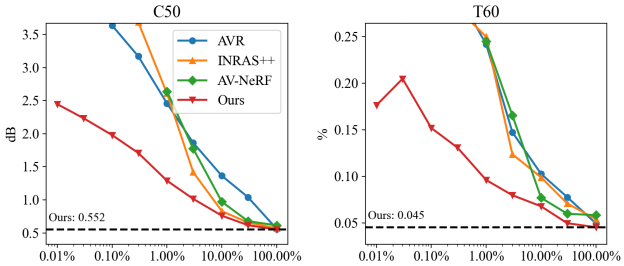

Figure 3. Performance comparison across training scales (from 0.01% to 100% of training data). Our model consistently outperforms baselines, particularly in few-shot scenarios (less than 3% data). See Supp. for EDT and Loudness metrics.

where $E = \int_0^\infty h^2(t)\,\mathrm{d}t$ is the energy of the signal $h(t)$.

Implementation Details. For multi-view images, we manually select a subset covering the scene, using 13–30 images for most scenes in the two datasets and 65 images for the RAF Furnished room to assess performance saturation (see Supp. for ablation details). In the HAA dataset, where multi-view images are not provided, we randomly sample cameras from the Polycam reconstructions to render images at 512×512 resolution. Note that Polycam is used only for visual rendering, while beam tracing is performed on the original coarse planar geometry. Additional architectural choices and training details are provided in Supp.

Baselines. We compare with a series of state-of-the-art approaches. NAF++ and INRAS++ [13] are improved versions of the implicit neural field methods NAF [40] and INRAS [59] in 3D settings, respectively. AV-NeRF [37] is an audio-visual method that also uses depth and RGB to learn implicit acoustic fields. AVR [33] is a physics-based method that uses NeRF-like volume rendering for RIR reconstruction. Diff-RIR [65] is a differentiable rendering method that uses the image-source method for forward rendering; due to its extensive pre-computation time (over 500 hours for just two reflections in one scene on the RAF dataset), we only include it in the HAA comparisons.

Method	Classroom			Complex Room			Dampened Room			Hallway		
	Loud (dB) ↓	C50 (dB) ↓	T60 (%) ↓	Loud (dB) ↓	C50 (dB) ↓	T60 (%)↓	Loud (dB) ↓	C50 (dB) ↓	T60 (%) ↓	Loud (dB) ↓	C50 (dB) ↓	T60 (%) ↓
NAF++ [13, 40]	8.27	1.62	134.0	4.43	2.25	44.8	3.88	4.24	306.9	8.71	1.36	21.4
INRAS++ [13, 59]	1.31	1.86	60.9	1.65	2.26	29.5	3.45	3.28	187.1	1.55	1.87	7.4
AV-NeRF[37]	1.51	1.43	50.0	2.01	1.88	36.6	2.40	3.05	107.9	1.26	1.03	9.5
AVR [33]	3.26	4.18	44.3	6.47	2.55	36.7	6.65	11.11	81.4	2.48	2.69	7.0
Diff-RIR [65]	2.24	2.42	39.7	1.75	2.23	18.5	1.87	1.56	44.9	1.32	3.13	6.8
Ours	0.99	1.02	24.3	0.98	1.44	10.8	1.11	1.45	31.9	0.85	1.15	6.3

Table 2. Results on the Hearing Anything Anywhere dataset [65] (2.0 s segments, 16 kHz), trained on 12 listener locations. Our method significantly outperforms all baseline methods in these scenes, demonstrating its effectiveness in accurately reconstructing room acoustics in few-shot settings. See Supp. for EDT error results.

4.2. Quantitative Results

Results on the RAF Dataset. To fully exploit the dense samples in RAF [13] and evaluate performance at various training scales, we split the original training set (80% of all data) into 9 nested subsets ranging from 0.01% to 100% of the data. The smallest subset contains only 3 samples, while the largest includes approximately 30K samples, with each larger subset including all samples from the smaller ones. All models are evaluated on the original test set from [13] to ensure comparability across scales. Table 1 reports results on the 1% dataset and on our model trained with 0.1% of the data. Notably, our method trained on only 0.1% of the data achieves comparable performance to state-of-the-art baselines trained on 10× more data. In addition, our method consistently outperforms all baselines when trained with the same amount of data, with improvements ranging from 16.6% (Loudness Error in RAF Empty Scene) to 50.9% (T60 Error in RAF Empty Scene). Figure 3 further demonstrates that our model outperforms existing baselines across all training scales, particularly in few-shot settings (below 3% training data).

Results on the HAA Dataset. Table 2 shows the performance on the HAA dataset [65]. Our method significantly outperforms all baseline methods in the four challenging real-world scenes, demonstrating its effectiveness in accurately reconstructing room acoustics. The only exception is the C50 metric in the *Hallway* scene, where AV-NeRF exhibits particularly strong performance. This is likely due to AV-NeRF uses depth as an input, which is especially beneficial in this simple, constrained hallway geometry.

Ablation Study. We further conduct an ablation study where we ablate key components in our framework to evaluate their individual contributions: 1) *Uni. Residual* denotes replacing the positional-dependent residual with a fixed learnable positional-independent vector; 2) *w/o Residual* sets Residual to zero; 3) *w/o Vision* sets vision features to zero; 4) *Ray-tracing* replaces beam tracing with specular ray tracing; and 5) *w/o IPE* uses fixed Fourier features in place of the multi-scale positional encoding. The results are summarized in Table 3. We can see that each component is essential for accurately rendering RIRs at novel locations.

Variant	C50	EDT	T60
Ours (full)	1.98	80.1	15.2
Uni. Residual	2.11	106.4	13.9
w/o Residual	3.82	142.8	49.0
w/o Vision	2.13	98.6	14.3
Ray-tracing	4.27	146.9	21.9
w/o IPE	2.10	101.2	15.0

Table 3. Ablation study results. See text for details.

4.3. Qualitative Results

Visualization of Signal Spatial Distribution. Figure 4 illustrates the spatial distribution of the sound signal at an unseen source location and orientation in the RAF Furnished Room. The top two rows display the phase and amplitude at a wavelength of 0.6 m, while the bottom row shows the loudness heatmap. Our model generates physically plausible wave distributions, as evidenced by the periodic patterns in both phase and amplitude, and successfully captures the source directivity and reflection decay with as little as 0.1% training data. In contrast, AVR, which is also a physics-based method, accurately models the source location but fails to capture the source directivity and produce physically consistent spatial distributions of phase and amplitude. Unsurprisingly, all learning-based methods (AV-NeRF, NAF++, INRAS++) struggle in this regard. While they can interpolate and predict RIRs by learning from, and sometimes overfitting to, the training data, they fail to produce a loudness heatmap with correct source localization or meaningful phase and amplitude distributions. For comparisons of predicted RIR signals, please refer to Supp.

Interpreting Learned Reflection Responses. In Figure 5, we visualize the reflection response on the image surfaces

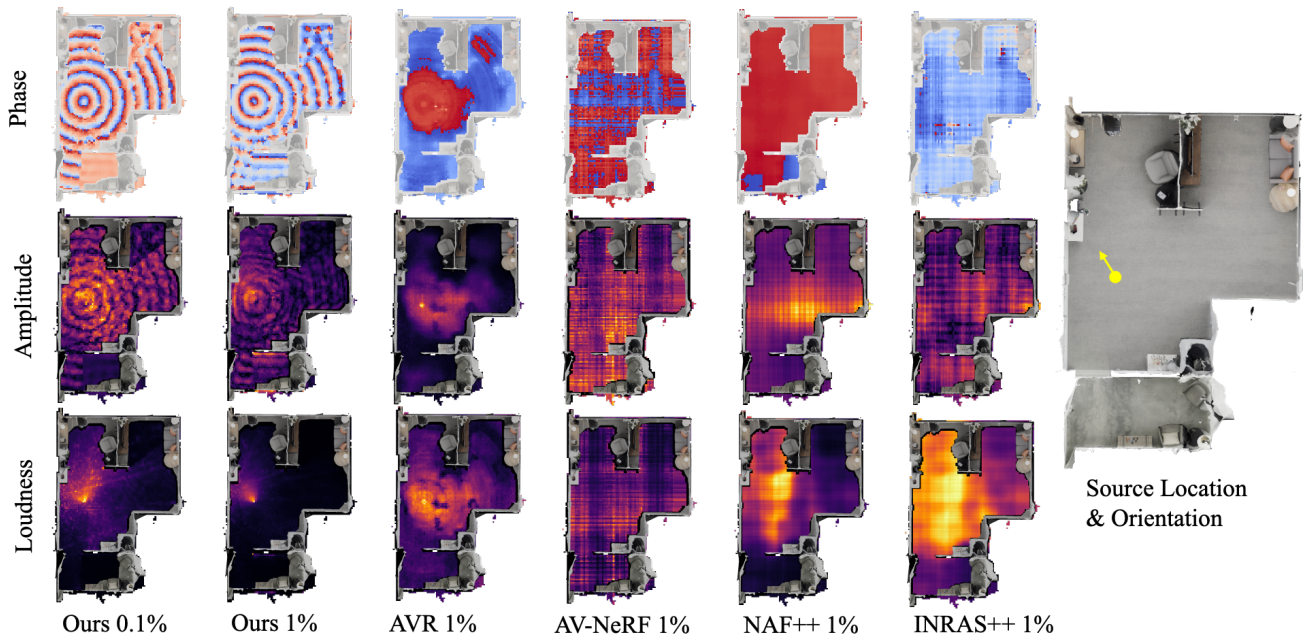

Figure 4. Signal spatial distribution visualization. **Top two rows:** Phase and amplitude maps at 0.6 m wavelength. **Bottom row:** Loudness heatmap. Our model, trained on only 0.1% of the data, accurately captures source directivity and localization, yielding plausible phase and amplitude distributions, while baseline methods fail to produce these patterns even with 10× training data.

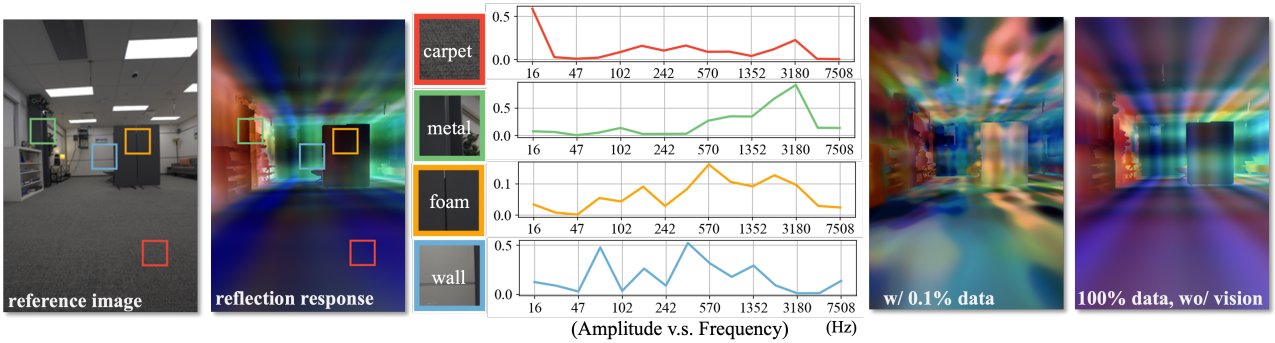

Figure 5. Reflection response visualization. The RGB color encodes frequency-dependent reflection response, with red indicating high-frequency and blue indicating low-frequency. Our method yields diverse, interpretable reflection patterns even with only 0.1% training data. In the middle, we visualize the reflection response curves. The results align with real-world observations—*e.g.*, carpet exhibits low high-frequency reflectivity, foam is generally absorptive, and metal reflects strongly at high frequencies.

by projecting RGB colors into the camera space. Red and blue indicate higher-frequency reflectivity and lower-frequency reflectivity, respectively. Our method yields diverse and interpretable reflection responses, even when trained on just 0.1% of the data. For example, the reflection responses for the carpet and metal areas align with their material properties—carpet effectively absorbs high frequencies, foam is generally absorptive, and metal is highly reflective in the high-frequency range. Moreover, an acoustic-only variant of our model (with vision features replaced by zeros) highlights the impact of incorporating visual cues, which significantly enhances the diversity and material relevance of the predicted reflection responses.

5. Conclusion

We presented AV-DAR, an audio-visual differentiable pipeline for synthesizing room impulse responses (RIRs). By combining beam tracing with visually-guided reflection modeling, our approach learns RIRs from sparse real-world measurements and outperforms state-of-the-art baselines while reducing data requirements. Our work opens new possibilities for immersive AR/VR applications. As future work, we plan to extend our framework to handle multi-scene scenarios for few-shot or zero-shot reflection response prediction. We also aim to explore implicit acoustic modeling from only raw audio data, leveraging much larger corpora for training more generalizable models.

References

[1] Byeongjoo Ahn, Karren D. Yang, Brian Hamilton, Jonathan Sheaffer, Anurag Ranjan, Miguel Sarabia, Oncel Tuzel, and Jen-Hao Rick Chang. Novel-view acoustic synthesis from 3d reconstructed rooms. *ArXiv*, abs/2310.15130, 2023. 2

[2] Andrew Allen and Nikunj Raghuvanshi. Aerophones in flatland: interactive wave simulation of wind instruments. *ACM Trans. Graph.*, 34(4), July 2015. 2

[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, 04 1979. 2, 3

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. 5

[5] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[6] Chunxiao Cao, Zhong Ren, Carl Schissler, Dinesh Manocha, and Kun Zhou. Interactive sound propagation with bidirectional path tracing. *ACM Trans. Graph.*, 35(6), Dec. 2016. 1

[7] Changan Chen, Ruohan Gao, Paul T Calamia, and Kristen Grauman. Visual acoustic matching. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18836–18846, 2022. 2

[8] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[9] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *CVPR*, 2023. 2

[10] 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. *arXiv*, 2022. 2

[11] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP*, 2023. 2

[12] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere - hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7819–7828, 2023. 2

[13] Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21886–21896, 2024. 2, 6, 7

[14] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *Conference on Robot Learning (CoRL)*, 2021. 2

[15] Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. 2023. 2

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

[17] Sanjoy Chowdhury, Sreyan Ghosh, Dasgupta Subhrajyoti, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. *ICCV*, 2023. 2

[18] Jesper Haahr Christensen, Sascha Hornauer, and X Yu Stella. Batvision: Learning to see 3d spatial layout with two ears. In *ICRA*. IEEE, 2020. 2

[19] Samuel Clarke, Negin Heravi, Mark Rau, Ruohan Gao, Jiajun Wu, Doug James, and Jeannette Bohg. Diffimpact: Differentiable rendering and identification of impact sounds. In *5th Annual Conference on Robot Learning*, 2021. 3

[20] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2020. 3

[21] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, page 21–32, New York, NY, USA, 1998. Association for Computing Machinery. 2, 3

[22] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 2

[23] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial visual representation learning through echolocation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[24] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[25] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for training household agents that see and hear. In *International Conference on Robotics and Automation (ICRA),*, 2023. 2

[26] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. In *British Machine Vision Conference (BMVC)*, 2021. 2

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

[28] Nail A. Gumerov and Ramani Duraiswami. A broadband fast multipole accelerated boundary element method for the three dimensional helmholtz equation. *The Journal of the Acoustical Society of America*, 125(1):191–205, 01 2009. 2

[29] John Kenneth Haviland and Balakrishna D. Thanedar. Monte carlo applications to acoustical field solutions. *The Journal of the Acoustical Society of America*, 54(6):1442–1448, 12 1973. 2, 3

[30] Xutong Jin, Chenxi Xu, Ruohan Gao, Jiajun Wu, Guoping Wang, and Sheng Li. Diffsound: Differentiable modal sound rendering and inverse rendering for diverse inference tasks. In *SIGGRAPH*, 2024. 3

[31] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[32] A. Krokstad, S. Strom, and S. Sørsdal. Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968. 2, 3

[33] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 3, 6, 7

[34] Christian Lauterbach, Anish Chandak, and Dinesh Manocha. Interactive sound rendering in complex and dynamic scenes using frustum tracing. *IEEE Transactions on Visualization and Computer Graphics*, 13:1672–1679, 2007. 2

[35] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2

[36] Tingle Li, Renhao Wang, Po-Yao Huang, Andrew Owens, and Gopala Anumanchipalli. Self-supervised audio-visual soundscape stylization. In *ECCV*, 2024. 2

[37] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 6, 7

[38] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *ArXiv*, abs/2309.15977, 2023. 1, 2

[39] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastià V. Amengual Garí, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2

[40] 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, 6, 7

[41] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[42] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 1

[43] 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, 12 2001. 4

[44] 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

[45] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 2018. 2

[46] Damian Murphy, Antti Kelloniemi, Jack Mullen, and Simon Shelley. Acoustic modeling using the digital waveguide mesh. *IEEE Signal Processing Magazine*, 24(2):55–66, 2007. 2

[47] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 5

[48] Vanessa Y Oviedo, Khia A Johnson, Madeline Huberth, and W Owen Brimijoin. Social connectedness in spatial audio calling contexts. *Computers in Human Behavior Reports*, 2024. 1

[49] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27164–27175, June 2024. 1, 2

[50] 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*, MM '22, page 924–933, New York, NY, USA, 2022. Association for Computing Machinery. 2

[51] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–575, 2022. 2

[52] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3209–3213, 2022. 2

[53] Lauri Savioja and U. Peter Svensson. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730, 08 2015. 2, 3

[54] Carl Schissler, Ravish Mehra, and Dinesh Manocha. High-order diffraction and diffuse reflections for interactive sound propagation in large environments. *ACM Trans. Graph.*, 33(4), July 2014. 2, 3

[55] Samuel Siltanen, Tapio Lokki, Sami Kiminki, and Lauri Savioja. The room acoustic rendering equation. *The Journal of the Acoustical Society of America*, 122:1624, 10 2007. 1

[56] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 286–295, October 2021. 2

[57] Julius O. Smith. Physical modeling using digital waveguides.

Computer Music Journal, 16(4):74–91, 1992. 2

[58] Meta Open Source. Aria gen 2 case study: Envision - spatial audio navigation and ai research, 2025. Accessed: 2025-04-28. 1

[59] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2, 6, 7

[60] Lonny L. Thompson. A review of finite-element methods for time-harmonic acoustics. *The Journal of the Acoustical Society of America*, 119(3):1315–1330, 03 2006. 2

[61] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *Computer Vision and Pattern Regognition (CVPR)*, 2017. 3

[62] Dirk van Maercke and Jacques Martin. The prediction of echograms and impulse responses within the epidaure software. *Applied Acoustics*, 38(2):93–114, 1993. 2, 3

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 5

[64] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for tasks in tracking and identifying humans from real room acoustics. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2023. 2

[65] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 1, 2, 3, 6, 7

[66] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Proceedings*, 2023. 6

[67] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1696–1704. Curran Associates, Inc., 2016. 3

[68] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021. 5