

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 WAVNAF: LEARNING WAVE PROPAGATION PRIORS FOR NEURAL ACOUSTIC FIELDS

Anonymous authors

Paper under double-blind review

## ABSTRACT

Room acoustics modeling requires capturing intricate wave phenomena such as reflections, refractions, and diffractions beyond direct sound propagation. Recent neural acoustic synthesis methods have improved acoustic realism but typically focus only on straight sound paths and coarse reverberation, missing detailed interactions like diffraction or multi-order reflections. We propose WavNAF, a neural framework that leverages physically-informed wave propagation priors to explicitly capture complex acoustic interactions. We generate these priors by numerically solving the wave equation with the Finite-Difference Time-Domain (FDTD) method, which directly simulates wave-based acoustic behavior that geometric methods cannot capture. Specifically, we extract essential acoustic parameters for FDTD, such as wave speed and density, from visual scene geometry encoded by Neural Radiance Fields (NeRF). We then generate physically-informed pressure maps and encode them via a feature extractor to learn wave propagation priors that capture intricate acoustic phenomena. To address the inherent computational cost issue of FDTD, we introduce a novel Neural Acoustic Scaling Module, inspired by traditional acoustic scale model. This module adaptively recalibrates encoded pressure map features from temporally compressed simulations to efficiently estimate accurate full-scale Room Impulse Responses. Experimental results demonstrate that WavNAF achieves substantial improvements in acoustic quality across various evaluation metrics compared to existing state-of-the-art methods.

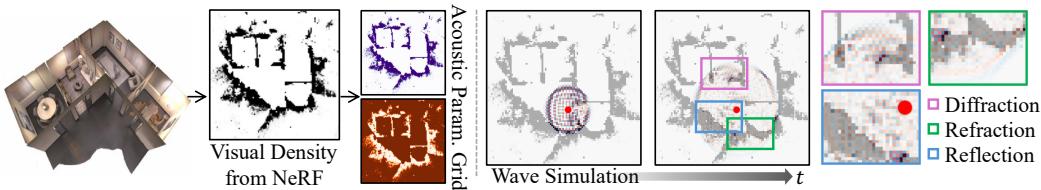


Figure 1: Our WavNAF framework uses Neural Radiance Fields (NeRF) to extract a normalized visual density grid encoding scene geometry, from which acoustic parameters, including sound speed and acoustic density, are directly derived. This enables wave-equation-based simulations that inherently capture complex acoustic phenomena, such as diffraction, refraction, and reflection, without explicit geometric modeling.

## 1 INTRODUCTION

Room acoustics modeling requires capturing intricate wave phenomena beyond direct sound propagation, including reflections, refractions, and diffractions. Traditional geometric methods (Krokstad et al., 1968; Cremer & Müller, 1948) efficiently handle direct paths but are inherently unable to model wave phenomena like diffraction due to their ray-based formulation. Recent neural approaches (Luo et al., 2022; Brunetto et al., 2024; Bhosale et al., 2024) have improved generalization and efficiency but remain limited to approximating coarse reverberation without capturing detailed wave physics.

We propose WavNAF, a framework that bridges physics-based simulation and neural acoustic modeling by leveraging wave propagation priors from Finite-Difference Time-Domain (FDTD) simulations (Botteldooren, 1995). FDTD numerically solves the wave equation on spatial grids, naturally capturing diffraction, interference, and multi-order reflections through physics-governed update rules.

054 However, integrating FDTD into neural acoustic modeling presents three fundamental challenges:  
 055 deriving scene geometry from visual data, effectively encoding wave physics for neural learning, and  
 056 managing computational bottlenecks under stringent stability constraints.

057 First, we derive essential acoustic parameters from visual scene geometry captured via NeRF (Milden-  
 058 hall et al., 2020). We extract volumetric density fields and transform them into 2D acoustic parameter  
 059 grids through vertical max-pooling, mapping visual opacity to spatially-varying sound speed and  
 060 acoustic density. This enables physics-informed wave simulations without explicit material annota-  
 061 tions, generating pressure maps that inherently encode complex wave phenomena.

062 Second, we effectively integrate FDTD-simulated pressure maps as robust wave propagation priors  
 063 into neural acoustic fields. Unlike previous methods that rely on geometric features or visual cues,  
 064 our approach directly learns from pressure maps through a dedicated feature extractor. By processing  
 065 these pressure maps that inherently capture diffraction, interference, and multi-order reflections  
 066 without requiring scene geometry, we demonstrate that the physics-governed wave patterns provide  
 067 strong inductive bias for acoustic field learning, achieving superior performance even with simplified  
 068 material parameters.

069 Third, we introduce a Neural Acoustic Scaling Module to alleviate computational bottlenecks in-  
 070 herent in standard FDTD methods. Inspired by traditional acoustic scale models (Suzuki & Hidaka,  
 071 2019; Barron, 2002; Rindel, 2011) where early reflections scale well but late reverberation deviates  
 072 systematically, our module learns adaptive transformations from temporally compressed simulations  
 073 to full-scale impulse responses, ensuring efficient and accurate representation while maintaining  
 074 acoustic fidelity.

## 076 2 RELATED WORK

077 **Implicit Neural Representations** Early neural methods for acoustic synthesis primarily utilized  
 078 implicit representations to model room acoustics. Neural Acoustic Fields (NAF) (Luo et al., 2022)  
 079 directly mapped emitter-listener positions to impulse responses, enabling continuous spatial interpola-  
 080 tion but providing limited geometry awareness. INRAS (Su et al., 2022) enhanced implicit modeling  
 081 by explicitly sampling bounce points from scene meshes, enabling geometry-aware inference of  
 082 reflections and reverberation. However, both methods typically have limitations to capture intricate  
 083 wave phenomena such as diffraction and complex reverberation patterns accurately.

084 **Image-guided Acoustic Modeling** Visual information has been effectively utilized to condition  
 085 acoustic synthesis models. AV-NeRF (Liang et al., 2023a) and Neural Acoustic Context  
 086 Field (NACF) (Liang et al., 2023b) represent prominent examples employing intermediate visual  
 087 renderings to inform acoustic predictions. Additionally, NACF explicitly leverages bounce points  
 088 sampled from 3D scene meshes, similar to INRAS (Su et al., 2022), to achieve geometry-aware  
 089 acoustic modeling. These approaches have significantly improved acoustic realism by integrating  
 090 visual scene context. Nevertheless, these visual rendering-based methods typically exhibit limitations  
 091 in fully capturing intricate wave interactions, particularly those involving acoustic interactions with  
 092 scene elements beyond rendered viewpoints.

093 **Explicit Spatial Acoustic Representations** Recent work such as AV-GS (Bhosale et al., 2024) and  
 094 NeRAF (Brunetto et al., 2024) leverage explicit 3D scene representations derived from advanced  
 095 visual modeling techniques to enrich acoustic modeling. AV-GS employs a Gaussian Splatting-  
 096 based (Kerbl et al., 2023) representation, using 3D Gaussian points augmented with audio-guidance  
 097 parameters to explicitly encode geometry and material characteristics. In contrast, NeRAF utilizes  
 098 volumetric features directly extracted from a NeRF (Mildenhall et al., 2020), capturing geometry  
 099 and appearance cues as voxel grids. These approaches encourage neural models to more explicitly  
 100 leverage local geometric information. However, they still lack explicit modeling of interactions  
 101 between scene geometry and acoustic wave propagation. Consequently, accurately capturing complex  
 102 wave phenomena remains challenging.

103 **Ray-casting and Acoustic Volume Rendering** Ray-based approaches like Scene Occlusion-Aware  
 104 Acoustic Fields (SOAF) (Gao et al., 2024) and Acoustic Volume Rendering (AVR) (Lan et al.,  
 105 2024) have enhanced acoustic realism by modeling acoustic occlusion and spatial variations within  
 106 continuous volumetric frameworks. These methods incorporate distance-based acoustic energy fields  
 107 and frequency-dependent transmittance characteristics to improve realism. Despite these advances,

108 there remains limitations in accurately modeling complex wave phenomena such as diffraction  
 109 around obstacles or interference patterns caused by multiple wavefront interactions, highlighting the  
 110 advantage of using wave-equation-based approaches that naturally capture these acoustic behaviors.  
 111

### 112 3 BACKGROUND

#### 114 3.1 WAVE SIMULATION WITH FINITE-DIFFERENCE TIME-DOMAIN (FDTD)

116 Acoustic wave propagation is formulated by the following wave equation (Kowalczyk & van Walstijn,  
 117 2009):

$$118 \quad \frac{\partial^2 p}{\partial t^2} = c^2 \nabla^2 p + f(X, t), \quad (1)$$

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

122 **Discretization and Computational Scheme** For computational efficiency, the above acoustic wave  
 123 equation can be solved in two dimensions using FDTD schemes (Hamilton, 2016). The simulation  
 124 domain is discretized into a spatial grid with indices  $(i, j)$  along the  $x$  and  $y$  axes, and time is  
 125 discretized into steps indexed by  $n$ . As a result, the simulation iteratively updates pressure  $p^n$  and  
 126 velocity  $v^n = (v_x^n, v_y^n)$  for each time step  $n$ , as follows (detailed derivation process in Appendix B.4):  
 127

$$128 \quad \text{Pressure Gradient: } \frac{\partial p}{\partial x} \approx \frac{p_{i+1,j} - p_{i,j}}{\Delta x}, \quad \frac{\partial p}{\partial y} \approx \frac{p_{i,j+1} - p_{i,j}}{\Delta y}, \quad (2)$$

$$130 \quad \text{Velocity Update: } v_x^{n+1} = v_x^n - \frac{\Delta t}{\rho_{i,j} \Delta x} \frac{\partial p}{\partial x}, \quad v_y^{n+1} = v_y^n - \frac{\Delta t}{\rho_{i,j} \Delta y} \frac{\partial p}{\partial y}, \quad (3)$$

$$132 \quad \text{Velocity Divergence: } \nabla \cdot v \approx \frac{v_{x,i,j} - v_{x,i-1,j}}{\Delta x} + \frac{v_{y,i,j} - v_{y,i,j-1}}{\Delta y}, \quad (4)$$

$$134 \quad \text{Pressure Update: } p^{n+1} = p^n - \rho_{i,j} c_{i,j}^2 \Delta t (\nabla \cdot v), \quad (5)$$

136 where terms  $\rho_{i,j}$  and  $c_{i,j}$  denotes acoustic density and speed of sound at the  $(i, j)$ -th grid cell.  
 137 By iteratively applying these update rules, the FDTD method naturally captures complex wave  
 138 phenomena including diffraction, interference, and scattering that arise from the underlying physics  
 139 of the wave equation, without requiring explicit modeling of each individual effect.

140 **CFL Condition for Stability** To ensure numerical stability and convergence, the Courant-Friedrichs-  
 141 Lewy (CFL) condition (Courant et al., 1928) must be satisfied. This condition constrains the time step  
 142  $\Delta t$  relative to spatial discretization  $\Delta x$  and  $\Delta y$ , preventing numerical instabilities and maintaining  
 143 physically accurate wave propagation. For anisotropic grids, where  $\Delta x \neq \Delta y$ , the CFL condition is:  
 144

$$145 \quad \Delta t \leq \frac{\Delta x \Delta y}{c \sqrt{\Delta x^2 + \Delta y^2}}. \quad (6)$$

147 We choose the simulation step size as the maximum permissible value under this CFL condition,  
 148 denoted as  $\Delta t_{\text{CFL}}$ .  
 149

#### 150 3.2 ACOUSTIC SCALE MODEL

151 The acoustic wave equation exhibits theoretical scale invariance under proportional scaling of spatial  
 152 and temporal dimensions. Specifically, consider scaling spatial and temporal variables by a factor  $\lambda$ ,  
 153 and define the scaled pressure field  $p'(X, t)$  as follows:  
 154

$$155 \quad p'(X, t) := p(X/\lambda, t/\lambda). \quad (7)$$

156 Under this definition, spatial and temporal derivatives transform as:  
 157

$$158 \quad \frac{\partial^2 p'}{\partial t'^2} = \frac{1}{\lambda^2} \frac{\partial^2 p}{\partial t^2}, \quad \nabla^2 p' = \frac{1}{\lambda^2} \nabla^2 p. \quad (8)$$

159 Substituting these relationships into the wave equation, we obtain:  
 160

$$161 \quad \frac{\partial^2 p'}{\partial t'^2} - c^2 \nabla^2 p' = \frac{1}{\lambda^2} \left( \frac{\partial^2 p}{\partial t^2} - c^2 \nabla^2 p \right) = 0. \quad (9)$$

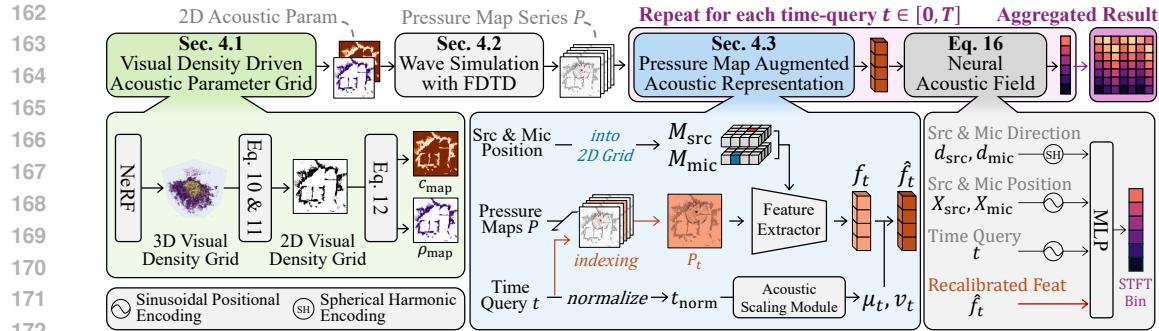


Figure 2: Overview of the WavNAF pipeline. A 3D voxel density grid is first generated via NeRF and converted into 2D acoustic parameter grids. Using these grids, we simulate acoustic pressure maps through FDTD wave simulation. For a given time query, the corresponding pressure map is combined with one-hot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

This demonstrates the theoretical invariance of the acoustic wave equation under simultaneous scaling of space and time. Such invariance motivates the use of acoustic scale models (Cremer & Müller, 1948; Beranek, 1992; Kuttruff & Vorländer, 2024), which are physically reduced versions of rooms or auditoriums designed to predict full-scale acoustic behavior by proportional scaling.

However, in practical scenarios, simple linear scaling fails to accurately predict full-scale acoustics due to frequency-dependent effects such as increased high-frequency air absorption, material differences, and measurement hardware limitations (Suzuki & Hidaka, 2019; Barron, 2002; Rindel, 2011). These practical constraints prevent straightforward linear scaling of acoustic responses, particularly affecting the accuracy of late reverberation. Therefore, accurately reconstructing full-scale acoustic responses from scaled-down simulations necessitates more sophisticated, non-linear acoustic scaling methods.

## 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 a 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  $[0, 1]$ , following NeRF’s scene-contraction method (Tancik et al., 2023). 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 visual density values  $\sigma(x, y, z)$  are first converted into alpha compositing values  $\alpha_{3D}$ :

$$\alpha_{3D}(x, y, z) = 1 - \exp(-\sigma(x, y, z) \cdot \delta), \quad (10)$$

216 where  $\delta$  is a small positive constant. To obtain a 2D alpha map at microphone height  $z_{\text{mic}}$ , we apply a  
 217 maximum pooling operation over a small vertical range  $\Delta z$  around the targeted  $z$ -coordinates:  
 218

$$\alpha_{2D}(x, y) = \max_{z \in [z_{\text{mic}} - \Delta z, z_{\text{mic}} + \Delta z]} \alpha_{3D}(x, y, z). \quad (11)$$

220 This vertical pooling procedure effectively smooths noisy surface details and implicitly captures  
 221 geometric influences along the  $z$ -axis. The resulting 2D alpha map  $\alpha_{2D} \in \mathbb{R}^{N \times N}$  is then mapped  
 222 onto acoustic parameters essential for FDTD simulations:  
 223

$$c_{\text{map}}(x, y) = c_{\text{min}} + (c_{\text{max}} - c_{\text{min}})(1 - \alpha_{2D}(x, y)), \quad \rho_{\text{map}}(x, y) = \rho_{\text{air}}(1 + \alpha_{2D}(x, y)), \quad (12)$$

225 where  $c_{\text{max}}$  is set to typical sound speed 343 m/s, and  $c_{\text{min}}$  is empirically chosen as  $0.9 \cdot c_{\text{max}}$ . The  
 226 parameter  $\rho_{\text{air}}$  represents the air density and set to  $1.21 \text{ kg/m}^3$ . While these mappings use empirically  
 227 chosen parameters, our method demonstrates robust performance across various parameter settings.  
 228 This robustness demonstrates that the physics-based update rules provide effective inductive bias  
 229 for learning wave propagation patterns, even when essential physical parameters are approximated.  
 230 Our physics-informed approach provides valuable neural guidance regardless of specific parameter  
 231 choices, consistently yielding superior results compared to baseline methods across a wide range of  
 232  $c_{\text{min}}/c_{\text{max}}$  ratios (detailed ablation studies in Appendix D).  
 233

## 4.2 FDTD WAVE SIMULATION

235 Based on the acoustic parameter grids from Section 4.1, we simulate wave propagation using the  
 236 Finite-Difference Time-Domain (FDTD) method. This simulation scheme iteratively updates the  
 237 pressure and velocity fields to capture complex acoustic behaviors, as detailed in Algorithm 1. The  
 238 spatial grid resolution is set to  $N \times N$ .  
 239

240 We adopt a scene normalization process based on NeRF’s scene contraction (Tancik et al., 2023).  
 241 This procedure maps the 3D scene geometry into a normalized coordinate space in the range  $[0, 1]^3$ ,  
 242 allowing the simulation to operate independently of the original room size or aspect ratio.  
 243

244 To determine the physical grid cell sizes  $\Delta x, \Delta y$ , we first empirically align audio sensor positions  
 245 (i.e., source and microphone) to the normalized scene geometry. This alignment yields a normalization  
 246 factor representing the full extent of the simulation domain in each spatial axis. We then compute  $\Delta x$   
 247 and  $\Delta y$  by dividing these spatial extents by the grid resolution  $N$ .  
 248

249 While the acoustic wave equation is theoretically scale-invariant under proportional spatial-temporal  
 250 scaling, directly simulating full-scale wave behavior is computationally expensive. Therefore, we  
 251 adopt a normalized, scaled-down simulation space that preserves physical fidelity through this  
 252 scale-invariance. This allows for more efficient simulation while maintaining acoustic fidelity.  
 253

254 To mitigate boundary artifacts and emulate realistic wave energy dissipation, we introduce two  
 255 damping mechanisms. First, a sponge layer (Cerjan et al., 1985) applies spatially varying damping  
 256 near the simulation boundaries to suppress spurious wave reflections. The damping coefficient  
 257  $\gamma_{\text{sponge}}(i, j)$  transitions smoothly from a minimum 0.9 at the boundary to 1.0 toward the interior.  
 258 Second, a global damping factor  $\gamma_{\text{global}} = 0.99$  is uniformly applied across the domain to model  
 259 natural air attenuation. These two damping mechanisms are combined at each timestep as follows:  
 260

$$p_{i,j}^{(n)} \leftarrow p_{i,j}^{(n)} \cdot \gamma_{\text{sponge}}(i, j) \cdot \gamma_{\text{global}}, \quad v_{i,j}^{(n)} \leftarrow v_{i,j}^{(n)} \cdot \gamma_{\text{sponge}}(i, j) \cdot \gamma_{\text{global}}. \quad (13)$$

261 The resulting sequence of pressure maps  $P \in \mathbb{R}^{T \times N \times N}$  explicitly encodes wave-based acoustic  
 262 interactions and serves as a physically grounded prior for the subsequent neural acoustic field model.  
 263

## 4.3 NEURAL ACOUSTIC FIELD

264 We leverage pressure maps from wave simulations as physically-informed priors for neural RIR  
 265 estimation, enabling the model to learn representations that capture intricate wave phenomena.  
 266

267 While FDTD effectively captures wave propagation, full-scale RIR simulation requires  $25\times$  more  
 268 steps overhead due to CFL constraints. Traditional acoustic scale modeling (Suzuki & Hidaka,  
 269 2019; Barron, 2002; Rindel, 2011) shows early reflections scale well but late reverberation deviates  
 270 systematically.

271 We introduce a neural acoustic scaling module that learns time-dependent transformations from  
 272 compressed simulations to full-scale responses. This adaptive approach bridges the gap between  
 273

270

**Algorithm 1:** Acoustic Wave Simulation with FDTD

271

**Data:** Acoustic parameter grids  $c_{\text{map}} \in \mathbb{R}^{N \times N}$ ,  $\rho_{\text{map}} \in \mathbb{R}^{N \times N}$ ,  $\gamma_{\text{sponge}} \in \mathbb{R}^{N \times N}$ ; Normalized source positions  $(x_{\text{src}}, y_{\text{src}}) \in [0, 1]^2$ ; Simulation parameters: grid cell size  $\Delta x, \Delta y$ , time step size  $\Delta t$ , initial source amplitude  $A_{\text{init}}$ , global damping coefficient  $\gamma_{\text{global}}$ , number of frames  $T$ , step interval  $m$  (default: 1)

272

**Result:** Series of pressure maps  $P[T, N, N]$

273

Initialize pressure field  $p^{(0)} \in \mathbb{R}^{N \times N}$ , velocity fields  $v_x^{(0)}, v_y^{(0)} \in \mathbb{R}^{N \times N} \leftarrow 0$ ;

274

Convert continuous source positions to discretized grid indices:;

275

$i \leftarrow \text{round}(x_{\text{src}} \times (N - 1)), j \leftarrow \text{round}(y_{\text{src}} \times (N - 1))$ ;

276

Set initial source amplitude:  $p^{(0)}[i, j] \leftarrow A_{\text{init}}$ ;

277

total\_steps  $\leftarrow T \times m$ ;

278

Initialize pressure map series  $P[T, N, N] \leftarrow 0$ ;

279

frame\_idx  $\leftarrow 0$ ;

280

**for**  $n = 1$  **to** total\_steps **do**

281

$\partial_x p, \partial_y p \leftarrow \text{Gradient}(p^{(n-1)}, \Delta x, \Delta y)$  (Eq. 2);

282

$v_x^{(n)}, v_y^{(n)} \leftarrow \text{UpdateVelocity}(v_x^{(n-1)}, v_y^{(n-1)}, \partial_x p, \partial_y p, \rho_{\text{map}}, \Delta t, \Delta x, \Delta y)$  (Eq. 3);

283

$\nabla \cdot v^{(n)} \leftarrow \text{Divergence}(v_x^{(n)}, v_y^{(n)}, \Delta x, \Delta y)$  (Eq. 4);

284

$p^{(n)} \leftarrow \text{UpdatePressure}(p^{(n-1)}, \nabla \cdot v^{(n)}, c_{\text{map}}, \rho_{\text{map}}, \Delta t)$  (Eq. 5);

285

$p^{(n)}, v_x^{(n)}, v_y^{(n)} \leftarrow \text{ApplyDamping}(p^{(n)}, v_x^{(n)}, v_y^{(n)}, \gamma_{\text{sponge}}, \gamma_{\text{global}})$  (Eq. 13);

286

**if**  $n \bmod m = 0$  **then**

287

$P[\text{frame\_idx}] \leftarrow p^{(n)}$ ;

288

    frame\_idx  $\leftarrow \text{frame\_idx} + 1$ ;

289

**end**

290

**end**

291

292

293

294

295

296

simulation intervals  $\Delta t_{\text{CFL}}$  and target STFT intervals  $\Delta t_{\text{STFT}}$ , enabling efficient and accurate RIR reconstruction.

297

**Wave Feature from Pressure Map for Learning Wave Propagation Prior** Given a time query  $t \in [0, T]$ , 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:

298

$$I_t = [P_t, M_{\text{src}}, M_{\text{mic}}]. \quad (14)$$

299

Crucially, the feature extraction network  $F_{\theta}$  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 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.

300

**Neural Acoustic Scaling Module for Temporal Alignment**

301

Due to the CFL stability condition, the FDTD time step  $\Delta t_{\text{CFL}}$  is much finer than the target STFT hop size  $\Delta t_{\text{STFT}}$ , which creates a systematic temporal mismatch between simulated and full-scale RIRs, especially in the late-reverberation region where cumulative energy-decay errors are amplified. To compensate for this, we introduce an adaptive neural scaling module that learns time-dependent transformations. Given a normalized time query  $t_{\text{norm}} \in [0, 1]$ , it recalibrates the extracted simulation features  $f_t$  as

302

$$\hat{f}_t = \mu_t \odot f_t + \nu_t, \quad (15)$$

303

where  $\mu_t = \text{MLP}_{\mu}(t_{\text{norm}})$  and  $\nu_t = \text{MLP}_{\nu}(t_{\text{norm}})$ , and  $\odot$  denotes element-wise multiplication. This time-varying affine transformation applies minimal correction to early reflections while allowing

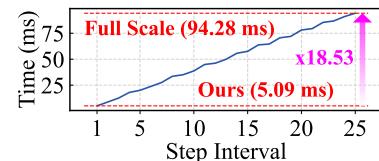


Figure 3: Wave simulation computational cost per step interval.

324 larger adjustment in the late tail, effectively aligning the temporal behavior of simulated features with  
 325 the target RIR representation and bridging the gap between  $\Delta t_{\text{CFL}}$  and  $\Delta t_{\text{STFT}}$  (see Appendix E.2).  
 326

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

$$329 \text{NAF}(\hat{f}_t, X_{\text{mic}}, d_{\text{mic}}, X_{\text{src}}, d_{\text{src}}, t_{\text{norm}}) \rightarrow \hat{s}_t \in \mathbb{R}^{C \times F}, \quad (16)$$

330 where  $\hat{s}_t$  represents the STFT magnitude spectrum at time  $t$ . Given microphone and source positions  
 331  $X_{\text{mic}}, X_{\text{src}}$  and orientations  $d_{\text{mic}}, d_{\text{src}}$ , we apply sinusoidal positional encoding for position  $X$  and  
 332 time  $t_{\text{norm}}$ , spherical harmonic encoding for direction  $d$ . Then, all elements including the pressure  
 333 map feature  $\hat{f}_t$  are concatenated into a single vector. This vector is processed by a feature-fusion MLP  
 334 architecture—similar to previous work (Brunetto et al., 2024)—to predict the single STFT time bin  
 335  $\hat{s}_t$ . 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  
 336 algorithm (Perraudin et al., 2013; Griffin & Lim, 1984) reconstructs the full-scale waveform RIR  
 337 form  $\hat{S}$ .  
 338

#### 339 4.4 LEARNING OBJECTIVE

340 Following prior works (Liang et al., 2023a; Brunetto et al., 2024), we train the visual model for  
 341 sufficient iterations to establish a stable geometric encoding before incorporating acoustic model  
 342 training. The loss function of visual model  $\mathcal{L}_{\text{V}}$  is the same as original NeRF (Mildenhall et al., 2020):  
 343

$$344 \mathcal{L}_{\text{V}} = \|C(r) - \hat{C}(r)\|_2^2, \quad (17)$$

345 where  $C(r)$  is the ground-truth pixel color and  $\hat{C}(r)$  is the rendered color at the ray  $r$ . Following the  
 346 previous work (Brunetto et al., 2024), we use a combination of spectral loss  $\mathcal{L}_{\text{SL}}$  (Défossez et al.,  
 347 2018a) and spectral convergence loss  $\mathcal{L}_{\text{SC}}$  (Arik et al., 2018) as our audio loss  $\mathcal{L}_{\text{A}}$ :  
 348

$$349 \mathcal{L}_{\text{SL}} = \|\log(|S| + \epsilon) - \log(|\hat{S}| + \epsilon)\|_2^2, \quad \mathcal{L}_{\text{SC}} = \frac{\| |S| - |\hat{S}| \|_F}{\| |S| \|_F}, \quad \mathcal{L}_{\text{A}} = \lambda_{\text{SL}} \mathcal{L}_{\text{SL}} + \lambda_{\text{SC}} \mathcal{L}_{\text{SC}}, \quad (18)$$

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

$$354 \mathcal{L} = \mathcal{L}_{\text{V}} + \lambda_{\text{A}} \mathcal{L}_{\text{A}}. \quad (19)$$

## 356 5 EXPERIMENTS

### 358 5.1 EXPERIMENTAL SETTINGS

#### 359 5.1.1 DATASET

361 Following prior works (Su et al., 2022; Brunetto et al., 2024; Bhosale et al., 2024), we compare our  
 362 method against baselines on top of representative benchmarks specified below.

363 **SoundSpaces** (Chen et al., 2020) is a synthetic 3D acoustic simulator that provides stereo RIR (for  
 364 discrete head orientations: 0, 90, 180, and 270) at the receiver positions of a 2D spatial grid. We  
 365 follow our baselines (Luo et al., 2022; Su et al., 2022; Brunetto et al., 2024; Liang et al., 2023a;  
 366 Bhosale et al., 2024) and employ 6 selected (Straub et al., 2019) indoor scenes with varying degrees of  
 367 complexity as: 2 scenes with a single rectangular room; 2 scenes with a single non-rectangular room;  
 368 and 2 scenes with a complex layout with multiple rooms. The train-validation splits are following the  
 369 conventions of prior works (Liang et al., 2023a; Bhosale et al., 2024; Brunetto et al., 2024).

370 **Real Acoustic Fields (RAF)** (Chen et al., 2024) is a multi-modal real-world 3D RIR dataset. It  
 371 consists of precise 6DoF pose tracking data paired with high-quality RIR annotations and dense  
 372 multi-view images. Acoustic data are collected via a pair of a loudspeaker and a microphone recording  
 373 system (earful tower) placed at different locations for RIR estimation, and visual data are captured  
 374 with a moving camera rig (eyeful tower) for multi-view image synthesis.

#### 375 5.1.2 EVALUATION METRIC

377 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 early reflections.

378  
 379 **Table 1: Quantitative comparison with state-of-the-art methods on SoundSpaces (left) and RAF**  
 380 **(right) datasets.** Lower values of T60, C50, EDT, and STFT error ( $\downarrow$ ) indicate higher RIR quality.  
 381 Checkmarks ( $\checkmark$ ) in the GT column denote methods using ground-truth mesh information.

382

SoundSpaces				RAF						
Methods	GT	T60 $\downarrow$	C50 $\downarrow$	EDT $\downarrow$	Methods	GT	T60 $\downarrow$	C50 $\downarrow$	EDT $\downarrow$	STFT error $\downarrow$
Opus-nearest		10.10	3.58	0.115	Opus-nearest		10.03	0.76	0.021	0.49
Opus-linear		8.64	3.13	0.097	Opus-linear		10.19	0.86	0.029	0.92
AAC-nearest		9.35	1.67	0.059	AAC-nearest		22.83	1.97	0.064	1.04
AAC-linear		7.88	1.68	0.057	AAC-linear		25.64	2.49	0.085	1.26
INRAS	$\checkmark$	3.14	0.60	0.019	INRAS	$\checkmark$	8.01	0.79	0.025	<b>0.36</b>
NACF	$\checkmark$	2.36	0.50	<b>0.014</b>	NACF	$\checkmark$	<b>6.62</b>	<b>0.59</b>	<b>0.017</b>	0.39
NACF w/ T	$\checkmark$	<b>2.17</b>	<b>0.49</b>	<b>0.014</b>	NACF w/ T	$\checkmark$	7.31	<b>0.59</b>	0.018	0.39
NAF		3.18	1.06	0.031	NAF		10.08	0.71	0.021	0.64
AV-NeRF		2.47	0.57	0.016	AV-NeRF		8.11	0.73	0.021	0.39
NeRAF		2.14	0.38	0.010	NeRAF		7.47	0.61	0.020	0.17
<b>WavNAF</b>		<b>1.95</b>	<b>0.33</b>	<b>0.009</b>	<b>WavNAF</b>		<b>7.43</b>	<b>0.61</b>	<b>0.019</b>	<b>0.16</b>

393 Specifically, T60 measures the time for the impulse response to decay by 60 dB, C50 quantifies the  
 394 energy ratio between the first 50ms and the remaining impulse response, and EDT assesses the initial  
 395 decay rate, closely aligning with human auditory perception. On the RAF dataset, we additionally  
 396 measures the **STFT error** (Luo et al., 2022; Défossez et al., 2018b), defined as the absolute error  
 397 between predicted and ground-truth log-magnitude STFTs.

## 399 5.2 IMPLEMENTATION DETAILS

400 Whole FDTD computations are optimized with CUDA kernels for GPU acceleration, ensuring  
 401 efficient memory usage and parallel execution. We use Nerfacto (Tancik et al., 2023) as our NeRF  
 402 backbone and adopt ResNet50 (He et al., 2016) as the feature extractor for pressure map features.  
 403 The spatial resolution for acoustic parameter grids is set to  $128 \times 128$ . We set the step interval  
 404  $m$  as 1, corresponding to the maximum permissible time step under the CFL condition  $\Delta t_{\text{CFL}}$   
 405 detailed in Section 3. The source amplitude  $A_{\text{init}}$  is initially set to 1.0 and subsequently adjusted  
 406 to be inversely proportional to the vertical distance between the source and microphone positions.  
 407 To minimize unnecessary reflections at simulation boundaries—outside of the room—we apply a  
 408 sponge layer (Cerjan et al., 1985) characterized by a minimum damping coefficient  $\gamma_{\text{min}} = 0.9$  and  
 409 additionally employ a global damping coefficient  $\gamma_{\text{global}} = 0.99$  to simulate realistic air attenuation  
 410 effect. We set batch size  $B = 1024$ .

## 412 5.3 RESULTS

413 We compare WavNAF with traditional audio encoding methods (AAC (International Organization for  
 414 Standardization, 2006), Opus (Xiph.Org Foundation, 2012)) and advanced neural acoustic synthesis  
 415 models (Luo et al., 2022; Su et al., 2022; Liang et al., 2023a;b; Brunetto et al., 2024). Neural models  
 416 are categorized into those utilizing ground-truth (GT) mesh information, such as INRAS and NACF,  
 417 and those without it.

418 On SoundSpaces, WavNAF consistently outperforms all baseline methods across evaluation metrics,  
 419 demonstrating superior acoustic realism.

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

## 424 5.4 ABLATION STUDY

426 **Effect of Wave Feature.** We analyze the impact of our key contributions: the pressure map feature as  
 427 a wave propagation prior and the neural acoustic scaling module (Table 2). The baseline *No condition*  
 428 includes only positional, directional, and temporal query information. The *3D grid feature* refers to the  
 429 static scene structure feature previously introduced in NeRAF (Brunetto et al., 2024). Introducing our  
 430 pressure map feature significantly enhances the RIR quality across all evaluation metrics. Furthermore,  
 431 the neural acoustic scaling module provides additional improvements, demonstrating the efficacy of  
 432 dynamically adapting simulation features to match full-scale acoustic conditions.

432

433  
434  
435  
Table 2: **Ablation study on feature conditioning**  
**method.** Performance difference between feature  
conditioning methods at SoundSpaces apartment 1

Methods	T60 (%) $\downarrow$	C50 (dB) $\downarrow$	EDT (sec) $\downarrow$
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
<b>+ acoustic scale module</b>	<b>2.306</b>	<b>0.446</b>	<b>0.0119</b>

436

**Effect of  $z$ -axis Max Pooling.** We investigate the influence of the vertical max pooling size  $\Delta z$  (Table 3). A setting  $\Delta z = 0$  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 alpha grid vertically to better capture acoustic interactions and reduce surface-level noise.

440

**Simulation Length.** We examine how varying the simulation length through different step intervals affects performance. Specifically, a step interval of 1 indicates that we utilize every simulation step, corresponding to a temporal resolution of  $\Delta t_{\text{CFL}}$ . Similarly, a step interval of  $m$  implies selecting every  $m$ -th pressure map, resulting in an effective temporal resolution of  $m \cdot \Delta t_{\text{CFL}}$ . The detailed simulation procedure for different step intervals is outlined in Algorithm 1. For the SoundSpaces apartment 1 scenario, a step interval of 25 corresponds to a full-length simulation. Due to the prohibitive computational cost associated with larger step intervals (longer simulation length), we limit our experiments to a maximum step interval of 3. Figure 4 compares the three main acoustic metrics across training iterations and simulation time for various step intervals. Increasing the step interval significantly raises the computational cost, by up to approximately 20 times (Figure 3). However, performance differences among these intervals remain minimal. These results indicate that our framework effectively learns the wave propagation prior, achieving comparable accuracy at substantially lower computational costs compared to conducting full-length simulations.

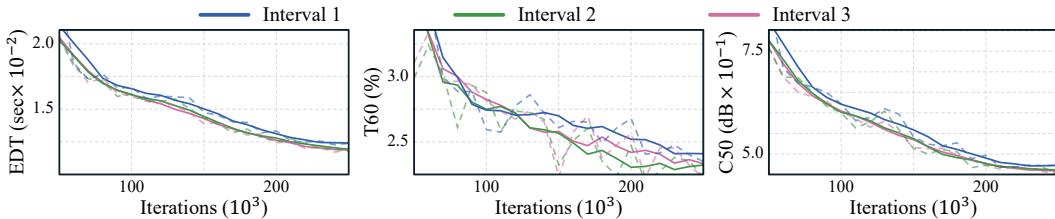


Figure 4: Comparison of metrics over iterations for different simulation lengths. This demonstrates that our method achieves robust results efficiently, even with temporally compressed simulations.

## 6 DISCUSSION

**Limitation** Despite the performance improvements achieved by our wave propagation prior learning strategy, there are several limitations. First, due to the computational demands, our current method utilizes only 2D wave simulations. Extending our framework to incorporate 3D simulations by including the vertical axis would likely yield additional performance gains. Second, WavNAF shows sensitivity to the alignment between visual and acoustic coordinate systems, and currently requires empirical tuning to ensure precise coordinate alignment. Future work should investigate systematic methods to mitigate these alignment sensitivities.

**Conclusion** In this paper, we presented WavNAF, a novel framework that effectively captures complex acoustic interactions by integrating FDTD wave simulations with neural acoustic fields. WavNAF leverages visual geometry extracted via NeRF to inform acoustic parameter grids, enabling modeling of intricate acoustic phenomena including reflection, refraction, and diffraction. Moreover, our neural acoustic scaling module significantly enhances computational efficiency, allowing accurate full-scale RIR estimation from temporally scaled simulations. Experimental results confirm that WavNAF achieves substantial improvements in acoustic quality, outperforming prior methods across multiple standard metrics. Future work includes expanding our framework to full 3D simulations and developing systematic methods to address visual-acoustic coordinate alignment sensitivities.

436  
437  
438  
439  
Table 3: **Ablation study on  $z$ -axis max pooling.**  
Performance difference between  $z$ -axis max  
pooling parameter at SoundSpaces apartment 1

$\Delta z$	T60 (%) $\downarrow$	C50 (dB) $\downarrow$	EDT (sec) $\downarrow$
0	2.352	0.470	0.0124
1	2.450	0.463	0.0121
2	<b>2.306</b>	<b>0.446</b>	<b>0.0119</b>
3	2.387	0.457	0.0121

## 486 7 REPRODUCIBILITY STATEMENT

488 Experimental settings for both training and evaluation are described in Sec 5.2. Detailed hyperparam-  
 489 eter settings and network configuration for each model variant are described in Appendix B.2.1 and  
 490 Appendix B.1. And detailed FDTD simulation parameters are described in Appendix B.3. We plan to  
 491 release the code for further reproducibility.

## 493 REFERENCES

495 Sercan Ö Arik, Heewoo Jun, and Gregory Diamos. Fast spectrogram inversion using multi-head  
 496 convolutional neural networks. *IEEE Signal Processing Letters*, 26(1):94–98, 2018.

497 Michael Barron. The accuracy of acoustic scale modelling at 1: 50 scale. In *Institute of Acoustics*,  
 498 2002.

500 Leo L Beranek. Music, acoustics, and architecture. *Bulletin of the American Academy of Arts and*  
 501 *Sciences*, 45(8):25–46, 1992.

502 Swapnil Bhosale, Haosen Yang, Diptesh Kanodia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learn-  
 503 ing material and geometry aware priors for novel view acoustic synthesis. *arXiv preprint*  
 504 *arXiv:2406.08920*, 2024.

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

508 Amandine Brunetto, Sascha Hornauer, and Fabien Moutarde. Neraf: 3d scene infused neural radiance  
 509 and acoustic fields. *arXiv preprint arXiv:2405.18213*, 2024.

511 Charles Cerjan, Dan Kosloff, Ronnie Kosloff, and Moshe Reshef. A nonreflecting boundary condition  
 512 for discrete acoustic and elastic wave equations. *Geophysics*, 50(4):705–708, 1985.

513 Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Kr-  
 514 ishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspace: Audio-visual navigaton in 3d  
 515 environments. In *ECCV*, 2020.

517 Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and  
 518 Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark.  
 519 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.  
 520 21886–21896, 2024.

521 Richard Courant, Kurt Friedrichs, and Hans Lewy. Über die partiellen differenzengleichungen der  
 522 mathematischen physik. *Mathematische annalen*, 100(1):32–74, 1928.

523 Lothar Cremer and Helmut A Müller. Die wissenschaftlichen grundlagen der raumakustik. 1948.

525 Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. Sing: Symbol-  
 526 to-instrument neural generator. *Advances in neural information processing systems*, 31, 2018a.

528 Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. Sing: Symbol-  
 529 to-instrument neural generator. *Advances in neural information processing systems*, 31, 2018b.

530 Huiyu Gao, Jiahao Ma, David Ahmedt-Aristizabal, Chuong Nguyen, and Miaomiao Liu. Soaf: Scene  
 531 occlusion-aware neural acoustic field. *arXiv preprint arXiv:2407.02264*, 2024.

533 Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE*  
 534 *Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

535 Brian Hamilton. *Finite difference and finite volume methods for wave-based modelling of room*  
 536 *acoustics*. PhD thesis, University of Edinburgh Edinburgh, UK, 2016.

538 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image  
 539 recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*,  
 pp. 770–778, 2016.

540 International Organization for Standardization. Advanced audio coding (aac). ISO/IEC 13818-7:2006,  
 541 2006.

542 Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting  
 543 for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

544 Konrad Kowalczyk and Maarten van Walstijn. Room acoustics simulation using 3-d compact explicit  
 545 fdtd schemes. *IEEE Transactions on Audio, Speech, and Language Processing*, 1(1):1–14, 2009.

546 Asbjørn Krokstad, Staffan Strom, and Svein Sørdsdal. Calculating the acoustical room response by  
 547 the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118–125, 1968.

548 Heinrich Kuttruff and Michael Vorländer. *Room acoustics*. Crc Press, 2024.

549 Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for  
 550 neural impulse response fields. *arXiv preprint arXiv:2411.06307*, 2024.

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

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

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

559 Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and  
 560 Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European  
 561 Conference on Computer Vision*, pp. 405–421, 2020.

562 Nathanaël Perraudin, Peter Balazs, and Peter L Søndergaard. A fast griffin-lim algorithm. In *2013  
 563 IEEE workshop on applications of signal processing to audio and acoustics*, pp. 1–4. IEEE, 2013.

564 Jens Holger Rindel. Room acoustic modelling techniques: A comparison of a scale model and a  
 565 computer model for a new opera theatre. *Building Acoustics*, 18(3-4):259–280, 2011.

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

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

571 Kazunori Suzuki and Takayuki Hidaka. A study on precise measurement of room impulse response  
 572 in a scale model and auralization. In *International Symposium on Room Acoustics, Amsterdam  
 573 (September 15–17, 2019)*, pp. 297–303, 2019.

574 Andrew Szot, Alex Clegg, Eric Underander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre,  
 575 Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus,  
 576 Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun,  
 577 Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange  
 578 their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

579 Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristof-  
 580 fersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for  
 581 neural radiance field development. In *ACM SIGGRAPH 2023 conference proceedings*, pp. 1–12,  
 582 2023.

583 Xiph.Org Foundation. Xiph opus. <https://opus-codec.org/>, 2012.

584 Lining Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel  
 585 Rota Bulò, Lorenzo Porzi, Peter Kortscheder, Aljaž Božič, et al. Vr-nerf: High-fidelity virtualized  
 586 walkable spaces. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–12, 2023.

594 **A TASK DEFINITION AND DATASET DETAILS**  
595596 **A.1 TASK DEFINITION**  
597598 The task of WavNAF, is to synthesize a high-fidelity Room Impulse Response (RIR) from a 3D scene  
599 represented by multi-view images. Given the visual information and arbitrary positional queries for  
600 a sound source and receiver, the model is trained to predict the corresponding RIR waveform that  
601 captures complex acoustic phenomena.602 **A.2 REQUIRED INPUTS**  
603604 The WavNAF model requires the following inputs to generate an RIR:  
605606 

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

  
614615 **A.3 OUTPUT**  
616617 The model’s final output is a time-domain waveform of the Room Impulse Response (RIR). Internally,  
618 the model first predicts the log-magnitude Short-Time Fourier Transform (STFT) of the RIR, which  
619 is then converted to a waveform using the Griffin-Lim algorithm (Perraudin et al., 2013; Griffin &  
620 Lim, 1984). The output format is dataset-specific:621 

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

  
625626 **A.4 DATASETS**  
627628 **A.4.1 SOUNDSPACES**629 **Audio Data** Following previous works (Su et al., 2022; Liang et al., 2023a;b; Brunetto et al., 2024),  
630 we use 90% of the audio data for training and 10% for testing. The Room Impulse Responses (RIRs)  
631 provided by SoundSpaces are binaural and resampled from the original 44.1 kHz to 22.05 kHz.  
632 For Short-Time Fourier Transform (STFT), we employ 512 FFT bins, with a Hann window of 512  
633 samples and a hop length of 128 samples. All the source and microphone are lying on the same height.  
634 The source is omnidirectional.635 **Visual Data** SoundSpaces provides visual data captured via Habitat Sim (Szot et al., 2021) with  
636 RGB images of resolution 512 × 512 and a field of view of 90 degrees. Images for training were  
637 sampled at varying numbers depending on the size of the environment: 45 images for small rooms,  
638 75 images for medium rooms, and 150 images for large rooms. Each training image is captured from  
639 positions randomly selected at the edges of the room, oriented towards the center with random offsets.  
640 Additionally, 50 randomly sampled test poses are used to evaluate visual performance.  
641642 **A.4.2 RAF**  
643644 **Audio Data** Following previous works (Chen et al., 2024; Brunetto et al., 2024), we use 80% of the  
645 audio data for training and the remaining 20% for testing. Audio clips are clipped to 0.32 seconds  
646 and sampled at 48 kHz. For STFT, we employ 1024 FFT bins, a Hann window of size 512, and  
647 a hop length of 256. The source and microphone can lie on different heights. The microphone is  
omnidirectional, and the source is directional.

648 **Visual Data** Visual data in RAF are multi-view images captured using the VR-NeRF (Xu et al.,  
 649 2023) camera rig Eyeful Tower. From the available 22 camera views, we select images from cameras  
 650 20 to 25 to ensure adequate overlap. Images have a resolution of  $684 \times 1024$  pixels. To limit the  
 651 amount of visual training data, one-third of these images are randomly subsampled, with 90% of the  
 652 selected images used for training and 10% reserved for evaluation.

## 654 B ADDITIONAL METHOD DETAILS

### 655 B.1 ARCHITECTURE DETAIL

#### 658 B.1.1 NEURAL ACOUSTIC SCALING MODULE

660 The Acoustic Scaling Layer adjusts simulation-derived features to account for temporal scaling  
 661 variations. It consists of two separate MLP branches,  $\mu$  and  $\nu$ , each comprising two fully-connected  
 662 layers with a hidden dimension of 128 and ReLU activation functions. Both branches receive the  
 663 normalized time query as input. The  $\mu$  branch's final linear layer is initialized to output ones,  
 664 effectively serving as a scaling factor. Conversely, the  $\nu$  branch's final layer is initialized to zeros,  
 665 initially producing no offset.

#### 666 B.1.2 FEATURE FUSION MLP

668 The Feature Fusion MLP architecture used in our model closely follows the design described in  
 669 NeRAF (Brunetto et al., 2024). Specifically, the network consists of two MLP blocks. The first MLP  
 670 block takes as input a concatenation of encoded positional information, directional embeddings, and  
 671 encoded wave features obtained via a ResNet. This block comprises 5 fully-connected layers, each  
 672 followed by Leaky ReLU activations with a slope 0.1, resulting in an intermediate 512-dimensional  
 673 representation. The second MLP block processes this intermediate representation to predict the STFT  
 674 bins corresponding to time queries, featuring separate output heads per audio channel. The final layer  
 675 utilizes a scaled tanh activation function to appropriately map the output to the log-magnitude STFT  
 676 range.

### 677 B.2 IMPLEMENTATION DETAIL

#### 679 B.2.1 HYPERPARAMETERS

681 We utilize the Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , and  $\epsilon = 10^{-15}$ . The initial learning rate  
 682 is set to  $10^{-4}$  and exponentially decreases to  $10^{-8}$ . For the first 2,000 iterations, only the NeRF  
 683 component is trained to allow adequate initialization. The training batch sizes are 4,096 rays for  
 684 NeRF and 1,024 STFT bins for the audio model. The total training duration is 500,000 iterations,  
 685 selecting the best-performing model across these iterations. Training is conducted on a single RTX  
 686 A6000 GPU. These hyperparameters remain consistent across both SoundSpaces and RAF datasets.

#### 688 B.2.2 LEARNING OBJECTIVE DETAIL

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

### 692 B.3 FDTD WAVE SIMULATION DETAIL

#### 693 B.3.1 SPONGE LAYER DETAIL.

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

#### 699 B.3.2 INITIAL AMPLITUDE SETTING.

701 The initial amplitude  $A_{\text{init}}$  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

702 energy with increasing vertical distance. Specifically, the amplitude  $A_{\text{init}}$  is scaled exponentially by  
 703  $\exp(-\alpha \cdot |z_{\text{src}} - z_{\text{mic}}|)$ , where  $\alpha$  is set to 0.01.  
 704

#### 705 B.4 FIRST-ORDER SYSTEM FORMULATION FOR FDTD

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

##### 711 B.4.1 FROM SECOND-ORDER TO FIRST-ORDER SYSTEM

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

715 **Conservation of momentum** (linearized Euler equation):

$$717 \rho_0 \frac{\partial \mathbf{v}}{\partial t} + \nabla p = 0 \Rightarrow \frac{\partial \mathbf{v}}{\partial t} = -\frac{1}{\rho} \nabla p \quad (20)$$

719 **Conservation of mass** with equation of state:

$$721 \frac{\partial p}{\partial t} = -\rho c^2 \nabla \cdot \mathbf{v} \quad (21)$$

723 where  $\mathbf{v}$  is the particle velocity vector,  $p$  is acoustic pressure,  $\rho$  is the medium density, and  $c$  is the  
 724 speed of sound.  
 725

##### 726 B.4.2 FINITE-DIFFERENCE DISCRETIZATION

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

730 **Step 1 - Pressure Gradient Computation:**

$$732 \frac{\partial p}{\partial x} \Big|_{i+1/2,j} \approx \frac{p_{i+1,j} - p_{i,j}}{\Delta x}, \quad \frac{\partial p}{\partial y} \Big|_{i,j+1/2} \approx \frac{p_{i,j+1} - p_{i,j}}{\Delta y} \quad (22)$$

734 **Step 2 - Velocity Update:**

$$736 v_x^{n+1} = v_x^n - \frac{\Delta t}{\rho_{i,j} \Delta x} \frac{\partial p}{\partial x}, \quad v_y^{n+1} = v_y^n - \frac{\Delta t}{\rho_{i,j} \Delta y} \frac{\partial p}{\partial y} \quad (23)$$

739 **Step 3 - Velocity Divergence:**

$$740 \nabla \cdot \mathbf{v} \Big|_{i,j} \approx \frac{v_{x,i,j} - v_{x,i-1,j}}{\Delta x} + \frac{v_{y,i,j} - v_{y,i,j-1}}{\Delta y} \quad (24)$$

742 **Step 4 - Pressure Update:**

$$743 p^{n+1} = p^n - \rho_{i,j} c_{i,j}^2 \Delta t (\nabla \cdot \mathbf{v}) \quad (25)$$

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

## 751 C ADDITIONAL EXPERIMENTAL RESULTS

### 753 C.1 PER-SCENE RESULTS

755 This section provides detailed per-scene evaluation results. Table 4 shows results for individual scenes  
 in the SoundSpaces dataset, Table 5 presents results for the RAF 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 4: Per-scene results on SoundSpaces dataset

Method	scene	T60 ↓	C50 ↓	EDT ↓
WavNAF	Furnished Room	6.895	0.613	0.020
	Empty Room	7.787	0.598	0.0180
Average		7.43	0.61	0.019

Table 5: Per-scene results on RAF dataset

## C.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.

### C.2.1 TRAINING TIME COMPARISON

Method	Time per Iteration	Relative Cost
No condition	~55ms	0.6×
3D grid feature	~94ms	1.0×
Only wave feature	~780ms	8.3×
+ Acoustic scale module	~1010ms	10.7×

Table 6: Training time comparison across different model variants

Table 6 breaks down the computational overhead of different components in our pipeline. While the full method requires 10.7× more computation than the baseline, this investment yields significant quality improvements across all metrics.

The majority of overhead comes from FDTD simulation and feature extraction rather than the neural network components themselves. This suggests opportunities for optimization through parallelization or preprocessing strategies.

### C.2.2 INFERENCE TIME BREAKDOWN

- **Total inference time** : 74.74 ms
  - Preparing FDTD (acoustic parameter grid generation): 53.33 ms (71.4%)
  - Pure FDTD CUDA kernel: 3.82 ms (5.1%)
  - Other operations (feature extraction): 17.59 ms (23.5%)

Compared to NeRAF (12.36ms), our method is approximately 6× slower during inference. However, the pure FDTD computation represents only 5.1% of the total inference time, indicating that the overhead comes primarily from data preparation and feature extraction rather than from the wave simulation itself.

### C.2.3 PRACTICAL CONSIDERATIONS

While our full method requires 10.7× more training time per iteration compared to the baseline, this computational investment yields substantial improvements across all acoustic quality metrics (as

810 shown in Table 1 of the main paper). The computational cost is primarily front-loaded during training,  
 811 with inference remaining efficient for practical deployment.  
 812

813 We note that our current implementation runs FDTD simulations on-the-fly during training for  
 814 implementational clarity. However, these computations can be preprocessed and cached when seeking  
 815 further training efficiency, potentially reducing the training overhead significantly.  
 816

### 817 C.3 FEW-SHOT LEARNING

818 To evaluate the data efficiency of our physics-informed approach, we conduct experiments with  
 819 reduced training data. Table 7 shows performance when training with only 25%, 50%, and 75% of  
 820 the available source-receiver pairs.  
 821

822 Our FDTD-based wave propagation priors enable significantly better generalization from limited data  
 823 compared to the baseline. With only 50% of training data, WavNAF achieves comparable performance  
 824 to NeRAF trained on 75% of data.  
 825

826 The consistent relative improvement at different data regimes indicates that our physics-informed fea-  
 827 tures complement rather than replace learned representations, providing valuable guidance regardless  
 828 of training set size.  
 829

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
100	<b>2.725</b>	<b>0.574</b>	<b>0.0149</b>	<b>2.306</b>	<b>0.446</b>	<b>0.0119</b>

834 Table 7: Few-shot learning performance comparison at SoundSpaces apartment 1  
 835  
 836

## 837 D HYPERPARAMETER ABLATION STUDY

838 We conduct comprehensive ablation studies to validate the robustness of our approach to various hy-  
 839 perparameter choices. These experiments demonstrate that our physics-informed framework provides  
 840 consistent improvements regardless of specific parameter settings.  
 841

### 842 D.1 $c_{min}/c_{max}$ RATIO

843 Table 8 examines the impact of varying the acoustic speed ratio in our density-to-acoustic parameter  
 844 mapping (equation 12). We test ratios from 0.6 to 0.9, where lower values create stronger acoustic  
 845 contrasts between air and solid boundaries. Despite these parameter variations representing different  
 846 degrees of approximation from traditional acoustic modeling perspectives, all settings consistently  
 847 outperform the baseline. This validates that the FDTD’s physics-governed update process where wave  
 848 phenomena like diffraction and interference naturally emerge from the wave equation solver provides  
 849 sufficient neural guidance for acoustic field learning, regardless of exact parameter values.  
 850

Method	$c_{min}/c_{max}$	T60 ↓	C50 ↓	EDT ↓
	0.6	2.278	0.459	0.0123
	0.7	<b>2.261</b>	0.467	0.0120
WavNAF	0.8	2.377	0.448	<b>0.0119</b>
	0.9 (paper setting)	2.306	<b>0.446</b>	<b>0.0119</b>

851 Table 8: Ablation study on  $c_{min}/c_{max}$  ratio. Results at the SoundSpace apartment 1.  
 852  
 853

### 861 D.2 GLOBAL DAMPING COEFFICIENT

862 Table 9 analyzes the effect of the global damping coefficient  $\gamma_{global}$ , which models air absorption  
 863 during wave propagation. Setting  $\gamma_{global} = 1.0$  represents no damping, while  $\gamma_{global} = 0.99$  introduces  
 864

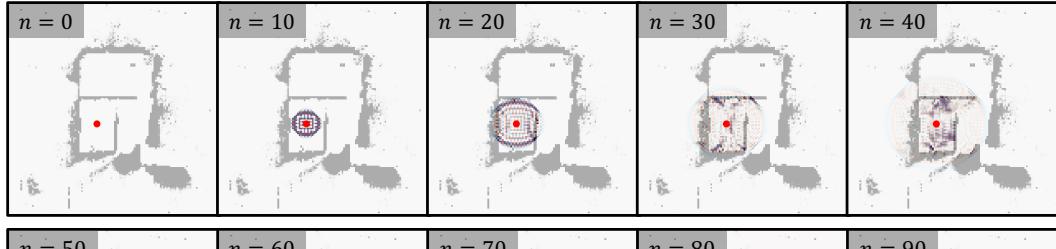
864 gradual energy dissipation. While the presence of damping slightly improves performance, our  
 865 method maintains substantial advantages over baselines even without damping, confirming that the  
 866 structural information from wave physics mainly drives our performance gains.  
 867

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

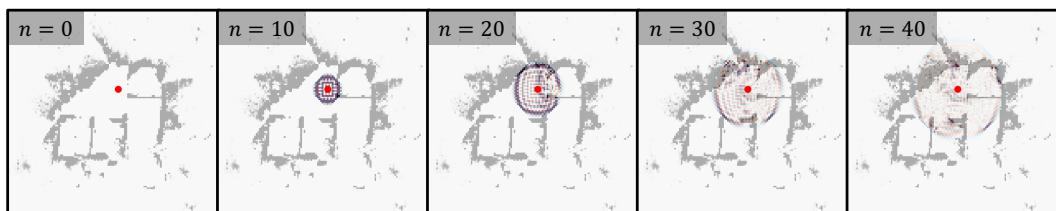
872 Table 9: Ablation study on global damping coefficient. Results at the SoundSpace apartment 1.  
 873

## 875 E VISUALIZATIONS

### 877 E.1 WAVE SIMULATION VISUALIZATION



885 (a) Apartment 1



908 (b) Apartment 2

910 Figure 5: Wave simulation visualization of SoundSpaces dataset. Our WavNAF framework uses  
 911 Neural Radiance Fields (NeRF) to extract a normalized visual density grid encoding scene geometry,  
 912 from which acoustic parameters, including sound speed and acoustic density, are directly derived.  
 913 This enables wave-equation-based simulations that inherently capture complex acoustic phenomena,  
 914 such as diffraction, refraction, and reflection, without explicit geometric modeling.  
 915

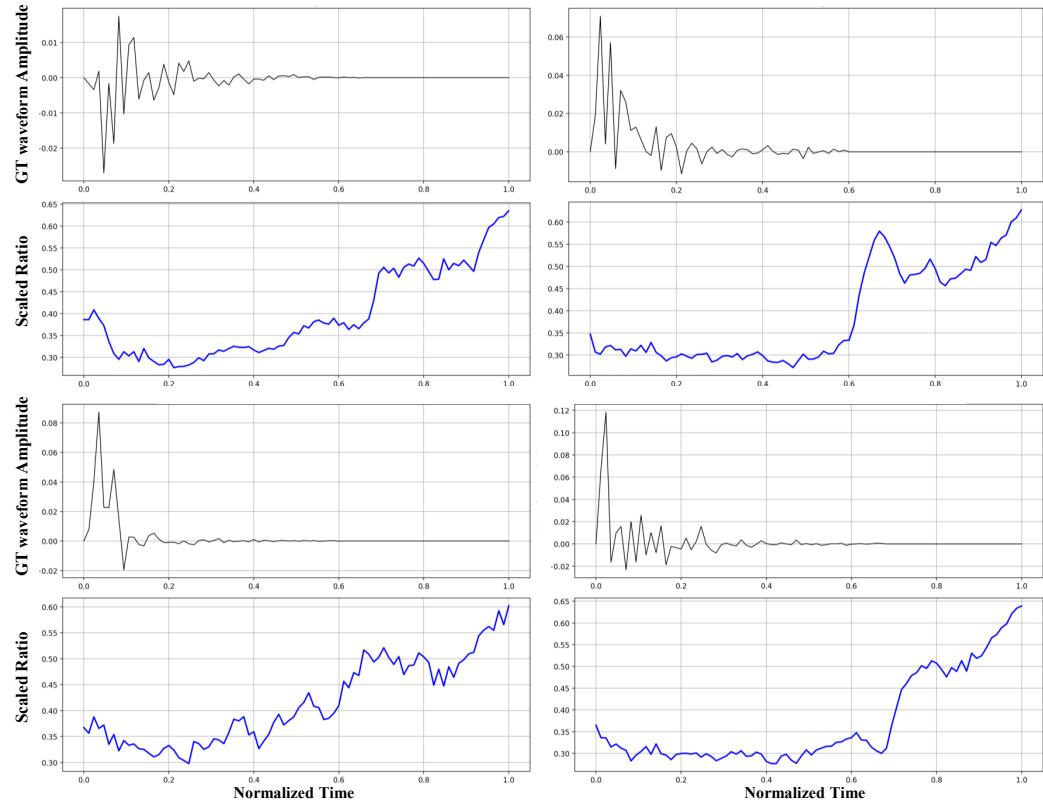
916 We visualized wave simulation results for scenes with complex, multi-room layouts from the  
 917 SoundSpaces dataset. A subset of frames from the full simulation sequence was selected to highlight  
 the temporal progression of wave phenomena. These visualizations clearly demonstrate intricate wave

918 behaviors such as reflection, refraction, and diffraction as the simulation advances over successive  
 919 time steps.  
 920

## 921 E.2 NEURAL ACOUSTIC SCALING MODULE VISUALIZATION

923 We visualize the adaptive scaling behavior of our Neural Acoustic Scaling Module by plotting the  
 924 scaling ratio  $\|\hat{f}_t\|_2/\|f_t\|_2$ , as a function of normalized time, where  $f_t$  represents the original feature  
 925 extracted from pressure map and  $\hat{f}_t = \mu_t \odot f_t + \nu_t$  represents the scaled feature.  
 926

927 This visualization demonstrates that our scaling module automatically learns appropriate corrections  
 928 for different temporal regions of the RIR, effectively bridging the gap between computationally  
 929 efficient compressed simulations and full-scale responses. The learned scaling patterns align with  
 930 known characteristics of acoustic scale models, where early reflections and late reverberation require  
 931 different degrees of adjustment.  
 932



944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 Figure 6: Visualization of the Neural Acoustic Scaling Module’s adaptive behavior. The plots show  
 966 ground-truth waveform amplitude (top row) and the scaling ratio  $\|\hat{f}_t\|_2/\|f_t\|_2$  (bottom row) over  
 967 normalized time. The scaling module learns to apply time-dependent transformations that vary  
 968 between early reflections and late reverberation, adapting the features from temporally compressed  
 969 simulations to match full-scale acoustic responses.  
 970  
 971

## 972 E.3 RIR VISUALIZATION

973 We visualize the predicted RIRs under various source-microphone configurations in the largest  
 974 scene of the SoundSpaces dataset, Apartment 1. This scene contains multiple interconnected rooms,  
 975 allowing us to evaluate challenging setups where the source and microphone are positioned in  
 976 different rooms. Even in a distant room separated by two doors, WavNAF accurately predicts the  
 977 RIR, capturing the early-reflection spike characteristics more faithfully than competing methods.  
 978 We additionally visualize the simulated pressure maps and the predicted loudness maps using the  
 979 same source positions as in Figure 7. The loudness map is obtained by fixing the source position and  
 980

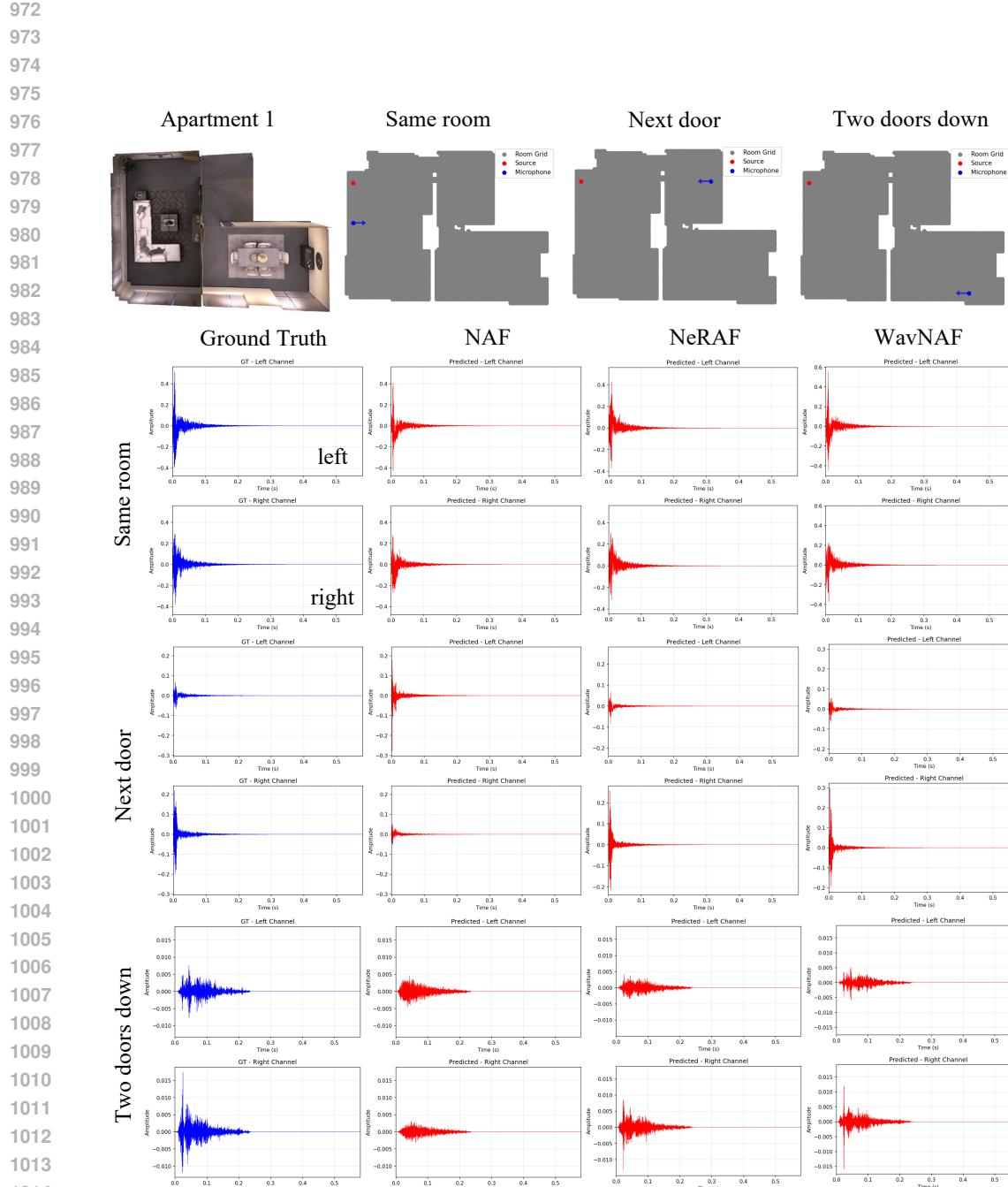


Figure 7: Predicted RIR visualization under various source-microphone configurations. Experiments are conducted in Apartment 1 of the SoundSpaces dataset. We visualize scenarios where the source and microphone are located in the same room as well as in different rooms. The farther and cross-room configurations 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.

1026  
1027  
1028  
1029  
1030  
1031  
1032  
1033  
1034  
1035  
1036  
1037  
1038  
1039  
1040  
1041  
1042  
1043  
1044  
1045  
1046  
1047  
1048  
1049  
1050  
1051  
1052  
1053  
1054  
1055  
1056  
1057  
1058  
1059  
1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079

performing inference at all navigable points as receivers. As shown in Figure 8 and Figure 9, the simulated pressure map and the predicted loudness map exhibit noticeable differences, primarily due to the temporal scale mismatch between the simplified simulation and the actual RIR.



Figure 8: Visualization of simulated pressure maps. Uses same source position with Figure 7.

#### E.4 AFFECT OF SOURCE–MICROPHONE HEIGHT DIFFERENCES

Figure 10 shows performance distribution with respect to source–microphone height differences in the Furnished Room scene of the RAF dataset.

## F THE USE OF LARGE LANGUAGE MODELS (LLMS)

In this study, LLMs were used for text editing, grammar correction, and coding assistance for visualization.

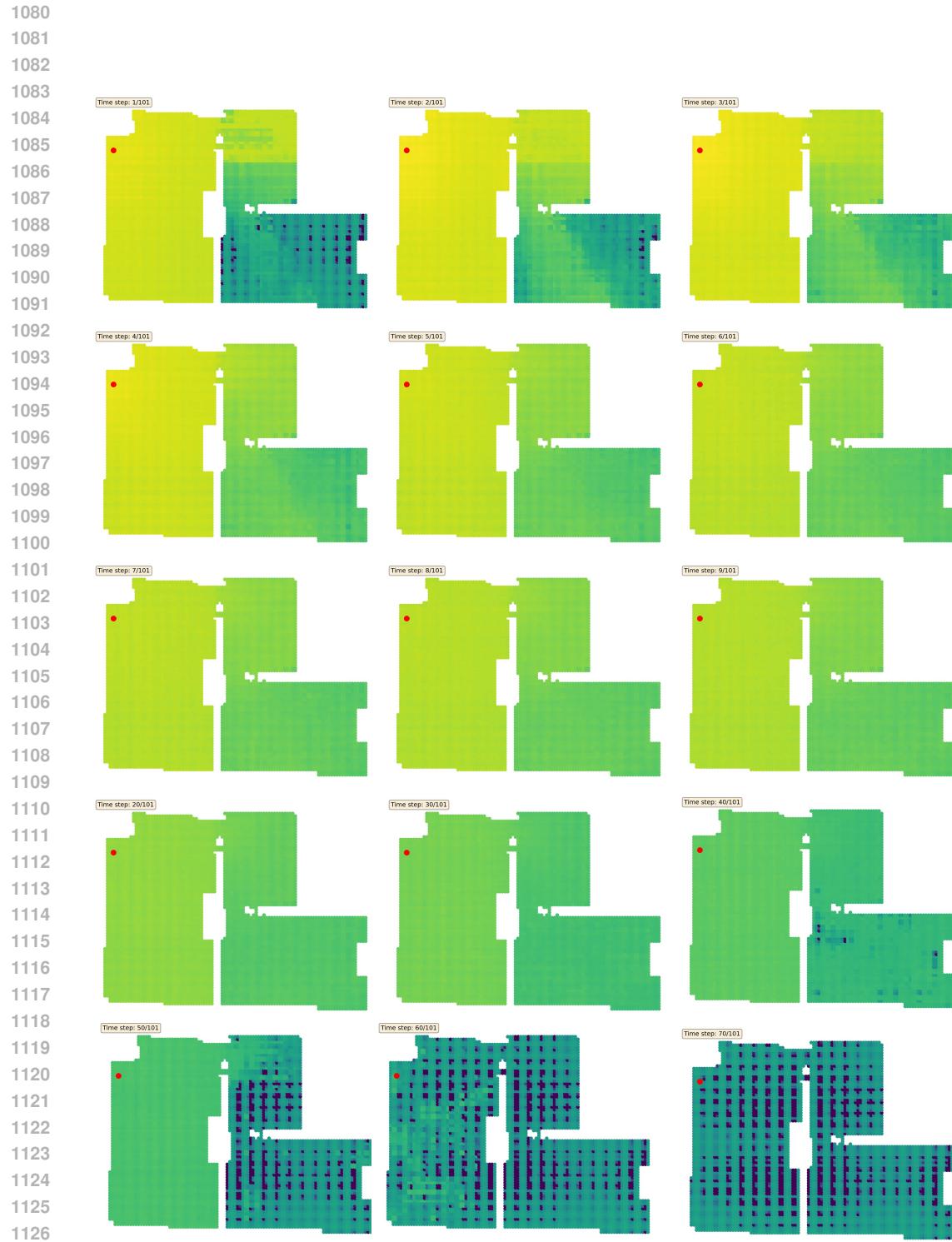


Figure 9: Visualization of loudness maps. Uses the same source position as Figure 7. The loudness map is obtained by fixing the source position and performing inference at all navigable points as receivers.

