NeuroAI

Be Everywhere - Hear Everything (BEE): Audio Scene Reconstruction by Sparse Audio-Visual Samples

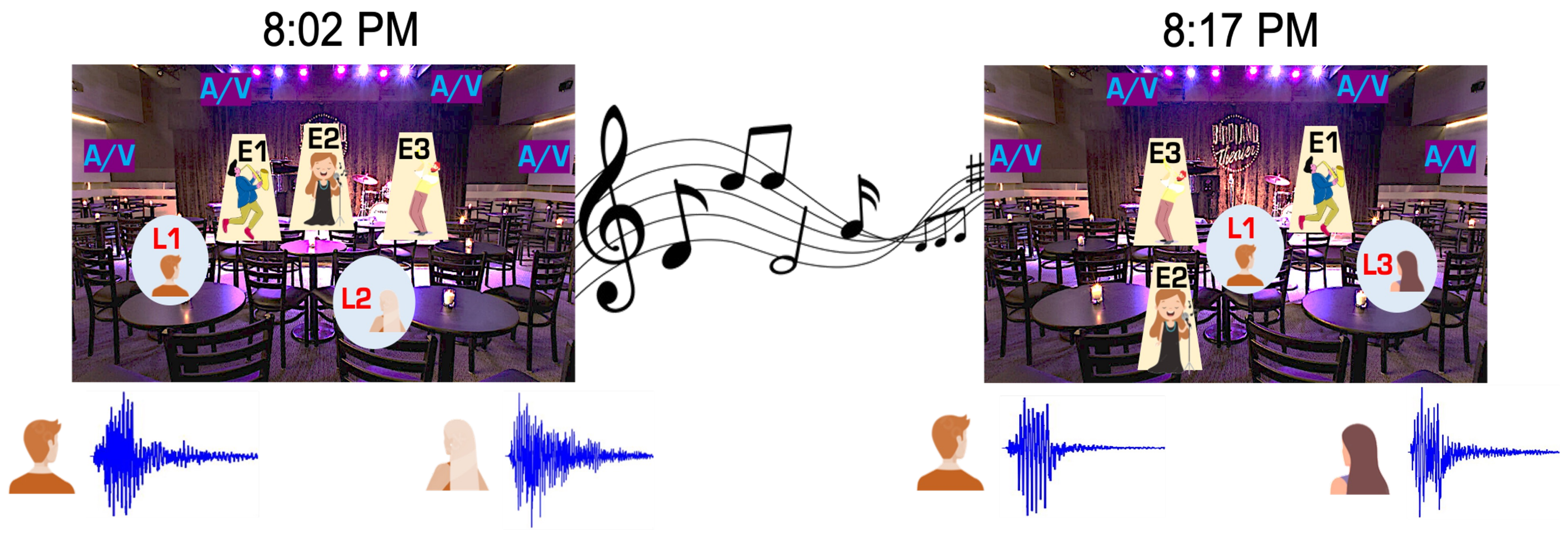
ICCV23 PARIS

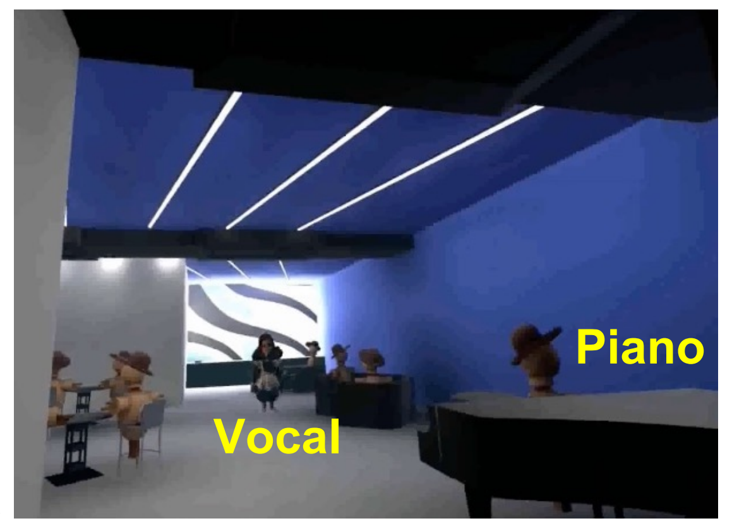
Mingfei Chen[1], Kun Su[1], Eli Shlizerman[1,2]

[1] Electrical & Computer Engineering, [2] Applied Mathematics, University of Washington, Seattle, USA

Be Everywhere - Hear Everything (BEE)

Audio reconstruction with **dynamic emitters** at **arbitrary listener locations**, leveraging inputs from **sparse A/V receivers**.





View of Listener 1

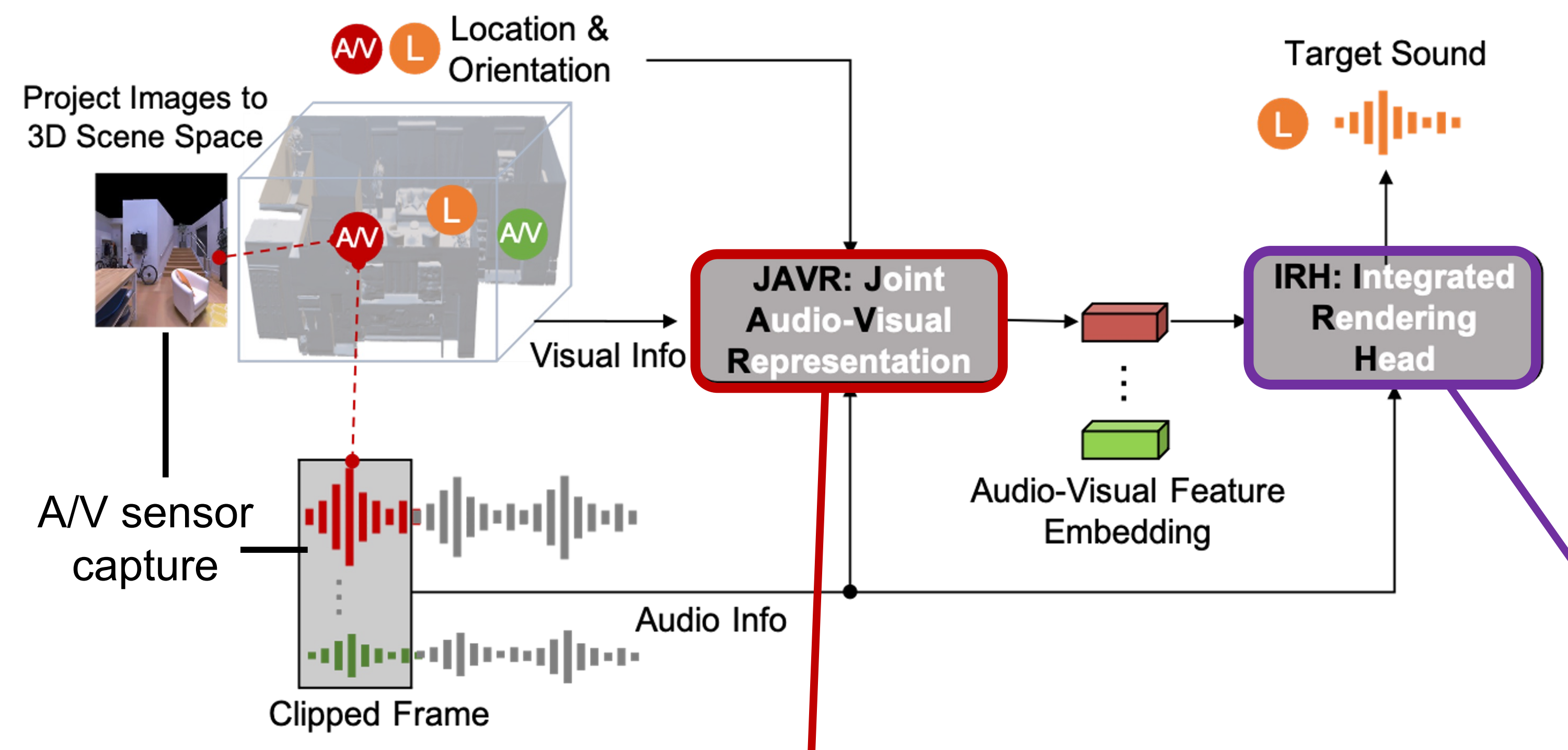
Input:
N A/V sensor captures (e.g. N=4)

Output:
Spatial audio waveform heard at Listener 1

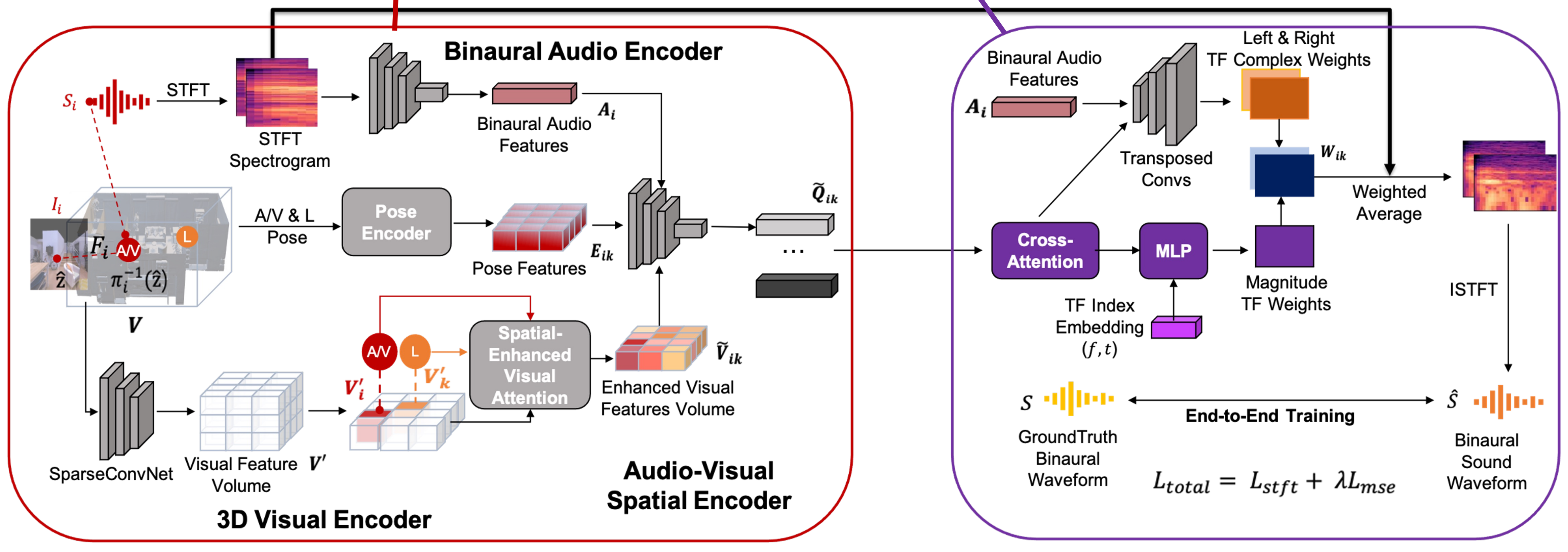
✅ No requirement of given/set emitters' locations

✅ No requirement of specific emitters' waveform

BEE Components


Location & Orientation
Project Images to 3D Scene Space
A/V
L
Visual Info
JAVR: Joint Audio-Visual Representation
Audio-Visual Feature Embedding
IRH: Integrated Rendering Head
Target Sound
A/V sensor capture
Audio Info
Clipped Frame

- **Joint Audio-Visual Representation** projects all images from reference A/V receivers to 3D scene space and integrates visual feature volume with audio clips.
- **Integrated Rendering Head** renders target sound at L by integrating audio-visual features of all receiver-listener pairs from JAVR and with learned time-frequency transformations.



(a) JAVR: Joint Audio-Visual Representation

(b) IRH: Integrated Rendering Head

Evaluations & Results

Method	Visual	Transform	Seen Scenes STFT ↓	Seen Scenes DPAM ↓	Seen Scenes ENV ↓	Unseen Scenes STFT ↓	Unseen Scenes DPAM ↓	Unseen Scenes ENV ↓
Replica: 12 seen scenes, 6 unseen scenes								
Nearest	✗	✓	1.614	0.992	0.257	1.686	0.993	0.277
Mean	✗	✓	1.600	1.039	0.265	1.618	1.036	0.275
Interpolation	✗	✓	1.575	1.039	0.256	1.614	1.033	0.267
AViTAR [6]	✓	✗	0.181	0.334	0.163	0.199	0.327	0.184
Few-shotRIR [17]	✓	✗	0.233	0.449	0.227	0.245	0.436	0.239
Mono2Binaural [10]	✓	✓	0.194	0.376	0.156	0.236	0.364	0.177
APNet [30]	✓	✓	0.164	0.263	0.154	0.185	0.253	0.176
BEE (Ours)	✓	✓	**0.151**	**0.215**	**0.133**	**0.177**	**0.221**	**0.160**
Matterport3D: 54 seen scenes, 25 unseen scenes								
Nearest	✗	✓	4.851	1.047	0.837	5.029	1.064	0.874
Mean	✗	✓	3.174	1.068	0.611	3.456	1.078	0.650
Interpolation	✗	✓	3.475	1.066	0.658	3.521	1.081	0.669
AViTAR [6]	✓	✗	0.516	0.610	0.595	0.509	0.625	0.548
Few-shotRIR [17]	✓	✗	0.597	0.476	0.731	0.591	0.500	0.694
Mono2Binaural [10]	✓	✓	0.533	0.440	0.545	0.582	0.492	0.529
APNet [30]	✓	✓	0.500	0.352	0.537	0.515	0.393	0.528
BEE (Ours)	✓	✓	**0.425**	**0.274**	**0.455**	**0.438**	**0.348**	**0.458**

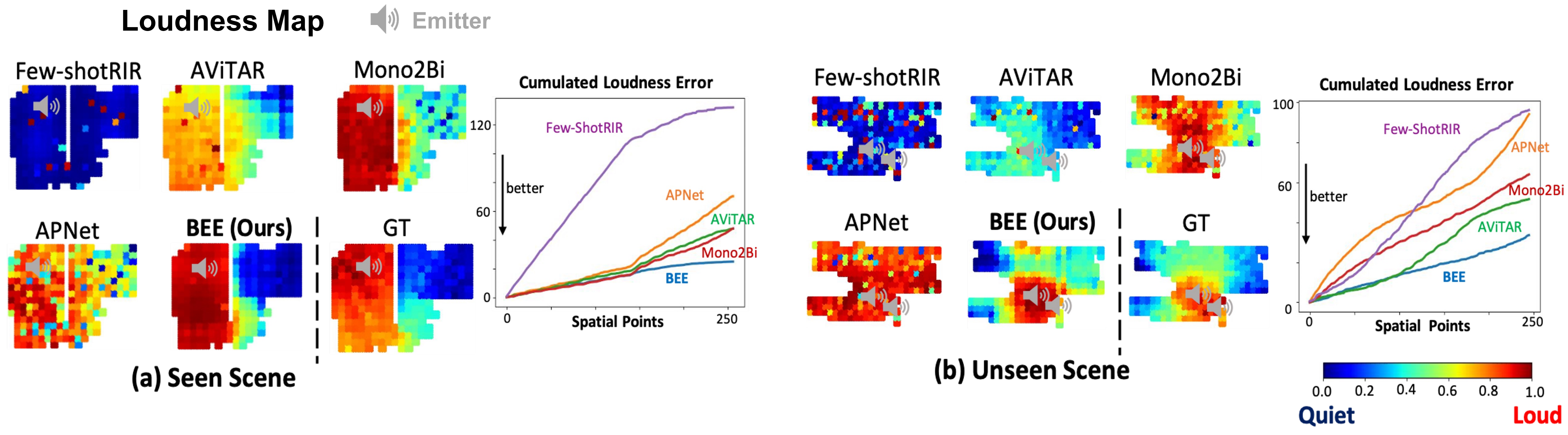
Components	3D Vis Enc	JAVR	IRH	Total
Speed (ms/sample)	16.00	18.40	11.94	30.34

Methods	Mono2Binaural [10]	APNet [30]	BEE (Ours)
Votes	29.6%	26.4%	**44%**

✅ SOTA Accuracy & Quality of Spatial Sound Reconstruction

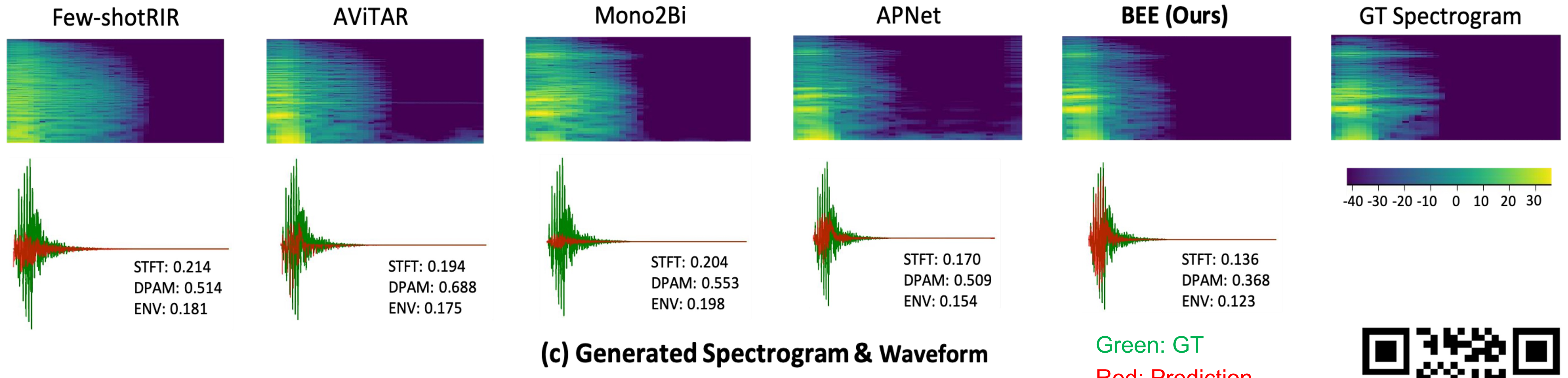
✅ Generalization Ability for Unseen Scenes

✅ Real-time Inference Speed



(a) Seen Scene

(b) Unseen Scene



(c) Generated Spectrogram & Waveform

Visit our demo video to see and listen to examples!