Hearing Anything Anywhere

Mason Wang*, Ryosuke Sawata*, Samuel Clarke, Ruohan Gao, and Jiajun Wu

Stanford University · Sony AI · University of Maryland

MOTIVATION

We want to **capture** and **reconstruct** the spatial acoustic characteristics of a **real** room, to synthesize **immersive auditory experiences.**

Existing methods require **hundreds** of measurements – ours outperforms with only:

- ~12 monoaural room impulse response (RIR) recordings
- A rough planar reconstruction of the room

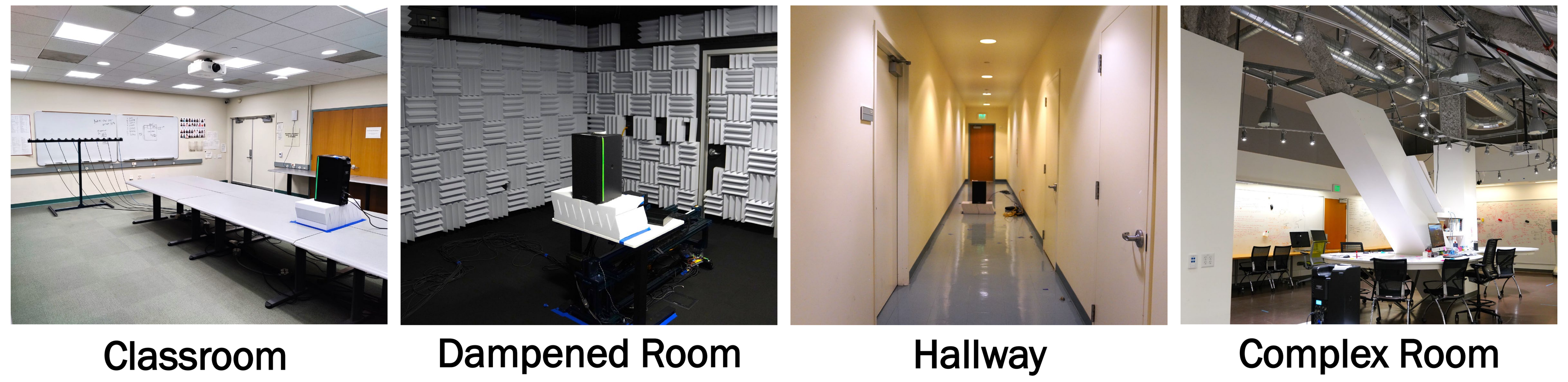
Using this data, we fit a differentiable acoustic inverse rendering framework containing **interpretable parametric models** of the scene's acoustic features, including **surface reflectivity** and **source directivity.**

DIFFRIR can:

- Render **accurate monoaural** and **binaural RIRs and music** at new listener locations
- Render **immersive** trajectories simulating the sonic experience of moving through the room
- Perform **zero-shot scene modification like virtual speaker rotation and translation**

DATASET

Base Datasets



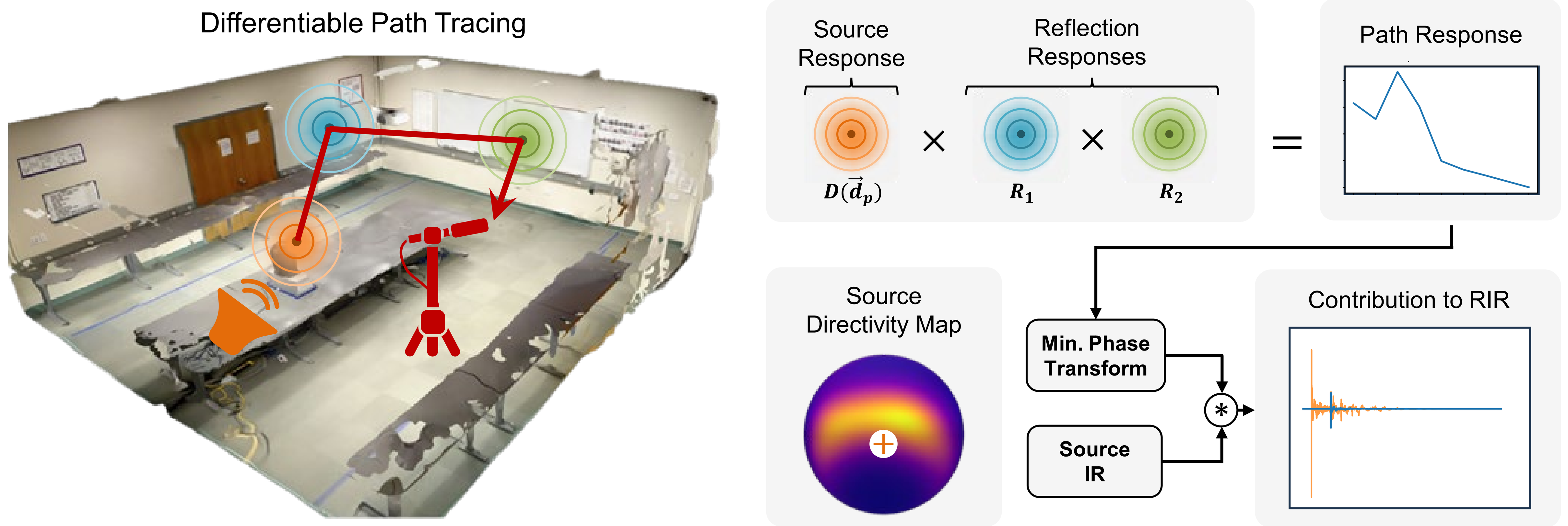
The dataset includes **monoaural** and **binaural RIRs** and **music** recordings from **over 3000** listener locations, in **four rooms** representing a wide range of room sizes, proportions, layouts, geometric complexities, materials, and reverberation effects.

	# Monoaural	# Binaural	Size (m)	N. Surfaces	RT60 (s)
Classroom	630	22	7.1 x 7.9 x 2.7	9	0.69
Dampened Room	768	64	4.9 x 5.2 x 2.7	6	0.14
Hallway	936	78	1.5 x 18.1 x 2.8	6	1.41
Complex Room	672	56	8.4 x 13.0 x 6.1	33	0.78

Additional Configurations

To evaluate zero-shot speaker rotation/translation, and panel insertion/relocation, we collect **10** additional **subdatasets** varying the speaker's location/orientation or the presence/number/location of whiteboard panels.

METHOD



We compute RIRs given a source-listener location. Each is a **sum** of **contributions from individual reflection paths.** After computing **reflection paths** between the source and listener, we characterize each by its outgoing **direction**, its **length**, and the **surfaces** it traverses. The **source** has a learned frequency response based on the path's outgoing direction, and each surface has a learned frequency response. These responses are **multiplied, inverted** to the time domain, **convolved** with a learned speaker response, and **time-shifted** to find the path's contribution to the RIR.

RESULTS

We compare **ground-truth RIRs** and **music recordings** from the **test set** with renderings from each method. Methods are given **12** training RIRs. "**Mag**" compares the **log-spectrograms** of ground-truth and rendered waveforms using the L1 distance at several time-frequency scales. **"Env"** is the **log-L1** distance between waveform energy envelopes.

	Classroom		Dampened Room		Hallway		Complex Room	
	Mag	Env	Mag	Env	Mag	Env	Mag	Env
NN	5.99	1.10	1.36	0.61	10.14	3.04	5.52	0.99
Linear	6.44	1.52	1.55	0.65	11.63	4.49	6.03	1.43
DeepIR	9.23	2.81	3.09	3.41	15.71	10.34	8.08	2.80
NAF	6.36	1.38	2.00	0.73	12.26	3.82	6.10	1.31
INRAS	9.99	4.52	4.20	2.48	14.52	9.19	9.02	2.58
DIFFRIR (Ours)	**5.22**	**0.94**	**1.21**	**0.56**	**9.13**	**2.95**	**4.86**	**0.92**

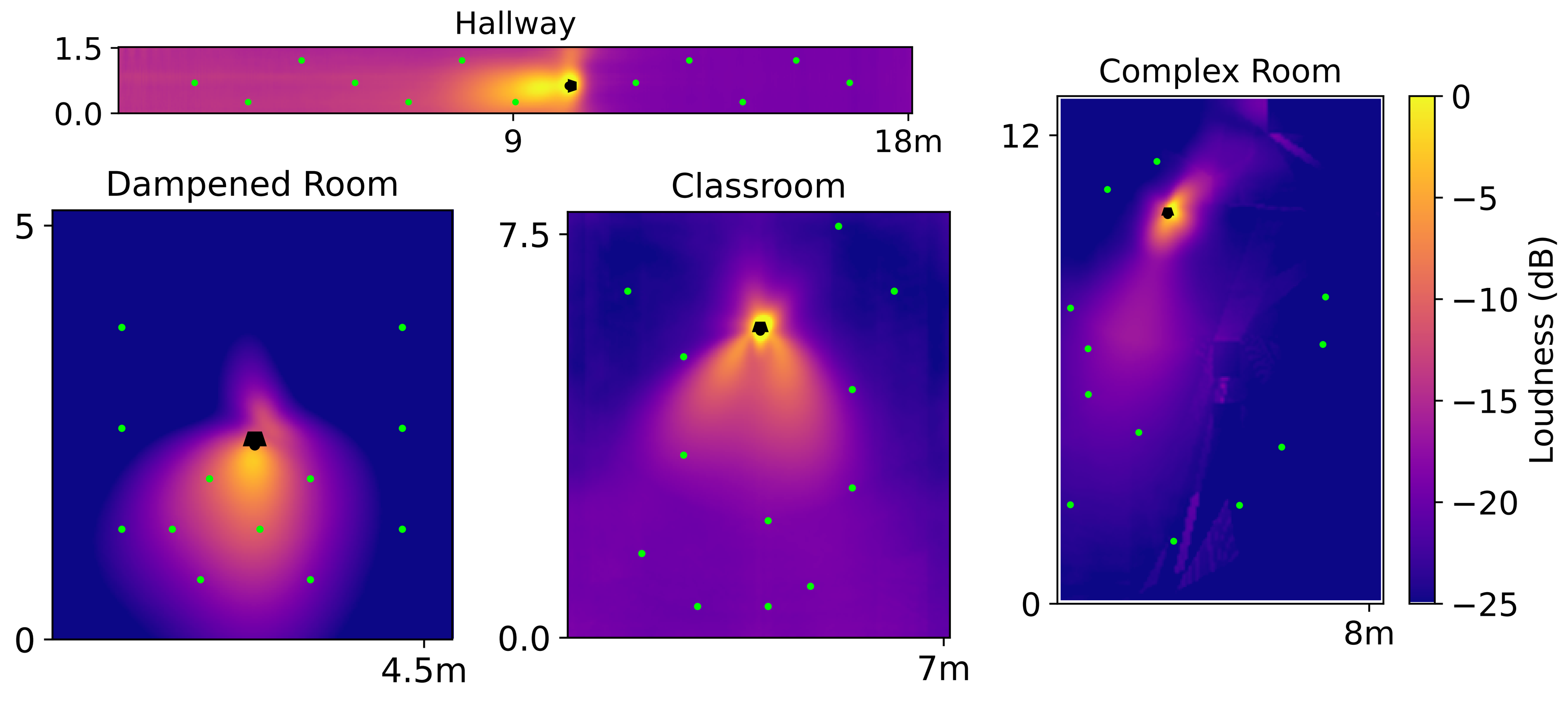
Table 1: Results comparing ground-truth **RIRs** with rendered **RIRs** from each baseline.

	Classroom		Dampened Room		Hallway		Complex Room	
	Mag	Env	Mag	Env	Mag	Env	Mag	Env
NN	2.95	1.42	1.99	1.36	2.62	1.32	2.39	1.42
Linear	3.34	1.82	2.43	1.66	3.11	1.75	2.74	1.74
DeepIR	3.15	1.65	3.39	2.22	2.97	1.47	2.62	1.65
NAF	3.32	1.75	3.38	1.54	3.13	1.46	2.87	1.71
INRAS	4.45	1.75	6.22	5.35	3.70	1.58	3.61	1.66
DIFFRIR (Ours)	**2.71**	**1.36**	**1.59**	**1.19**	**2.59**	**1.25**	**2.25**	**1.41**

Table 2: Results comparing ground-truth **music** with rendered **music** from each baseline.

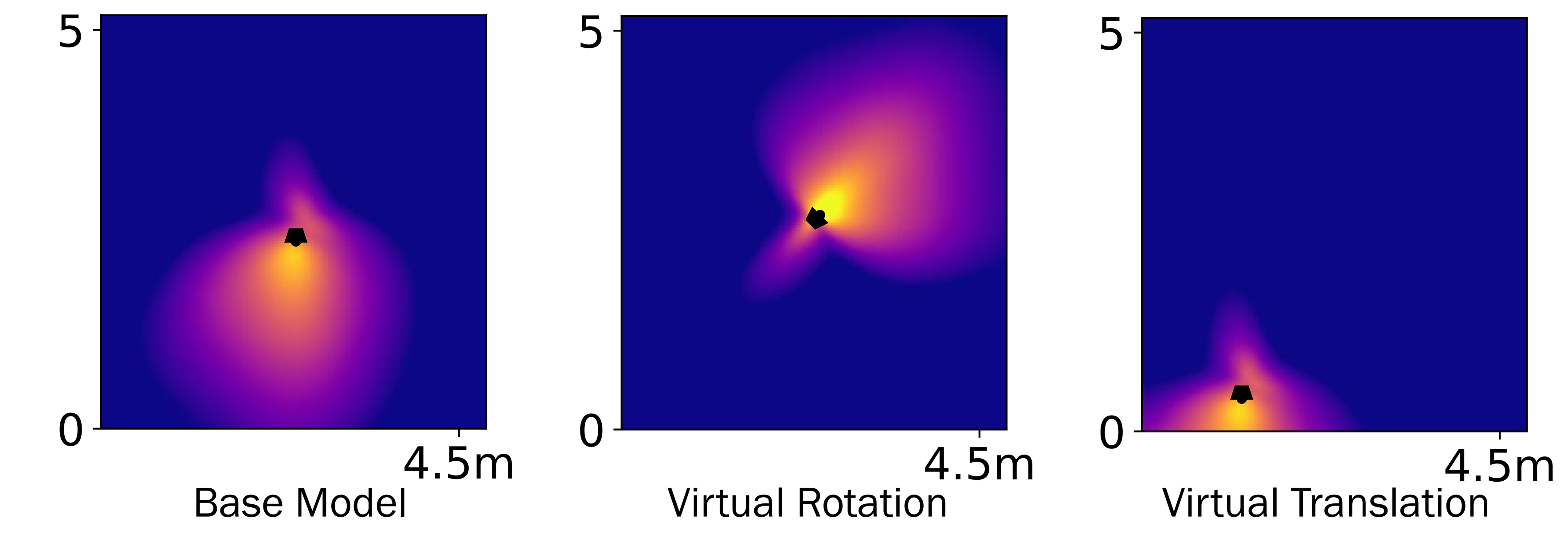
VISUALIZATIONS

RIR Heatmaps



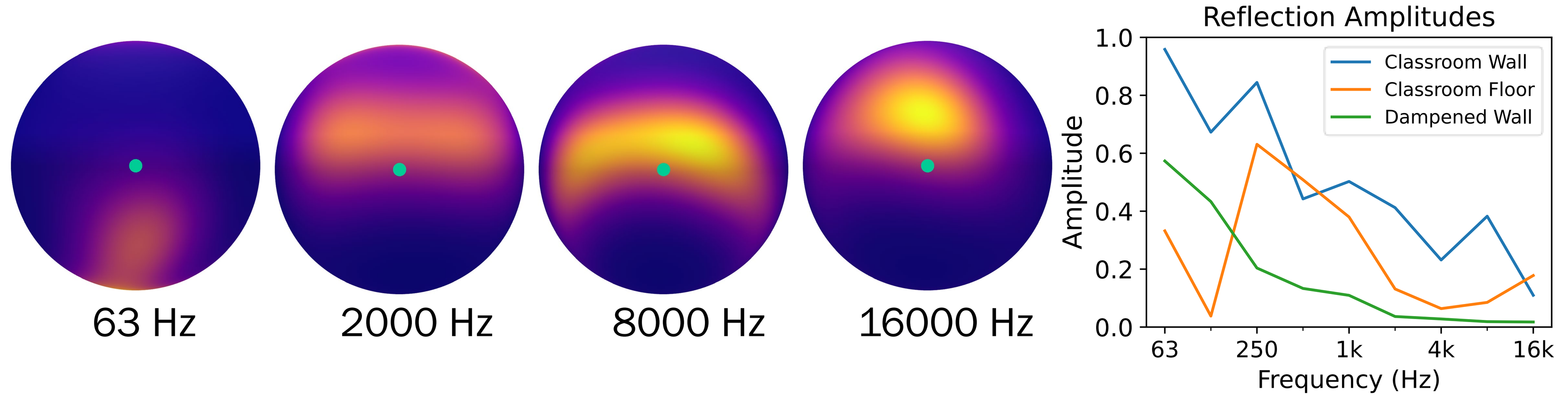
Visualization of RIR loudness maps generated from DIFFRIR trained in each of the four base subdatasets. 12 points were used to train DIFFRIR in each room, shown in green.

Zero-Shot Speaker Rotation and Translation



DIFFRIR fits interpretable parameters to the speaker, so we can train it on a static room configuration (Dampened Base), then simulate virtual speaker transformations.

Interpretability



Left: Speaker directivity maps we fit to 12 points from the Classroom subdataset.
Right: Reflection amplitude responses learned by our model for various surfaces.