Hearing Anywhere in Any Environment

Supplementary Material

7. Summary of Supplementary Materials

In our supplementary materials, we provide:

1. Supplementary video. 8
2. Additional details including explanation of reference RIR setup, implementations of xRIR and baselines as well as experiments setup, see Section 9.
3. More details about AcousticRooms, see Section 10.
4. Single-room RIR prediction results compared to prior works, see Section 11.
5. Additional experiments and ablation studies of the xRIR, see Section 12.
6. Additional qualitative samples, RIR prediction comparison 13.

8. Supplementary Video

In the supplementary video, we provide a brief summary of our work and qualitative samples of audio rendered with different predicted RIRs in simulation as well as real environments. For best perceptual experience, please turn the Audio ON and use headphone when watching.

For demos in the supplementary video that demonstrate audio rendering along trajectories in real scenes, we convolve our predicted single-channel Room Impulse Response (RIR) at each point in the trajectory with a Head-Related Impulse Response (HRIR) from a predefined Head-Related Transfer Function (HRTF). This process yields binaural RIRs, which capture spatial effects. We then convolve these binaural RIRs with the source audio to obtain the binaural audio along these trajectories.

9. Additional Details

Further explanation on reference RIR setup: Our reference RIRs setup is practically useful. Model trained on this setup is able to predict new RIRs without any *reference RIRs re-measurement* when either source or receiver freely moves in the target room. Our method is capable of addressing both scenarios below:

i) *Fixed reference receiver, multiple sources*: Measure reference RIRs between a fixed receiver and sources at various locations, then the model predicts the RIR for the fixed receiver and a new source location (AcousticRooms setup).

ii) *Fixed reference source, multiple receivers*: Measure reference RIRs between a fixed source and receivers at various locations, then the model predicts the RIR for the fixed source and a new receiver location (HearAnythingAnywhere setup).

A model trained on i) can be directly applied to ii) by switching source and receiver subscripts (due to symmetry of wave equation for single-channel RIR), as shown in our sim-to-real transfer experiment. We adopt scenario i) in our task, where the receiver always matches the reference receiver location.

xRIR: For xRIR, we implement a Vision Transformer block F_{vt} with 6 multi-head attention layers (8 heads, hidden size 512). For panorama depth map, the center pixel corresponds to the receiver location. Two spherical angles maps are initialized for equirectangular projection from the depth map to 3D coordinates map. In the vision transformer module, the 3D coordinates map is divided into 16×32 patches, resulting in all reflection-based features such as $g_{\text{r,rf}}$ and $g_{\text{s,rf}}$ of dimension 256×512. Direct path features are calculated using sinusoidal positional encoding on each 3D coordinate with 20 frequency bins, and are then projected into 256-dimensional vectors via MLP. Similarly, the time basis vector T_b is calculated by sinusoidal positional encoding with 10 frequency bins for each time index, where the length of T_b is same as the length of spectrogram, 310. Before performing weight combination, we further preprocess the reference RIRs by time-shifting them based on the distance difference between the target and reference source-receiver pairs, divided by the speed of sound. For loss calculation, we set $\lambda = 0.01$ to balance the STFT loss and the energy decay loss.

Few-Shot RIR: Unlike the approach and the problem setup in [32], which use binaural echoes where the source and receiver are co-located to predict a target binaural RIR, we use reference RIRs measured with the source at different locations from the receiver location to predict single-channel RIR at a target source. This is very important since the echo input used by [32] are infeasible to obtain under the single-channel RIR scenario, because it is not reasonable in physics to co-locate source and receiver at the same location to measure the single-channel RIR. In addition, we also omit the RGB image for the visual input and use only depth maps as inputs to the vision branch of the Few-shot RIR model, due to the weak correspondence between room semantics and material properties in AcousticRooms dataset. We also emphasize that we use panorama depth images captured from each reference source location instead of egocentric depth images as the depth inputs. For all depth observations, we rendered at a resolution of 128×256. Except for the adaptations above, all other implementation details follow the Few-Shot RIR model [32].

Diff-RIR: We use their released Github code and model

checkpoints to perform evaluations on all rooms in the Hearing-Anything-Anywhere Dataset [56]. We strictly followed the inference and evaluation settings in the paper, and obtained the same results in terms of the metric errors reported in the paper (Mag and Env error metrics) to make sure there are no implementation issues. And then we evaluate their inference results on our three acoustic metrics which are more related to perceptual quality of the RIRs: EDT error, C50 error and T60 error.

Experiments Setup: For cross-room RIR prediction experiments on ACOUSTICROOMS, we manually select 10 sources in each simulated room as candidate reference sources to make sure that their spatial locations are evenly distributed within the scene as much as possible. For the seen setting, we split training and test set within each room by receivers, where the RIRs of 90% of receivers belong to training split and remaining 10% belong to test split. For the unseen setting, we split the data by rooms. For each room category, we use 90% of rooms for training and 10% rooms for testing. During both training and testing, we randomly select $K = 1, 4, 8$ reference RIRs from these candidate sources. And for each K, we train a separate model on the dataset.

For experiments on the Hearing-Anything-Anywhere Dataset [56], we sample reference $K = 8$ RIRs from their selected 12 reference RIRs in each room. In this real-world dataset, the source has specific directivity patterns which are not captured in our pretrained model using ACOUSTICROOMS. Therefore, we further finetune the pretrained XRIR model on these 12 selected RIRs to make sure a fair comparison with Diff-RIR [56]. Specifically, in each iteration during finetuning, we randomly sample 8 of 12 as reference RIRs and predict a target RIR sampled from remaining 4 RIRs. We also use the same validation set as Diff-RIR to select the model checkpoint for testing. We select the one with lowest validation loss to evaluate on the test split of the dataset. It is note-worthy that even though XRIR is finetuned, compared to the training time of Diff-RIR on these few shot samples (6 hours / scene), XRIR converges much faster than Diff-RIR, with a matter of minutes. This helps us to quickly perform sim-to-real transfer across different environments efficiently.

10. Dataset Details

In this section, we present details about our large-scale simulated RIR dataset, ACOUSTICROOMS, used for cross-room RIR prediction task. ACOUSTICROOMS contains 260 rooms from 10 different categories, simulating a total of 30,000 RIRs from different source-receiver pairs. ACOUSTICROOMS features professional room architecture designs of high quality, covering a wide range of room categories, including: apartment, auditorium, bathroom, bedroom, cafe, listening room, living room, meeting room, office and restaurant, as shown in Figure 6. The area of rooms ranges from $20m^3$ to $1000m^3$, with diverse range of sizes and geometries. ACOUSTICROOMS uses a realistic and commercial acoustics simulation platform, Treble, to perform single-channel RIR simulation. The platform supports a wide variety of simulation methods along with specific settings. To obtain more realistic RIR data as well as simulate large-scale data, we adopt the hybrid-based simulation. At low frequency bands, an advanced wave-based method is used to capture more subtle wave interaction effects such as diffraction and resonance. At high frequency bands, we use geometric-based simulation that combines two simulation techniques: image-source technique and stochastic ray-tracing technique.

Source Receiver Placement: To set up the simulation for each room, we first choose the sources and receivers and place them at different locations. For sources, we use omnidirectional source devices without particular directivity patterns since our goal is not to overfit the model to a particular device pattern. Similarly, for receivers, we use monaural receivers such that they do not model specific HRTF patterns. Depending on the size of rooms, we place 10 to 100 sources and 25 to 100 receivers per room, to ensure they sufficiently cover the whole area of the rooms. To determine the location of each source and receiver, we apply a set of placement rules to avoid interference among devices and room surfaces when the distance becomes too small to cause issue in the simulation quality. We require that: i) Sources should be at least $0.5m$ away from each room surface, $1.0m$ away from other sources and $1.0m$ away from receivers. ii) Receivers should be at least $0.5m$ away from room surface and at least $0.5m$ away from the sources. Given these rules, we apply a point-picking algorithm to randomly sample valid source and receiver locations within each room at different height level from $0.5m$ to $2.5m$.

Material Assignment: Once the source and receivers are determined, we assign materials to room surfaces by associating their semantic labels with particular material category. Treble platform provides a large-scale material database with 332 specific materials from 11 material categories, with each category containing 30 different material coefficients on average. We define the mapping between each semantic labels of room surfaces and the 11 material categories. In each room, each semantic label of a particular surface gets randomly mapped to one of the specific material with a set of acoustics coefficients under the material category. Different from existing RIR datasets [3, 6, 52], this random assignment ensures enough diversity in the material properties of the room surfaces. Even two rooms share similar geometries and semantic objects, they could have very different acoustics behavior due to differences in their specific acoustics coefficients of materials.

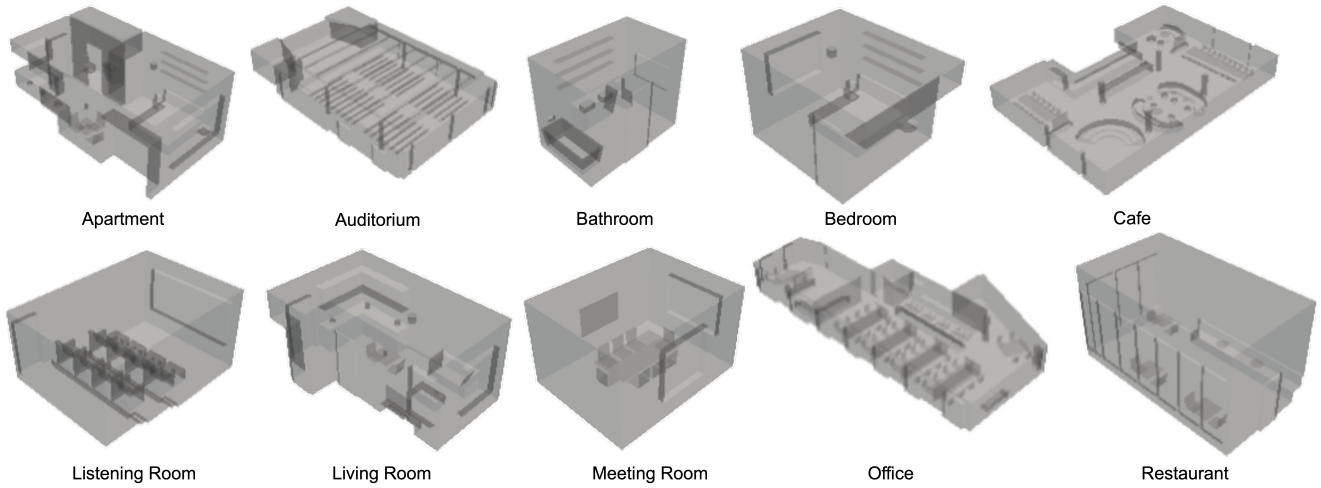

Figure 6. A visualization of different room categories in ACOUSTICROOMS.

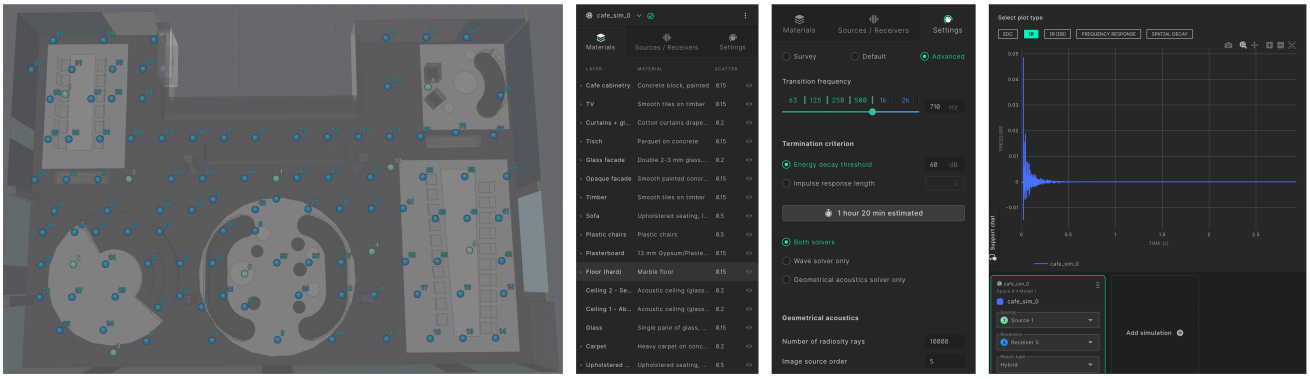

Figure 7. An overview of procedures to simulate RIRs in ACOUSTICROOMS.

Simulation Setup: Once sources and receivers are in place and room materials get assigned, we set up the simulation by specifying the hybrid mode to split the geometric-based and wave-based method. We choose the crossover frequency to be $f_{\text{cross}} = 710$Hz between two methods such that wave-based effects could be sufficiently captured across different rooms and objects of different size. For geometric-based simulation, we use image source method up to reflection order of 4 with $50k$ rays emitting from the source. For refection orders higher than 4, we apply stochastic ray-tracing method with 5000 rays. With this configuration, we simulate the RIR of each room until $60dB$ energy decay to ensure all possible acoustics effects are sufficiently captured.

11. Comparison on Single-Room RIR Prediction

Adaptation of xRIR: Although xRIR focuses on solving cross-room RIR prediction task, it could be easily adapted to single-room RIR prediction task as well. By removing all the components related to reference RIRs, we extract the geometric and spatial features related to only target source and receiver positions. We then follow [50] to use their implicit neural decoder to perform RIR waveform synthesis. Specifically, given a time basis vector $\mathbf{B}$, and the outputs from Geometric feature extractor g_{dir}, $g_{r,\text{rf}}$ and $g_{s,\text{rf}}$, we learn a implicit neural mapping function to synthesize time domain RIR waveform from the outer product between $\mathbf{B}$ with the three features: $\hat{A}_t = F_{\text{inr}}(g_{dir}\mathbf{B^T}, g_{r,\text{rf}}\mathbf{B^T}, g_{s,\text{rf}}\mathbf{B^T})$. We train the adapted model with same loss function as in [50], the multi-resolution STFT loss combined with waveform L2 loss.

Model	Apartment 1			Apartment 2			FRL Apartment 2		
	EDT↓	C50 ↓	T60 ↓	EDT ↓	C50 ↓	T60 ↓	EDT ↓	C50 ↓	T60 ↓
NAF	0.077	0.426	7.508	0.066	0.453	7.925	0.088	0.420	6.308
INRAS	0.027	1.036	6.514	0.025	0.843	5.816	0.022	0.634	2.224
xRIR	0.026	1.000	6.192	0.031	0.932	5.755	0.021	0.587	1.972

Model	Room 2			FRL Apartment 4			Office 4			Mean		
	EDT ↓	C50 ↓	T60 ↓	EDT ↓	C50 ↓	T60 ↓	EDT ↓	C50 ↓	T60 ↓	EDT ↓	C50 ↓	T60 ↓
NAF	0.056	0.407	4.969	0.085	0.421	7.475	0.081	0.337	6.760	0.076	**0.411**	6.824
INRAS	0.020	0.555	1.990	0.022	0.625	2.145	0.014	0.610	3.251	0.022	0.717	3.657
xRIR	0.019	0.541	1.910	0.021	0.561	2.140	0.013	0.502	2.767	**0.022**	0.687	**3.456**

Table 3. Performance comparison on single-room RIR prediction task on six rooms in SoundSpaces 1.0 - Replica dataset in terms of EDT (s), C50 (dB), and T60 (%) error metrics.

Experiment Setup: In single-room RIR prediction task, the goal is to the fit scene acoustic with dense RIR observations. Therefore, we use the standard dense RIR dataset, SoundSpaces 1.0 Replica, to perform the experiment. Following prior works [31, 50], we use the six scenes from Replica. But instead of using binaural RIR data, we use single-channel RIR data by extracting the first channel of ambisonic RIR data of these scenes. For each scene, we split the RIR data into training and test set with a ratio of 9:1. We cut the RIR to the maximum length of 8000 samples (0.363s) at sampling rate 22,050Hz for all six rooms. And for panorama image at each receiver location, we render it by setting the orientation to 0 in the habitat simulator [43].

Baselines: We compare xRIR with two prior works of the state-of-the-art performance on the dataset, NAF [31] and INRAS [50]. For both methods, we remove the orientation conditioning vector to adjust for single-channel RIR prediction, while keeping the remaining implementations the same. For evaluation, we use the same metrics in the cross-room RIR prediction task for comparison.

Quantatitive Results: We report individual results for each of six scenes and their average results and show them in Table 3. As could be seen, when compared to NAF, xRIR outperforms on both EDT and T60 metrics, while slightly underperforming in C50 error. When compared to INRAS, xRIR outperform INRAS on 5 out of 6 scenes, except "Apartment 2". The reason is due to the fact that the mesh of "Apartment 2" shows significant amount of holes in a region, which leads to degraded quality of panorama depth inputs. While, INRAS does not suffer from this degradation since the method samples mesh points instead of rendering images. Overall, xRIR, when adapted to single-room RIR prediction task, shows on par or even better performance when compared to these prior arts, demonstrating the effectiveness of Geometric Feature Extractor to learn the scene acoustics from local geometric observations (around receiver).

Method	Classroom		Dampened		Hallway		Complex	
	MAG	ENV	MAG	ENV	MAG	ENV	MAG	ENV
Random Across	1.98	4.45	3.571	7.573	3.415	7.571	1.762	6.791
Random Same	0.710	2.182	0.213	1.635	1.104	6.582	0.685	2.831
Linear Interp	0.725	1.890	0.110	0.908	1.082	5.566	0.637	2.370
Nearest Neigh	0.600	2.003	0.108	0.916	0.793	5.589	0.542	2.498
Diff-RIR (K=12)	0.486	1.826	**0.085**	**0.883**	0.724	**5.173**	**0.442**	2.197
xRIR (K=8)	**0.456**	**1.824**	0.093	0.892	**0.718**	5.320	0.466	**2.142**

Table 4. **Sim-to-Real Transfer Results** using MAG and ENV metrics on HearingAnythingAnywhere dataset.

Method	EDT	C50	T60
xRIR w.o Reference RIRs	0.166	3.925	32.69
xRIR w.o Direct Path Module	0.061	1.596	12.33
xRIR w.o Reflection Module	0.059	1.498	11.93
xRIR (full)	**0.055**	**1.457**	**10.53**

Table 5. Ablation Study: Comparison of different ablated xRIR components on AcousticRooms with unseen splits. We report EDT (s), C50 (dB), and T60 (%) error metrics

12. Additional Experiments and Ablation Studies

Additional Comparisons on HearingAnythingAnywhere dataset: We use MAG from [8] and ENV from [24] to further evaluate waveform similarity. For fair comparison, we evaluate our model on the first 0.435s of RIRs due to model length constraints. As shown in Table 4, our method performs similarly to DiffRIR on these metrics but outperforms in perceptual metrics like EDT and C50 (Table 2 main paper), which better reflect perceptual

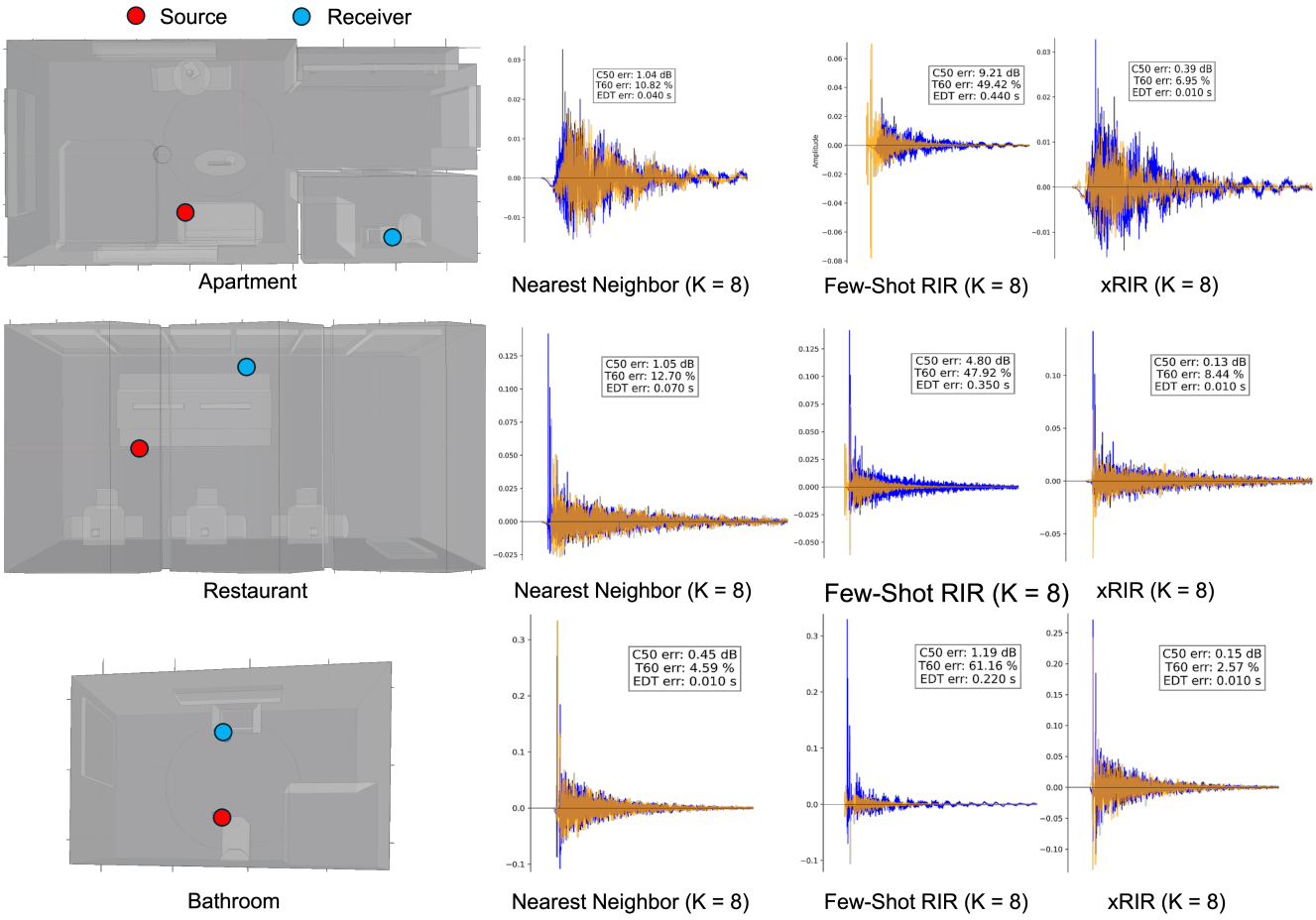

Figure 8. Additional qualitative comparisons on RIR waveform predictions in AcousticRooms.

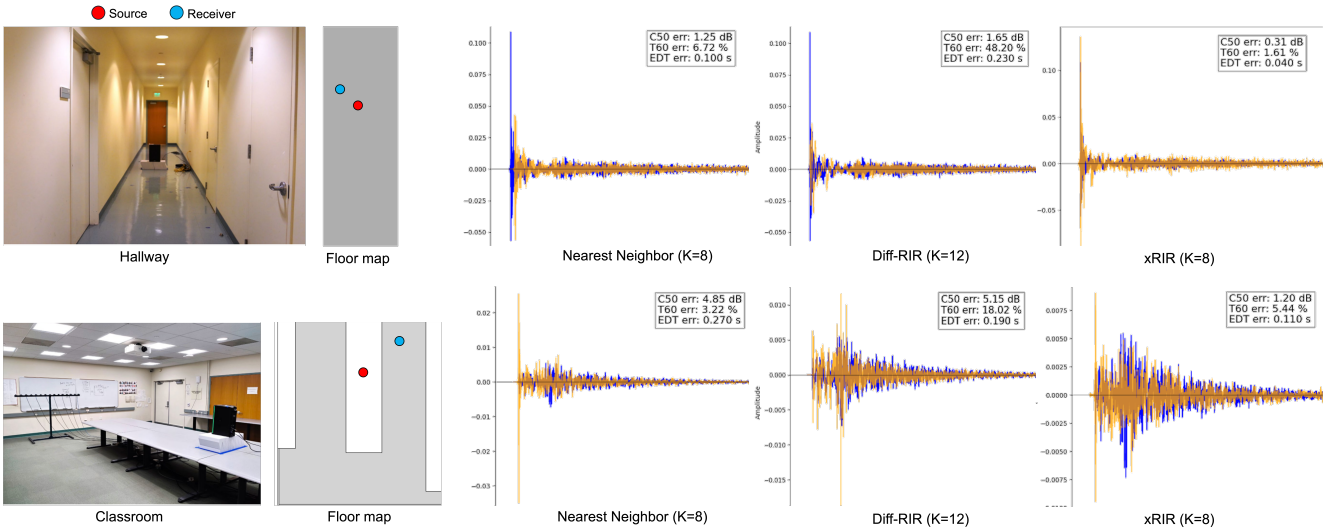

Figure 9. Additional qualitative comparisons on RIR waveform predictions on the Hearing-Anything-Anywhere Dataset.

quality of rendered RIRs.

Ablation Studies:

We perform ablation studies on the components of xRIR by considering the followings:

- xRIR w.o Reference RIRs: This ablated variant is the same as our model's adaptation to single-room RIR pre-

diction task. Since there are no reference RIRs as inputs, the model utilizes the implicit neural function to synthesize the target RIR.

- xRIR w.o Direct Path Module: By removing the Direct Path Module, the model only considers the relationship between sources / receivers and room geometry information, without computing the direct path features.
- xRIR w.o Reflection Module: By removing the Reflection Module, the model only considers the spatial relationship between sources and receivers without taking local geometry information into account.

For the above ablation variants, we perform experiments on AcousticRooms under unseen settings. We set the number of reference RIRs $K = 8$ for models that take reference RIRs as inputs. As shown in Table 5, our full model outperforms all ablated variants across all metrics. It is note-worthy that without providing reference RIR as inputs to the model, the model is not able to synthesize reasonable RIRs by just relying on geometric and positional inputs. Also, removing either Direct Path module or Reflection module will lead to degraded performance across all acoustics metrics, demonstrating the importance of capturing full spatial and geometric information for accurate RIR predictions.

Furthermore, we study the importance of finetuning as well as the pretraining on our simulation dataset. In general, we find that simulation data helps the model capture general acoustic properties, such as geometry and material effects. Finetuning on just 12 real samples allows the model to adapt to specific factors like the source's directivity, improving EDT and C50 metrics compared to training from scratch or without finetuning, as shown in Table 6.

Setting	EDT (s)	C50 (dB)	T60 (%)
Scratch	0.322	4.322	7.381
Pretrained	0.204	3.427	**4.685**
Finetuned	**0.092**	**1.614**	6.020

Table 6. Importance of finetuning and pretraining on HearingAnythingAnywhere dataset (classroom).

In addition, to study the impact of scale of AcousticRooms on the performance of model, we retrain our model on different number of rooms using xRIR (8-shot) in the unseen setting, while keeping the test split the same. As shown in Table 7, performance improves with more data but diminishes as room count increases.

Rooms	EDT (s)	C50 (dB)	T60 (%)
65	0.088	1.813	13.79
130	0.062	1.578	11.46
260	**0.055**	**1.457**	**10.53**

Table 7. Impact of data scale on model performance.

13. Additional Qualitative Samples

We provide additional qualitative results of comparisons between our model xRIR and the baseline methods on the predicted RIRs on both the simulation dataset and real dataset.

As shown in Figure 8 and 9, we visualize predicted RIR waveforms versus ground truth RIRs on three simulation environments (apartment, restaurant and bathroom) in AcousticRooms as well as two real environments (hallway and classroom) in the Hearing-Anything-Anywhere dataset. At same location in these environments, RIRs predicted by xRIR align more closely with the ground truth RIRs, demonstrating the effectiveness of xRIR in RIR predictions under both simulated and real settings.

References

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

[2] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiankang Deng, and Xiatian Zhu. Av-gs: Learning material and geometry aware priors for novel view acoustic synthesis. *arXiv preprint arXiv:2406.08920*, 2024. 2

[3] 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, 5

[4] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[5] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868, 2022. 3

[6] 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. *Advances in Neural Information Processing Systems*, 35:8896–8911, 2022. 2

[7] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6409–6419, 2023. 3

[8] Mingfei Chen and Eli Shlizerman. AV-cloud: Spatial audio rendering through audio-visual cloud splatting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 4

[9] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. *Conference on Robot Learning (CoRL)*, 2021. 3

[10] Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21886–21896, 2024. 1, 2

[11] Sanjoy Chowdhury, Sreyan Ghosh, Dasgupta Subhrajyoti, Anton Ratnarajah, Utkarsh Tyagi, and Dinesh Manocha. Adverb: Visually guided audio dereverberation. *ICCV*, 2023. 3

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

[13] Elke Deckers, Onur Atak, Laurens Coox, Roberto D'Amico, Hendrik Devriendt, Stijn Jonckheere, Kunmo Koo, Bert Pluymers, Dirk Vandepitte, and Wim Desmet. The wave based method: An overview of 15 years of research. *Wave Motion*, 51(4):550–565, 2014. 2

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[15] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 3

[16] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[17] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial visual representation learning through echolocation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[18] Ruohan Gao, Hao Li, Gokul Dharan, Zhuzhu Wang, Chengshu Li, Fei Xia, Silvio Savarese, Li Fei-Fei, and Jiajun Wu. Sonicverse: A multisensory simulation platform for training household agents that see and hear. In *International Conference on Robotics and Automation (ICRA),*, 2023. 3

[19] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. In *British Machine Vision Conference (BMVC)*, 2021. 3

[20] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. In *International Journal of Computer Vision (IJCV)*, 2023. 3

[21] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. 5

[22] Reinhold Haeb-Umbach, Jahn Heymann, Lukas Drude, Shinji Watanabe, Marc Delcroix, and Tomohiro Nakatani. Far-field automatic speech recognition. *Proceedings of the IEEE*, 109(2):124–148, 2020. 2

[23] Brian Hamilton and Craig J. Webb. Room acoustics modelling using gpu-accelerated finite difference and finite volume methods on a face-centered cubic grid. 2013. 5

[24] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. *arXiv preprint arXiv:2411.06307*, 2024. 2, 4

[25] Tobias Lentz, Dirk Schröder, Michael Vorländer, and Ingo Assenmacher. Virtual reality system with integrated sound field simulation and reproduction. *EURASIP journal on advances in signal processing*, 2007:1–19, 2007. 2

[26] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 3

[27] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490, 2023. 2

[28] 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

[29] Xiulong Liu, Sudipta Paul, Moitreya Chatterjee, and Anoop Cherian. Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3765–3773, 2024. 3

[30] Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see - video to audio generation through text. In *Advances in Neural Information Processing Systems*, pages 101337–101366. Curran Associates, Inc., 2024. 8

[31] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022. 1, 2, 3, 6, 4

[32] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 35:2522–2536, 2022. 2, 5, 6, 7, 1

[33] 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. 2

[34] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 2018. 3

[35] Brady Peters. Integrating acoustic simulation in architectural design workflows: the fabpod meeting room prototype. *Simulation*, 91(9):787–808, 2015. 2

[36] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. Irgan: Room impulse response generator for far-field speech recognition. *arXiv preprint arXiv:2010.13219*, 2020. 2

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

[38] 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

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

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

[41] W H Reed and T R Hill. Triangular mesh methods for the neutron transport equation. 1973. 2, 6

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

[43] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 4

[44] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 351–355. IEEE, 2018. 2

[45] Carl Schissler, Gregor Mückl, and Paul Calamia. Fast diffraction pathfinding for dynamic sound propagation. 40 (4), 2021. 2

[46] Manfred R Schroeder. New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(6_Supplement):1187–1188, 1965. 6

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

[48] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems*, 33:7462–7473, 2020. 2

[49] Arjun Somayazulu, Changan Chen, and Kristen Grauman. Self-supervised visual acoustic matching. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[50] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 1, 2, 3, 4, 6

[51] Kun Su, Xiulong Liu, and Eli Shlizerman. From vision to audio and beyond: A unified model for audio-visual representation and generation. *arXiv preprint arXiv:2409.19132*, 2024. 8

[52] Zhenyu Tang, Rohith Aralikatti, Anton Jeran Ratnarajah, and Dinesh Manocha. Gwa: A large high-quality acoustic dataset for audio processing. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 5

[53] Po-Yao Huang Andrew Owens Gopala Anumanchipalli Tingle Li, Renhao Wang. Self-supervised audio-visual soundscape stylization. In *ECCV*, 2024. 3

[54] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4, 5

[55] Mason Wang, Samuel Clarke, Jui-Hsien Wang, Ruohan Gao, and Jiajun Wu. Soundcam: a dataset for finding humans using room acoustics. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[56] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11790–11799, 2024. 2, 6, 7, 8