Real Acoustic Fields: An Audio-Visual Room Acoustics Dataset and Benchmark

Ziyang Chen[1,2*] Israel D. Gebru[2] Christian Richardt[2] Anurag Kumar[3]
William Laney[2] Andrew Owens[1] Alexander Richard[2]

[1]University of Michigan [2]Codec Avatars Lab, Pittsburgh, Meta [3]Reality Labs Research, Meta

https://facebookresearch.github.io/real-acoustic-fields

Abstract

We present a new dataset called Real Acoustic Fields (RAF) that captures real acoustic room data from multiple modalities. The dataset includes high-quality and densely captured room impulse response data paired with multi-view images, and precise 6DoF pose tracking data for sound emitters and listeners in the rooms. We used this dataset to evaluate existing methods for novel-view acoustic synthesis and impulse response generation which previously relied on synthetic data. In our evaluation, we thoroughly assessed existing audio and audio-visual models against multiple criteria and proposed settings to enhance their performance on real-world data. We also conducted experiments to investigate the impact of incorporating visual data (i.e., images and depth) into neural acoustic field models. Additionally, we demonstrated the effectiveness of a simple sim2real approach, where a model is pre-trained with simulated data and fine-tuned with sparse real-world data, resulting in significant improvements in the few-shot learning approach. RAF is the first dataset to provide densely captured room acoustic data, making it an ideal resource for researchers working on audio and audio-visual neural acoustic field modeling techniques. Demos and datasets are available on our project page.

1. Introduction

Sound waves reflect off objects in a scene before reaching a listener's ears. These reflections change the sound waves in complex ways and convey the objects' size, shape, and material properties. Accurately modeling these changes is crucial for spatial audio rendering, and plays a key role in adding the sense of immersion in a variety of application domains, such as 3D games, virtual and augmented reality [22, 70].

The goal of a sound propagation model is typically to estimate a room impulse response (RIR) for a given emitter and listener pose. RIRs are linear filters that, when convolved with an input sound, simulate the sound that would be perceived by the listener, performing changes like adding reverb or dampening certain frequencies. Estimating RIRs for novel emitter and listener poses from sparsely sampled RIRs acquired from a scene has been a major focus of recent work in audio [5, 21, 26, 28, 42] and audio-visual learning [2, 9, 13, 37, 54, 55]. Inspired by novel-view synthesis [32, 38, 40], an emerging line of work has proposed learning-based models based on neural fields [34, 36, 57].

Despite the recent interest in sound propagation in the audio-visual community, existing methods have been developed and evaluated on highly simplified datasets with artificially generated impulse responses. This is due to the fact that collecting real-world RIRs is a challenging process that requires both playing and recording sounds from densely sampled positions throughout a scene. Many different data collection efforts have each made different compromises between the conflicting factors in terms of realism, ground truth, and costs. Existing datasets [11, 29, 33] thus make highly restrictive assumptions, such as by having only a single sound emitter at a fixed pose, by having limited (2D-only) spatial coverage of the scenes, or by having only simple planar geometry. Consequently, these datasets do not fully capture the complexities of real-world room geometry, material variations, and source directivity. The lack of a real "gold standard" benchmark makes it challenging to effectively analyze existing approaches under real-world assumptions and to drive research on audio-visual informed sound propagation toward its true potential.

In this paper, we propose an audio-visual sound propagation dataset and benchmark that addresses the shortcomings of previous approaches. Our *Real Acoustic Fields (RAF)* dataset is a multimodal real acoustic room dataset with dense 3D audio captures of a large space filled with and without furniture. To capture dense and calibrated audio in the rooms, we used a custom-built microphone tower system and robotic loudspeaker stand. The microphone tower contains 36 omnidirectional microphones placed at different heights and positions. The robotic stand can rotate and position the loud-

[*] Work done during an internship at Meta.

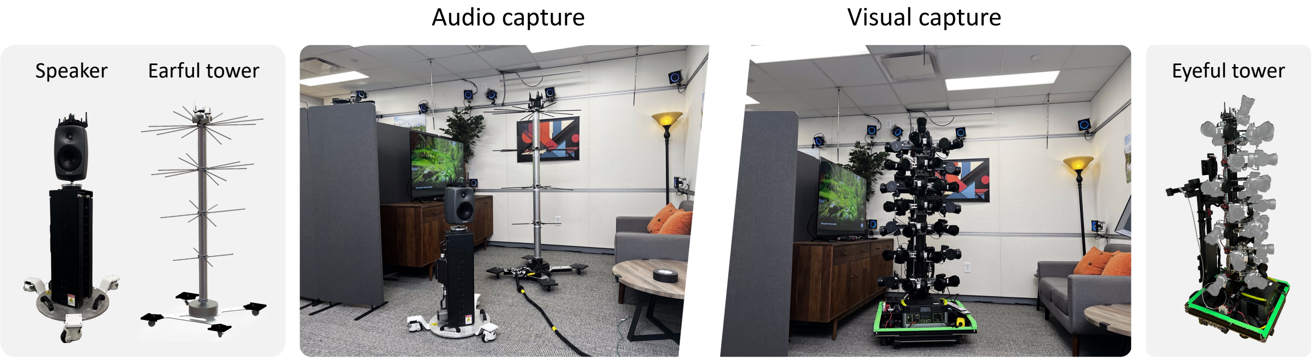

Figure 1. **Data capturing setup.** (a) Audio capture (left): the loudspeaker and microphone recording system (Earful Tower) are placed at different locations within the room to measure and capture RIRs. (b) Visual capture (right): the camera rig (Eyeful Tower) moves around rooms to capture multi-view images for visual reconstruction and novel-view synthesis.

speaker at different heights, enabling us to capture sound source directivity data. We used a motion capture system to precisely track the pose of the microphones and the loudspeaker throughout the scene. Moreover, we pair our RIR recordings with captured high-fidelity images and geometry [66] to enable more potential research in the audio-visual direction. The resulting dataset contains high-fidelity dense RIRs, speech recordings from existing speech datasets, position annotations, and visual reconstructions.

Using this dataset, we conduct the first systematic study of recent audio and audio-visual sound propagation models, including: 1) extension of common 2D approaches into 3D scenes, 2) how perceptual similarity metrics proposed by other models [34, 37] change the generated sounds, and 3) the role of visual information in audio-visual models. Finally, our dataset also allows us to evaluate few-shot training. We propose a simple, yet highly effective "sim2real" approach that begins by pretraining on synthetic data and then refining the result with a small number of real-world samples. We will release the dataset and benchmark upon acceptance.

2. Related Work

Novel-view acoustic datasets. Many RIR datasets are collected for acoustic research [26, 30, 31, 63] while they are not applicable for novel-view acoustic propagation modeling. MeshRIR [29] recorded real-world monaural impulse responses from a three-dimensional cuboidal room, with microphones at a fixed height. The room was empty and lacked visual information about the scene. Two previous methods by Liang et al. [33] and Chen et al. [11] collected real-world audio-visual datasets for novel-view acoustic synthesis tasks. Nevertheless, Liang et al. [33] only features a single stationary sound source and Chen et al. [11] only has sparse receiver positions, which might not represent the entire acoustic environment for arbitrary speaker-receiver pairs. SoundSpaces 1.0 [7] and SoundSpaces 2.0 [10] generated large-scale synthetic acoustic datasets based on room mesh from existing 3D scene datasets [6, 46, 56, 65, 67]. However, these synthetic datasets lack the complexities of real-world room geometry, material variations, and source directivity. To address this, we have gathered a real-world multimodal acoustic room dataset to further research in the field of neural acoustics.

RIR synthesis. Synthesizing RIR has been a longstanding research topic. Simulated approaches for RIR synthesis primarily rely on wave-based [19, 61] or geometric methods [8, 10, 52]. While these methods effectively simulate sound propagation in space, they often struggle to reproduce all wave-based sound effects. Geometric models do not account for interference and diffraction. While wave-based models are theoretically applicable to all frequencies, they face difficulty in accurately modeling the frequency-dependent directional characteristics of sound sources, receivers, and rooms with complex geometries. Recent methods have leveraged machine learning techniques to create more realistic RIRs. Ratnarajah et al. [48] use a generative adversarial network (GAN) to synthesize RIRs. Later work extended this approach by conditioning on scene meshes [47] and visual signals [49]. Few other works focus on learning continuous implicit neural representations for audio scenes, which target generating high-fidelity impulse responses at any arbitrary emitter-listener positions for a single scene, such as NAF [36], INRAS [57] and NACF [34]. Nevertheless, prior studies have primarily focused on simulated data owing to the absence of suitable real-world datasets. Our novel dataset presents a path to extend these approaches toward real-world modeling of neural acoustic fields.

Audio-visual acoustic learning. Recent works have explored learning acoustic information from both audio and vision. Chen et al. [12] and Chowdhury et al. [15] propose de-reverberating audio signals using visual environment encoding. Some researchers investigate the visual acoustic matching problem [9, 54, 55], aiming to synthesize audio that matches target acoustic properties based on images.

Table 1. **Dataset comparison.** We compare the attributes of our dataset with previously proposed datasets.

Dataset	Modality	Real-world	Visual source	Dimension	Scenes	Density
SoundSpaces 1.0 [7]	$\mathcal{A}$ & $\mathcal{V}$	✗	Mesh	2D	103	16 samples/m^2
SoundSpaces 2.0 [10]	$\mathcal{A}$ & $\mathcal{V}$	✗	Mesh	3D	1600+	–
MeshRIR [29]	$\mathcal{A}$	✓	–	2.5D	1	18 samples/m^3
RAF (ours)	$\mathcal{A}$ & $\mathcal{V}$	✓	NeRF & Mesh	3D	2	372 samples/m^3

Some works learn to generate sounds at the arbitrary speaker and listener positions via sparse audio-visual observations of scenes [13, 37]. Others focus on the novel-view acoustic synthesis task, synthesizing binaural sound from audio and visual information at a new viewpoint [2, 11, 14, 33]. We use our dense 3D audio-visual dataset to evaluate these methods' effectiveness and the role of vision.

Visual scene capture and view synthesis. There is a rich literature on capturing static scenes to reconstruct them in 3D and/or to render novel viewpoints; see recent surveys for a comprehensive overview [50, 60]. Many approaches that focus on 3D scene reconstruction use representations such as (truncated) signed distance fields to combine multiple observations from RGB-D sensors [43, 44, 58, 69] or standard color videos [20, 25, 41, 59]. These approaches tend to sacrifice rendering fidelity in favor of better 3D reconstruction accuracy. On the other hand, when the visual quality of novel views is paramount, approaches building on image-based rendering [4, 24, 45, 53] or, more recently, neural radiance fields [3, 39, 64, 66] have achieved the highest visual fidelity, even while compromising the quality of the reconstructed 3D geometry. To maximize the visual fidelity of our dataset, we capture and reconstruct it using the VR-NeRF approach [66].

3. The RAF Dataset

We present RAF, a dataset of densely recorded real-world room impulse responses (RIR) paired with dense multi-view images of the scenes. To the best of our knowledge, this is the first multi-modal 3D RIR dataset with dense audio and visual measurements paired with precise 6DoF tracking data. In this section, we will introduce the hardware setup used for data collection and our data collection pipeline.

3.1. Audio Capturing

Our goal is to collect dense RIR samples that cover the entire scene with paired transmitter and receiver locations.

Hardware. To facilitate the audio data collection process, we developed a novel microphone tower system called *Earful Tower*, as shown in Figure 1. The tower features 36 omnidirectional microphones. These microphones were placed at different height levels on the tower, arranged in the shape of an inverted pine cone. We positioned more microphones at the average human ear height level and used fewer microphones at lower levels. The microphones are integrated with three RME 12Mic-D units, daisy-chained and phase-locked to record synchronized multi-channel audio signals.

For generating room excitation signals during RIR measurements, we used a Genelec 8030C speaker mounted on a robotic stand. This stand offers remote control, programmable height adjustment, and speaker axis rotation.

Capturing procedure. We uniformly distribute the microphone tower at walkable positions in the room that might be occupied by a human listener (*i.e.*, open areas). We used the robotic stand to automate the rotation of the loudspeaker every 120° on its axis at each position, to obtain differently oriented sound sources. During the recording process, we played logarithmic sine-sweep signals and simultaneously recorded the resulting reverberated signals using the microphones on the tower. After completing a full-circle rotation, the speaker stand would adjust its height, and we would repeat the measurements for each 120° turn. Then we relocated the microphone tower to a new position and repeated the measurements. We shifted the speaker to a new location after the microphone tower had swept through the entire scene. Meanwhile, after each sine-sweep, we played and recorded 6-second long speech utterances randomly sampled from the VCTK dataset [62].

To accurately track the orientation and positions of the loudspeaker and microphones in the room, we used the OptiTrack motion capture system. We placed reflective markers on the loudspeaker and the microphone tower, allowing us to precisely estimate their 6 degrees of freedom (6DoF) poses.

Captured data. With the setup described above, we collected dense data from one room under two different configurations. In the first setting, the room was empty and only contained the essential equipment necessary for capturing impulse response data. In the second setting, we furnished the room to resemble a simple studio or living room. We collected 47K RIRs for the empty room and 39K RIRs for the furnished room. The collected RIRs are **4 seconds long**, which comprehensively captures the acoustic information. Our room has two parts: a large room with soft material walls to absorb sounds, and a smaller room with concrete walls for increased reverberation. We show our RIR distribution and room measurements in Figure 2. Please see Appendix A.3 for more details.

3.2. Visual Capturing

To provide a high-fidelity visual reconstruction of the scenes and synthesize the appearance from any viewpoint, we follow the VR-NeRF approach [66] to capture dense multi-view images for the scenes, using the *Eyeful Tower* multi-camera rig shown in Figure 1. We move the Eyeful Tower rig to cover the available floor area for a dense capture of our static scenes, resulting in 3,388 images for the furnished room and 8,030 images for the empty room. We use Agisoft Metashape [1] to estimate the poses of cameras within the rig using structure-from-motion and reconstruct a textured mesh of each scene. Lastly, we use ground control points to align the cameras and RIR data to the same coordinate system. The audio and visual captures are performed separately to prevent any interference between audio and visual devices (*e.g*., speakers and microphones appearing in the images) and to eliminate the impact of camera devices on audio capture (*e.g*., cameras creating reflections).

To generate the views at each microphone or speaker position, we train NeRF models using the Instant NGP architecture [66] for each scene. This enables us to examine the effectiveness of incorporating visual signals, such as RGB and depth information, into acoustic field modeling.

3.3. Comparison to Prior Datasets

We compare our dataset to several prior acoustic datasets collected from real scenes or through simulators in Table 1. In comparison to the real MeshRIR dataset [29], our dataset offers 20 times denser and more extensive coverage of room impulse response data from different height levels. Furthermore, our dataset features more complex room geometry and materials with furniture, going beyond the limitations of a single box-shaped room. Compared to simulated datasets such as SoundSpaces 1.0 [8] and SoundSpaces 2.0 [10], our dataset stands out for its high-quality real impulse responses and high-fidelity visual rendering from NeRF. In contrast, the simulated datasets fall behind in terms of both audio and visual quality, resulting in a less realistic representation of real-world acoustics.

4. Learning 3D Neural Acoustic Fields

Modeling acoustic fields can be formulated as: given the speaker's spatial position $\mathbf{s} = (x_\text{s}, y_\text{s}, z_\text{s}) \in \mathbb{R}^3$, the speaker orientation $\boldsymbol{\theta} \in \mathbb{R}^2$, and the receiver spatial position $\mathbf{r} = (x_\text{r}, y_\text{r}, z_\text{r}) \in \mathbb{R}^3$ in a room, a function $\mathcal{F}$ predicts the corresponding impulse response h:

$$\mathcal{F} : (\mathbf{s}, \mathbf{r}, \boldsymbol{\theta}) \mapsto h \in \mathbb{R}^T. \tag{1}$$

Previous studies have proposed various methods to learn $\mathcal{F}$, but they rely on synthetic data. We investigate and improve those existing models in real-world scenarios using our dataset. Additionally, we evaluate the effectiveness of using geometry or visual cues to model acoustic fields. Moreover, we introduce a simple yet effective sim2real approach for synthesizing RIRs in few-shot scenarios, which can significantly enhance performance.

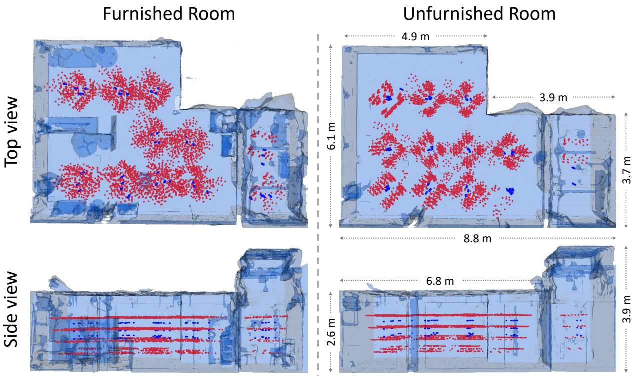

Figure 2. **Data distribution of RAF.** Blue dots represent speaker positions and red dots represent microphone positions. The room dimensions are shown on the right.

4.1. Models

We adopted several existing state-of-the-art 2D acoustic field and audio-visual models to our 3D setup, with some modifications. These models are briefly described below.

NAF. The neural acoustic field [36] models the room acoustics using an implicit representation. NAF learns a grid of local geometric features $\mathbf{G}$ to encode the spatial information of speakers and receivers at different positions, and queried speakers and receivers grid features $\mathbf{G}(\mathbf{s})$ and $\mathbf{G}(\mathbf{r})$ will be provided to the NAF $\mathcal{F}$ as additional context.

NAF represents impulse response h in the time-frequency domain $H = |\text{STFT}(h)| \in \mathbb{R}^{F \times K}$ using short-time Fourier transform (STFT), where F is the numbers of frequency bin and K is the number of time frames. Given the frequency bin f and time frame k, NAF predicts the log magnitude of the spectrogram $\hat{H}(k, f)$:

$$\hat{H}(k, f) = \mathcal{F}\left(\mathbf{G}(\mathbf{s}), \mathbf{G}(\mathbf{r}), \boldsymbol{\theta}, k, f\right), \tag{2}$$

and it minimizes the $L1$ loss between predicted and ground-truth impulse response in log scale:

$$\mathcal{L}_\text{NAF} = \| \log \hat{H}(k, f) - \log H(k, f) \|_1. \tag{3}$$

To obtain the time-domain impulse response h, NAF performs inverse STFT on predicted spectrogram magnitude $|\hat{H}|$ with random phase.

INRAS. The implicit neural representation for audio scenes [57] is inspired by interactive acoustic radiance transfer, where sound energy first scatters from the emitter to the boundaries of the scene, then propagates through the scene by bouncing between the surfaces, and finally gathers at the listener position. INRAS defines a set of bounce points

$\{\mathbf{b}_i\}_{i=1}^N \subset \mathbb{R}^3$, which are uniformly sampled from the scene surface. It represents the position of speakers or receivers using the relative distance to those bounce points, which provides more information about the scene geometry:

$$\{\mathbf{d}_i^{\mathbf{s}}\}_{i=1}^N = \{\mathbf{s} - \mathbf{b}_i\}_{i=1}^N, \quad \{\mathbf{d}_i^{\mathbf{r}}\}_{i=1}^N = \{\mathbf{r} - \mathbf{b}_i\}_{i=1}^N. \quad (4)$$

INRAS encodes speaker relative distance $\{\mathbf{d}_i^{\mathbf{s}}\}_{i=1}^N$, receiver relative distance $\{\mathbf{d}_i^{\mathbf{r}}\}_{i=1}^N$, and bounce point positions $\{\mathbf{b}_i\}_{i=1}^N$ into latent features $\mathbf{S}, \mathbf{R}, \mathbf{B} \in \mathbb{R}^{N \times D}$, where D is the feature dimension size. The time embedding $\mathbf{M} \in \mathbb{R}^{T \times D}$ is introduced for the whole time sequence and obtains spatial-time features via matrix multiplication. INRAS generates time-domain RIRs directly by decoding given features:

$$\hat{h} = \mathcal{F}\left(\mathbf{M}\mathbf{S}^\top, \mathbf{M}\mathbf{R}^\top, \mathbf{M}\mathbf{B}^\top, \boldsymbol{\theta}\right). \quad (5)$$

The INRAS model minimizes the STFT loss in the time-frequency domain $H = |\mathrm{STFT}(h)|$, including spectral convergence loss $\mathcal{L}_{\text{sc}}$ and magnitude loss $\mathcal{L}_{\text{mag}}$:

$$\mathcal{L}_{\text{INRAS}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{mag}} = \frac{\|\hat{H} - H\|_2}{\|H\|_2} + \|\hat{H} - H\|_1. \quad (6)$$

We used multi-resolution STFT loss [68], which involves computing an STFT loss at multiple time-frequency scales.

NACF. Neural Acoustic Context Field (NACF) [34] is a multimodal extension of INRAS which uses additional *context* from other modalities. Specifically, NACF uses RGB images v_{rgb} and depth images v_{depth} for each predefined bounce point $\mathbf{b}_i$ to extract local geometric and semantic information, and material properties. Similar to INRAS, RGB and depth images are encoded into latent *context* embeddings $\mathbf{C}_{\text{rgb}}, \mathbf{C}_{\text{depth}} \in \mathbb{R}^{N \times D}$ via nonlinear projection and converted into space-time features.[1] NACF decodes provided features with additional context to generate the impulse response in the time domain:

$$\hat{h} = \mathcal{F}\left(\mathbf{M}\mathbf{S}^\top, \mathbf{M}\mathbf{R}^\top, \mathbf{M}\mathbf{B}^\top, \mathbf{M}\mathbf{C}_{\text{rgb}}^\top, \mathbf{M}\mathbf{C}_{\text{depth}}^\top, \boldsymbol{\theta}\right). \quad (7)$$

NACF minimizes the same loss as INRAS (Equation 6), as well as the *energy decay loss* proposed by Majumder et al. [37], which encourages the energy decay curves of the predicted and target RIRs to be similar. Given the magnitude spectrogram $H \in \mathbb{R}^{F \times K}$, we calculate the decay curve $\mathcal{D}(H)$:

$$\mathcal{D}(H)[k] = 1 + \frac{E_k}{\sum_{i=k+1}^{K} E_i}, \quad (8)$$

where $E_k = \sum_f H(f,k)^2$ is the energy of time frame k. We minimize the $L1$ distance between the predicted and ground-truth decay curves in log space:

$$\mathcal{L}_{\text{decay}} = \|\log \mathcal{D}(\hat{H}) - \log \mathcal{D}(H)\|_1, \quad (9)$$

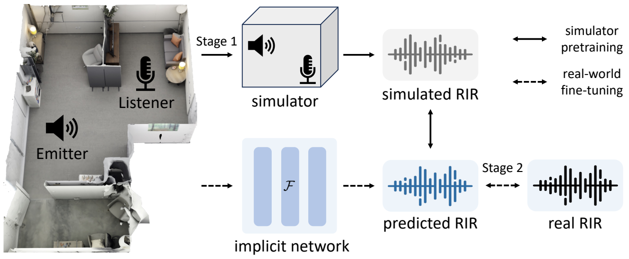

Figure 3. **Sim2real method overview.** First, we train the implicit network on simulated data with densely sampling emitter–listener position pairs. We then fine-tune it on sparse real-world data.

resulting in the overall loss with multi-resolution STFT:

$$\mathcal{L}_{\text{NACF}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{mag}} + \lambda \mathcal{L}_{\text{decay}}, \quad (10)$$

where λ is the weight of the decay loss.

AV-NeRF. We also consider the very recent AV-NeRF model [33]. Unlike NACF, which uses fixed visual contexts independent from speaker-receiver positions, AV-NeRF provides local visual information $\mathbf{C}_{\text{rgb}}^{\mathbf{r}}, \mathbf{C}_{\text{depth}}^{\mathbf{r}}$ that depends on the listener position. It predicts impulse responses:

$$\hat{h} = \mathcal{F}\left(\mathbf{s}, \mathbf{r}, \mathbf{C}_{\text{rgb}}^{\mathbf{r}}, \mathbf{C}_{\text{depth}}^{\mathbf{r}}, \boldsymbol{\theta}\right). \quad (11)$$

We also minimize the losses in Equation 10.

NAF++ and INRAS++. We observed that the energy decay loss (Equation 9) from Majumder et al. [37] used in NACF can also improve the results of the other models. We therefore introduce improved models NAF++ and INRAS++ that have these losses:

$$\mathcal{L}_{\text{NAF++}} = \mathcal{L}_{\text{mag}} + \lambda \mathcal{L}_{\text{decay}}, \quad (12)$$

$$\mathcal{L}_{\text{INRAS++}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{mag}} + \lambda \mathcal{L}_{\text{decay}}. \quad (13)$$

4.2. Sim2real for Few-Shot RIR Synthesis

Real-world impulse responses can be expensive to acquire in large quantities, and obtaining a dense capture of such data for scenes can be particularly challenging. For example, in comparison to a visual NeRF, there are comparatively fewer geometric constraints. To address this limitation, we propose a two-stage training approach that leverages simulated audio data to enhance the synthesis of real-world audio with a limited amount of training samples. Our method comprises two key stages: pretraining on dense synthetic data and fine-tuning on sparse real-world samples, as shown in Figure 3.

Pretraining on dense synthetic data. In the first stage, we pretrain our audio neural field $\mathcal{F}$, *e.g.*, INRAS, on the rich synthetic impulse responses generated from an acoustic simulator with diverse emitter and listener positions. We use the room's geometry and acoustic properties (reverberation) observed from limited real examples to create the simulator.

[1] We remove the acoustic coefficient context due to the unavailability of material coefficient annotations for our dataset and our objective of modeling real-world captures without additional annotations.

Table 2. **Evaluation on RAF with 48 kHz high-fidelity impulse responses.** We evaluate each method with the quality of generated impulse response, storage requirements, and inference speed. The results are averaged across two scenes. Original denotes uncompressed audio. The best results are in **bold**.

	Method	Variation	STFT error (dB) ↓	C_{50} error (dB) ↓	EDT error (sec) ↓	T_{60} error (%) ↓	Parameters (Million) ↓	Storage (MB) ↓	Speed (ms) ↓
Classical	Linear	AAC	1.26	2.49	0.085	25.64	–	2,033.81	–
		Opus	0.92	0.86	0.029	10.19		2,033.81	
		original	0.88	0.83	0.027	7.82		9,518.32	
	Nearest	AAC	1.04	1.97	0.064	22.83	–	2,033.81	–
		Opus	0.49	0.76	0.021	10.03		2,033.81	
		original	0.38	0.71	0.020	7.67		9,518.32	
Neural	NAF [36]	vanilla	0.64	0.71	0.021	10.08	5.51	22.04	11.98
		+ decay loss	0.64	**0.53**	**0.017**	8.19			
	INRAS [57]	vanilla	**0.36**	0.79	0.025	8.01	**1.33**	**5.31**	3.36
		+ decay loss	0.39	0.57	**0.017**	**6.17**			
	NACF [34]	vanilla	0.39	0.59	**0.017**	6.62	1.52	6.05	**3.17**
		+ temporal	0.39	0.59	0.018	7.31	1.75	7.00	3.41
	AV-NeRF [33]	vanilla	0.39	0.73	0.021	8.11	12.99	51.98	6.48

By exposing the model to a diverse range of simulated audio data, we enable it to learn general impulse response patterns and spatial information, which serve as a strong foundation for subsequent fine-tuning.

Fine-tuning on sparse real-world samples. We use sparse real-world audio samples for fine-tuning the neural field $\mathcal{F}$. By fine-tuning it on real-world data, the model adapts to the specifics of real-world audio while retaining the knowledge gained from the simulator data. By combining the strengths of the simulator and real-world data, our method achieves high-quality audio synthesis with sparse real-world data and strikes a balance between data collection cost and synthesis performance.

5. Experiments

We use our real-world dataset to evaluate audio and audio-visual acoustic field modeling methods. Also, we showcase our sim2real method that boosts them in the few-shot setting.

5.1. Evaluation of 3D Neural Acoustic Fields

We evaluate the methods on the 3D acoustic field modeling task using our full real-world dataset.

Models. We consider both state-of-the-art neural field models and classical models. To adapt to the 3D domain, we extend NAF [36], INRAS [57], NACF [34], AV-NeRF [33], and their variants, introducing an extra dimension for neural acoustic field modeling, which was not feasible with other existing datasets. Following [36, 57], we also compare with traditional signal processing methods using linear and nearest-neighbor interpolation on the training data. To improve the storage efficiency, we also apply audio encoding methods such as Advanced Audio Coding (AAC) and Opus to compress audio with low bit rates; see Appendix A.2 for details.

Metrics. Following Su et al. [57], we use several metrics to assess the quality of the predicted impulse responses, including Clarity (C_{50}), Reverberation Time (T_{60}), and Early Decay Time (EDT). C_{50} quantifies the clarity of acoustics by measuring the ratio of initial sound energy to subsequent reflections within a room, with higher values indicating clearer acoustics. T_{60} reflects the overall sound decay within a room, while EDT focuses on the early portion of the sound decay curve. We also evaluate STFT error, the absolute error between the predicted and the ground-truth log-magnitude spectrograms [16, 36]. Additionally, we measure the computational efficiency of each method by evaluating storage requirements for saving audio scenes and the inference time needed for rendering an impulse response.

Experimental setup. For each scene, we use 80% of the data for training and hold out 5% and 15% for validation and testing, respectively. The impulse responses are resampled to 48 kHz or 16 kHz sampling rate and are cut to 0.32s for training and evaluation based on the average reverberate time of the room. For all experiments, we use the AdamW optimizer [27, 35] with a learning rate of 10^{-3}, an exponential decay learning rate scheduler with a rate of 0.98, and a batch size of 128. We train all the models on an NVIDIA A100 GPU for 200 epochs and evaluate the last epoch. For NACF [34] and AV-NeRF [33], we use the visual NeRF model to render the corresponding RGB and depth images for novel views. We test inference time on the same NVIDIA A100 GPU for all the methods to ensure fair comparison.

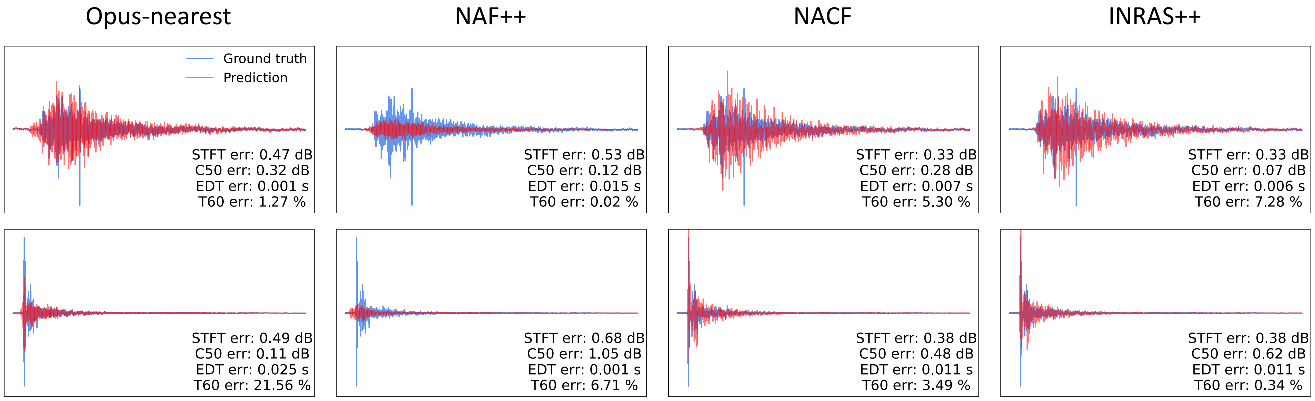

Figure 4. **Visualization of generated RIRs from different methods.** We visualize the ground-truth (in blue) and predicted (in red) impulse responses of several methods for qualitative comparison.

Results. We show our quantitative results with 48K sampling rate RIRs in Table 2. We found that INRAS++, the version of INRAS with a decay loss, performs best on most metrics, has a lightweight architecture, and fast inference speed. Opus audio encoding with nearest-neighbor interpolation is on par with several learning-based methods, suggesting the dense distribution of our captured RIRs throughout the scenes. INRAS++ and NAF++ outperform their vanilla models by a large margin on the C_{50}, EDT, and T_{60} metrics which indicates that the energy decay loss helps models learn the energy attenuation significantly better. We show the qualitative result in Figure 4. We see that NACF and INRAS++ can generate impulse responses closer to ground truth while NAF++ fails even despite having good metric results. We also visualize loudness maps in Figure 6, where we obtain 3D occupancy grids from our visual NeRF model with a resolution of 0.1 m. We can see that both INRAS++ and NACF learn a continuous acoustic field. When comparing with the acoustic fields in furnished and unfurnished rooms, we can see the models have successfully captured the phenomenon of sound occlusion. See Appendix A.1 for more results.

5.2. Evaluation of Few-Shot RIR Synthesis

We next conduct experiments to explore how model performance varies with different training data scales. Additionally, we benchmark our Sim2Real method against other baselines in challenging few-shot scenarios.

Experimental setup. We trained each model on the furnished room using various training data scales, ranging from 0.3% to 100%, where the former comprises ~100 samples and the latter has 31.3K samples. To prevent overfitting, we reserved 10% of the training samples for early stopping at each scale. For our sim2real model, we created a geometric-based Pyroomacoustics Scheibler et al. [51] shoebox acoustic simulator for pretraining, using the parameters of room bounding box and average T_{60} calculated from real-world examples. The training involved a dense pretraining stage on simulated data, followed by fine-tuning using real-world examples with a learning rate of 5×10^{-4}.

Results. We present our results in Figure 5 and Table 3. Our Sim2Real model demonstrates substantial performance improvements in few-shot setups, specifically with 0.3%, 1%, and 5% of the total training samples (approximately 100, 300, and 1500 samples, respectively). As the number of training samples increases, the advantages of using simulated data for training become smaller. While our sim2real model lags behind the model trained with the complete dense dataset, which exhibits a 55% improvement over our model trained with 5% of the data, it's worth noting that we only use a basic shoebox simulator. We believe this performance gap will

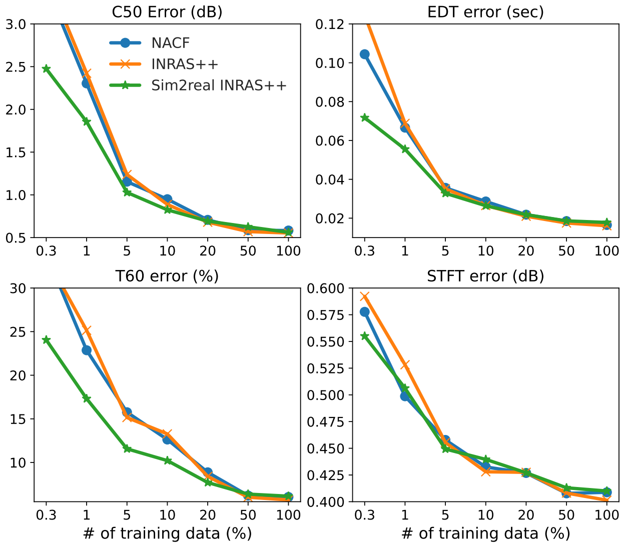

Figure 5. **Few-shot RIR synthesis results.** We evaluate the performances of models with different numbers of training data. The results are reported in the furnished room. Our Sim2Real method can improve the performance in cases of limited training data.

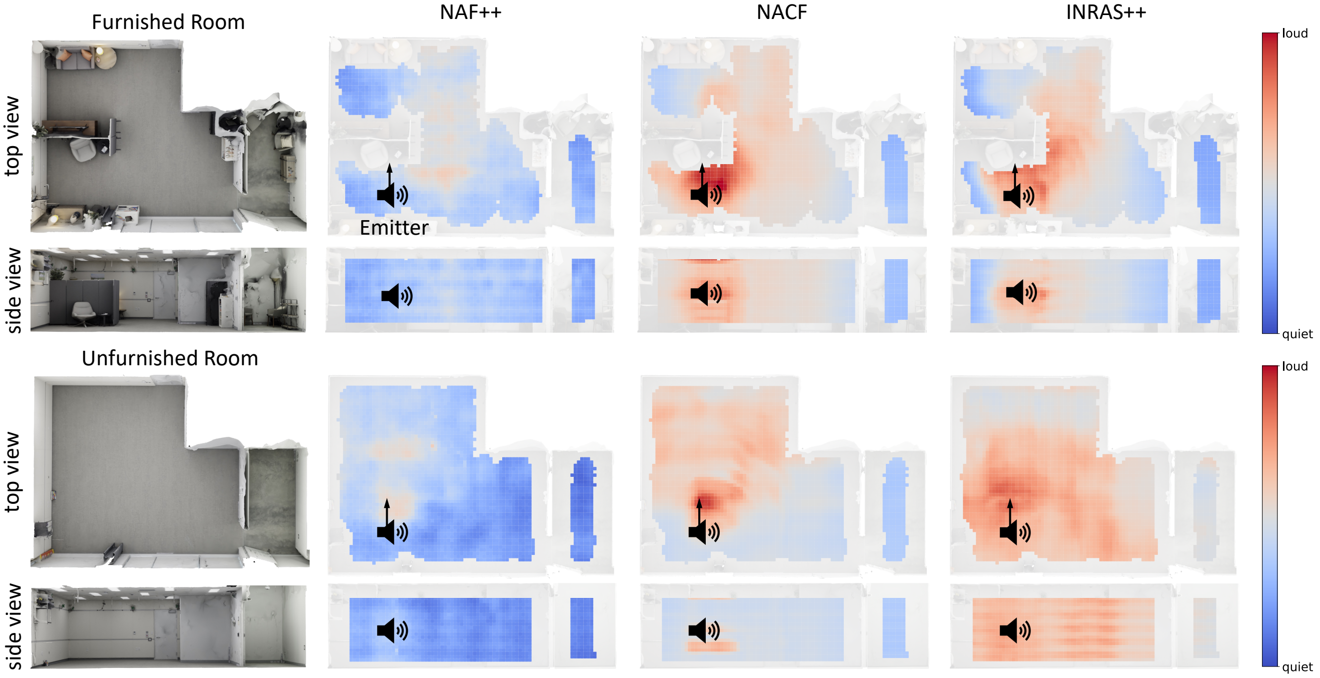

Figure 6. **Loudness map visualization.** Given an emitter position and its orientation, we visualize the intensity of predicted impulse responses over the spaces from the top view and side view, for the furnished and unfurnished room. Red means loud and blue means quiet. The arrow denotes the speaker's orientation.

Table 3. **Few-shot experiments with 1% training data.**

Method	STFT Err (dB) ↓	C_{50} Err (dB) ↓	EDT Err (sec) ↓	T_{60} Err (%) ↓
Simulator [51]	1.47	5.14	0.170	44.00
NAF++	0.69	2.38	0.072	21.77
INRAS++	0.53	2.42	0.098	25.15
NACF	**0.50**	2.30	0.067	22.87
sim2real INRAS++	0.51	**1.86**	**0.056**	**17.31**

Table 4. **Ablation experiments on the bounce point sampling strategy.** We perform the experiments on the furnished room.

Method	Modality	Bnc point strategy	STFT (dB) ↓	C_{50} (dB) ↓	EDT (sec) ↓	T_{60} (%) ↓
INRAS++	$\mathcal{A}$	2D	0.42	0.57	0.017	6.21
NACF	$\mathcal{A}$ & $\mathcal{V}$	2D	**0.40**	0.58	0.017	6.04
INRAS++	$\mathcal{A}$	3D	0.41	**0.53**	**0.016**	**5.84**

be further narrowed down if we can apply more advanced simulators, such as Chen et al. [7, 10].

5.3. Ablation Study

Bounce point sampling. We investigate the impact of our bounce point sampling strategy on model performance. To do this, we compare our 3D bounce point sampling method with the original 2D sampling method at a fixed height. As shown in Table 4, our 3D bounce point sampling enhances the model's performance. Additionally, INRAS++ with 2D bounce points shows a comparable performance against the audio-visual model NACF, suggesting that the audio modality alone suffices for the current setup.

Table 5. **Ablation experiments on speaker orientation.**

Method	STFT Err (dB) ↓	C_{50} Err (dB) ↓	EDT Err (sec) ↓	T_{60} Err (%) ↓
INRAS++ w/o ori.	0.42	0.64	0.018	6.57
INRAS++	**0.40**	**0.55**	**0.016**	**5.59**

Speaker orientation. We use a directional speaker during our data capture, exhibiting directivity patterns that affect the acoustic experience of receivers We explore how neural models use orientation information by removing speaker orientation embeddings from the inputs. As demonstrated in Table 5, the quality of the generated RIRs significantly improves when orientation information is included. Please see Appendix A.3 for RIR visualization for different orientations.

Energy decay loss. We study how model performance varies with the weights of the energy decay loss. We conduct experiments for INRAS++ on the furnished room and set λ to $\{0.1, 0.2, 0.3, 0.5\}$. We present our results in Table 6. It shows that increasing the weights of decay loss improves metrics like C_{50}, EDT, and T_{60} errors, though it comes with a tradeoff in the STFT error metric. For our primary experiments, we choose $\lambda = 2.0$ for balanced performance.

Table 6. **Ablation experiments on the energy decay loss.**

	λ	STFT Err (dB) ↓	C_{50} Err (dB) ↓	EDT Err (sec) ↓	T_{60} Err (%) ↓
INRAS++	1.0	**0.39**	0.58	0.017	5.98
	2.0	0.40	0.55	0.016	5.59
	3.0	0.41	**0.49**	0.016	5.48
	5.0	0.43	**0.49**	**0.015**	**5.34**

6. Conclusion

This paper introduces RAF, a multimodal real-world acoustic room dataset collected for facilitating research on novel-view acoustic synthesis and neural acoustic field modeling techniques. RAF includes dense 3D room impulse response captures of a large space, both with and without furniture. It also include visual data captured from multiple viewpoints and precise tracking of sound sources and receivers in the room. We systematically evaluated existing techniques for audio and audio-visual novel-view acoustic synthesis using this real-world data. We provided insights into the performance of individual models and proposed new improvements. Furthermore, we conducted experiments to investigate the impact of incorporating visual data (*i.e.*, images and depth) into neural acoustic field models. This dataset fills a gap in existing research by providing real-world data for evaluating and benchmarking novel-view acoustic synthesis models and impulse response generation techniques. In the future, we plan to expand the dataset to more room configurations.

Limitations and Broader Impacts. Collecting real-world room impulse data is expensive and time-consuming, which makes scaling up data collection for multiple rooms or scenes challenging. Our dataset only have RIRs data from a single physical room, although with two different configurations. Thus, its utility is limited for research aiming to generalize across different rooms and scenes. Using RIR data can produce audio recordings that mimic real recordings from a specific room. However, this capability can lead to the creation of deceptive and misleading media.

Acknowledgements. We would like to thank Jake Sandakly, Steven Krenn, and Todd Keebler for their assistance with data collection, as well as Linning Xu and Vasu Agrawal for their support with VR-NeRF rendering. Our gratitude also extends to Kun Sun, Mingfei Chen, Susan Liang, and Chao Huang for their insightful discussions.

References

[1] Agisoft, LLC. Metashape 2.0, 2023. 4

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

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-NeRF: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 3

[4] Tobias Bertel, Mingze Yuan, Reuben Lindroos, and Christian Richardt. OmniPhotos: Casual 360° VR photography. *ACM Trans. Graph.*, 39(6):267:1–12, 2020. 3

[5] Diego Di Carlo, Pinchas Tandeitnik, Cedrić Foy, Nancy Bertin, Antoine Deleforge, and Sharon Gannot. dechorate: a calibrated room impulse response dataset for echo-aware signal processing. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021:1–15, 2021. 1

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *3DV*, pages 667–676, 2017. 2

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *ECCV*, 2020. 2, 3, 8

[8] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *ECCV*, pages 17–36. Springer, 2020. 2, 4

[9] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *CVPR*, pages 18858–18868, 2022. 1, 2

[10] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS*, pages 8896–8911, 2022. 2, 3, 4, 8

[11] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *CVPR*, pages 6409–6419, 2023. 1, 2, 3

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

[13] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere-hear everything (BEE): Audio scene reconstruction by sparse audio-visual samples. In *ICCV*, pages 7853–7862, 2023. 1, 3

[14] Ziyang Chen, Shengyi Qian, and Andrew Owens. Sound localization from motion: Jointly learning sound direction and camera rotation. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[15] Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Ad-Verb: Visually guided audio dereverberation. In *ICCV*, pages 7884–7896, 2023. 2

[16] Alexandre Défossez, Neil Zeghidour, Nicolas Usunier, Léon Bottou, and Francis Bach. Sing: Symbol-to-instrument neural generator. In *NeurIPS*, 2018. 6

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 12

[18] Angelo Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio engineering society convention 108*. Audio Engineering Society, 2000. 12

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

[20] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3D scene reconstruction with the Manhattan-world assumption. In *CVPR*, 2022. 3

[21] Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 313–317, 2014. 1

[22] Dorte Hammershøi and Henrik Møller. Binaural technique—basic methods for recording, synthesis, and reproduction. *Communication Acoustics*, pages 223–254, 2005. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 12

[24] Peter Hedman, Tobias Ritschel, George Drettakis, and Gabriel Brostow. Scalable inside-out image-based rendering. *ACM Trans. Graph.*, 35(6):231:1–11, 2016. 3

[25] Hyeonjoong Jang, Andréas Meuleman, Dahyun Kang, Donggun Kim, Christian Richardt, and Min H. Kim. Egocentric scene reconstruction from an omnidirectional video. *ACM Trans. Graph.*, 41(4):100:1–12, 2022. 3

[26] Marco Jeub, Magnus Schafer, and Peter Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *International Conference on Digital Signal Processing*, 2009. 1, 2

[27] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015. 6

[28] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *ICASSP*, pages 5220–5224, 2017. 1

[29] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021. 1, 2, 3, 4

[30] Jonathan Le Roux and Emmanuel Vincent. A categorization of robust speech processing datasets. 2014. 2

[31] Jonathan Le Roux, Emmanuel Vincent, John R Hershey, and Daniel PW Ellis. Micbots: collecting large realistic datasets for speech and audio research using mobile robots. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5635–5639. IEEE, 2015. 2

[32] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, pages 5521–5531, 2022. 1

[33] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. AV-NeRF: Learning neural fields for real-world audio-visual scene synthesis. *arXiv preprint arXiv:2302.02088*, 2023. 1, 2, 3, 5, 6, 12, 13

[34] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *arXiv preprint arXiv:2309.15977*, 2023. 1, 2, 5, 6, 13

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[36] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. In *NeurIPS*, pages 3165–3177, 2022. 1, 2, 4, 6, 12, 13

[37] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. In *NeurIPS*, pages 2522–2536, 2022. 1, 2, 3, 5

[38] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021. 1

[39] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 3

[40] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[41] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3D scene reconstruction from posed images. In *ECCV*, 2020. 3

[42] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada. Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition. In *LREC*, pages 965–968, 2000. 1

[43] Richard A. Newcombe, Andrew J. Davison, Shahram Izadi, Pushmeet Kohli, Otmar Hilliges, Jamie Shotton, David Molyneaux, Steve Hodges, David Kim, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 3

[44] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6):169:1–11, 2013. 3

[45] Ryan Styles Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, compressing, and rendering panoramic light field stills for virtual reality. *ACM Trans. Graph.*, 37(6):197:1–15, 2018. 3

[46] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X.

Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. arXiv:2109.08238, 2021. 2

[47] Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. MESH2IR: Neural acoustic impulse response generator for complex 3D scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 924–933, 2022. 2

[48] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. In *ICASSP*, pages 571–575, 2022. 2

[49] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. *arXiv preprint arXiv:2312.00834*, 2023. 2

[50] Christian Richardt, James Tompkin, and Gordon Wetzstein. Capture, reconstruction, and representation of the visual real world for virtual reality. In *Real VR – Immersive Digital Reality: How to Import the Real World into Head-Mounted Immersive Displays*, pages 3–32. Springer, 2020. 3

[51] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *ICASSP*, pages 351–355, 2018. 7, 8

[52] Carl Schissler and Dinesh Manocha. Interactive sound propagation and rendering for large multi-source scenes. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2016. 2

[53] Heung-Yeung Shum, Shing-Chow Chan, and Sing Bing Kang. *Image-Based Rendering*. Springer, 2007. 3

[54] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *ICCV*, pages 286–295, 2021. 1, 2

[55] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *arXiv preprint arXiv:2307.15064*, 2023. 1, 2

[56] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2

[57] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In *NeurIPS*, 2022. 1, 2, 4, 6, 13

[58] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. iMAP: Implicit mapping and positioning in real-time. In *ICCV*, 2021. 3

[59] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *CVPR*, 2021. 3

[60] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. *Comput. Graph. Forum*, 41(2):703–735, 2022. 3

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

[62] Christophe Veaux, Junichi Yamagishi, and Simon King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013. 3

[63] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: A dataset for finding humans using room acoustics. *arXiv preprint arXiv:2311.03517*, 2023. 2

[64] Xiuchao Wu, Jiamin Xu, Zihan Zhu, Hujun Bao, Qixing Huang, James Tompkin, and Weiwei Xu. Scalable neural indoor scene rendering. *ACM Trans. Graph.*, 41(4):98:1–16, 2022. 3

[65] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *CVPR*, pages 9068–9079, 2018. 2

[66] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Proceedings*, 2023. 2, 3, 4, 12

[67] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. Habitat-Matterport 3D semantics dataset. In *CVPR*, pages 4927–4936, 2023. 2

[68] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020. 5

[69] Sheng Yang, Beichen Li, Yan-Pei Cao, Hongbo Fu, Yu-Kun Lai, Leif Kobbelt, and Shi-Min Hu. Noise-resilient reconstruction of panoramas and 3D scenes using robot-mounted unsynchronized commodity RGB-D cameras. *ACM Trans. Graph.*, 2020. 3

[70] Wen Zhang, Parasanga N Samarasinghe, Hanchi Chen, and Thushara D Abhayapala. Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences*, 7 (5):532, 2017. 1

A.1. Additional Experimental Results

Benchmark on 16 KHz impulse responses. We also evaluate each method on our benchmark with impulse responses of 16 kHz sampling rate. We show the results in Table 7. We can see that INRAS++ performs best overall, which matches with the conclusion in Section 5.1.

More qualitative results. We provide more predicted RIR visualization for qualitative comparison in Figure 9. We also provide more loudness map visualization on the different scenes for qualitative comparison in Figure 10.

Empty versus furnished room. One advantage of our dataset is that it contains a scene in two conditions — empty and furnished, which allows studying the difference in acoustic fields introduced by furniture. Due to a lack of ground-truth comparison, we visualize the generated impulse responses from INRAS++ trained on each scene individually as an approximation of the acoustic field. We show our results in Figure 7, where we can see that generated impulse responses with different acoustic properties.

A.2. Implementation Details

In this section, we will demonstrate the implementation of each baseline in detail.

AAC and Opus. We convert the raw waveform (`.wav`) into AAC (`.m4a`) and Opus (`.opus`) encoding and reverse the compression using FFmpeg commands as shown below:

```
# AAC compression
encode_command = f"ffmpeg -i audio.wav -t {audio_length} -c:a aac -b:a 24k temp.m4a"
decode_command = f"ffmpeg -i temp.m4a -c:a pcm_f32le -ar {sampling_rate} audio_aac.wav"

# Opus compression
encode_command = f"ffmpeg -i audio.wav -t {audio_length} -c:a opus -strict -2 -b:a 24k temp.opus"
decode_command = f"ffmpeg -i temp.opus -c:a pcm_f32le -ar {sampling_rate} audio_opus.wav"
```

Listing 1. FFmpeg commands for audio compression

We cut the audio to be the same length (0.32s) and corresponding sampling rate (16K or 48K) for fair evaluation comparison.

Note that we use a different Opus encoder which can achieve better compression performance than NAF used [36]. Due to the heavy computation of constructing a high-dimensional interpolation engine, we modify the baseline algorithm by first matching the nearest neighbor of the emitter in the training distribution and then performing the nearest neighbor or linear interpolation to generate impulse responses for given listener positions.

NAF. We follow the official implementation of NAF [2], and create 3D grid features based on the bounding boxes of scenes. For experiments with 16 kHz sampling rate, we use an STFT with an FFT size of 512, a window length of 256, and a hop length of 128. For 48 kHz sampling rate, we use an STFT with an FFT size of 1024, a window length of 512, and a hop length of 256. We perform inverse STFT on the predicted magnitude of the RIR spectrogram with random spectrogram phase to obtain time-domain RIR. We set $\lambda = 1.0$ for the weight of energy decay loss when training NAF++.

INRAS. We follow the implementation of INRAS provided by the authors in their supplementary material[3], and add an extra dimension for the emitter, listener, and bounce point position. We changed the original bounce point sampling method, which only sampled points with a specific height. Instead, we apply Poisson sampling on the scene meshes to obtain 256 bounce points in 3D to represent scene geometry in a better way. To optimize multi-resolution STFT loss, we set FFT size as $\{128, 512, 1024, 2048\}$, window length as $\{80, 240, 600, 1200\}$, and hop length as $\{16, 50, 120, 240\}$. We set $\lambda = 2.0$ for the weight of the energy decay loss.

NACF. We use the same architecture as INRAS for NACF. We keep the original bounce point sampling strategy in the paper and render visual context using VR-NeRF [66]. We render 256×256 pixel RGB, and depth images with a field of view of 90°. We use the surface normal of each bounce point to determine the look-at view of the virtual camera. Following the original paper, RGB and depth images are down-sampled to 16×16 and are encoded with an MLP as visual contexts. We set $\lambda = 2.0$ for the weight of energy decay loss. We optimize the multi-resolution STFT loss with the same hyperparameters as INRAS.

AV-NeRF. Because we have a different setup from AV-NeRF [33] where we have omnidirectional microphones instead of orientated binaural receivers, we adopt their method with several changes. We use VR-NeRF [66] to render 4 perspective views of 256×256 RGB and depth maps with a field of view of 90° for each receiver's position, and encode them with frozen ResNet18 [23] trained on ImageNet [17]. We removed the relative angle because it does not fit our setup. We set $\lambda = 2.0$ for the weight of energy decay loss.

A.3. Dataset

Impulse response data processing. We followed the sine-sweep deconvolution process as described by Farina [18] to extract the impulse response from the signals recorded by the microphones. For each extracted impulse response, we saved the 3D location of the receiver, as well as the 3D location and

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

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

Table 7. **Benchmark with 16 kHz sampling rate.**

	Method	Variation	STFT error (dB) ↓	C_{50} error (dB) ↓	EDT error (sec) ↓	T_{60} error (%) ↓	Parameters (Million) ↓	Storage (MB) ↓	Speed (ms) ↓
Classical	Linear	AAC	1.14	1.09	0.040	8.79	–	680.45	–
		Opus	1.06	0.80	0.032	7.48		680.45	
		original	1.02	0.82	0.032	6.82		3,172.77	
	Nearest	AAC	0.72	0.83	0.027	8.08	–	680.45	–
		Opus	0.58	0.61	0.020	6.96		680.45	
		original	0.48	0.71	0.020	7.68		3,172.77	
Neural	NAF [36]	vanilla	0.77	0.69	0.025	8.15	5.51	22.04	5.57
		+ decay loss	0.77	0.63	0.023	7.43			
	INRAS [57]	vanilla	**0.44**	0.65	0.024	6.15	**1.33**	**5.31**	**2.10**
		+ decay loss	0.45	**0.54**	**0.019**	**5.34**			
	NACF [34]	vanilla	0.45	0.58	0.020	5.47	1.52	6.05	2.39
		+ temporal	0.48	0.60	0.022	6.59	1.75	7.00	2.78
	AV-NeRF [33]	vanilla	0.46	0.58	0.021	6.12	12.99	51.98	5.80

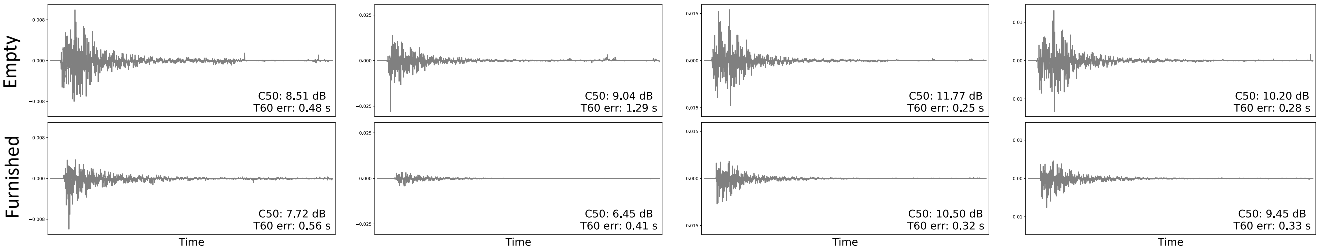

Figure 7. **Visualization comparison of generated RIRs from different scenes.** We present visualizations of four pairs of generated impulse responses, each sharing the same emitter-receiver position in both the empty room and the furnished room. These visualizations highlight the variations in the acoustic fields between the two distinct scenes.

orientation of the sound source. The length of the impulse response is 4 seconds and all audio data was recorded at a sampling rate of 48 kHz and stored at a resolution of 32 bits. We show the RT60 distribution of our collected RIRs in Figure 8

Visual rendering. We provide renderings of room meshes as a simple overview in Figure 11.

Speaker orientation. In Figure 12, we provide visualizations of impulse response pairs from our captured dataset. These pairs share the same emitter-listener position but differ in emitter orientations. The orientations of directional speakers impact the resulting impulse responses.

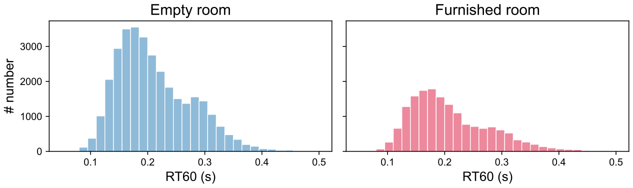

Figure 8. **RT60 distribution.**

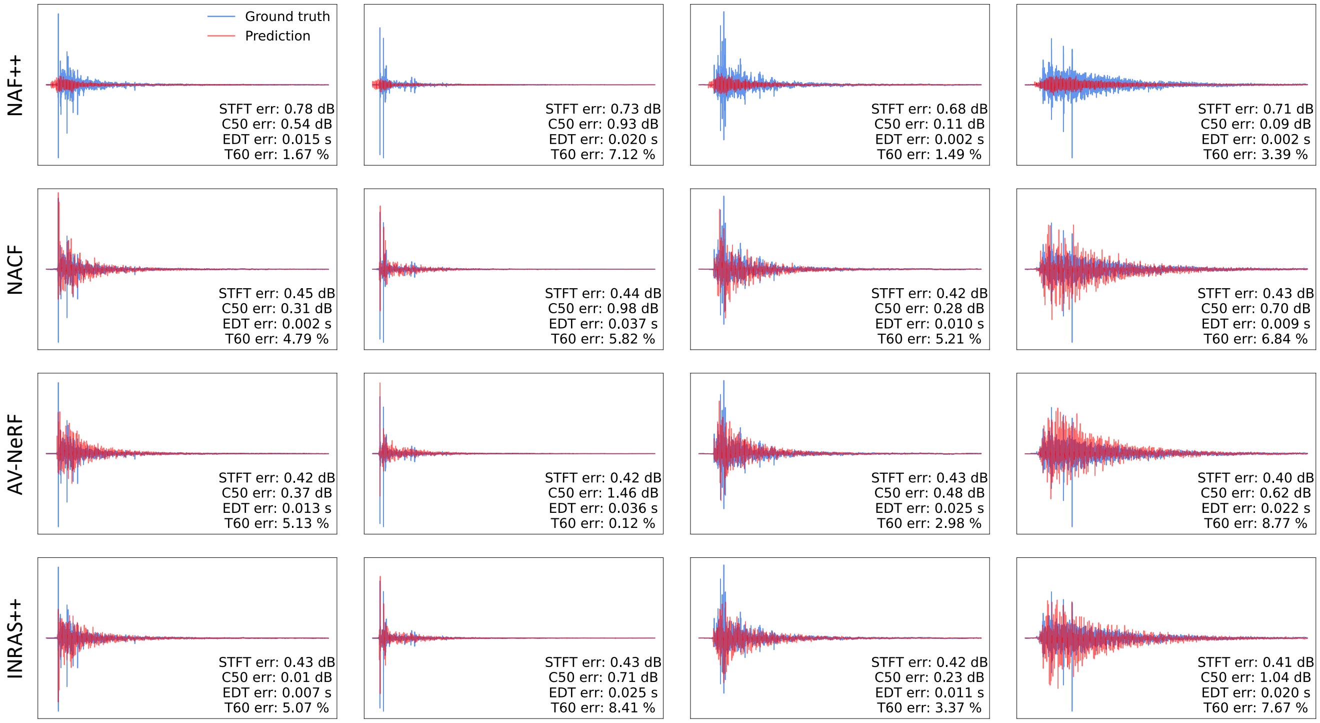

Figure 9. **Visualization of generated RIRs.** We visualize the ground truth (in blue) and predicted (in red) impulse responses of several methods for qualitative comparison.

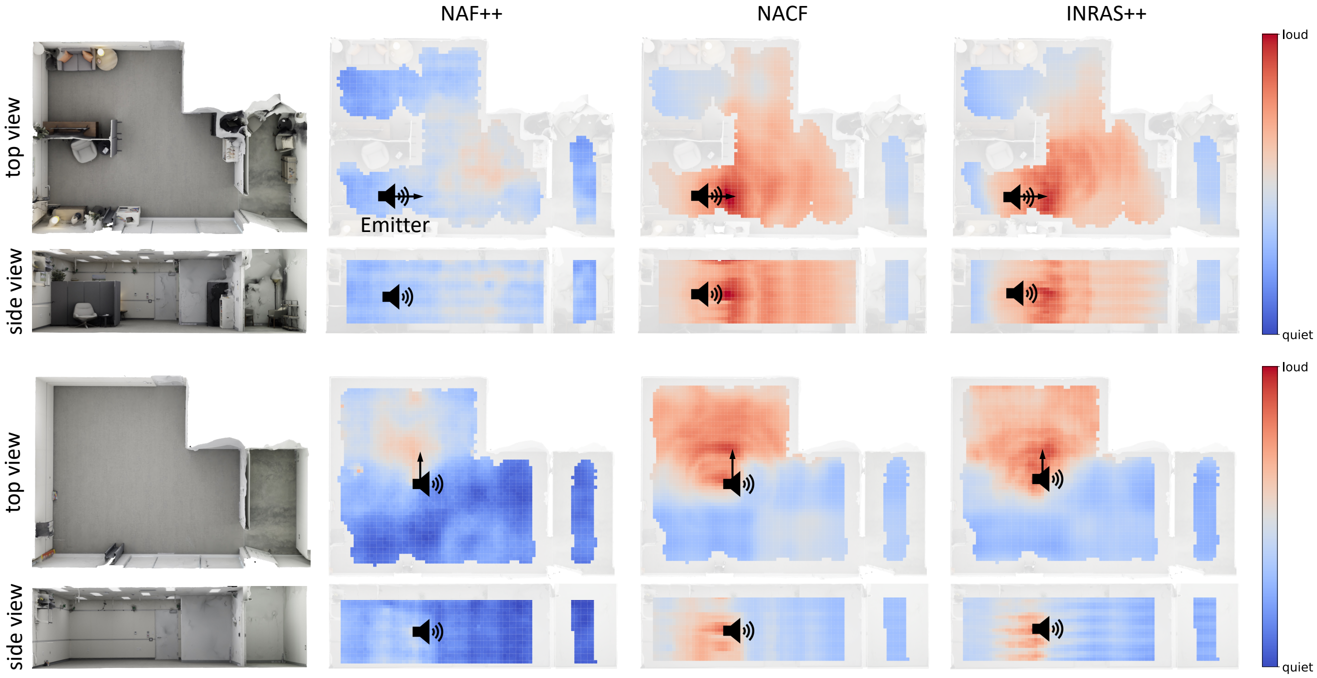

Figure 10. **Loudness map visualization.** We visualize the intensity of predicted impulse responses over the spaces from the top view and side view given an emitter position and its orientation. Red means loud and blue means quiet.

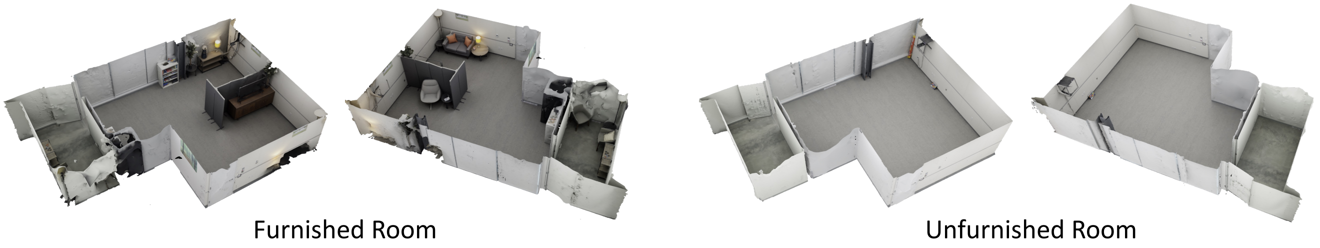

Figure 11. **Scene overview of RAF.**

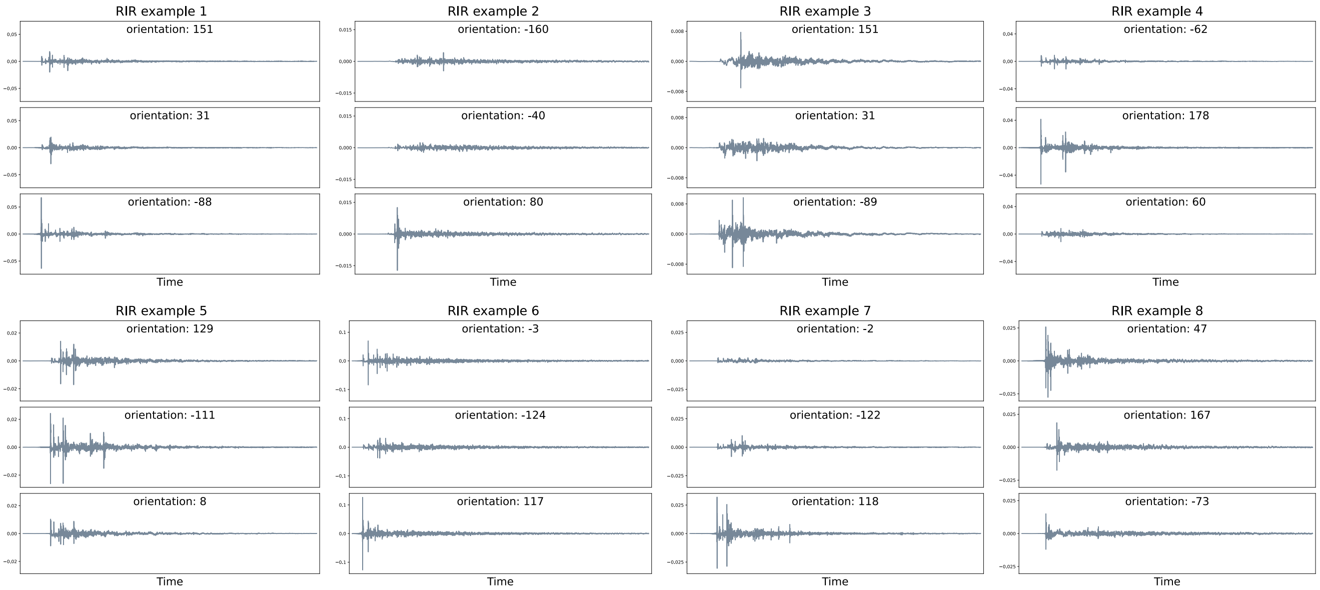

Figure 12. **Visualization of ground-truth RIRs with different orientations.**