Differentiable Room Acoustic Rendering with Multi-View Vision Priors (Supplementary Material)

Anonymous ICCV submission

Paper ID 4

Contents

1. Method Details

1.1. Acoustic Beam Tracing Algorithm

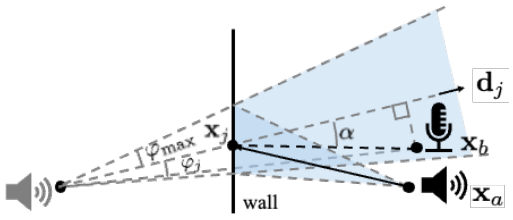

Figure 1. Acoustic beam tracing: in acoustic beam tracing the source and listener are considered as two point, the sound is propagate via a cone-shape beam in space. Acoustic beam tracing handles reflection the same as ray tracing does. The key difference is that acoustic beam tracing enumerate a reflection path if the listener is contatined in the beam volume but not necessarily being hitted by the sampled ray

Given the source location $\mathbf{x}_a$ and listener location $\mathbf{x}_b$, we adopt acoustic beam tracing [2, 4, 7, 12] to sample specular beams in a source-to-listener manner. First we cast N_d beams from the source, using a Fibonacci lattice [3] to approximate uniform coverage of directions. A small apex angle $2\varphi_{\max}$ is selected to ensure the cone-shape beams remain disjoint. Next, each beam's center ray intersects with room geometry to find reflection points (e.g. via Open3D [14]), and after each reflection, we check if the reflected beam can hit the listener. To determine whether a reflected beam at j-th reflection point $\mathbf{x}_j$ (with out-going direction $\mathbf{d}_j$) reaches the listener before hitting another surface, we check if the listener is within the reflected cone (as show in Figure 1). Denote l_j as the distance traveled by reaching $\mathbf{x}_j$, and α_j as the angle between $\mathbf{d}_j$ and the line from $\mathbf{x}_j$ to $\mathbf{x}_b$ and φ_j as the sampled half-apex angle:

$$\varphi_j = \arctan\left(\frac{\|\mathbf{x}_b - \mathbf{x}_j\| \sin\alpha}{\|\mathbf{x}_b - \mathbf{x}_j\| \cos\alpha + l_j}\right). \tag{1}$$

The listener is considered "hit" if α is acute, $\varphi_j < \varphi_{\max}$, and $\mathbf{x}_j$ is visible by $\mathbf{x}_b$. In addition, the time-of-arrival is by:

$$\mathrm{toa}_j = \frac{\|\mathbf{x}_b - \mathbf{x}_j\| \sin\alpha}{v_{\text{sound}} \cdot \sin\varphi_j}. \tag{2}$$

Algorithm 1 summarizes our beam-tracing procedure.

1.2. Local Variance Derivation

As shown in Figure 2, consider a beam traveling distance l before hitting the surface at $\mathbf{x}$, with half-apex angle φ and local surface normal $\mathbf{z}$. Let θ be the angle between the reflected direction $\mathbf{d}$ and $\mathbf{z}$. In a local coordinate system whose axes are $\{\mathbf{t}_1, \mathbf{t}_2, \mathbf{z}\}$, where we requires $\mathbf{t}_1$ aligns with the projection of $\mathbf{d}$ in the tangent surface, the beam's cross-section at distance l is approximately an ellipse with semi-major and semi-minor axes proportional to $l\sin\varphi$, modulated by θ. A simple way to encode this elliptical patch is to use a diagonal covariance at local coordinate

$$\Sigma_{\text{local}} = \mathrm{diag}\Big(\sigma_1^2,\ \sigma_2^2,\ 0\Big), \tag{3}$$

where σ_1^2 and σ_2^2 grow with $l\sin\varphi$, adjusted by $\cos\theta$. In the case when φ is small:

$$\sigma_1^2 \approx \big(l\sin\varphi\big)^2/\cos^2\theta, \quad \sigma_2^2 \approx \big(l\sin\varphi\big)^2/\cos\theta.$$

Algorithm 1: Acoustic Beam Tracing

Input: Source $\mathbf{x}_a$, Listener $\mathbf{x}_b$, Geometry $\mathcal{M}$
Output: Specular paths $\{\tilde{\mathbf{x}}_k\}_{k=1}^{N}$
for $i = 1$ *to* N_d **do**
  $\mathbf{x}_{i,0} \leftarrow \mathbf{x}_a$; $l_{i,0} \leftarrow 0$;
  $\mathbf{d}_{i,0} \leftarrow$ `SampleFib`(N_d, i)
end
ANS $\leftarrow \{\}$
if *IsVisible*$(\mathbf{x}_a, \mathbf{x}_b)$ **then**
  ANS.add$(\varnothing)$ // direct path
end
for $j = 1$ *to* $\mathrm{MAX}_{\mathrm{depth}}$ **do**
  for $i = 1$ *to* N_d **do**
    $[\mathbf{x}_{i,j}, \mathbf{z}] =$ `HitPoint`$(\mathcal{M}, \mathbf{x}_{i,j-1}, \mathbf{d}_{i,j-1})$
    $\mathbf{d}_{i,j} = \mathbf{d}_{i,j-1} - 2\,(\mathbf{z}^\top \mathbf{d}_{i,j-1})\,\mathbf{z}$
    $l_{i,j} = l_{i,j-1} + \|\mathbf{x}_{i,j} - \mathbf{x}_{i,j-1}\|$
    if *BeamHit*$(\mathbf{x}_b, \mathbf{x}_{i,j}, \mathbf{d}_{i,j}, l_{i,j})$ **then**
      ANS.add$\big([\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,j}]\big)$
    end
  end
end
return ANS

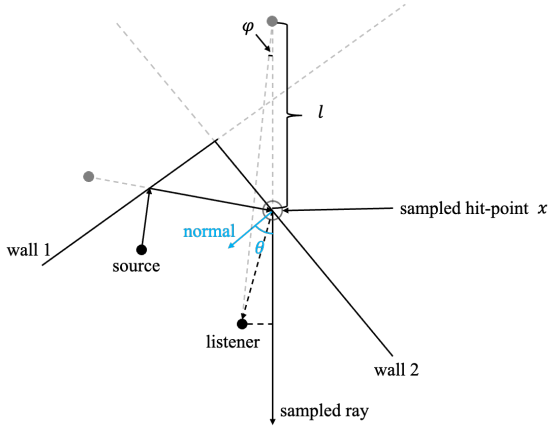

Figure 2. Local covariance derivation: as the traveling space l increases, the region of the contact area expand linearly in terms of radius. In addition, since the half-apex angle is assumed to be small, the contact region is considered an ellipse, which motivates use model the region information with a gaussian distribution.

These terms capture how the beam's ellipse "stretches" along $\mathbf{t}_1$ and $\mathbf{t}_2$. In world coordinates, the final covariance Σ is simply

$$\Sigma \;=\; Q\,\Sigma_{\mathrm{local}}\,Q^\top,$$

where $Q = [\,\mathbf{t}_1\;\mathbf{t}_2\;\mathbf{z}\,]$ rotates from local axes to world axes.

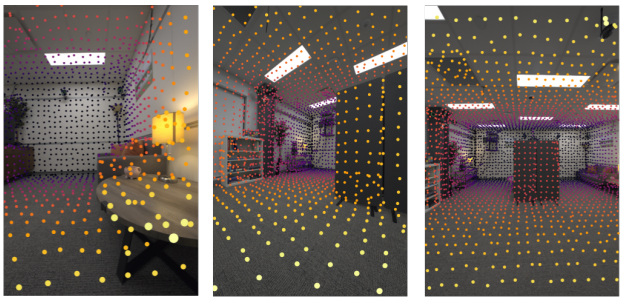

Figure 3. Visualization of surface basis samples for extracting multi-view images features.

1.3. Basis Points Sampling

we sample the basis point in two steps, first we densely sample 100,000 points on the room geometry, then, we downsample them with voxel size 0.2m and use the median point (closest to mean point) as the basis samples for vision features, as shown in Figure 3, in this way, we ensures the distances between samples are stable.

1.4. Hyperparameters

Following [13], we use a spherical Gaussian weighting function with a sharpness parameter of 8 for source directional response. We decode the image feature using a 4-layer MLP and sample frequencies from 12 to 7800 Hz with 16 logarithmically spaced samples, linearly interpolating the frequency response.

1.5. Optimization

We optimize the network using the AdamW optimizer with a fixed learning rate of 5×10^{-4} (and 1×10^{-4} for the residual component). Our loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MAG}} + \lambda_{\mathrm{pink}}\mathcal{L}_{\mathrm{pink}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{MAG}}$ is a multi-scale log L1 loss, and $\mathcal{L}_{\mathrm{pink}}$ is the pink noise supervision loss. We adopt a progressive training strategy, starting with a reflection order $N = 1$ and increasing by 1 every 100 epochs until $N = 6$. During training, we sample 16,384 points from Fibonacci lattices for beam tracing, reducing this to 8,192 points per RIR during inference. Training is performed with a batch size of 1.

2. Additional Results on RAF [1] Dataset and HAA [13] Dataset

2.1. Wave Comparison

Figure 4 shows wave visualizations on the Hearing Anything Anywhere dataset. All models were trained on only 12 data points. Our model significantly outperforms the baselines in preserving the wave structure, producing a wave front that closely matches the ground truth in terms of peak

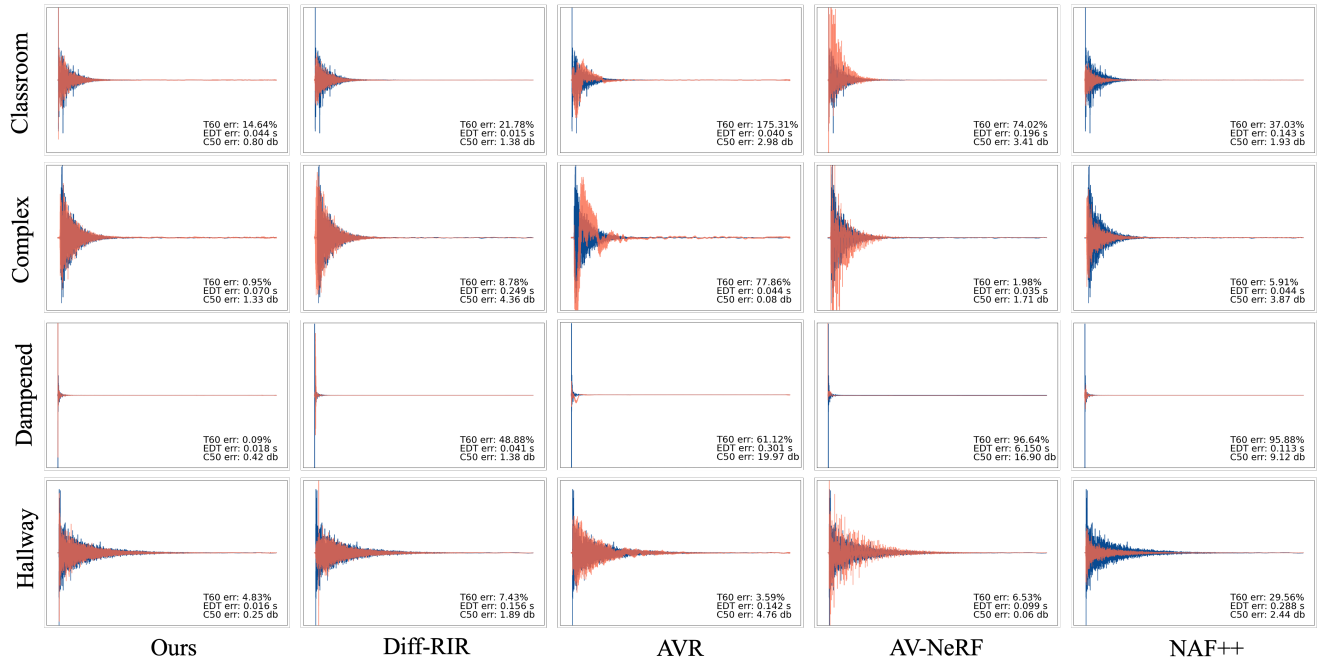

Figure 4. Wave visualization on the Hearing Anything Anywhere dataset [13]. All models are trained on 12 data points. Our model significantly outperforms all baselines in preserving the wave structure—producing the most faithful wave front with accurate peak locations and magnitudes. Note that quantitative metrics do not always capture these perceptual details; some methods may have low error values despite producing distorted wave patterns.

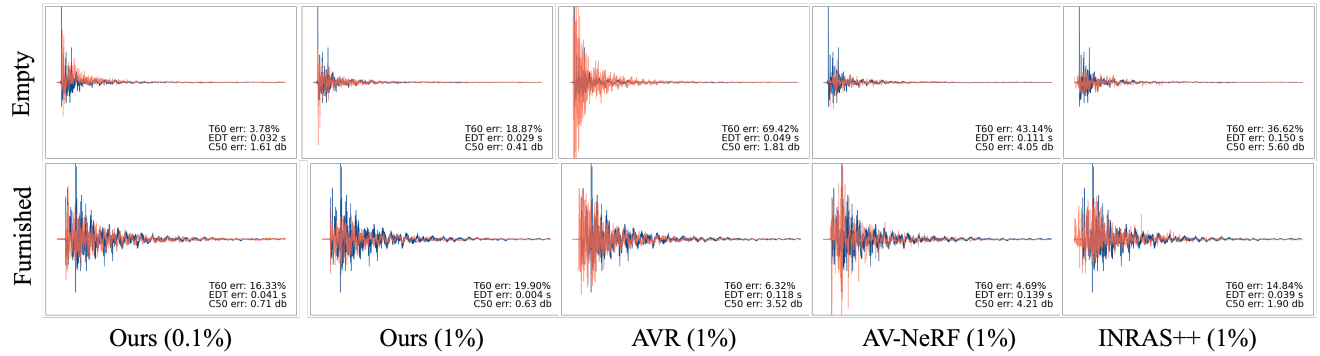

Figure 5. Wave visualization on the Real Acoustic Field dataset [1]. We show results from three baseline models trained on 1% of the data alongside our model trained on 1% and 0.1% of the data. Our model exhibits better peak alignment and magnitude than baseline methods—even when trained on only 0.1% of the data—and significantly outperforms all baselines when using the same amount of training data.

locations and magnitudes. Note that quantitative metrics do not always capture these perceptual differences; some methods may achieve low error values despite generating distorted wave patterns. This comparison highlights the superior capability of our approach in modeling acoustic dynamics in few-shot settings.

Figure 5 presents wave visualizations on the Real Acoustic Field dataset. Here, we compare three baseline models trained on 1% of the data with our model trained on both 1% and 0.1% of the data. Our results demonstrate that, in terms of wave structure, our model achieves better peak alignment and peak magnitude than the baselines—even when our model is trained on only 0.1% of the data. When trained on 1% of the data, our method further outperforms the baselines.

2.2. Multi-scale Performance Comparison

Figure 6 extend the multi-scale performance comparison in main paper by evluating on two more metrics, i.e., Loudness and EDT. The result shows that our model performs consistantly better than baselines in all training data scale, which is aligned with our observation in the main paper.

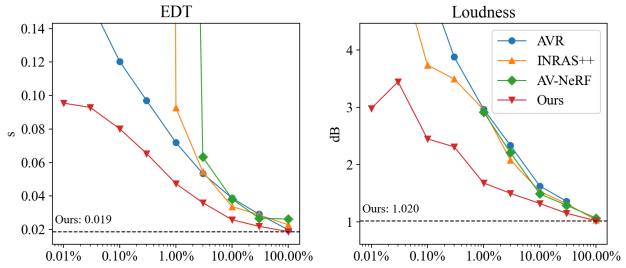

Figure 6. Performance comparison across training scales (from 0.01% to 100% of training data). In addition to the metrics reported in the main paper, our model consistently outperforms the baselines in terms of both EDT and Loudness.

2.3. Full Metric on HAA Dataset

Tables 2 and 3 present the complete evaluation metrics on the HAA dataset, including Loudness, C50, EDT, and T60 across four scenes. Our results show that our method outperforms state-of-the-art baselines across almost all metrics, confirming the trends observed in the main paper. The only exception is the C50 metric and EDT metric in the *Hallway* scene, where AV-NeRF performs particularly well, likely due to its effective use of depth information in this constrained geometry. These comprehensive results validate the robustness and effectiveness of our model in diverse real-world acoustic environments.

2.4. Ablations on Vision Features

We investigate the impact of vision features by varying two aspects: the number of multi-view images used for training and the choice of the pretrained encoder. Both experiments are conducted on the RAF Furnished scene using only 0.1% of the training data.

In our vision feature saturation experiment, we initially use 65 images to cover the entire scene, then reduce the number to 20, 10, and 5 views (see top four rows of Table 1). Reducing from 65 to 20 views incurs less than a 1% drop in C50 and EDT, but further reduction from 20 to 10 views causes a marked performance decline, indicating that adequate view redundancy is essential for effective visual guidance. Performance remains stable when further reduced from 10 to 5 views, suggesting that with only 10 views the model nearly abandons visual feature learning and relies primarily on acoustic cues.

We also replace the DINO-v2 [10] encoder with ResNet18 [5], which results in a noticeable drop in EDT, demonstrating that DINO-V2 is better suited for our model. Notably, all vision ablations have minimal impact on T60, indicating that vision features primarily contribute to modeling early reflection phenomena rather than late reverberation.

Variant	C50	EDT	T60
65 views	1.98	80.1	15.2
20 views	2.01	80.9	15.7
10 views	2.13	97.9	15.2
5 views	2.12	97.2	15.3
ResNet18	1.96	89.4	15.3

Table 1. Ablation study on vision features. “65 views” denotes using 65 images for training; “20 views”, “10 views”, and “5 views” denote reduced image sets. “ResNet18” indicates replacing the DINO-V2 encoder with ResNet18.

Method	Classroom				Complex Room			
	Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓	Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓
NAF++ [9]	8.27	1.62	162.3	134.0	4.43	2.25	203.5	44.8
INRAS++ [11]	1.31	1.86	110.0	60.9	1.65	2.26	150.7	29.5
AV-NeRF[8]	1.51	1.43	77.8	50.0	2.01	1.88	107.9	36.6
AVR [6]	3.26	4.18	135.6	44.3	6.47	2.55	138.3	36.7
Diff-RIR [13]	2.24	2.42	139.7	39.7	1.75	2.23	129.5	18.5
Ours	0.99	1.02	55.5	24.3	0.98	1.44	86.5	10.8

Table 2. Result on Diff-RIR [13] dataset, 2.0s, 16K sample rate

Method	Dampened Room				Hallway			
	Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓	Loudness (dB) ↓	C50 (dB) ↓	EDT (ms) ↓	T60 (%) ↓
NAF++ [9]	3.88	4.24	360.0	306.9	8.71	1.36	148.3	21.4
INRAS++ [11]	3.45	3.28	187.1	382.9	1.55	1.87	157.9	7.4
AV-NeRF [8]	2.40	3.05	242.1	107.9	1.26	1.03	89.9	9.5
AVR [6]	6.65	11.11	305.3	81.4	2.48	2.69	195.4	7.0
Diff-RIR [13]	1.87	1.56	153.0	44.9	1.32	3.13	188.1	6.8
Ours	1.11	1.45	139.0	31.9	0.85	1.15	96.5	6.3

Table 3. Result on Diff-RIR [13] dataset, 2.0s, 16K sample rate

References

[1] 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, 3

[2] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '98, page 21–32, New York, NY, USA, 1998. Association for Computing Machinery. 1

[3] Douglas P. Hardin, Timothy Michaels, and Edward B. Saff. A comparison of popular point configurations on $\mathbb{S}^2$. *Dolomites Research Notes on Approximation*, 9, 2016. 1

[4] John Kenneth Haviland and Balakrishna D. Thanedar. Monte carlo applications to acoustical field solutions. *The Journal of the Acoustical Society of America*, 54(6):1442–1448, 12 1973. 1

[5] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 4

[6] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 5

[7] Christian Lauterbach, Anish Chandak, and Dinesh Manocha. Interactive sound rendering in complex and dynamic scenes using frustum tracing. *IEEE Transactions on Visualization and Computer Graphics*, 13:1672–1679, 2007. 1

[8] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 5

[9] 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. 5

[10] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[11] Kun Su, Mingfei Chen, and Eli Shlizerman. INRAS: Implicit neural representation for audio scenes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 5

[12] Dirk van Maercke and Jacques Martin. The prediction of echograms and impulse responses within the epidaure software. *Applied Acoustics*, 38(2):93–114, 1993. 1

[13] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 1, 2, 3, 5

[14] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 1