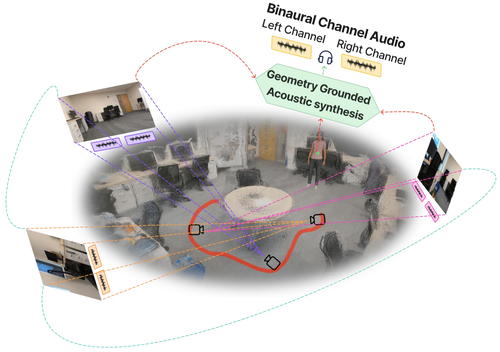
Visual Geometry Grounded Novel-View Acoustic Synthesis

Jay Polra[1]; Dhwanil Chauhan[1]; Wenjun Huang[2] ; Kyle Toth[2] ; Xianhui Wang[4] ; Yang Ni[1]
[1]Purdue University Northwest, [2]University of California, Irvine ,
[3]Center for Innovation Through Visualization and Simulation (CIVS) , [4]San Diego State University

CVPR JUNE 3-7, 2026 DENVER COLORADO

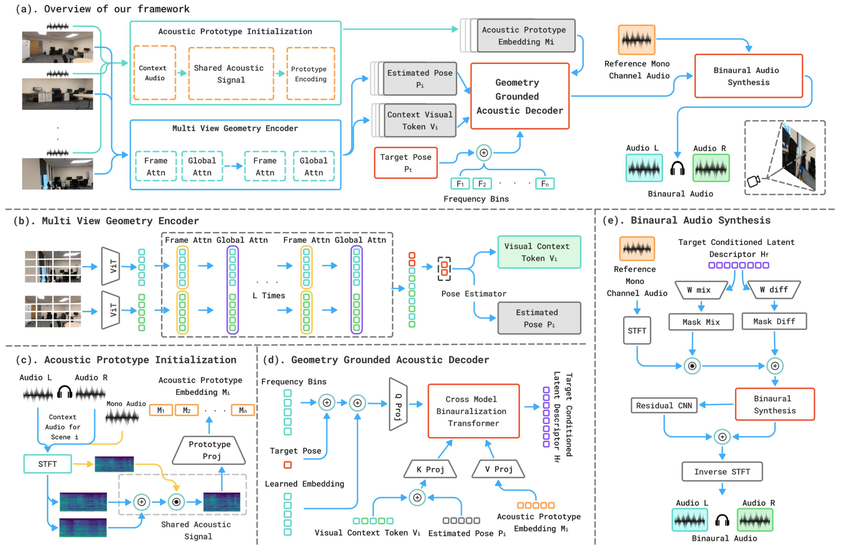
Geometry Grounded Acoustic Synthesis Introduction



- **Novel-view acoustic synthesis (NVAS)** generates spatial audio for unseen listener viewpoints, enabling immersive AR/VR and 3D scene experiences.
- **Existing methods rely on costly 3D reconstruction, target-view rendering, or dense point maps**, while our approach uses feed-forward visual geometry to synthesize binaural audio directly from reference video and audio.

NVAS Pipeline

- **Multi-view geometry encoder** extracts visual context and estimated pose features from sparse reference video frames.
- **Acoustic prototype initialization** encodes reference mono and binaural audio into scene-specific acoustic transfer features.
- **Geometry-Grounded Acoustic Decoder (GGAD)** attends to visual, geometric, and acoustic features to predict target-view binaural transfer fields.
- **Spectral binaural synthesis** applies the predicted transfer fields to the mono audio and reconstructs left-right binaural audio for the queried viewpoint.



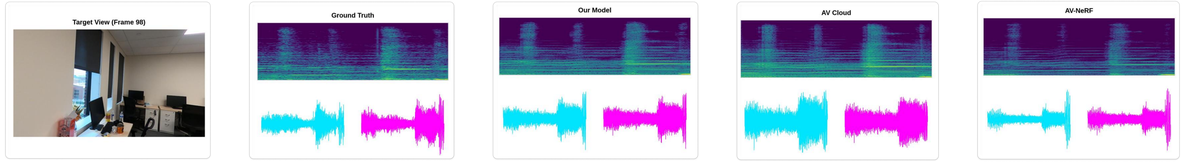
Evaluation and results

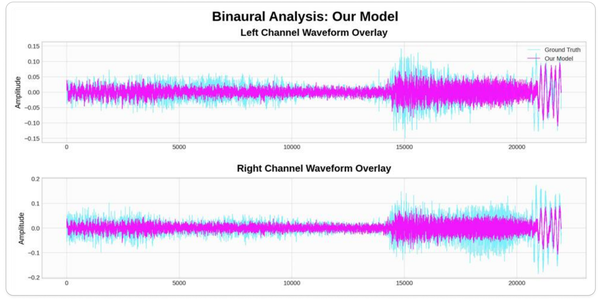
RWAVS								
Methods	**# Params**	**FPS**	**Image**	**Point**	**MAG ↓**	**ENV ↓**	**LRE ↓**	**DPAM ↓**
Mono-Mono	–	–	✗	✗	1.460	0.445	1.328	0.756
Mono-Energy	–	–	✗	✗	0.532	0.156	1.328	0.510
Stereo-Energy	–	–	✗	✗	0.560	0.160	–	0.535
DSP [8]	163M	–	✗	✗	1.016	0.274	3.468	0.588
VAM [5]	46.7M	174	✓	✗	0.390	0.156	0.996	0.459
ViGAS [6]	13.1M	90	✓	✗	0.370	0.147	1.089	0.357
AVNeRF [15]	12.0M	314	✓	✗	0.370	0.145	1.013	0.381
NACF [16]	0.44M	108	✗	✓	0.459	0.176	1.364	0.506
INRAS [27]	0.31M	475	✗	✓	0.455	0.179	1.503	0.485
NAF [18]	0.22M	261	✗	✓	0.448	0.522	1.204	0.353
AV-Cloud [7]	3.91M	219	✗	✓	0.3652	0.1509	1.0297	0.2776
Our Method	3.24M	189	✗	✗	**0.3485**	**0.1424**	**0.9589**	**0.2705**

Split	Group/Method	MAG ↓	ENV ↓	LRE ↓	DPAM ↓
Random	Office / AV-Cloud	0.3157	0.1374	1.5822	0.2768
	Office / Ours	**0.2886**	**0.1254**	**1.2628**	**0.2563**
	Outdoor / AV-Cloud	0.2563	0.1121	0.8557	0.3325
	Outdoor / Ours	**0.2405**	**0.1071**	**0.7608**	**0.3066**
	Avg. / AV-Cloud	0.3640	0.1502	1.1054	0.2856
	Avg. / Ours	**0.3485**	**0.1424**	**0.9589**	**0.2705**
50/50	Office / AV-Cloud	0.3749	0.1485	3.0037	0.4096
	Office / Ours	**0.3061**	**0.1312**	**1.8340**	**0.2975**
	Outdoor / AV-Cloud	0.3127	0.1310	1.0781	0.3609
	Outdoor / Ours	**0.2897**	**0.1228**	**0.9333**	**0.3281**
	Avg. / AV-Cloud	0.4089	0.1620	1.7099	0.3658
	Avg. / Ours	**0.3799**	**0.1555**	**1.2425**	**0.3165**

- **Results comparison on RWAVS and Replay-NVAS:** Average metrics are reported using MAG, ENV, LRE, and DPAM, where **lower** values indicate **better** binaural audio synthesis.
- **Our method improves RWAVS performance and shows stronger viewpoint generalization**, outperforming AV-Cloud under both the original random split and the stricter 50/50 train/test split.

Visualizations





- **Qualitative comparison of target-view binaural synthesis,** Spectrogram and waveform results show that our model produces audio closer to the ground truth than AV-NeRF and AV-Cloud.
- **Our method better preserves stereo structure and viewpoint-conditioned audio patterns**, demonstrating more accurate binaural rendering for the queried listener viewpoint.

This work is supported by the Steel Manufacturing Simulation and Visualization Consortium (SMSVC).