WavNAF: Learning Wave Propagation Priors for Neural Acoustic Fields

Taeho Kim
Sungkyunkwan University
kth0522@skku.edu

Hyunjun Kim
Sungkyunkwan University
arith@skku.edu

MinKyu Lee
Sungkyunkwan University
bluelati98@skku.edu

SuBeen Lee
Sungkyunkwan University
leesb7426@skku.edu

Inkyu An*
Kookmin University
inkyu.an@kookmin.ac.kr

Jae-Pil Heo*
Sungkyunkwan University
jaepilheo@skku.edu

Abstract

Room acoustics modeling requires capturing intricate wave phenomena beyond direct sound propagation, including reflections, refractions, and diffractions. Recent neural acoustic synthesis methods have progressively incorporated richer scene representations and physics-informed priors, yet typically learn acoustic behavior without considering wave propagation dynamics, which inherently capture diffraction and interference phenomena. We propose WavNAF, a framework that bridges wave dynamics and neural acoustic modeling by leveraging Finite-Difference Time-Domain (FDTD) simulations, which solve the wave equation on discrete spatiotemporal grids to capture these phenomena. However, full-scale FDTD is computationally prohibitive under CFL stability constraints. Inspired by traditional acoustic scale models, WavNAF constructs a geometry-preserving, scaled digital replica from NeRF-derived scene geometry and acoustically probes it with temporally compressed FDTD simulation. The resulting pressure maps are encoded not as direct full-scale Room Impulse Response (RIR) estimates, but as compressed-domain wave signatures that provide the neural acoustic field with a structured inductive bias beyond static geometry. To bridge the mismatch between these scaled simulation features and full-scale acoustic responses, we introduce a Neural Acoustic Scaling Module that adaptively recalibrates pressure-map features for full-scale RIR prediction. Experiments show that WavNAF yields consistent gains over prior neural baselines across standard acoustic metrics. [2]

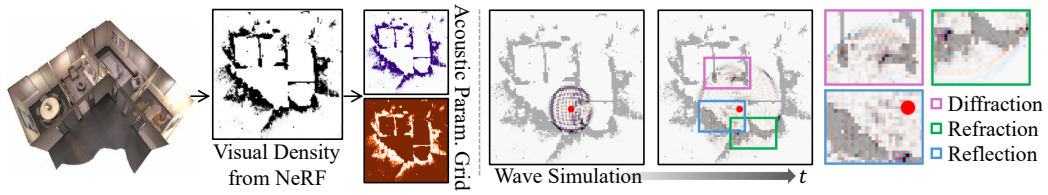


Figure 1: WavNAF converts a NeRF-derived visual density grid into approximate 2D acoustic parameter grids for scaled FDTD probing. The resulting pressure maps provide source-dependent wave-propagation signatures beyond static geometry and are used as conditioning signals rather than direct full-scale RIR estimates.

*Co-corresponding authors
[2]project page: https://kth0522.github.io/WavNAF_project_page/

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

1 Introduction

Room acoustics modeling requires capturing wave phenomena beyond direct sound propagation, including reflections, refractions, and diffractions. Recent neural acoustic synthesis methods have incorporated richer scene representations and physics-informed priors to improve acoustic realism [23, 26, 33, 6, 18, 22, 25]. However, they typically rely on static scene representations, visual observations, or geometric propagation paths, without explicitly exposing source-dependent wave-propagation dynamics such as diffraction-like spreading and interference.

To address this, we propose WavNAF, a framework that augments neural acoustic fields with wave-propagation cues from Finite-Difference Time-Domain (FDTD) simulation [5]. FDTD evolves pressure and velocity fields according to the wave equation, producing pressure maps that reveal structured propagation patterns such as wavefront evolution, interference, and multi-order boundary interactions. Since full-scale FDTD is computationally prohibitive under CFL stability constraints [10], WavNAF instead follows the logic of acoustic scale modeling [34, 3, 31]: it constructs a geometry-preserving scaled digital replica from NeRF-derived scene geometry [28] and probes it with short FDTD rollouts. The resulting pressure maps are not used as direct full-scale RIR estimates, but as compressed-domain wave-propagation signatures that provide source-dependent conditioning beyond static geometry.

WavNAF consists of three components: a NeRF-derived scaled acoustic replica with approximate acoustic parameters, a pressure-map encoder that extracts wave-propagation features, and a Neural Acoustic Scaling Module that recalibrates these compressed-domain features for full-scale RIR prediction. This formulation allows WavNAF to exploit structured wave cues from tractable simulations while leaving scale-, material-, and dimensionality-related discrepancies to the learned acoustic field.

2 Related Work

Implicit Neural Representations Early neural methods utilized implicit representations to model room acoustics. Neural Acoustic Fields (NAF) [26] directly mapped emitter-listener positions to impulse responses, enabling continuous spatial interpolation but with limited geometry awareness. INRAS [33] enhanced this by sampling bounce points from scene meshes for geometry-aware inference. However, these methods remain limited to approximating coarse reverberation without capturing wave physics such as diffraction and interference.

Image-guided Acoustic Modeling Visual information has been utilized to condition acoustic synthesis models. AV-NeRF [23] and NACF [24] employ intermediate visual renderings and bounce points from 3D meshes to inform acoustic predictions. While these approaches improve acoustic realism through visual scene context, relying on observations from specific viewpoints limits their ability to capture acoustic interactions beyond the rendered views.

Explicit Spatial Acoustic Representations Recent work leverages explicit 3D scene representations for acoustic modeling. AV-GS [4] employs Gaussian Splatting [19] with audio-guidance parameters, while NeRAF [6] utilizes volumetric features from NeRF [28]. However, these approaches encode static scene geometry without considering dynamic sound-geometry interactions, limiting their ability to capture complex wave phenomena.

Ray-based Acoustic Modeling Ray-based approaches model sound as rays using geometrical acoustics principles. SOAF [12] and AVR [22] model acoustic occlusion within volumetric frameworks, while AV-DAR [18] and π-AVAS [25] explicitly compute reflection paths through beam/ray tracing. However, the ray assumption inherently limits capturing wave phenomena such as diffraction and interference, motivating wave-based approaches where these behaviors naturally arise from the underlying physics.

3 Background

3.1 FDTD Wave Simulation

The wave equation is the fundamental partial differential equation governing how acoustic pressure evolves through space and time. Crucially, solving this equation inherently captures complex wave phenomena, including diffraction around obstacles, interference between wavefronts, and multi-order reflections, without requiring explicit modeling of each effect. The acoustic wave equation is formulated as [20]:

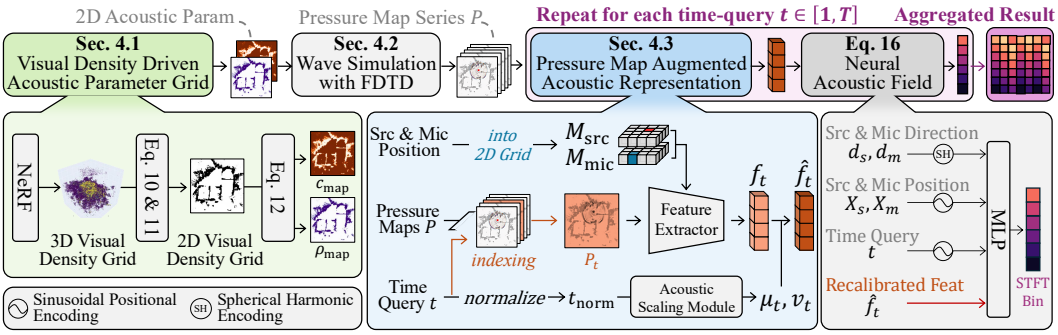


Figure 2: **Overview of WavNAF.** A 3D voxel density grid is first generated via NeRF and converted into 2D acoustic parameter grids. Using these grids, we simulate acoustic pressure maps through FDTD wave simulation. For a given time query, the corresponding pressure map is combined with one-hot encoded source and microphone positions, then processed by a feature extractor and scaled by our neural acoustic scaling module. Finally, an MLP predicts the STFT magnitude of the room impulse response from these scaled features.

$$\frac{\partial^2 p}{\partial t^2} = c^2 \nabla^2 p + f(X, t), \tag{1}$$

where $p(X, t)$ is acoustic pressure at spatial position $X = (x, y, z)$ and time t, c is the speed of sound, and $f(X, t)$ is the acoustic source term.

Discretization and Computational Scheme For computational efficiency, the wave equation can be solved in two dimensions using FDTD schemes [14]. The simulation domain is discretized into a spatial grid with indices (i, j) along the x and y axes, and time is discretized into steps indexed by n. The simulation iteratively updates pressure p^n and velocity $\mathbf{v}^n = (v_x^n, v_y^n)$ through the following steps (detailed derivation in Appendix B.4).

First, pressure gradients are computed via finite differences:

$$\nabla p \approx \left(\frac{p_{i+1,j} - p_{i,j}}{\Delta x}, \frac{p_{i,j+1} - p_{i,j}}{\Delta y} \right). \tag{2}$$

The velocity field is then updated using these gradients:

$$v_x^{n+1} = v_x^n - \frac{\Delta t}{\rho_{i,j}} \frac{\partial p}{\partial x}, \quad v_y^{n+1} = v_y^n - \frac{\Delta t}{\rho_{i,j}} \frac{\partial p}{\partial y} \tag{3}$$

Next, the velocity divergence is computed:

$$\nabla \cdot \mathbf{v} \approx \frac{v_{x,i,j} - v_{x,i-1,j}}{\Delta x} + \frac{v_{y,i,j} - v_{y,i,j-1}}{\Delta y}. \tag{4}$$

Finally, the pressure field is updated:

$$p^{n+1} = p^n - \rho_{i,j} c_{i,j}^2 \Delta t (\nabla \cdot \mathbf{v}), \tag{5}$$

where $\rho_{i,j}$ and $c_{i,j}$ denote acoustic density and speed of sound at grid cell (i, j), enabling spatially varying acoustic properties.

CFL Condition for Stability To ensure numerical stability and convergence, the Courant-Friedrichs-Lewy (CFL) condition [10] must be satisfied. This condition constrains the time step Δt relative to spatial discretization Δx and Δy, preventing numerical instabilities and maintaining physically accurate wave propagation. For anisotropic grids, where $\Delta x \neq \Delta y$, the CFL condition is:

$$\Delta t \leq \frac{\Delta x \Delta y}{c\sqrt{\Delta x^2 + \Delta y^2}}. \tag{6}$$

We choose the simulation step size as the maximum permissible value under this CFL condition, denoted as Δt_{CFL}.

3.2 Acoustic Scale Model

An acoustic scale model is a reduced physical replica of a full-scale room or auditorium used to study its acoustic behavior under controlled scaled conditions [3, 31, 21]. Its foundation is ideal

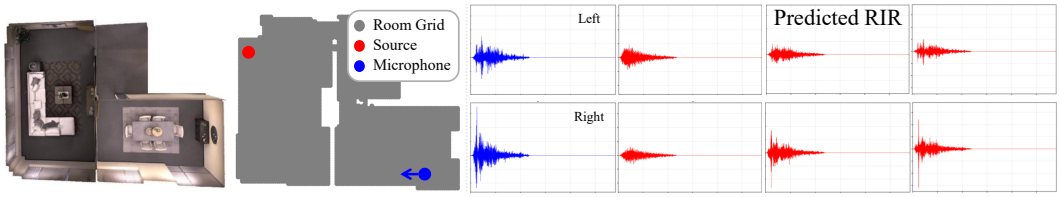


Figure 3: **Visualization of predicted RIRs in Apartment 1 of SoundSpaces.** Distant room setups are particularly challenging as sound must diffract through doorways and undergo multiple reflections before reaching the receiver. WavNAF better preserves the early stage spike characteristics, reflecting more accurate wave propagation behavior. Additional examples are provided in Appendix F.3

acoustic similarity: for the lossless wave equation, proportional scaling of space and time, e.g., $p_\lambda(X,t) = p(X/\lambda, t/\lambda)$, preserves the form of wave propagation when the source and boundary conditions are scaled consistently. This principle allows measurements in a miniature room to approximate the behavior of the original full-scale space after appropriate scale conversion [3, 31].

In practice, however, a physical scale model is better understood as a geometry-preserving acoustic replica than as a perfectly material-identical copy. Room layout and major wave-structure interactions are preserved under a controlled reduction ratio, while scale-dependent effects such as surface absorption, air attenuation, source/receiver responses, and frequency-dependent material behavior are only approximately matched through material selection, calibration, or correction [34, 3, 31]. These residual mismatches are not uniform over time: early propagation is largely governed by geometric similarity, whereas late reverberation accumulates repeated boundary interactions, absorption, scattering, and attenuation errors. Consequently, practical scale-model studies cannot rely on raw proportional scaling alone, especially for the reverberant tail [34, 31].

WavNAF transfers this scale-model perspective to a digital and learning-based setting. Instead of fabricating a physical miniature, we construct a NeRF-derived digital acoustic scale model: a geometry-preserving, normalized replica whose approximate acoustic parameters induce spatial contrasts for wave propagation. A short FDTD probe in this scaled domain produces pressure maps, which we encode as compressed-domain wave-propagation signatures rather than using them as direct full-scale RIR estimates [5, 14, 20]. These signatures provide structured cues about source-scene interactions, such as wavefront evolution, occlusion, boundary interaction, and diffraction-like spreading. This use of imperfect but structured simulation signals is aligned with physics-guided machine learning [17]. WavNAF later introduces a learned acoustic scaling stage to relate the encoded compressed-domain signatures to full-scale RIR representations under supervision.

4 Method

We propose **WavNAF**, a framework that integrates physics-informed wave propagation priors into neural RIR estimation through FDTD simulations (Figure 2). WavNAF derives acoustic parameters from visual scene geometry (Section 4.1), performs FDTD simulations to generate pressure maps capturing complex wave phenomena (Section 4.2), and employs an efficient Neural Acoustic Scaling Module to transform compressed simulations into full-scale RIRs (Section 4.3). This approach enables high-fidelity acoustic modeling while maintaining computational efficiency.

4.1 Acoustic Parameter Grid

To perform acoustic wave simulations with the FDTD method, each grid cell must define essential acoustic parameters describing local sound speeds and densities. However, these acoustic parameters are generally not directly available. To address this, we leverage NeRF to obtain spatially varying acoustic density information from visual data.

Neural Radiance Field The scene is represented by an N^3 voxel grid, with coordinates normalized to the range $[0, 1]$ using scene contraction [2], as implemented in Nerfacto [36]. NeRF encodes scene structures through a learned volumetric density function $\sigma(X)$. Querying this visual density at discrete spatial coordinates produces a density grid, which is used to define local acoustic parameters within our FDTD simulation.

Conversion from Visual Density Field to Acoustic Parameters. The NeRF density $\sigma(\mathbf{u})$ is defined over continuous normalized scene coordinates $\mathbf{u} \in [0, 1]^3$. To construct a discrete acoustic grid, we



Figure 4: **Loudness maps and learned acoustic representations.** (a) Nearest-receiver ground-truth and WavNAF-predicted loudness maps (brighter means louder). (b) Recalibrated wave features $\hat{f}_t$. (c) Last hidden-layer features of the NAF with wave or static 3D grid conditioning. In (c), wave conditioning reveals diffraction-like patterns resembling those in the ground-truth loudness map (a), whereas static grid conditioning shows horizontal and vertical stripe patterns. Colors in (b,c) are assigned independently using 1D t-SNE [27] and are not comparable across panels. Stars mark the sound source.

first query this density at voxel locations $\mathbf{u}_{i,j,k}$, yielding $\sigma_{i,j,k} = \sigma(\mathbf{u}_{i,j,k})$. We then convert the sampled density grid into alpha values:

$$\alpha_{3\mathrm{D}}[i,j,k] = 1 - \exp(-\sigma_{i,j,k} \cdot \delta), \tag{7}$$

where δ is a small positive constant set to 10^{-2}.

To obtain a 2D alpha map around the microphone height, we first convert the continuous microphone height z_{mic} into a discrete voxel index $k_{\mathrm{mic}} = \mathrm{round}(z_{\mathrm{mic}}(N-1))$. We then apply max pooling over a vertical index window:

$$\alpha_{2\mathrm{D}}[i,j] = \max_{k \in \mathcal{K}_{\mathrm{mic}}} \alpha_{3\mathrm{D}}[i,j,k], \quad \mathcal{K}_{\mathrm{mic}} = \{k : |k - k_{\mathrm{mic}}| \le \Delta k_z\}, \tag{8}$$

where indices are clipped to the valid range $[0, N-1]$. We use $\Delta k_z = 2$ voxels in all experiments. This operation captures dominant local geometric boundaries around the receiver plane, yielding a compact 2D layout for the subsequent scaled wave simulation.

The resulting 2D alpha grid $\alpha_{2\mathrm{D}} \in \mathbb{R}^{N \times N}$ is then mapped to acoustic parameter grids:

$$c_{\mathrm{map}}[i,j] = c_{\min} + (c_{\max} - c_{\min})(1 - \alpha_{2\mathrm{D}}[i,j]), \quad \rho_{\mathrm{map}}[i,j] = \rho_{\mathrm{air}}(1 + \alpha_{2\mathrm{D}}[i,j]), \tag{9}$$

where $c_{\max} = 343$ m/s, $c_{\min} = 0.9 \cdot c_{\max}$, and $\rho_{\mathrm{air}} = 1.21$ kg/m^3.

We do not intend this mapping to recover physically exact material parameters from visual density. Instead, following the acoustic scale-modeling perspective in Section 3.2, it provides an approximate acoustic parameterization for generating geometry-conditioned wave-propagation signatures in a tractable scaled domain. Table 5 further shows that WavNAF remains robust across a range of $c_{\min}/c_{\max}$ ratios, suggesting that the benefit mainly arises from the wave-structured prior induced by the FDTD update rules rather than from exact parameter fidelity.

4.2 FDTD Wave Simulation

Based on the acoustic parameter grids from Section 4.1, we simulate wave propagation using the Finite-Difference Time-Domain (FDTD) method. This simulation scheme iteratively updates the pressure and velocity fields to capture complex acoustic behaviors, as detailed in Algorithm 1. The spatial grid resolution is set to $N \times N$. The normalized coordinate space from Section 4.1 enables the simulation to handle diverse room sizes and aspect ratios in a unified manner. To determine the physical grid cell sizes $\Delta x, \Delta y$, we first empirically align audio sensor positions (i.e., source and microphone) to the normalized scene geometry. This alignment yields a normalization factor representing the full extent of the simulation domain in each spatial axis. We then compute Δx and Δy by dividing these spatial extents by the grid resolution N.

While the acoustic wave equation is theoretically scale-invariant under proportional spatial-temporal scaling, directly simulating full-scale wave behavior is computationally expensive. We therefore use the normalized scene as a tractable scaled probing domain, preserving coarse geometry while leaving scale-, material-, and dimensionality-related discrepancies to the learned acoustic field.

To mitigate boundary artifacts and emulate realistic wave energy dissipation, we introduce two damping mechanisms. Our simulation domain is finite, yet the boundaries should ideally represent

Algorithm 1 Acoustic Wave Simulation with FDTD

Input: Acoustic parameters $c, \rho, \gamma_s \in \mathbb{R}^{N\times N}$; source position $(x_{\text{src}}, y_{\text{src}}) \in [0,1]^2$; grid size $\Delta x, \Delta y$; time step Δt; initial amplitude A_0; global damping γ_g; frames T; step interval m
Output: Pressure map series $P \in \mathbb{R}^{T\times N\times N}$
Initialize $p^{(0)}, v_x^{(0)}, v_y^{(0)} \in \mathbb{R}^{N\times N} \leftarrow 0$
$i \leftarrow \text{round}(x_{\text{src}} \cdot (N-1)), \quad j \leftarrow \text{round}(y_{\text{src}} \cdot (N-1))$
$p^{(0)}[i,j] \leftarrow A_0, \quad k \leftarrow 0$
for $n = 1$ **to** $T \times m$ **do**
  $\nabla p \leftarrow \textsc{Gradient}(p^{(n-1)})$ ▷ Eq. 2
  $v^{(n)} \leftarrow \textsc{UpdateVelocity}(v^{(n-1)}, \nabla p)$ ▷ Eq. 3
  $\nabla \cdot v^{(n)} \leftarrow \textsc{Divergence}(v^{(n)})$ ▷ Eq. 4
  $p^{(n)} \leftarrow \textsc{UpdatePressure}(p^{(n-1)}, \nabla \cdot v^{(n)})$ ▷ Eq. 5
  $p^{(n)}, v^{(n)} \leftarrow \textsc{ApplyDamping}(p^{(n)}, v^{(n)})$ ▷ Eq. 10
  if $n \mod m = 0$ **then**
    $P[k] \leftarrow p^{(n)}, \quad k \leftarrow k+1$
  end if
end for
return P

open space extending to infinity rather than physical walls. Without proper treatment, waves reaching these computational boundaries produce spurious reflections that are artifacts of the finite domain, not real acoustic phenomena. To address this, we apply a sponge layer [7] that introduces spatially varying damping near the simulation boundaries. Waves entering the sponge layer experience rapid attenuation, effectively absorbing them before they can reflect back into the simulation domain. The damping coefficient $\gamma_{\text{s}}(i,j)$ transitions smoothly from a minimum 0.9 at the boundary to 1.0 toward the interior. Second, a global damping factor $\gamma_{\text{g}} = 0.99$ is uniformly applied across the domain to model natural air attenuation. These two damping mechanisms are combined at each timestep as follows:

$$p_{i,j}^{(n)} \leftarrow p_{i,j}^{(n)} \cdot \gamma_{\text{s}}(i,j) \cdot \gamma_{\text{g}}, \qquad v_{i,j}^{(n)} \leftarrow v_{i,j}^{(n)} \cdot \gamma_{\text{s}}(i,j) \cdot \gamma_{\text{g}}. \tag{10}$$

The resulting sequence of pressure maps $P \in \mathbb{R}^{T\times N\times N}$ provides source-dependent wave-structured signatures, which are used as physics-guided conditioning signals for the neural acoustic field.

4.3 Neural Acoustic Field

We leverage pressure maps from wave simulations as physically-informed priors for neural RIR estimation, enabling the model to learn representations that capture intricate wave phenomena.

The FDTD probe is intentionally short and scaled: its pressure maps are not direct samples of the target RIR timeline, but compressed-domain signatures of source-scene wave interactions. We therefore introduce a Neural Acoustic Scaling Module that learns a time-conditioned feature recalibration, mapping these low-fidelity wave signatures into conditioning features suitable for full-scale RIR prediction.

Wave Feature from Pressure Map for Learning Wave Propagation Prior

Given a time query $t \in [1, T]$, where T is the total number of STFT time bins, we select the pressure map $P_t \in \mathbb{R}^{N\times N}$ from generated pressure map series P. Additionally, one-hot encoded positional maps $M_{\text{src}}, M_{\text{mic}} \in \mathbb{R}^{N\times N}$ for source and microphone positions are concatenated as follows:

$$I_t = [P_t, M_{\text{src}}, M_{\text{mic}}]. \tag{11}$$

Crucially, the feature extraction network F_θ processes this input to learn representations $f_t = F_\theta(I_t)$ that serve as wave propagation priors, derived purely from pressure patterns without any explicit scene geometry information. The network learns these acoustic priors solely

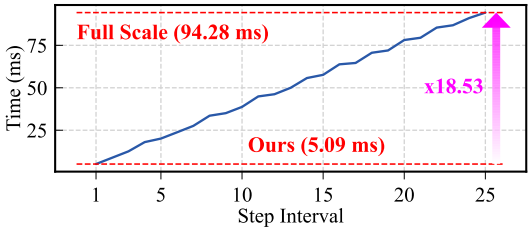


Figure 5: **Wave simulation computational cost per step interval.** Full-scale simulation leads to significant time overhead, while our acoustic scaling enables an order of magnitude faster inference.

from the pressure map that inherently encodes wave phenomena such as diffraction, interference, and multi-order reflections. This design choice is intentional: we aim to demonstrate that wave physics itself, as captured in pressure distributions, provides valuable priors for acoustic field learning beyond simple geometric structure.

Neural Acoustic Scaling Module. Following Section 3.2, FDTD pressure-map features are treated as compressed-domain wave signatures rather than direct full-scale RIR estimates. We recalibrate them using a FiLM-style time-conditioned affine modulation [29]:

$$\hat{f}_t = \mu_t \odot f_t + \nu_t, \tag{12}$$

where $\mu_t = \mathrm{MLP}_\mu(t_{\text{norm}})$ and $\nu_t = \mathrm{MLP}_\nu(t_{\text{norm}})$. Here, the modulation learns a cross-scale correction from compressed FDTD wave signatures to features suitable for full-scale RIR prediction.

Neural Impulse Response Estimation We predict impulse responses by directly estimating their STFTs with a unified feature-fusion MLP.

$$\mathrm{NAF}(\hat{f}_t, X_{\text{mic}}, d_{\text{mic}}, X_{\text{src}}, d_{\text{src}}, t_{\text{norm}}) \rightarrow \hat{s_t} \in \mathbb{R}^{C\times F}, \tag{13}$$

where $\hat{s_t}$ represents the STFT magnitude spectrum at time t, with C denoting the number of audio channels and F the number of frequency bins. Given microphone and source positions $X_{\text{mic}}, X_{\text{src}}$ and orientations $d_{\text{mic}}, d_{\text{src}}$, we apply sinusoidal positional encoding to position X and time t_{norm}, spherical harmonic encoding for direction d. Then, all elements including the pressure map feature $\hat{f}_t$ are concatenated into a single vector. This vector is processed by a feature-fusion MLP, similar to previous work [6], to predict the single STFT time bin $\hat{s_t}$. By iterating over all T queries, we assemble the full STFT $\hat{S} \in \mathbb{R}^{C\times F\times T}$. Finally, the Griffin-Lim algorithm [30, 13] reconstructs the full-scale waveform RIR form $\hat{S}$. Figure 4 visualizes the recalibrated wave features, the resulting hidden representations of the NAF, and the predicted loudness maps. Their qualitative characteristics are discussed in Section 5.4.

4.4 Learning Objective

Following prior works [23, 6], we train the visual model for sufficient iterations to establish a stable geometric encoding before incorporating acoustic model training. The loss function of visual model $\mathcal{L}_{\text{V}}$ is the same as original NeRF [28]:

$$\mathcal{L}_{\text{V}} = ||C(r) - \hat{C}(r)||_2^2, \tag{14}$$

where $C(r)$ is the ground-truth pixel color and $\hat{C}(r)$ is the rendered color at the ray r. Following the previous work [6], we use a combination of spectral loss $\mathcal{L}_{\text{SL}}$ [11] and spectral convergence loss $\mathcal{L}_{\text{SC}}$ [1] as our audio loss $\mathcal{L}_{\text{A}}$:

$$\begin{gathered}\mathcal{L}_{\text{SL}} = \left\|\log(|S|+\epsilon) - \log(|\hat{S}|+\epsilon)\right\|_2^2, \quad \mathcal{L}_{\text{SC}} = \frac{\||S| - |\hat{S}|\|_F}{\||S|\|_F}, \\ \mathcal{L}_{\text{A}} = \lambda_{\text{SL}}\mathcal{L}_{\text{SL}} + \lambda_{\text{SC}}\mathcal{L}_{\text{SC}},\end{gathered} \tag{15}$$

where S is the ground-truth STFT, $\hat{S}$ is the predicted STFT, $\|.\|_F$, $\|.\|_2$ denotes Frobenius norm and L2 norm and $\epsilon = 10^{-3}$. The final learning objective $\mathcal{L}$ can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{V}} + \lambda_{\text{A}}\mathcal{L}_{\text{A}}. \tag{16}$$

5 Experiments

5.1 Experimental Settings

5.1.1 Dataset

Following prior works [33, 6, 4], we compare our method against baselines on top of representative benchmarks specified below.

SoundSpaces [8] is a synthetic 3D acoustic simulator that provides stereo RIR (for discrete head orientations: 0, 90, 180, and 270) at the receiver positions of a 2D spatial grid. We follow our baselines [26, 33, 6, 23, 4] and employ 6 selected [32] indoor scenes with varying degrees of complexity: 2 scenes with a single rectangular room; 2 scenes with a single non-rectangular room;

Table 1: **Quantitative comparison with state-of-the-art methods on SoundSpaces and RAF datasets.** Lower values of T60, C50, EDT, and STFT error (↓) indicate higher RIR quality. Checkmarks (✓) in the GT column denote methods using ground-truth mesh information.

Methods	SoundSpaces				RAF				
	GT	T60 ↓	C50 ↓	EDT ↓	GT	T60 ↓	C50 ↓	EDT ↓	STFT error ↓
Opus-nearest		10.10	3.58	0.115		10.03	0.76	0.021	0.49
Opus-linear		8.64	3.13	0.097		10.19	0.86	0.029	0.92
AAC-nearest		9.35	1.67	0.059		22.83	1.97	0.064	1.04
AAC-linear		7.88	1.68	0.057		25.64	2.49	0.085	1.26
INRAS	✓	3.14	0.60	0.019	✓	8.01	0.79	0.025	**0.36**
NACF	✓	2.36	0.50	**0.014**	✓	**6.62**	**0.59**	**0.017**	0.39
NACF w/ T	✓	**2.17**	**0.49**	**0.014**	✓	7.31	**0.59**	0.018	0.39
NAF		3.18	1.06	0.031		10.08	0.71	0.021	0.64
AV-NeRF		2.47	0.57	0.016		8.11	0.73	0.021	0.39
NeRAF		2.14	0.38	0.010		7.47	0.61	0.020	0.17
WavNAF		**1.95**	**0.33**	**0.009**		**7.22**	**0.59**	**0.018**	**0.16**

and 2 scenes with a complex layout with multiple rooms. We use the same train/test splits as prior work [23, 4, 6].

Real Acoustic Fields (RAF) [9] is a multimodal real-world 3D RIR dataset consisting of 6DoF pose tracking data with quality RIR annotations and dense multiview images. Acoustic data are collected via a loudspeaker and a microphone system (earful tower) placed at multiple locations for RIR estimation, and visual data are captured with a camera rig (eyeful tower) for multiview image synthesis. RAF has two scenes: Furnished Room and Empty Room, which share the same room but differ in furniture.

5.1.2 Evaluation Metric

We evaluate each method using three standard acoustic metrics: **reverberation time (T60)** for late reverberation, **clarity (C50)** for acoustic clarity, and **early decay time (EDT)** for initial energy decay. We report errors with respect to ground-truth descriptors rather than raw descriptor values: T60 is reported as mean absolute percentage error (%), while C50 and EDT are reported as mean absolute errors in dB and seconds, respectively, averaged over test RIRs and channels. On RAF, we additionally report the **STFT error** [26, 11], computed as the mean absolute error between predicted and ground-truth log-magnitude STFTs. Lower is better for all metrics.

5.2 Implementation Details

All FDTD computations are optimized with CUDA kernels for GPU acceleration, ensuring efficient memory usage and parallel execution. We use Nerfacto [36] as our NeRF backbone and adopt ResNet50 [15] as the feature extractor for pressure map features (architecture details in Appendix B.1). The spatial resolution for acoustic parameter grids is set to 128×128. We set the step interval $m = 1$, corresponding to the maximum permissible time step under the CFL condition Δt_{CFL} detailed in Section 3. The source amplitude A_0 is initialized to 1.0 and exponentially attenuated according to the source–microphone height difference. (details in Appendix B.3.2). To minimize unnecessary reflections at simulation boundaries—outside of the room—we apply a sponge layer [7] characterized by a minimum damping coefficient $\gamma_{\text{min}} = 0.9$ (details in Appendix B.3) and additionally employ a global damping coefficient $\gamma_{\text{g}} = 0.99$ to simulate air attenuation effect. We set batch size $B = 1024$ STFT bins. Additional hyperparameters are provided in Appendix B.2.1.

5.3 Results

We compare WavNAF with traditional audio encoding methods (AAC [16], Opus [37]) and advanced neural acoustic synthesis models [26, 33, 23, 24, 6], grouping the latter according to whether they use ground-truth (GT) mesh information.

SoundSpaces WavNAF consistently outperforms all baselines across evaluation metrics, demonstrating superior acoustic realism (per-scene results in Appendix D.1). Notably, WavNAF achieves competitive or superior performance compared to methods that rely on ground-truth mesh information (INRAS, NACF), despite using only visual observations without explicit geometric annotations.

Table 2: Few-shot performance on RAF, averaged over Furnished and Empty Room scenes.

Training Fraction (%)	NeRAF				WavNAF			
	T60 ↓	C50 ↓	EDT ↓	STFT error ↓	T60 ↓	C50 ↓	EDT ↓	STFT error ↓
1	29.2295	3.1740	0.0919	0.2457	26.5940	2.7225	0.0811	0.2580
25	9.8157	0.8696	0.0266	0.1694	9.5127	0.7511	0.0236	0.1727
50	8.5270	0.7226	0.0241	0.1737	8.3505	0.6607	0.0206	0.1726
75	8.0687	0.6540	0.0210	0.1702	7.7141	0.6201	0.0192	0.1667

Table 3: **Feature conditioning ablation study** on SoundSpaces apartment 1.

Methods	T60 ↓	C50 ↓	EDT ↓
no condition	2.776	0.590	0.0151
3D grid feature	2.725	0.574	0.0149
only wave feature	2.406	0.467	0.0122
+ acoustic scale module	**2.306**	**0.446**	**0.0119**

Table 4: **Ablation study on z-axis max pooling.**

Δk_z	T60 ↓	C50 ↓	EDT ↓
0	2.352	0.470	0.0124
1	2.450	0.463	0.0121
2	**2.306**	**0.446**	**0.0119**
3	2.387	0.457	0.0121

Table 5: **Ablation study on $c_{\text{min}}/c_{\text{max}}$ ratio.**

$c_{\text{min}}/c_{\text{max}}$	T60 ↓	C50 ↓	EDT ↓
0.6	2.278	0.459	0.0123
0.7	**2.261**	0.467	0.0120
0.8	2.377	0.448	**0.0119**
0.9	2.306	**0.446**	**0.0119**

RAF WavNAF improves T60, EDT, and STFT error compared to neural methods without GT mesh data. These results confirm that WavNAF effectively leverages physically-informed wave propagation priors for acoustic synthesis.

Few-shot learning. Table 2 shows that WavNAF improves data efficiency under reduced supervision, consistently outperforming NeRAF on T60, C50, and EDT at the same training fraction. Notably, with 50% of RAF data, WavNAF achieves a lower EDT than NeRAF trained with 75% of the data (0.0206 vs. 0.0210). These results support our hypothesis that pressure-map features act as structured inductive biases that complement learned neural acoustic representations, particularly in low-data regimes.

5.4 Ablation Study

Effect of Wave Feature We examine the contributions of pressure-map features as wave propagation priors and the neural acoustic scaling module (Table 3). The *No condition* baseline uses only positional, directional, and temporal query information, while *3D grid feature* uses the static scene representation introduced in NeRAF [6]. The wave feature conditioning improves RIR quality across all evaluation metrics, and the neural acoustic scaling module provides further gains through time-conditioned feature recalibration. Figure 4 complements this ablation with qualitative visualizations of the predicted loudness and learned representations. WavNAF captures the broad spatial loudness pattern of the nearest-receiver ground truth (a), and the recalibrated wave features exhibit spatially organized variation (b). In the projected features of the last hidden layer (c), wave conditioning reveals diffraction-like spatial patterns that qualitatively resemble those in the ground-truth loudness map (a), whereas static 3D grid conditioning shows clear horizontal and vertical stripe patterns.

Effect of z-axis Max Pooling We investigate the influence of the vertical max pooling size Δk_z (Table 4). A setting $\Delta k_z = 0$ corresponds to using a single slice at the exact query height without any pooling. The results clearly demonstrate that employing an appropriate pooling size along the z-axis significantly enhances RIR quality, highlighting the benefit of aggregating the alpha grid along the vertical axis to better capture acoustic interactions and reduce surface-level noise.

Effect of Acoustic Parameters We investigate the impact of varying the acoustic speed ratio $c_{\text{min}}/c_{\text{max}}$ in our density-to-acoustic parameter mapping (Equation 9). Table 5 tests ratios from 0.6 to 0.9, where lower values create stronger acoustic contrasts between air and solid boundaries. Despite these variations representing different degrees of approximation, all settings consistently outperform the baseline. The results demonstrate that FDTD's physics-governed update process provides sufficient neural guidance for acoustic field learning, regardless of exact parameter values.

Simulation Length We examine how varying the simulation length through different step intervals affects performance. A step interval of m implies selecting every m-th pressure map, resulting in an effective temporal resolution of $m \cdot \Delta t_{\text{CFL}}$. A larger step interval increases the effective FDTD rollout horizon while keeping the number of stored pressure maps fixed. In Apartment 1, m=25 would roughly match the target STFT horizon, but is computationally expensive. We therefore evaluate short

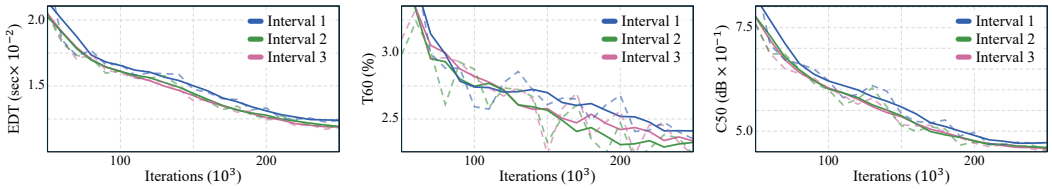


Figure 6: Comparison of metrics over iterations for different simulation lengths. This demonstrates that our method achieves robust results efficiently, even with temporally compressed simulations.

rollouts up to m=3. Figure 6 shows that performance is stable across these short rollouts, suggesting that early compressed probes already provide useful wave-structured conditioning.

6 Discussion

Limitation Despite the improvements from wave-propagation priors, several limitations remain. First, WavNAF currently uses 2D FDTD probes for computational efficiency, which cannot fully capture vertical propagation, floor/ceiling reflections, or height-dependent scattering. Extending the framework to efficient 3D or multi-plane wave probing is an important future direction. Second, our visual-to-acoustic mapping uses approximate, frequency-independent parameters derived from NeRF density, rather than explicit material absorption or scattering properties. Future work could incorporate material-aware or frequency-dependent acoustic parameterization to generate richer wave signatures while preserving the tractability of scaled probing.

Conclusion In this paper, we presented WavNAF, a neural acoustic synthesis framework that augments neural acoustic fields with wave-propagation priors from FDTD simulations. Rather than using the simulated pressure maps as physically exact full-scale RIR estimates, WavNAF treats them as compressed-domain acoustic signatures that encode source-dependent propagation patterns beyond static scene geometry. By deriving approximate acoustic parameter grids from NeRF-based visual geometry, WavNAF enables scaled wave probing of the scene and provides the acoustic field with structured cues such as boundary interactions, diffraction-like spreading, and multi-order propagation effects. Our neural acoustic scaling module further recalibrates these temporally compressed wave features for full-scale RIR prediction, reducing the need for direct long-horizon FDTD simulation. Experiments and ablations show that the resulting wave features provide more effective conditioning than static geometry features and lead to consistent improvements over prior neural acoustic field baselines.

Acknowledgements

This work was supported in part by MSIT/IITP (No. RS-2022-II220680, RS-2020-II201821, RS-2019-II190421, RS-2024-00459618, RS-2024-00360227, RS-2025-25442569, RS-2025-02219317), MSIT/NIPA (No. RS-2026-25621569), MSIT/NRF (No. RS-2024-00357729), and KNPA/KIPoT (No. RS-2025-25393280).

References

[1] S. Ö. Arık, H. Jun, and G. Diamos. Fast spectrogram inversion using multi-head convolutional neural networks. *IEEE Signal Processing Letters*, 26(1):94–98, 2018.

[2] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.

[3] M. Barron. The accuracy of acoustic scale modelling at 1: 50 scale. In *Institute of Acoustics*, 2002.

[4] S. Bhosale, H. Yang, D. Kanojia, J. Deng, and X. Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. *arXiv preprint arXiv:2406.08920*, 2024.

[5] D. Botteldooren. Finite-difference time-domain simulation of low-frequency room acoustic problems. *The Journal of the Acoustical Society of America*, 98(6):3302–3308, 1995.

[6] A. Brunetto, S. Hornauer, and F. Moutarde. Neraf: 3d scene infused neural radiance and acoustic fields. *arXiv preprint arXiv:2405.18213*, 2024.

[7] C. Cerjan, D. Kosloff, R. Kosloff, and M. Reshef. A nonreflecting boundary condition for discrete acoustic and elastic wave equations. *Geophysics*, 50(4):705–708, 1985.

[8] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *ECCV*, 2020.

[9] Z. Chen, I. D. Gebru, C. Richardt, A. Kumar, W. Laney, A. Owens, and A. 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.

[10] R. Courant, K. Friedrichs, and H. Lewy. Über die partiellen differenzengleichungen der mathematischen physik. *Mathematische annalen*, 100(1):32–74, 1928.

[11] A. Défossez, N. Zeghidour, N. Usunier, L. Bottou, and F. Bach. Sing: Symbol-to-instrument neural generator. *Advances in neural information processing systems*, 31, 2018.

[12] H. Gao, J. Ma, D. Ahmedt-Aristizabal, C. Nguyen, and M. Liu. Soaf: Scene occlusion-aware neural acoustic field. *arXiv preprint arXiv:2407.02264*, 2024.

[13] D. Griffin and J. Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[14] B. Hamilton. *Finite difference and finite volume methods for wave-based modelling of room acoustics*. PhD thesis, University of Edinburgh, UK, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] International Organization for Standardization. Advanced audio coding (aac). ISO/IEC 13818-7:2006, 2006.

[17] X. Jia, Y. Xie, S. Li, S. Chen, J. Zwart, J. Sadler, A. Appling, S. Oliver, and J. Read. Physics-guided machine learning from simulation data: An application in modeling lake and river systems. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 270–279. IEEE, 2021.

[18] D. Jin and R. Gao. Differentiable room acoustic rendering with multi-view vision priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[19] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[20] K. Kowalczyk and M. van Walstijn. Room acoustics simulation using 3-d compact explicit fdtd schemes. *IEEE Transactions on Audio, Speech, and Language Processing*, 1(1):1–14, 2009.

[21] H. Kuttruff and M. Vorländer. *Room acoustics*. Crc Press, 2024.

[22] Z. Lan, C. Zheng, Z. Zheng, and M. Zhao. Acoustic volume rendering for neural impulse response fields. *arXiv preprint arXiv:2411.06307*, 2024.

[23] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36: 37472–37490, 2023.

[24] S. Liang, C. Huang, Y. Tian, A. Kumar, and C. Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023.

[25] S. Liang, C. Huang, Y. Tang, Z. Zhang, and C. Xu. π -AVAS: Can physics-integrated audio-visual modeling boost neural acoustic synthesis? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13942–13951. IEEE, 2025.

[26] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022.

[27] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[28] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421, 2020.

[29] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[30] N. Perraudin, P. Balazs, and P. L. Søndergaard. A fast griffin-lim algorithm. In *2013 IEEE workshop on applications of signal processing to audio and acoustics*, pages 1–4. IEEE, 2013.

[31] J. H. Rindel. Room acoustic modelling techniques: A comparison of a scale model and a computer model for a new opera theatre. *Building Acoustics*, 18(3-4):259–280, 2011.

[32] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[33] K. Su, M. Chen, and E. Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022.

[34] K. Suzuki and T. Hidaka. A study on precise measurement of room impulse response in a scale model and auralization. In *International Symposium on Room Acoustics, Amsterdam (September 15–17, 2019)*, pages 297–303, 2019.

[35] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[36] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pages 1–12, 2023.

[37] Xiph.Org Foundation. Xiph opus. `https://opus-codec.org/`, 2012.

[38] L. Xu, V. Agrawal, W. Laney, T. Garcia, A. Bansal, C. Kim, S. Rota Bulò, L. Porzi, P. Kontschieder, A. Božič, et al. Vr-nerf: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

A Task Definition and Dataset Details

A.1 Task Definition

The task of WavNAF is to synthesize a high-fidelity Room Impulse Response (RIR) from a 3D scene represented by multi-view images. Given the visual information and arbitrary positional queries for a sound source and receiver, the model is trained to predict the corresponding RIR waveform that captures complex acoustic phenomena.

A.2 Required Inputs

The WavNAF model requires the following inputs to generate an RIR:

- **Multi-view RGB images with known camera poses:** A set of images capturing the 3D scene from various viewpoints, used for visual reconstruction via Neural Radiance Fields (NeRF).
- **Query source position** $(x_{\textbf{src}}, y_{\textbf{src}}, z_{\textbf{src}})$**:** The 3D coordinates of the sound source.
- **Query receiver position** $(x_{\textbf{mic}}, y_{\textbf{mic}}, z_{\textbf{mic}})$**:** The 3D coordinates of the receiver.
- **Query directions (Dataset-specific):** Directional information for the source or receiver, which varies by dataset. Specifically, source direction d_{src} is used for the RAF dataset, and receiver direction d_{mic} is used for the SoundSpaces dataset.

A.3 Output

The model's final output is a time-domain waveform of the Room Impulse Response (RIR). Internally, the model first predicts the log-magnitude Short-Time Fourier Transform (STFT) of the RIR, which is then converted to a waveform using the Griffin-Lim algorithm [30, 13]. The output format is dataset-specific:

- **SoundSpaces:** Predicts binaural (stereo) RIR with receiver directional information (d_{mic}) for discrete head orientations (0°, 90°, 180°, 270°).
- **RAF:** Predicts monaural RIRs with source directional information (d_{src}).

A.4 Datasets

A.4.1 SoundSpaces

Audio Data Following previous works [33, 23, 24, 6], we use 90% of the audio data for training and 10% for testing. The Room Impulse Responses (RIRs) provided by SoundSpaces are binaural and resampled from the original 44.1 kHz to 22.05 kHz. For Short-Time Fourier Transform (STFT), we employ 512 FFT bins, with a Hann window of 512 samples and a hop length of 128 samples. All sources and microphones lie at the same height. The source is omnidirectional.

Visual Data SoundSpaces provides visual data captured via Habitat Sim [35] with RGB images of resolution 512×512 and a field of view of 90 degrees. Images for training were sampled at varying numbers depending on the size of the environment: 45 images for small rooms, 75 images for medium rooms, and 150 images for large rooms. Each training image is captured from positions randomly selected at the edges of the room, oriented towards the center with random offsets. Additionally, 50 randomly sampled test poses are used to evaluate visual performance.

A.4.2 RAF

Audio Data Following previous works [9, 6], we use 80% of the audio data for training and the remaining 20% for testing. Audio clips are clipped to 0.32 seconds and sampled at 48 kHz. For STFT, we employ 1024 FFT bins, a Hann window of size 512, and a hop length of 256. The source and microphone can be located at different heights. The microphone is omnidirectional, and the source is directional.

Visual Data Visual data in RAF are multi-view images captured using the VR-NeRF [38] camera rig Eyeful Tower. From the available 22 camera views, we select images from cameras 20 to 25 to ensure adequate overlap. Images have a resolution of 684×1024 pixels. To limit the amount of visual training data, one-third of these images are randomly subsampled, with 90% of the selected images used for training and 10% reserved for evaluation.

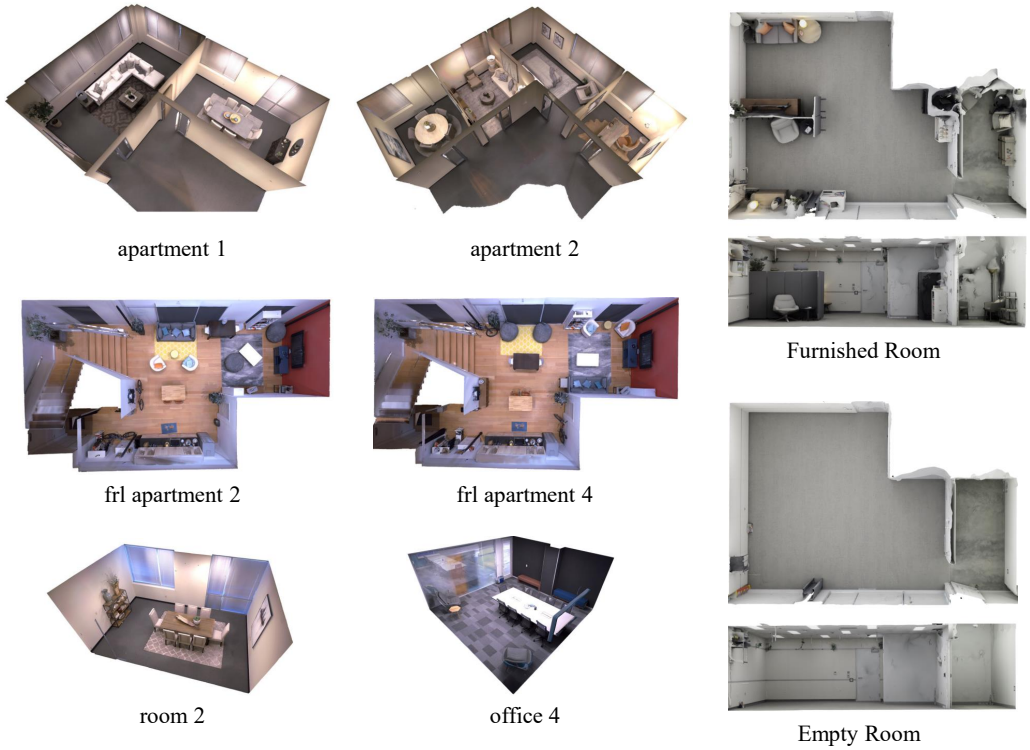


Figure 7: Overview of the used datasets. Our benchmarks span both synthetic and real-world domains, covering diverse room scales and layouts.

B Additional Method Details

B.1 Architecture Detail

B.1.1 Neural Acoustic Scaling Module

The Acoustic Scaling Layer adjusts simulation-derived features to account for temporal scaling variations. It consists of two separate MLP branches, μ and ν, each comprising two fully-connected layers with a hidden dimension of 128 and ReLU activation functions. Both branches receive the normalized time query as input.

B.1.2 Feature Fusion MLP

The Feature Fusion MLP architecture used in our model closely follows the design described in NeRAF [6]. Specifically, the network consists of two MLP blocks. The first MLP block takes as input a concatenation of encoded positional information, directional embeddings, and encoded wave features obtained via a ResNet. This block comprises 5 fully-connected layers, each followed by Leaky ReLU activations with a slope 0.1, resulting in an intermediate 512-dimensional representation. The second MLP block processes this intermediate representation to predict the STFT bins corresponding to time queries, featuring separate output heads per audio channel. The final layer utilizes a scaled tanh activation function to appropriately map the output to the log-magnitude STFT range.

B.2 Implementation Detail

B.2.1 Hyperparameters

We utilize the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-15}$. The initial learning rate is set to 10^{-4} and exponentially decreases to 10^{-8}. For the first 2,000 iterations, only the NeRF component is trained to allow adequate initialization. The training batch sizes are 4,096 rays for NeRF and 1,024 STFT bins for the audio model. The total training duration is 500,000 iterations, selecting the best-performing model across these iterations. These hyperparameters remain consistent across both SoundSpaces and RAF datasets.

B.2.2 Learning Objective Detail

We set $\lambda_{\text{A}} = 10^{-3}$, $\lambda_{\text{SC}} = 10^{-1}$, and empirically set λ_{SL} to 1 for SoundSpaces and 3 for RAF.

B.3 FDTD Wave Simulation Detail

B.3.1 Sponge Layer Detail.

We implement a sponge layer γ_{s} to mitigate boundary reflections during simulation. This sponge layer has a thickness of 8 grid points, where damping coefficients gradually increase from $\gamma_{\text{min}} = 0.9$ at the outer boundary to 1.0 towards the interior, following a quadratic profile.

B.3.2 Initial Amplitude Setting.

The initial amplitude A_0 for each batch is adjusted based on the vertical distance between the source and microphone positions. This adjustment accounts for the natural attenuation of sound energy with increasing vertical distance. Specifically, the amplitude A_0 is scaled exponentially by $\exp(-\alpha \cdot |z_{\text{src}} - z_{\text{mic}}|)$, where α is set to 0.01.

B.4 First-Order System Formulation for FDTD

The FDTD scheme presented in Equations (1)-(5) originates from the physically-motivated first-order system of the wave equation rather than direct discretization of the second-order form. We provide the complete mathematical derivation connecting the continuous wave equation to our discrete FDTD formulation.

B.4.1 From Second-Order to First-Order System

Starting from the acoustic wave equation in Equation (1), we reformulate it as a first-order system based on fundamental conservation laws:

Conservation of momentum (linearized Euler equation):

$$\rho_0 \frac{\partial \mathbf{v}}{\partial t} + \nabla p = 0 \quad \Rightarrow \quad \frac{\partial \mathbf{v}}{\partial t} = -\frac{1}{\rho} \nabla p \tag{17}$$

Conservation of mass with equation of state:

$$\frac{\partial p}{\partial t} = -\rho c^2 \nabla \cdot \mathbf{v} \tag{18}$$

where $\mathbf{v}$ is the particle velocity vector, p is acoustic pressure, ρ is the medium density, and c is the speed of sound.

B.4.2 Finite-Difference Discretization

We discretize this first-order system using staggered grids with spatial steps $\Delta x, \Delta y$ and time step Δt:

Step 1 - Pressure Gradient Computation:

$$\left.\frac{\partial p}{\partial x}\right|_{i+1/2,j} \approx \frac{p_{i+1,j} - p_{i,j}}{\Delta x}, \quad \left.\frac{\partial p}{\partial y}\right|_{i,j+1/2} \approx \frac{p_{i,j+1} - p_{i,j}}{\Delta y} \tag{19}$$

Step 2 - Velocity Update:

$$v_x^{n+1} = v_x^n - \frac{\Delta t}{\rho_{i,j}} \frac{\partial p}{\partial x}, \quad v_y^{n+1} = v_y^n - \frac{\Delta t}{\rho_{i,j}} \frac{\partial p}{\partial y} \tag{20}$$

Step 3 - Velocity Divergence:

$$\nabla \cdot \mathbf{v}|_{i,j} \approx \frac{v_{x,i,j} - v_{x,i-1,j}}{\Delta x} + \frac{v_{y,i,j} - v_{y,i,j-1}}{\Delta y} \tag{21}$$

Step 4 - Pressure Update:

$$p^{n+1} = p^n - \rho_{i,j} c_{i,j}^2 \Delta t (\nabla \cdot \mathbf{v}) \tag{22}$$

The spatially-varying parameters $c_{i,j}$ and $\rho_{i,j}$ arise naturally from discretizing the continuous parameters $c(\mathbf{X})$ and $\rho(\mathbf{X})$ onto the computational grid, enabling heterogeneous media modeling. This first-order system is mathematically equivalent to the original second-order wave equation but provides a computationally stable framework for time-stepping with proper boundary condition handling.

C Downsampled RIR Training

All RAF results in the main paper follow the original protocol of NeRAF [6], which uses the full-band 48 kHz RIRs, the 80/20 train/test split, and the NeRAF evaluation implementation (Appendix A.4.2). In contrast, recent geometry- and physics-aware methods evaluated on RAF, namely AVR [22] and AV-DAR [18], train and evaluate on RIRs downsampled to 16 kHz, use their own train/test split, and compute acoustic metrics with different definitions. Their published numbers are therefore not directly comparable to Table 1. We retrain WavNAF under the official AV-DAR protocol: we use AV-DAR's 16 kHz preprocessing, its released train/test split, and its original evaluation code. All results in this section follow this protocol and are separate from the 48 kHz results reported in the main paper.

Under the 16 kHz protocol, each RIR is clipped to 0.32 s (5,120 samples). WavNAF predicts the log-magnitude STFT with 512 FFT bins, a Hann window of 256 samples, and a hop length of 128 samples, resulting in 40 STFT frames per RIR. Unless noted otherwise, the architecture follows the 48 kHz setting (Appendix B.1).

C.1 Evaluation Metrics

We use the four metrics of the AV-DAR evaluation code: Loudness, C50, EDT, and T60 errors. All metrics are computed per test RIR between the predicted RIR $\hat{h}$ and the ground-truth RIR h sampled at $f_s = 16$ kHz, and then averaged over the test set. Lower is better for all metrics. Let $E(t) = \sum_{\tau \geq t} h^2(\tau)$ denote the Schroeder backward-integrated energy decay curve of h, and let $L(t) = 10 \log_{10}\big(E(t)/E(0)\big)$ denote its level in dB.

- **Loudness error (dB).** The loudness of an RIR is its total energy in dB, $\mathcal{L}(h) = 10 \log_{10}\big(\frac{1}{f_s} \sum_t h^2(t)\big)$, and the error is $|\mathcal{L}(\hat{h}) - \mathcal{L}(h)|$. This metric measures whether the overall energy, *i.e.*, the attenuation between the source and the receiver, is correctly predicted.
- **C50 error (dB).** Clarity is the early-to-late energy ratio with a 50 ms boundary, $\mathrm{C50}(h) = 10 \log_{10}\big(\frac{E(0) - E(t_{50})}{E(t_{50})}\big)$, where t_{50} is the sample index corresponding to 50 ms. The error is $|\mathrm{C50}(\hat{h}) - \mathrm{C50}(h)|$.
- **EDT error (s).** The early decay time is the time for $L(t)$ to decay from 0 dB to -10 dB, extrapolated to a 60 dB decay (*i.e.*, multiplied by 6). The error is $|\mathrm{EDT}(\hat{h}) - \mathrm{EDT}(h)|$.
- **T60 error (relative).** The reverberation time is measured from the interval in which $L(t)$ decays from -5 dB to -25 dB, extrapolated to a 60 dB decay (*i.e.*, multiplied by 3). The error is the relative error $|\mathrm{T60}(\hat{h}) - \mathrm{T60}(h)|/\mathrm{T60}(h)$, reported as a fraction rather than a percentage.

Differences from the 48 kHz protocol. The main difference lies in T60. The NeRAF evaluation used in the main paper applies a 200 Hz high-pass filter to the RIR, estimates T60 from the -5 dB to -15 dB decay range, and reports the relative error as a percentage, whereas the AV-DAR evaluation uses the unfiltered RIR, the -5 dB to -25 dB decay range, and reports the relative error as a fraction. Hence, a T60 error of 0.0398 in this section corresponds to 3.98% but is not numerically comparable to the T60 errors in the main paper. C50 and EDT measure the same acoustic descriptors in both protocols, but their implementations differ in numerical thresholding and, for EDT, the treatment of the decay onset. The STFT error used in the main paper is not part of the AV-DAR evaluation and is replaced by the Loudness error.

C.2 Comparison with AVR and AV-DAR

Table 6 compares WavNAF with AVR [22] and AV-DAR [18] on the two RAF scenes using 100% of the official AV-DAR training split. WavNAF outperforms AVR and both AV-DAR variants on all four metrics in both scenes. The gap is particularly large against AVR, which, like WavNAF, does not require any explicit scene geometry: WavNAF reduces the C50, EDT, and T60 errors of AVR by 41–43%, 28–37%, and 43–51%, respectively, across the two rooms. The two AV-DAR variants are discussed in Appendix C.5.

Table 6: Comparison with AVR and AV-DAR on RAF under the 16 kHz AV-DAR protocol with 100% of the training split. All values are errors (lower is better); T60 error is relative (fraction). Checkmarks (✓) in the GT column denote methods using ground-truth mesh information.

Scene	Method	GT	Loudness ↓	C50 ↓	EDT ↓	T60 ↓
Furnished	AVR		1.4568	0.9738	0.0323	0.0696
	AV-DAR (polished GT mesh)	✓	1.3715	0.5807	0.0209	0.0510
	AV-DAR (NeRF-derived mesh)		1.3104	0.6872	0.0239	0.0544
	WavNAF		**1.1021**	**0.5780**	**0.0204**	**0.0398**
Empty	AVR		1.4463	1.0105	0.0328	0.0742
	AV-DAR (polished GT mesh)	✓	1.9019	0.6914	0.0278	0.0634
	AV-DAR (NeRF-derived mesh)		1.8977	0.7062	0.0294	0.0621
	WavNAF		**0.8130**	**0.5720**	**0.0237**	**0.0365**

C.3 Wave Feature Fidelity Control

The ablation in Section 5.4 shows that pressure-map conditioning outperforms static geometry conditioning, but it does not fully separate meaningful wave-propagation structure from the mere addition of a source-conditioned feature or extra model capacity. Note that the pressure-map encoder receives no geometry map directly; the scene geometry influences the wave feature only through the FDTD-evolved pressure field and the source/microphone indicator maps. To isolate the contribution of the physical alignment of the wave feature, we introduce two controls that perturb only the input of the FDTD simulation while keeping the network architecture, the feature dimensionality, and all other inputs unchanged:

- **Wrong source.** The pressure map is generated from a deterministic incorrect source position, while the correct source and microphone positions are retained in the remaining network inputs. This preserves a structured, non-static, source-dependent feature but breaks its physical alignment with the target RIR.
- **Free space.** The pressure map is generated from the correct source position after replacing the acoustic parameter grid derived from the NeRF density with an empty grid. This preserves direct free-space propagation from the correct source while removing geometry-dependent reflections and other room interactions.

Both controls are trained and evaluated under the same 16 kHz protocol on the RAF Furnished Room, and the default WavNAF in Table 7 is the same model as in Table 6.

As shown in Table 7, both controls degrade C50 and EDT, and the wrong-source control additionally degrades T60. Thus, providing a structured, non-static pressure map is insufficient by itself: the wave feature must remain aligned with the queried source, and removing the room geometry from

Table 7: Wave feature fidelity control on the RAF Furnished Room under the 16 kHz AV-DAR protocol. Only the pressure map fed to the encoder is altered; the network architecture and all other inputs are unchanged.

Pressure-map condition	Loudness ↓	C50 ↓	EDT ↓	T60 ↓
Wrong source	0.8701	0.6485	0.0231	0.0429
Free space (correct source, empty grid)	**0.8209**	0.6316	0.0225	**0.0386**
Correct source and room geometry (Ours)	1.1021	**0.5780**	**0.0204**	0.0398

the simulation also reduces clarity and early-decay accuracy. The consistent EDT degradation is particularly informative, since EDT reflects the initial energy decay of the RIR. This early region is shaped by the direct arrival and the earliest source–scene interactions, including low-order reflections and diffraction around nearby boundaries. When the pressure map is generated from an incorrect source or without the room geometry, the model loses these physically aligned early-propagation cues, which cannot be fully recovered from the remaining positional inputs. While these controls do not isolate diffraction or interference individually, they indicate that WavNAF benefits from source-aligned and geometry-conditioned wave-propagation structure, rather than merely from additional feature dimensionality or a generic source-conditioned spatial input.

C.4 Few-shot Learning

We further compare WavNAF with AV-DAR in a reduced-data setting. We randomly sample 10% of the official AV-DAR training split and use the same subset for both methods; the test split is unchanged. As in Appendix D.3, training and evaluation are performed within the same scene, so this setting measures learning from reduced acoustic supervision in a known room rather than generalization to unseen rooms.

Table 8: Few-shot learning on RAF under the 16 kHz AV-DAR protocol with 10% of the training split. Both methods use the same training subset.

Scene	Method	Loudness ↓	C50 ↓	EDT ↓	T60 ↓
Furnished	AV-DAR	1.3562	0.9228	**0.0332**	0.0742
	WavNAF	**1.1122**	**0.9036**	0.0336	**0.0724**
Empty	AV-DAR	2.3317	1.1270	0.0462	0.0955
	WavNAF	**1.0985**	**0.8675**	**0.0379**	**0.0647**

As shown in Table 8, WavNAF outperforms AV-DAR in seven of the eight scene–metric comparisons. The only exception is the EDT error in the Furnished Room, where the difference is marginal (0.0336 vs. 0.0332). The gains are most pronounced in the Empty Room, where WavNAF reduces the Loudness, C50, EDT, and T60 errors of AV-DAR by 53%, 23%, 18%, and 32%, respectively. This is consistent with the 48 kHz few-shot results in Appendix D.3, indicating that the wave-propagation prior provides useful guidance under limited supervision.

C.5 Dependency on Explicit Scene Geometry

AV-DAR, similarly to INRAS [33] and NACF [24], assumes access to the ground-truth scene geometry. Especially, AV-DAR requires a polished mesh to perform beam tracing. WavNAF, in contrast, only uses multi-view images and derives its acoustic parameter grid from the NeRF density. To assess how much AV-DAR relies on this privileged geometry, we additionally evaluate AV-DAR with a mesh extracted from the NeRF learned by the visual component of WavNAF, instead of the polished ground-truth mesh. In this way, both methods rely on the same image-derived scene representation, and the comparison isolates how each method exploits it.

As shown in Table 6, replacing the polished ground-truth mesh with the NeRF-derived mesh degrades the C50 and EDT errors of AV-DAR in both rooms (Furnished: 0.5807 → 0.6872 and 0.0209 → 0.0239; Empty: 0.6914 → 0.7062 and 0.0278 → 0.0294), and the T60 error in the Furnished Room (0.0510 → 0.0544). This indicates that the explicit beam tracing of AV-DAR is sensitive to

the accuracy of the input geometry, especially for clarity and the early part of the RIR. WavNAF outperforms both AV-DAR variants on all metrics without requiring an explicit or polished mesh. We attribute this to the use of the reconstructed geometry as the medium of a short wave simulation rather than as exact reflecting surfaces: the pressure maps serve as compressed-domain source–scene interaction signatures, and the neural acoustic field together with the Neural Acoustic Scaling Module learns the residual mapping to the full-scale RIR, so that local geometric inaccuracies of the reconstruction do not directly translate into errors in the predicted reflection paths.

D Additional Experimental Results

D.1 Per-scene Results

This section provides detailed per-scene evaluation results. Table 9 shows results for individual scenes in the SoundSpaces dataset, Table 10 presents results for the RAF dataset.

Table 9: Per-scene results on SoundSpaces dataset

Method	scene	T60 ↓	C50 ↓	EDT ↓
WavNAF	Apartment 1	2.306	0.446	0.0119
	Apartment 2	2.400	0.478	0.0111
	FRL Apartment 2	2.294	0.323	0.0105
	FRL Apartment 4	2.717	0.262	0.0085
	Room 2	0.928	0.254	0.0084
	Office 4	1.056	0.239	0.0063
	Average	1.95	0.33	0.009

Table 10: Per-scene results on RAF dataset

Method	scene	T60 ↓	C50 ↓	EDT ↓	STFT error ↓
WavNAF	Furnished Room	6.850	0.594	0.0179	0.1693
	Empty Room	7.585	0.577	0.0180	0.1586
	Average	7.22	0.59	0.018	0.16

D.2 Computational Cost Analysis

This section provides a detailed computational cost analysis of our method, focusing on the FDTD simulation overhead and its practical implications for training and deployment.

D.2.1 Inference Time Breakdown

Figure 8 reports the per-RIR inference latency measured during demo audio generation. Each sample on the horizontal axis corresponds to one microphone pose along the trajectory, i.e., one predicted room impulse response (RIR). The measurements were collected on a workstation with an NVIDIA RTX PRO 6000 GPU and an Intel(R) Xeon(R) Gold 5416S CPU. The corresponding video demo is included in the supplementary material, with a representative RAF Furnished Room frame shown in Figure 9.

The breakdown decomposes the end-to-end RIR prediction time into the following components. `total` denotes the wall-clock time required to predict one RIR, from constructing the input batch to reconstructing the waveform. `batch_setup` measures the construction of audio batch tensors, including the source position, microphone position, source rotation for the current trajectory sample. `audio_forward` measures the wall-clock time for the WavNAF audio forward pass, including neural audio inference and the CUDA wave simulation. The `audio_forward_cuda_event` reports CUDA-event timing for the same forward section, excluding part of the Python-side overhead included in `audio_forward`. `raw_to_cpu` measures the transfer and conversion of the raw predicted STFT tensor from GPU memory to a CPU NumPy array. `magnitude` measures conversion from the predicted log-STFT representation to a clipped linear magnitude representation and transfer back to the selected device. Finally, `istft` measures the Griffin–Lim inverse STFT step used to reconstruct the waveform RIR from the predicted magnitude.

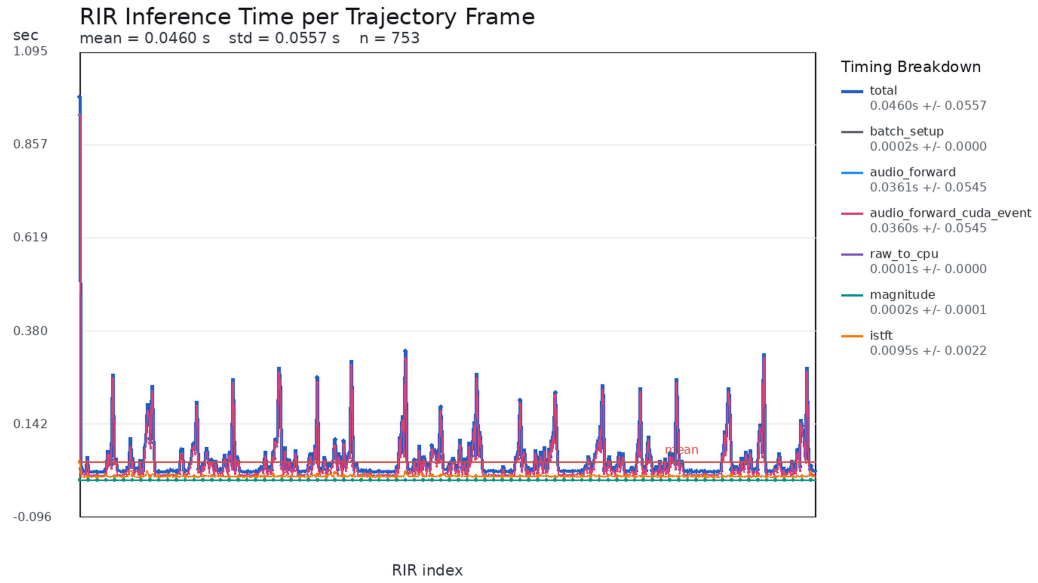


Figure 8: Inference cost analysis of supplementary demo at furnished room.

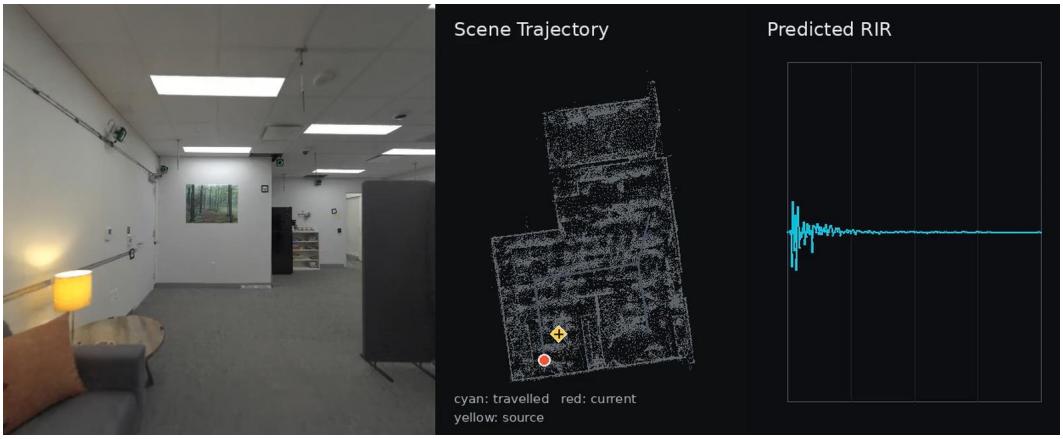


Figure 9: Representative frame from our RAF Furnished Room video demo. In the supplementary material, we provide demos for both Furnished Room and Empty Room scenes, each with two audio variations: a simple sound effect version and a music version. Each trajectory sample corresponds to one microphone pose and one predicted RIR.

D.3 Few-shot Learning

To evaluate the data efficiency of our physics-informed approach, we conduct experiments under reduced training data using both the SoundSpaces and RAF datasets. Table 11 presents the SoundSpaces results, while Table 2 reports the performance averaged over the two RAF scenes (Furnished Room and Empty Room).

Across both datasets, our model leads to consistently improved generalization in low-data regimes. On SoundSpaces, WavNAF trained with only 50% of the data achieves performance comparable to NeRAF trained with 75%. A similar trend is observed on RAF: with 50% of the data, WavNAF attains an EDT of 0.0206, slightly outperforming NeRAF even when the latter is trained with 75% of the data (EDT = 0.0210).

These consistent improvements across datasets and room configurations indicate that our physics-informed features complement, rather than replace, learned representations. The wave-based priors provide robust guidance under limited supervision, enabling stable reconstruction of both early and higher-order reflections even in challenging few-shot settings.

Table 11: Few-shot learning performance comparison at SoundSpaces apartment 1

Training Fraction (%)	NeRAF			WavNAF		
	T60 ↓	C50 ↓	EDT ↓	T60 ↓	C50 ↓	EDT ↓
25	5.378	0.853	0.0226	4.247	0.726	0.0200
50	3.869	0.686	0.0180	3.074	0.553	0.0152
75	2.954	0.610	0.0158	2.407	0.501	0.0137

E Hyperparameter Ablation Study

We conduct comprehensive ablation studies to validate the robustness of our approach to various hyperparameter choices. These experiments demonstrate that our physics-informed framework provides consistent improvements regardless of specific parameter settings.

E.1 Global Damping Coefficient

Table 12 analyzes the effect of the global damping coefficient γ_{g}, which models air absorption during wave propagation. Setting $\gamma_{\mathrm{g}} = 1.0$ represents no damping, while $\gamma_{\mathrm{g}} = 0.99$ introduces gradual energy dissipation. While the presence of damping slightly improves performance, our method maintains substantial advantages over baselines even without damping, confirming that the structural information from wave physics mainly drives our performance gains.

Table 12: Ablation study on global damping coefficient. Results at Apartment 1 of SoundSpaces.

Method	T60 ↓	C50 ↓	EDT ↓
without global damping ($\gamma_{\text{global}} = 1.0$)	2.416	0.456	**0.0118**
with global damping ($\gamma_{\text{global}} = 0.99$)	**2.306**	**0.446**	0.0119

F Visualizations

F.1 Wave Simulation Visualization

We visualize wave simulation results for scenes with complex, multi-room layouts from the SoundSpaces dataset. A subset of frames from the full simulation sequence is selected to illustrate the temporal evolution of the simulated pressure field. These visualizations qualitatively show structured wavefront propagation and boundary interactions over successive time steps, including multi-path propagation, reflections, and diffraction-like spreading around scene structures.

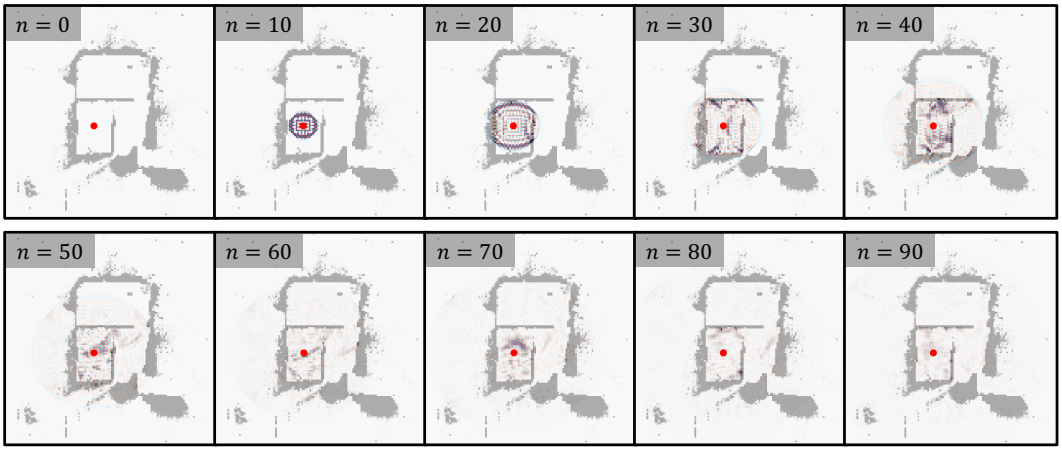
F.2 Neural Acoustic Scaling Module Visualization

We visualize the adaptive scaling behavior of our Neural Acoustic Scaling Module by plotting the scaling ratio $||\hat{f}_t||_2/||f_t||_2$, as a function of normalized time, where f_t represents the original feature extracted from the pressure map and $\hat{f}_t = \mu_t \odot f_t + \nu_t$ represents the scaled feature.

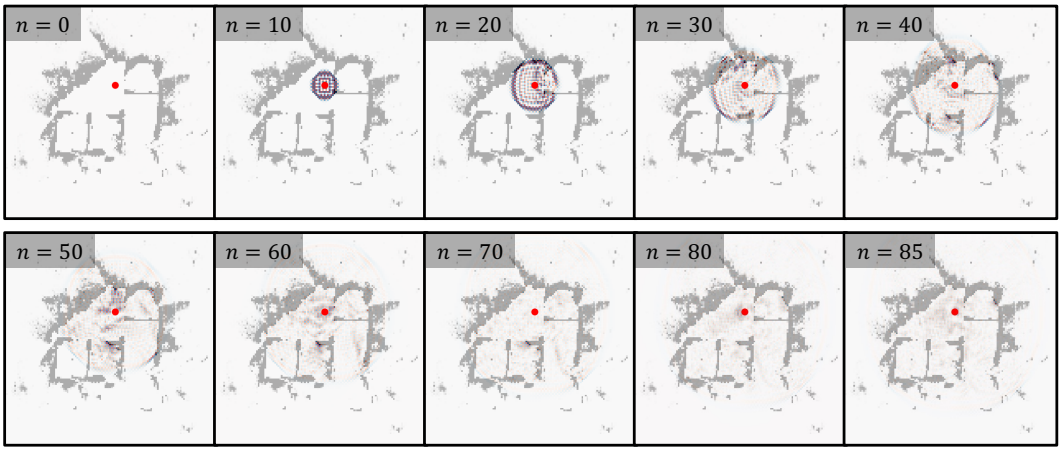
This visualization shows that the module applies time-dependent feature recalibration to compressed wave signatures. Rather than directly aligning FDTD timesteps with RIR samples, the module learns how much the wave-derived conditioning should be adjusted across the predicted RIR timeline.

F.3 RIR Visualization

We visualize the predicted RIRs under various source-microphone configurations in the largest scene of the SoundSpaces dataset, Apartment 1. This scene contains multiple interconnected rooms, allowing us to evaluate challenging setups where the source and microphone are positioned in different rooms. Even in a distant room separated by two doors, WavNAF accurately predicts the RIR, capturing the early-reflection spike characteristics more faithfully than competing methods. We additionally visualize the simulated pressure maps and the predicted loudness maps using the same



(a) Apartment 1



(b) Apartment 2

Figure 10: The gray regions visualize the scene geometry used to build the acoustic parameter grid, but are included only for readability. They are not fed to the wave-feature extractor, which receives only the pressure map and source/microphone indicator maps.

source positions as in Figure 12. The loudness map is obtained by fixing the source position and performing inference at all navigable points as receivers. As shown in Figures 13 and 14, noticeable differences between the simulated pressure maps and the predicted loudness maps are expected because the pressure maps are low-fidelity probing signals, while the loudness maps are decoded predictions after feature extraction, neural scaling, and RIR supervision.

F.4 Effect of source–microphone height differences

Figure 15 shows performance distribution with respect to source–microphone height differences in the Furnished Room scene of the RAF dataset.

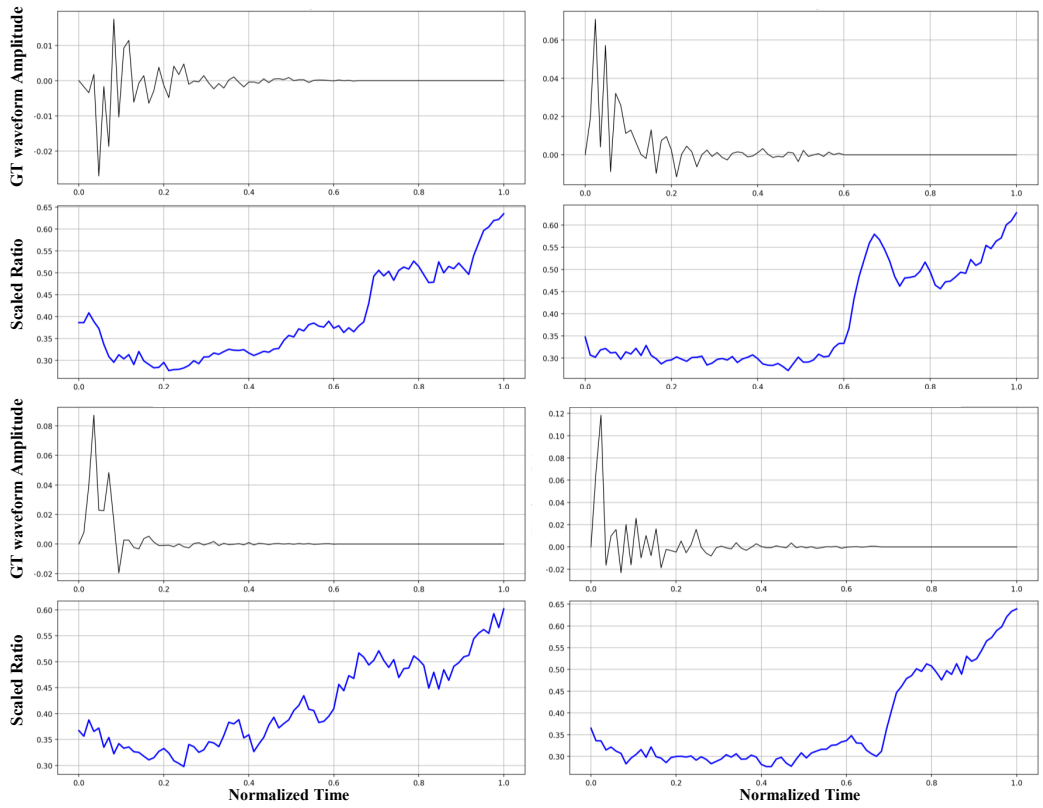


Figure 11: Visualization of the Neural Acoustic Scaling Module's adaptive behavior. The plots show ground-truth waveform amplitude (top row) and the scaling ratio $||\hat{f}_t||_2/||f_t||_2$ (bottom row) over normalized time. The scaling module learns to apply time-dependent transformations that vary between early reflections and late reverberation, adapting the features from temporally compressed simulations to match full-scale acoustic responses.

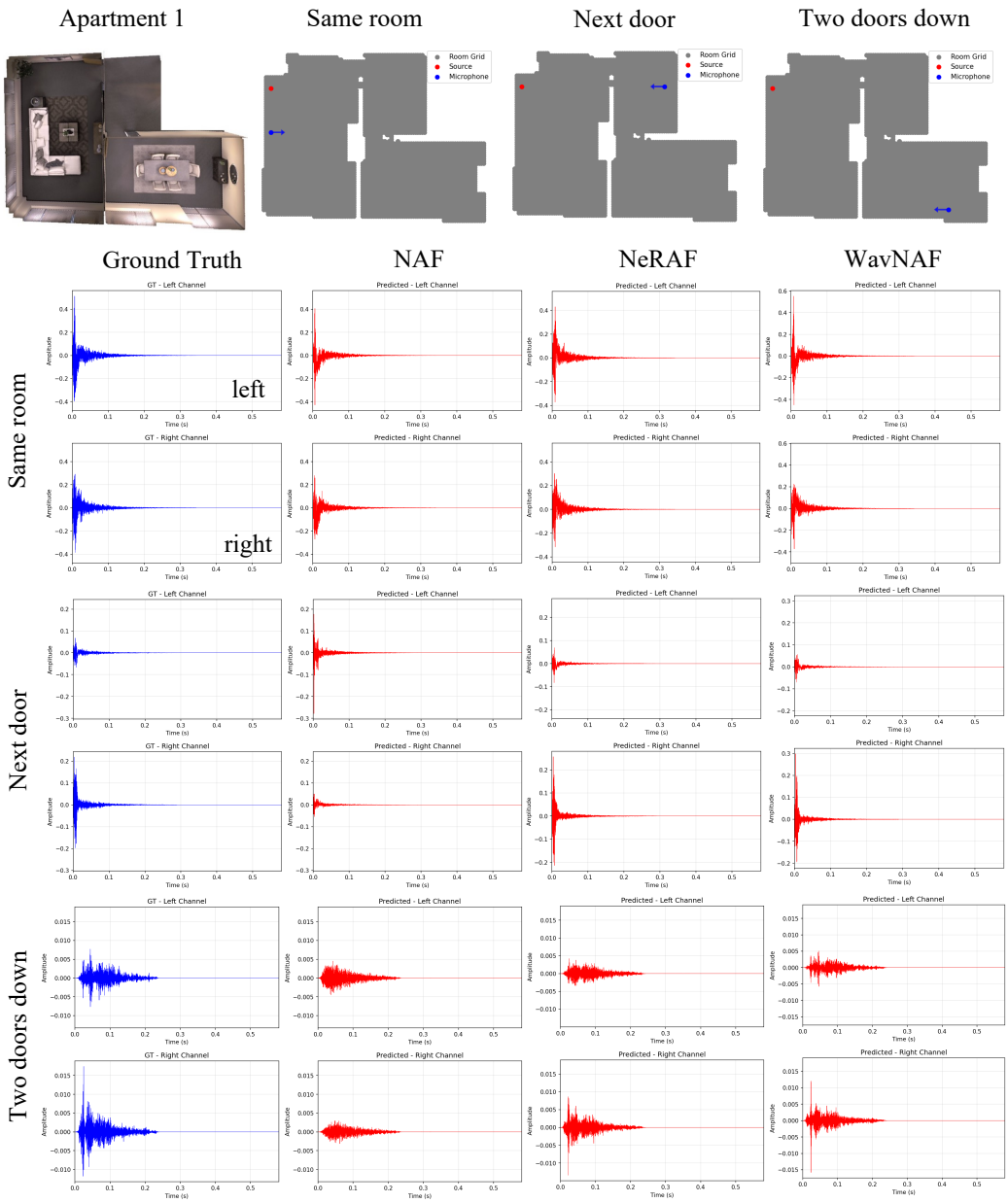


Figure 12: Predicted RIR visualization under various source-microphone configurations. Experiments are conducted in Apartment 1 of the SoundSpaces dataset. We visualize scenarios where the source and microphone are located in the same room as well as in different rooms. The farther and cross-room configurations are particularly challenging because the early part of the RIR involves multiple reflections through doors and walls. We compare WavNAF with NAF and NeRAF, and observe that WavNAF consistently outperforms both baselines across all settings. Notably, in distant room setups, WavNAF better preserves the early reflection spike characteristics, reflecting more accurate wave propagation behavior.

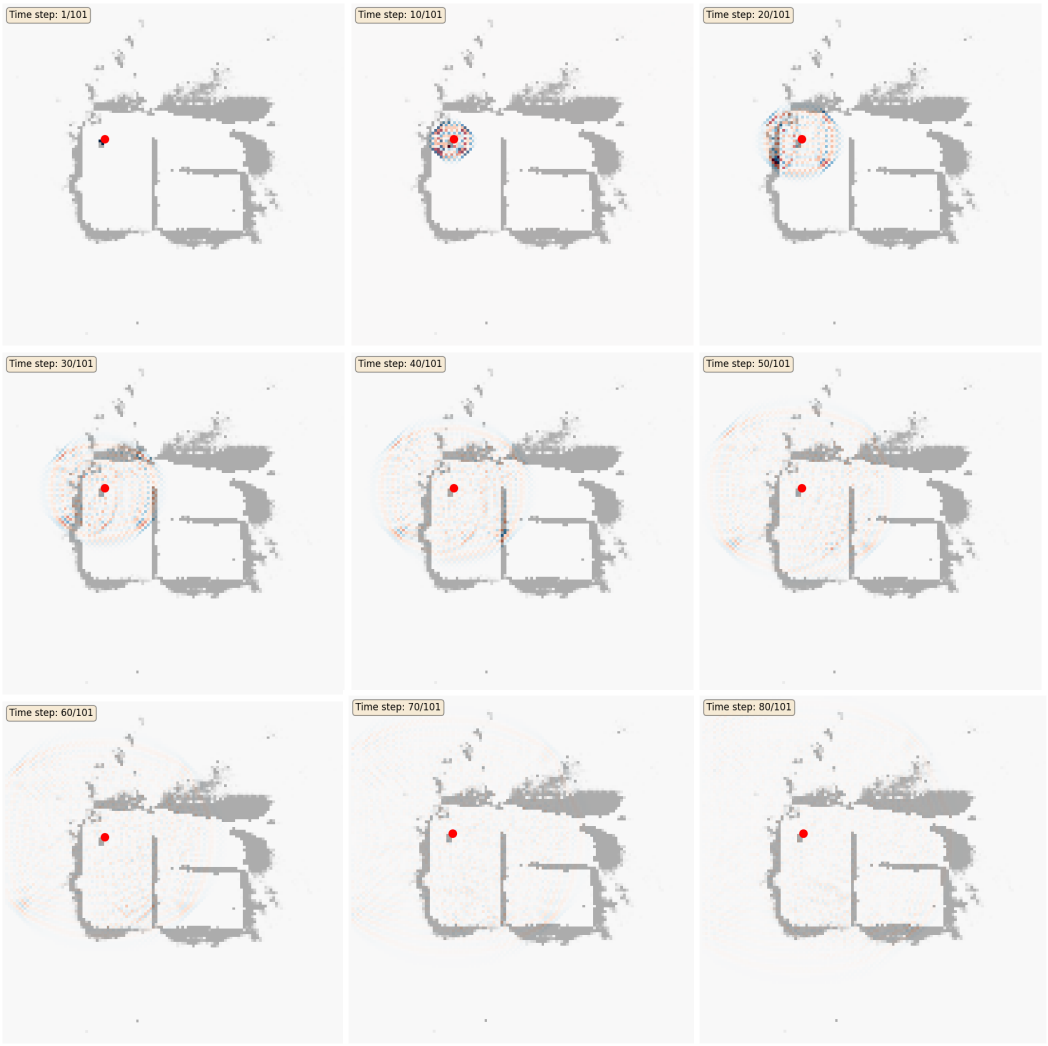


Figure 13: Visualization of simulated pressure maps. The source position is the same as in Figure 12.

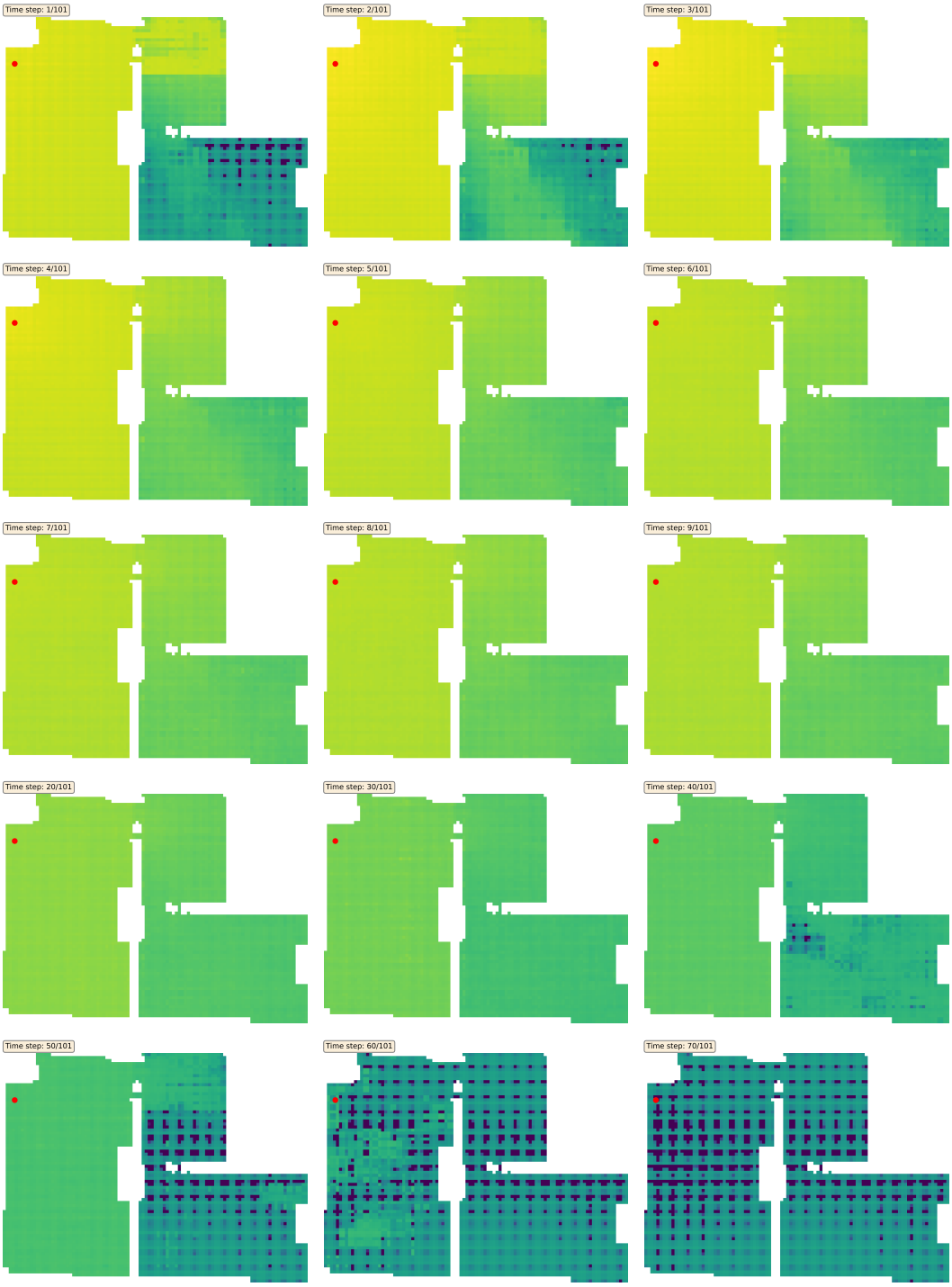


Figure 14: Visualization of loudness maps. The source position is the same as in Figure 12. The loudness map is obtained by fixing the source position and performing inference at all navigable points as receivers.

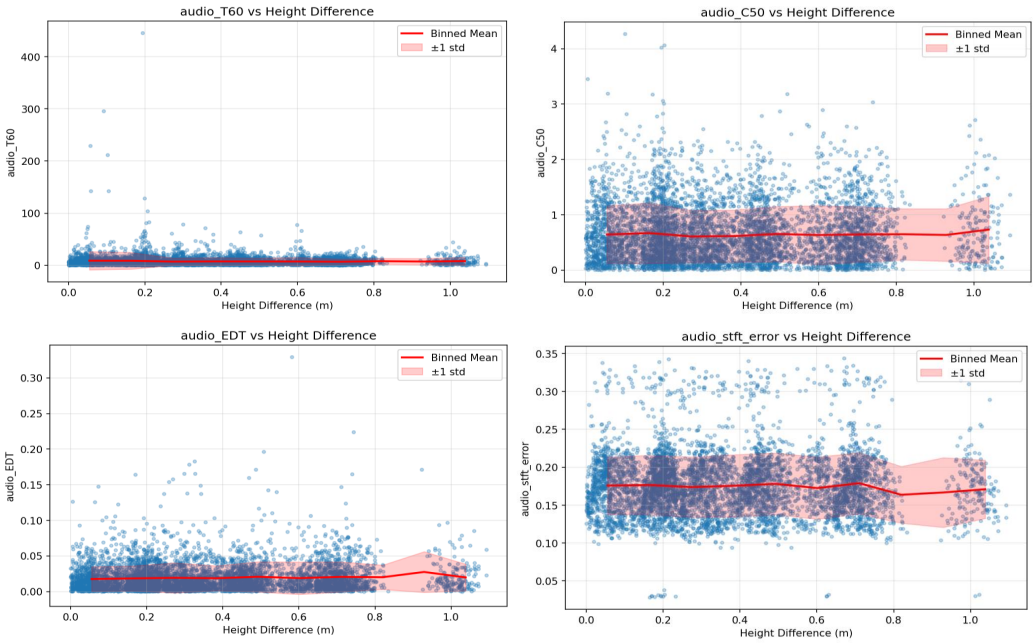


Figure 15: Performance distribution with respect to source–microphone height differences in the Furnished Room scene of the RAF dataset.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state that WavNAF uses compressed FDTD pressure maps as wave-propagation priors, rather than physically exact full-scale RIR estimates, and clearly scope the contribution to neural RIR prediction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses key limitations, including the use of 2D FDTD probes, missing vertical propagation and floor/ceiling effects, approximate visual-to-acoustic parameter mapping, and the need for future 3D or material-aware extensions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper provides the FDTD update formulation, CFL condition, and acoustic scale-modeling motivation in the main text, with detailed derivations and implementation details in the appendix.

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The paper describes the datasets, evaluation protocol, model components, acoustic simulation setup, training details, and key hyperparameters needed to reproduce the main experiments, with additional details in the appendix.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The paper uses public datasets, but the full training pipeline, model weights, and complete code are not released at submission time. We plan to release the full implementation after publication.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the datasets, train/test protocols, evaluation metrics, FDTD grid resolution, feature extractor, batch size, simulation parameters, and additional hyperparameters in the main text and appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: The main comparison follows the standard reporting protocol of prior neural acoustic field baselines, which report single benchmark results under fixed dataset splits. This preserves direct comparability with published numbers, but we do not report full multi-seed confidence intervals or formal statistical significance tests for all main metrics.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper reports implementation details and computational cost analysis, including FDTD simulation overhead and inference-time measurements, with additional details in the appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work uses public datasets and standard simulation/training procedures and does not involve human subjects.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: The work advances neural acoustic modeling for scene-conditioned RIR estimation. We do not anticipate direct harmful applications such as content manipulation or surveillance, since the method predicts room acoustic responses rather than generating semantic media content.

Guidelines: Our framework is fitted to a specific scene to predict its acoustic response and does not generate or alter media content, so we do not anticipate direct misuse risks such as content manipulation.

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release a high-risk generative model, scraped dataset, or dual-use system requiring special safeguards.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper cites the existing datasets, models, and software components used in the experiments, including SoundSpaces, RAF, NeRF/Nerfacto-related components, and prior neural acoustic field baselines.

Guidelines:

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

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce a new dataset or benchmark asset. The proposed method and implementation will be documented upon code release.

Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

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

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

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

Answer: [N/A]

Justification: The work does not involve human-subject research, so IRB approval is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were used only for editing, polishing, and visualization assistance, not for generating experimental results or making scientific decisions.

Guidelines:

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