Visual Geometry Grounded Novel-View Acoustic Synthesis

Jay Polra[1] Dhwanil Chauhan[1] Wenjun Huang[2] Kyle Toth[1,3] 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

yangni@purdue.edu

Abstract

We present the first unified framework for novel-view acoustic synthesis that entirely bypasses explicit 3D visual rendering and costly photogrammetry by directly grounding spatial audio generation in feed-forward visual geometry. We show its capability to synthesize accurate and immersive spatial audio in 3D spaces without requiring viewpoint images, dense point maps, or any ground-truth poses for input video. Our motivation stems from the observation that existing methods suffer from limited geometry cues, requirements on simulated acoustic environments, inefficient multimodal visual-audio learning, and reliance on costly and unstable photogrammetry pipelines. Our proposed approach overcomes these challenges collectively by blending the learned visual representation and geometry from feed-forward scene encoding and jointly conditioning on visual and audio features in geometry-aware binauralization. In particular, we design the Geometry Grounded Acoustic Decoder to dynamically attend to cross-modal features, which embed local and global geometries in audio and visual modalities. Extensive experiments show that our framework outperforms prior work across various benchmarks in high-quality, viewpoint-accurate spatial audio synthesis, without requiring time-consuming explicit rendering of novel-view images or dense point maps.

1. Introduction

For immersive media, augmented or virtual reality (AR/VR) applications, and interactive embodied systems, maintaining spatially consistent sound as the viewpoint shifts is essential for realism and immersion. Novel-view acoustic synthesis (NVAS) addresses this need by rendering spatial and binaural audio at unseen listener positions from mono source audio [6, 18], resembling the novel-view synthesis (NVS) task in visual 3D reconstruction. Both require a deep understanding and precise capture of the global 3D geometric relations.

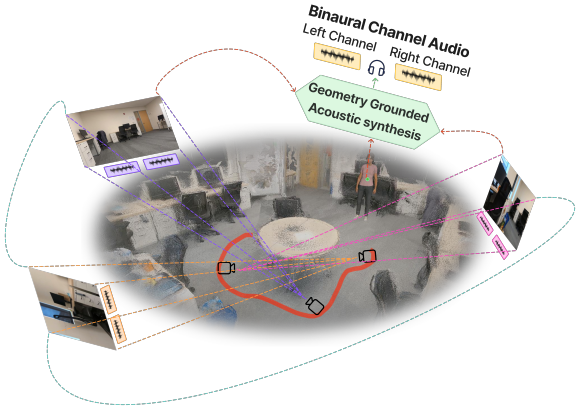

Figure 1. Given a set of reference frames and recorded audio from a video clip, our method renders binaural audio at the queried viewpoint via geometry-grounded NVAS.

Despite the rapid progress in visual 3D modeling and NVS, which has made realistic scene navigation increasingly practical [20], acoustically correct rendering remains challenging. Spatial sound must vary coherently with scene geometry, listener motion, and environmental context. This challenge is further compounded by the inherently more complex spatio-temporal dynamics of sound compared to visual signals, as well as the limited geometric information contained in audio recordings [17].

Early works on NVAS mainly focused on modeling the acoustic field directly from audio signals, which limits their geometric understanding to the audio modality [18, 23, 26, 27]. Many approaches also rely on impulse responses generated in simulated environments, restricting their applicability in real-world scenarios [4]. These limitations motivate the integration of visual information to provide additional geometric cues. With the success of multimodal audio-visual learning, some visually informed NVAS approaches have emerged. These methods leverage geometric cues and acoustic signals derived from images to guide spa-

tial audio synthesis [5, 6, 21, 32]. However, a crucial challenge lies in acquiring the corresponding visual observations for novel viewpoints. This observation has motivated recent efforts to jointly model spatial audio synthesis with NVS and 3D reconstruction [1, 2, 15]. By utilizing neural rendering techniques developed for NVS, these approaches learns geometry-aware acoustic representations from reconstructed scenes. However, such methods typically require training multiple per-scene neural networks or large sets of 3D Gaussians priors to acoustic synthesis, which introduces significant runtime and memory overheads. Chen *et al*. [7] improve efficiency by replacing neural rendering with anchor-based conditioning initialized by Structure-from-Motion (SfM). Nevertheless, SfM is still computationally expensive as the number of images increases and becomes unreliable when images are sparse or limitedly overlapped.

Recent progress in 3D reconstruction has shifted from iterative optimization towards feed-forward inference with foundation-scale models [13, 28–30]. Instead of relying on costly scene-by-scene reconstruction and post-processing, these models directly estimate camera poses and predict globally consistent scene structures from image collections, even under sparse settings. In addition, they simultaneously support depth estimation, dense 3D point maps, point tracks, and semantic segmentation, etc [14, 25, 29, 33]. Such feed-forward reconstruction pipelines offer an efficient way to obtain rich geometric and semantically plausible representations of scenes at both global and local scales.

The widespread yet inefficient reliance on SfM in previous NVAS methods motivates us to rethink how geometry should be incorporated into spatial audio synthesis. Feed-forward reconstruction models provide an opportunity to extract richer geometric representations more efficiently. However, improved geometric reconstruction does not guarantee the yield of high-fidelity or geometrically calibrated spatial audio. As prior works have observed [7, 15], even accurate geometric relations between the listener and sound source do not directly translate into high-fidelity spatial audio. It is necessary to jointly learn how to extract and represent 3D geometry together with the acoustic properties of the environment in order to accurately model sound propagation.

In this paper, we propose **the first unified framework for NVAS with feed-forward visual geometry grounding**. As shown in Figure1, our framework accepts short video clips as input, capturing a scene from multiple viewpoints in a natural and dynamic way. And the output is a viewpoint-consistent binaural audio at any specified location and direction, as if the user were a listener at that exact position.

We begin with an implicit modeling of the whole scene. Our framework samples a sparse set of reference video frames and applies alternating transformer attention to image patches. We extract learned representations that not only contain per-frame visual content but also global spatial understanding across all image frames. To aid the viewpoint-consistent rendering of spatial audio, we introduce the Geometry-Grounded Acoustic Decoder (GGAD). Its transformer-based architecture features a cross-modal conditioning of local visual semantics and global geometry. GGAD queries with frequency-aware tokens initialized from a novel view pose, then computes cross-attention between the listener's pose and the reference views. The latter includes multimodal features covering per-view visual semantics, spatial acoustic prototypes, and global geometry. Finally, we apply a learnable binaural audio synthesis based on the decoded frequency-aware acoustic masks. The main contributions of our paper are listed as follows:

- We propose a novel framework that unifies novel-view acoustic synthesis with feed-forward 3D reconstruction paradigms. We leverage the learned representation from visual geometry grounding to address the gap of limited spatial geometry in audio references, facilitating high-quality binaural audio rendering from arbitrary viewpoints.
- Our framework provides a more viable solution in the real world, as it is free of dependencies on the rendered image for the novel view, ground-truth poses for input video frames, or costly per-scene 3D visual modeling. Most importantly, as a unified pipeline, we also do not require explicit reconstruction of 3D dense point maps, allowing dense reconstruction, visual rendering, and NVAS to run in parallel.
- We introduce a cross-modal acoustic decoder that learns the acoustic primitives necessary for viewpoint-accurate audio binauralization by conditioning on a sparse set of references. The transformer learn to jointly attend to per-view local semantics, global spatial geometry, and prototypical acoustic properties.
- Extensive experiments on NVAS benchmarks demonstrate that our framework outperforms prior work in the quality of novel-view audio binauralization and synthesis, as well as in the robustness and efficiency for the support of sparse reference video.

2. Methods

Our goal is to render binaural audio at a target viewpoint by combining visual and spatial geometric understanding from sampled viewpoints with acoustically grounded transfer cues from source audio as reference. As shown in Figure 2, the framework proceeds in three stages. We first construct a multimodal context in which each reference viewpoint is represented by visual descriptors, pose-aware geometric cues, and an acoustic prototype embedding. We then query this context with frequency-conditioned target-pose

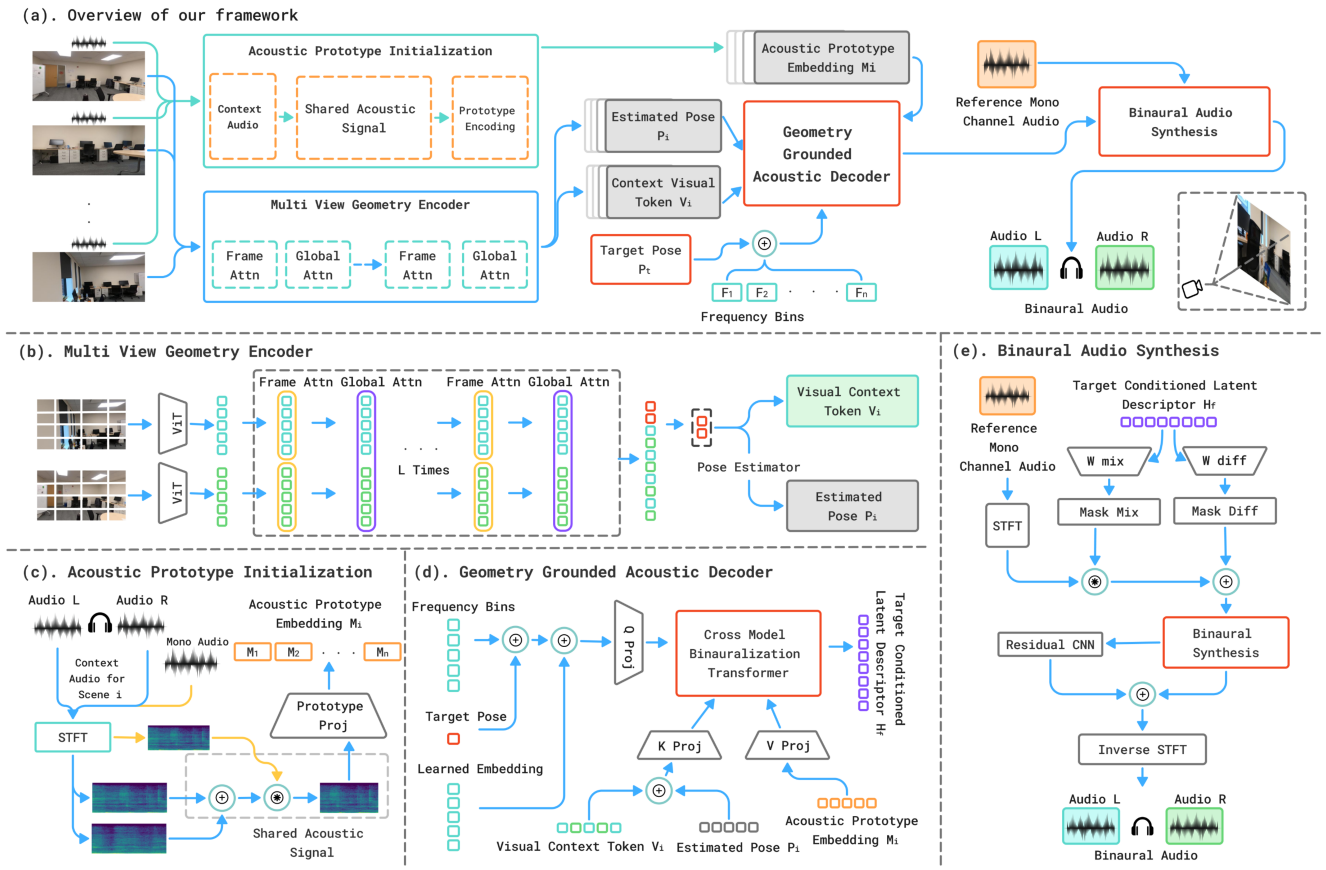

Figure 2. Overview of the proposed framework. We first construct a multimodal context from reference views by combining visual context tokens, estimated pose descriptors, and acoustically grounded prototype embeddings computed from aligned reference mono and binaural audio. The Geometry-Grounded Acoustic Decoder (GGAD) then performs listener-conditioned retrieval over this context using frequency-aware target queries to predict shared transfer and binaural contrast fields. Finally, these fields are applied to the reference mono spectrogram and refined through spectral binaural synthesis to reconstruct target-view binaural audio.⊛: represents element-wise multiplication ⊕: represents element-wise addition

tokens to infer listener-dependent transfer fields in the time-frequency domain. Finally, these transfer fields are applied to the reference mono spectrogram and refined through a residual synthesis module to reconstruct the target-view binaural waveform.

2.1. Multimodal Context Construction

Novel-view acoustic synthesis requires more than a target pose. The model must build a context that captures scene geometry across the reference views while also preserving the acoustic transfer behavior observed at those views. Visual context alone describes how the scene is arranged, but it does not specify what acoustic transfer content should be retrieved later. We therefore construct a multimodal context in which each reference view contributes visual-geometric descriptors together with an acoustically grounded prototype.

Given a set of reference images $\mathcal{C} = \{I_i\}_{i=1}^{N}$, a feed-forward multi-view encoder alternates between scene-level aggregation and frame-level refinement to produce a visual descriptor and a pose-aware geometric descriptor for each view,$(v_i, \rho_i) = \Phi(I_i; \mathcal{C}), i = 1, \ldots, N$. We collect these outputs as:

$$V = [v_1, \ldots, v_N] \qquad P = [\rho_1, \ldots, \rho_N],$$

where $v_i \in \mathbb{R}^{d_v}$ encodes the visual content of the i-th reference view and $\rho_i \in \mathbb{R}^{d_p}$ encodes its geometry-aware pose descriptor. Crucially, this feed-forward estimation directly infers camera poses and global geometry from $\mathcal{C}$, bypassing the need for ground-truth pose annotations or slow photogrammetry pipelines. The alternating global-frame attention also interpolates the missing geometric relationships even among sparse frames.

To complement this visual-geometric context, we con-

struct an acoustic prototype from the aligned reference audio at the same viewpoint. Let x_i^m denote the mono reference audio, and let x_i^L and x_i^R denote the corresponding binaural channels. We first transform them into the spectral domain, $S_i^m = |\mathrm{STFT}(x_i^m)|$, $S_i^L = \left|\mathrm{STFT}(x_i^L)\right|$, $S_i^R = \left|\mathrm{STFT}(x_i^R)\right|$, and define the shared spectral transfer at that reference view as $G_i = [S_i^L + S_i^R]/2S_i^m$. This transfer is then encoded as an acoustic prototype embedding,

$$M_i = \phi(G_i), \qquad M_{\mathrm{ctx}} = [M_1, \ldots, M_N]$$

Before decoding, the visual-geometric context and target pose are projected into a common representation space:

$$c_i = W_v v_i, \qquad e_i = W_p \rho_i, \qquad e_t = W_t p_t,$$

$$C = [c_1, \ldots, c_N] \in \mathbb{R}^{N\times d}, E_{\mathrm{ctx}} = [e_1, \ldots, e_N] \in \mathbb{R}^{N\times d}$$

Together, C, E_{ctx}, and M_{ctx} form the multimodal context passed to GGAD, C and E_{ctx} define the scene-aware addressing structure, while M_{ctx} provides acoustically grounded transfer prototypes aligned with the same reference views.

2.2. Geometry-Grounded Acoustic Decoding

Once the multimodal context has been constructed, the remaining challenge is to convert it into transfer fields that are specific to the queried listener pose. Scene context alone does not determine how spectral energy should change at a new viewpoint, and acoustic prototypes alone do not indicate which reference views are relevant to that query. We therefore use the Geometry-Grounded Acoustic Decoder (GGAD) to retrieve target-dependent acoustic transfer under visual-geometric guidance.

GGAD operates in the time-frequency domain, assigning one query to each STFT frequency bin. For each frequency bin $f \in \{1, \ldots, F\}$, we compute a log-frequency embedding $\gamma(f)$ and project it to the decoder space,

$$q_f = W_q\, \gamma(f), \qquad Q = [q_1, \ldots, q_F] \in \mathbb{R}^{F\times d}.$$

To make retrieval explicitly listener-dependent, each frequency query is combined with a learned frequency vector and the target-pose embedding,

$$p_f = u_f + e_t, \qquad P_q = [p_1, \ldots, p_F].$$

On the context side, the projected visual content and geometric terms define where GGAD should attend, while the acoustic prototype embeddings define what acoustic information can be retrieved: $K = C + E_{\mathrm{ctx}}, V = M_{\mathrm{ctx}}$. We then form the target-conditioned queries $\hat{Q} = Q + P_q$, and apply cross-attention,

$$H = \mathrm{Attn}(\hat{Q}, K, V) = \mathrm{softmax}\left(\frac{\hat{Q}K^\top}{\sqrt{d}}\right) V,$$

where $H = [h_1, \ldots, h_F] \in \mathbb{R}^{F\times d}$ and each h_f is a target-conditioned latent descriptor for one frequency bin. In this way, By setting the keys (K) to the fused visual-geometric context and values (V) to the acoustic prototypes, this cross-attention effectively realize geometry-aware binauralization. Our proposed GGAD dynamically retrieves acoustic features from the reference views, strictly grounded by the visual and geometric relevance to the target pose, i.e., the extracted per-frame context token from the multi-view geometry encoder and the aligned acoustic prototypes.

The decoded latent descriptors are mapped to two complementary acoustic fields,

$$\alpha_f = W_{\mathrm{mix}} h_f, \qquad \beta_f = 2\,\sigma(W_{\mathrm{diff}} h_f) - 1,$$

where α_f models shared spectral scaling and $\beta_f \in [-1, 1]$ models binaural asymmetry. To interface with synthesis, these per-frequency coefficients are expanded across the time axis of the source spectrogram,

$$\mathcal{M}_{\mathrm{mix}}(f,t) = \alpha_f, \qquad \mathcal{M}_{\mathrm{diff}}(f,t) = \beta_f, \quad t = 1, \ldots, T.$$

The resulting fields $\mathcal{M}_{\mathrm{mix}}$ and $\mathcal{M}_{\mathrm{diff}}$ are then passed to the spectral binaural synthesis stage.

2.3. Spectral Binaural Synthesis

Once the target-conditioned transfer fields have been inferred, the remaining task is to render binaural audio at the queried viewpoint. Rather than predicting waveform samples directly, we transform the mono reference signal in the time-frequency domain, so that the learned transfer fields control shared spectral scaling and binaural asymmetry while the reference audio preserves the underlying temporal structure.

Given mono reference audio s_{mono}, we first compute

$$S_{\mathrm{ref}} = \mathrm{STFT}(s_{\mathrm{mono}}), \qquad A = |S_{\mathrm{ref}}|, \qquad \phi = \angle S_{\mathrm{ref}},$$

where $A \in \mathbb{R}^{F\times T}$ and $\phi \in \mathbb{R}^{F\times T}$ denote the reference magnitude and phase, respectively. Using the inferred transfer fields, we construct a binaural decomposition,

$$A_{\mathrm{mono}} = A \odot \mathcal{M}_{\mathrm{mix}}, \qquad A_{\mathrm{diff}} = A_{\mathrm{mono}} \odot \mathcal{M}_{\mathrm{diff}},$$

$$A_L^{\mathrm{base}} = A_{\mathrm{mono}} - A_{\mathrm{diff}}, \qquad A_R^{\mathrm{base}} = A_{\mathrm{mono}} + A_{\mathrm{diff}}.$$

Here, $\mathcal{M}_{\mathrm{mix}}$ controls the shared spectral energy after viewpoint transfer, while $\mathcal{M}_{\mathrm{diff}}$ introduces the left-right imbalance required for binaural perception.

We refine these estimates with a shared residual CNN. The per-ear inputs are:

$$x_L = [A, \mathcal{M}_{\mathrm{mix}}, -\mathcal{M}_{\mathrm{diff}}, A_L^{\mathrm{base}}],$$

$$x_R = [A, \mathcal{M}_{\mathrm{mix}}, \mathcal{M}_{\mathrm{diff}}, A_R^{\mathrm{base}}],$$

and the network predicts residual corrections $\Delta_L = \mathcal{R}(x_L), \Delta_R = \mathcal{R}(x_R)$. The final magnitudes are then

$$\hat{A}_L = \mathrm{ReLU}(A_L^{\mathrm{base}} + \Delta_L), \qquad \hat{A}_R = \mathrm{ReLU}(A_R^{\mathrm{base}} + \Delta_R).$$

Finally, we reconstruct the complex spectra using the reference phase, $\hat{S}_L = \hat{A}_L \cdot e^{j\phi}, \hat{S}_R = \hat{A}_R \cdot e^{j\phi}$, and apply inverse STFT to obtain the rendered binaural waveform,

$$\hat{s} = \mathrm{iSTFT}(\hat{S}_L, \hat{S}_R).$$

This gives a stable synthesis path in which GGAD predicts viewpoint-dependent transfer, and the spectral rendering stage converts that transfer into target-view binaural audio.

Because this synthesis directly modulates spectral scaling and binaural asymmetry of the reference audio, allowing our design to circumvent the need to render dense 3D scenes or novel-view images.

2.4. Training Objective

We supervise the final synthesized binaural waveform directly against the ground-truth binaural target. Let $\hat{s} = [\hat{s}_L, \hat{s}_R]$ denote the predicted binaural waveform and $s = [s_L, s_R]$ denote the ground truth. Rather than imposing a waveform-phase objective, we optimize a log-magnitude STFT loss on the two output channels,

$$\mathcal{L}_{\mathrm{train}} = 20\,\mathcal{L}_{\mathrm{stft}}(\hat{s}, s),$$

where

$$\mathcal{L}_{\mathrm{stft}}(\hat{s}, s) = \frac{1}{2} \sum_{c \in \{L,R\}} \Big\| \log\big(1 + |\mathrm{STFT}(\hat{s}_c)|\big) - \log\big(1 + |\mathrm{STFT}(s_c)|\big) \Big\|_2^2$$

In implementation, this loss is computed with a Hamming window using FFT size 512, hop size 128, and window length 512. This objective encourages the synthesized binaural output to match the target spectral structure while remaining aligned with the viewpoint-conditioned transfer predicted by GGAD.

3. Experiments

3.1. Experimental Setup

We evaluate on RWAVS and ReplayNVAS, following prior audio-visual spatial audio rendering work [7, 15]. RWAVS is a real-world scene-level benchmark with synchronized visual observations, listener poses, mono reference audio, and binaural targets, and we use it at 22.05 kHz. ReplayNVAS is a complementary indoor benchmark with multiview visual observations and spatial audio targets, and we use it at 16 kHz. For each scene, we sample 256 reference frames and process them with VGGT [29] to obtain per-view visual context tokens and estimated camera descriptors. In parallel, we construct aligned acoustic prototype embeddings from the corresponding reference mono and binaural audio. Together, these visual-geometric descriptors and acoustic prototypes define the multimodal context used for target-view spatial audio rendering.

3.2. Comparison with Baselines

We report four standard metrics, MAG, ENV, LRE, and DPAM. MAG measures magnitude-spectrogram error, ENV measures envelope distance, LRE measures the absolute error in left-right energy ratio, and DPAM measures perceptual distance between predicted and ground-truth audio. Lower is better for all metrics.

We compare our method with heuristic, image-conditioned, and geometry-conditioned NVAS baselines. The heuristic baselines, Mono-Mono, Mono-Energy, and Stereo-Energy, provide simple energy-based references without scene understanding. DSP uses hand-crafted spatial audio processing. VAM, ViGAS, and AVNeRF are image-conditioned methods, while NACF, INRAS, NAF, and AV-Cloud rely on geometry- or point-cloud-based scene representations. Table 1 reports the main comparison results on RWAVS and Replay-NVAS.

Our results in Table 1 show that the proposed framework is highly competitive across both datasets while not requiring target-view images and explicit dense 3D rendering at inference time. Overall, our method achieves stronger or more competitive performance than prior approaches across the reported benchmarks while avoiding rendered target-view images and explicit point-cloud reconstruction at inference time. Relative to AV-Cloud, it improves MAG from 0.3652 to 0.3485, ENV from 0.1509 to 0.1424, LRE from 1.0297 to 0.9589, and DPAM from 0.2776 to 0.2705 on RWAVS, while reducing the parameter count from 3.91M to 3.24M, on Replay-NVAS, it attains the best ENV and DPAM scores and runs faster at 398 FPS versus 319 FPS. We attribute these gains to the proposed multimodal context construction and geometry-grounded acoustic decoding, which retrieve listener-conditioned acoustic information from aligned visual, geometric, and acoustic cues without depending on rendered target-view images or explicit point-cloud reconstruction.

3.3. Results Under Stricter Train/Test Split

We first evaluate under the original random split, where temporally adjacent views from the same scene may appear across training and test sets. Since this setting reduces viewpoint novelty at test time, we additionally consider a stricter 50/50 split, where each scene is partitioned into disjoint train and test subsets to provide a stronger test of viewpoint generalization.

Table 1. Main results comparison. Best values are **bolded** and second-best are underlined. Lower values indicate better performance for MAG, ENV, LRE, and DPAM. Grey shading highlights our method's results.

Dataset	Methods	# Params	FPS	Image	Point-based	MAG ↓	ENV ↓	LRE ↓	DPAM ↓
RWAVS	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**
Replay-NVAS	Mono-Mono	–	–	✗	✗	0.313	0.127	0.934	0.521
	Mono-Energy	–	–	✗	✗	0.191	0.050	0.934	0.496
	Stereo-Energy	–	–	✗	✗	0.196	0.054	–	0.473
	DSP [8]	163M	–	✗	✗	0.228	0.066	6.186	0.482
	VAM [5]	46.5M	204	✓	✗	0.239	0.062	0.824	0.458
	ViGAS [6]	12.7M	105	✓	✗	0.193	0.054	0.698	1.177
	AVNeRF [15]	11.8M	368	✓	✗	0.214	0.055	0.773	0.290
	NACF [16]	0.54M	139	✗	✓	0.298	0.079	0.722	0.544
	INRAS [27]	0.32M	501	✗	✓	0.211	0.058	0.928	0.807
	NAF [18]	0.23M	309	✗	✓	0.208	0.059	0.820	0.565
	AV-Cloud [7]	2.47M	319	✗	✓	**0.1560**	0.0450	**0.6080**	0.2280
	Our Method	3.24M	398	✗	✗	0.1590	**0.0400**	0.8060	**0.2240**

Table 2. Comparison under the original random split and the revised 50/50 split. Lower is better for all metrics.

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**

Table 3. Ablation study on reducing the number of frames for the RWAVS dataset. Best values are **bolded**, second-best are underlined. Grey shading highlights our method.

Frames	MAG ↓	ENV ↓	LRE ↓	DPAM ↓
AV-Cloud (RWAVS)				
4	X	X	X	X
8	X	X	X	X
16	X	X	X	X
32	X	X	X	X
64	0.3579	0.1469	1.1120	0.2853
128	0.3585	0.1472	1.0958	0.2825
256	0.3640	0.1502	1.1054	0.2856
Our Method (RWAVS)				
4	0.3729	0.1490	1.1326	0.3347
8	0.3542	0.1449	1.0487	0.2786
16	0.3745	0.1503	1.0217	0.3246
32	0.3515	0.1439	1.0191	0.2821
64	0.3498	0.1436	1.0493	0.2737
128	0.3516	0.1442	1.0430	0.2802
256	**0.3487**	**0.1431**	**0.9890**	**0.2742**

Table 2 shows that our method achieves the best overall average under both protocols, with gains visible in both Office and Outdoor scenes. The improvement becomes larger under the stricter split, especially in LRE and DPAM, indicating stronger robustness when train and test viewpoints are more strongly separated. We attribute this to our multimodal context construction, which combines visual-geometric structure with aligned acoustic prototypes and provides a more stable basis for target-view acoustic retrieval than methods relying more heavily on local view similarity.

3.4. Robustness to Sparse Reference Frames

Table 3 studies how performance changes as the number of reference frames is reduced on RWAVS. This experiment is intended to test robustness under sparse visual context, where overlap between reference views becomes increasingly limited.

A key difference appears at low frame counts. When AV-Cloud is given 32 or fewer frames, COLMAP fails to produce a valid unified point cloud. In practice, the reduced overlap between reference images causes the reconstruction to fragment into multiple disconnected point clouds rather than a single coherent scene representation. We mark these settings as unavailable in Table 3. In contrast, our method does not depend on explicit point-cloud reconstruction and therefore remains fully operational even in these sparse-reference settings.

Among the valid settings, our method remains competitive and degrades more gracefully as the number of reference frames is reduced. Even with only 8 or 32 frames, it

Table 4. Running time comparison between SfM via COLMAP and VGGT across different numbers of point clouds. Results are for a single NVIDIA RTX 6000 Blackwell at 428×270 resolution. All values are in seconds (s).

SfM via COLMAP					VGGT		
N	**Feat. Ext.**	**Matching**	**Recon.**	**Total**	**N**	**Img. Read**	**Total**
8	10.21	0.27	0.55	11.03	8	0.02	**0.09**
16	10.26	0.74	0.29	11.29	16	0.06	**0.20**
32	30.37	2.84	3.65	36.86	32	0.11	**0.62**
64	40.60	11.17	7.27	59.04	64	0.25	**1.98**
128	21.00	46.78	47.13	114.91	128	0.52	**7.64**
256	21.69	143.89	124.55	290.13	256	1.05	**28.70**
616	22.14	248.74	302.99	573.88	–	–	–

Table 5. Ablation study on the RWAVS validation set. Best value in each column is shown in **bold**. Lower is better for all metrics.

Variant	MAG ↓	ENV ↓	LRE ↓	DPAM ↓
Full model (Ours)	**0.349**	**0.142**	**0.959**	**0.271**
No acoustic prototype init.	0.349	0.143	0.989	0.274
No estimated pose P_q	0.361	0.147	1.178	0.299
Q : target register token	0.359	0.146	1.148	0.298
32 reference frames	0.363	0.147	1.127	0.299
V : register tokens & 32 reference frames				
Q : ground truth target pose	0.401	0.157	1.245	0.487
Add target image→reference frames	0.385	0.153	1.215	0.458

maintains results close to the full 256-frame setting, while 64–256 frames yield consistently strong performance across all metrics. These results support the claim that the proposed feed-forward visual geometry grounding and acoustic prototype construction provide a robust basis for novel-view acoustic synthesis than reconstruction-dependent pipelines when references become sparse.

3.5. Ablation Studies

We conduct ablation studies on the RWAVS validation set to evaluate the contribution of the main components in the proposed model. Table 5 reports the results, with the full system as the reference. In addition to the visual context token and estimated pose used in the final model, the multi-view geometry encoder also produces intermediate register tokens that summarize scene-level semantic information. These register tokens are not used in the final architecture, but we include them in the ablations below as alternative query or value representations to test whether semantic descriptors can replace explicit geometric conditioning and acoustic prototype initialization.

The first group isolates the design choices of the final model. Removing acoustic prototype initialization leaves MAG nearly unchanged but degrades spatial and perceptual quality, increasing LRE from 0.959 to 0.989 and DPAM from 0.271 to 0.274. This indicates that initializing the value branch with acoustically grounded prototypes provides useful scene-specific bias beyond visual geometry

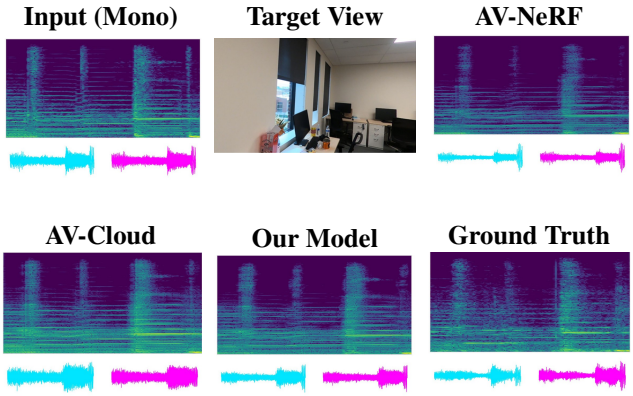

Figure 3. Visual comparison of binaural synthesis at a representative target view. Given a mono source signal and the queried target frame, we compare various methods against the ground-truth binaural audio. Our prediction is visually closest to the ground truth in both spectrogram structure and stereo waveform shape, indicating more accurate viewpoint-conditioned binaural rendering.

alone. Removing the estimated target pose embedding (P_q) causes a much larger drop, increasing LRE to 1.178 and DPAM to 0.299, which confirms that explicit listener-pose conditioning is essential for target-view acoustic retrieval. Replacing the query pose with a target register token partially recovers performance relative to removing P_q altogether, but still remains clearly worse than the full model (LRE 1.148 vs. 0.959), showing that semantic target descriptors cannot substitute for direct geometric target-pose conditioning. Reducing the reference set from 256 to 32 frames also degrades performance, with MAG rising to 0.363 and LRE to 1.127, demonstrating that wider scene coverage remains important for robust novel-view rendering.

The second group studies an alternative value design in which (V) is formed from visual register tokens using only 32 reference frames. Under this setup, using the ground-truth target pose in the query performs poorly, yielding MAG 0.401, LRE 1.245, and DPAM 0.487. Adding the target image to the reference set improves the result to MAG 0.385, LRE 1.215, and DPAM 0.458, but it still remains substantially worse than the final model. These results show that simply enriching the decoder with target-view semantics does not replace the role of acoustically grounded prototypes. Overall, the ablations support the final design choice of combining acoustic prototype initialization, explicit query-side target pose conditioning, and broad reference context in GGAD.

3.6. Qualitative Results Comparison

Figure 3 shows a representative visual comparison of binaural synthesis results. The input is a mono source signal, and the target frame specifies the queried listener viewpoint.

Compared with the ground truth, AV-NeRF shows a noticeably larger mismatch in spectral energy distribution and stereo waveform shape. AV-Cloud is closer, but still misses part of the target structure and binaural balance. Our model remains the closest to the ground truth in both the spectrogram and the left-right waveform patterns. These visual comparisons are consistent with the quantitative trends and indicate that the proposed multimodal context construction and geometry-grounded acoustic decoding produce more accurate viewpoint-conditioned binaural synthesis.

3.7. Reconstruction Runtime Comparison

Table 4 compares scene preprocessing time between SfM via COLMAP and feed-forward inference with VGGT as the number of reference views increases. COLMAP becomes increasingly expensive because matching and reconstruction dominate the runtime, whereas VGGT predicts visual geometry directly in a single forward pass without iterative reconstruction. The gap grows rapidly with scene size: at 64, 128, and 256 views, VGGT reduces preprocessing time from 59.04 s to 1.98 s, from 114.91 s to 7.64 s, and from 290.13 s to 28.70 s, respectively. These results support the practical motivation of our framework, namely that replacing reconstruction-dependent initialization with feed-forward visual geometry grounding substantially reduces preprocessing cost and improves scalability.

4. Related Works

Deep Acoustic Fields Encoding Historically, 3D spatial audio and sound-field representation methods have relied on handcrafted priors, sacrificing fidelity for efficiency [3, 10, 11, 19]. With the advent of deep learning, data-driven acoustic modeling began to take over. Prior works mainly focused on modeling, estimating, and generating Room Impulse Responses (RIRs) [22–24, 26], which characterizes how sound propagates in an enclosed space, including direct sound, reflections, and reverberation. For example, deep generative models are applied to learn RIR [22, 23]. However, RIR-based methods were often constrained in their spatial flexibility, typically requiring either a stationary listener or a stationary emitter, making them unsuitable for dynamic, free-roaming virtual environments.

Audio-Visual Spatial Audio Generation As researchers have noted the close connection between visual and audio modalities, visually informed audio spatialization methods have been proposed to jointly utilize information from both modalities [9, 21, 31, 32]. For example, Gao et al. [9] proposed injecting video frame features into the audio spatialization process to leverage spatial cues. Similarly, Zhou et al. [32] further improved visual-spatial cue injection using the proposed pyramid network. However, they generally focused on the matching between pairs of image and spatial audio, which resulted in limited modeling for the global acoustic field and the 3D scene.

Coupling Visual and Acoustic Rendering To cope with these limitations, methods featuring neural acoustic rendering are proposed to either explicitly or implicitly model the entire acoustic field [1, 6, 27], thereby enabling a new task known as novel-view acoustic synthesis, or NVAS. As VR and 3D vision applications are commercialized, breakthroughs in 3D reconstruction and rendering from 2D images [12, 20] are prompting a rethinking of spatial acoustic synthesis. Recent works have proposed similar acoustic rendering processes inspired by novel-view image rendering [2, 7, 15, 18]. Luo et al. [18] introduced neural acoustic fields as an implicit representation that captures sound propagation. Liang et al. [15] proposed AV-NeRF, which is reconstruction-driven by explicitly coupling novel-view image synthesis with acoustic synthesis. However, both [15, 18] require a known emitter or sound source location, making the process much less flexible. Prior works [2, 7] leverage explicit point-based 3D scene representations to model sound propagation. Among them, AV-Cloud [7] leverages a sparse anchor set to minimize the reliance on explicit visual rendering and avoid the heavy computations associated with full reconstruction of dense 3D Gaussians in work [2]. However, both still require a large number of continuous or overlapping visual frames as references to initialize point-based representations via SfM. AV-Cloud also suffers from limited visual semantics as the anchors only contain point-wise RGB values.

These works establish strong baselines for viewpoint-aware binaural rendering, but they still suffer from limited geometry cues with inefficient multimodal learning and photogrammetry pipelines. Our method addresses these challenges by leveraging fused visual semantics and geometry from feed-forward scene encoding and jointly attending to multimodal features during binauralization.

5. Conclusion

We presented a unified framework for novel-view acoustic synthesis with feed-forward visual geometry grounding. Our method builds a multimodal context from reference views, estimated scene geometry, and acoustically grounded prototype embeddings, and uses the Geometry-Grounded Acoustic Decoder to retrieve listener-conditioned acoustic information for target-view binaural rendering. Experiments on RWAVS and Replay-NVAS show that our approach achieves strong spatial-audio quality, favorable efficiency, and improved robustness compared with prior methods. These results suggest that feed-forward visual geometry grounding provides a practical and scalable foundation for high-quality novel-view binaural synthesis.

Acknowledgement

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

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, 8

[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. *Advances in Neural Information Processing Systems*, 37:28920–28937, 2024. 2, 8

[3] Jeroen Breebaart, Jürgen Herre, Christof Faller, Jonas Rödén, Francois Myburg, Sascha Disch, Heiko Purnhagen, Gerard Hotho, Matthias Neusinger, C Kjörling, et al. Mpeg spatial audio coding/mpeg surround: Overview and current status. *Preprint 119th Conv. Aud. Eng. Soc.*, 2005. 8

[4] 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 *European conference on computer vision*, pages 17–36. Springer, 2020. 1

[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. 2, 6

[6] 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. 1, 2, 6, 8

[7] Mingfei Chen and Eli Shlizerman. Av-cloud: Spatial audio rendering through audio-visual cloud splatting. *Advances in Neural Information Processing Systems*, 37:141021–141044, 2024. 2, 5, 6, 8

[8] Corey I Cheng and Gregory H Wakefield. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *Journal-Audio Engineering Society*, 49(4):231–249, 2001. 6

[9] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 8

[10] Michael A Gerzon. Periphony: With-height sound reproduction. *J. Audio Eng. Soc*, 21(1):2–10, 1973. 8

[11] Jürgen Herre, Johannes Hilpert, Achim Kuntz, and Jan Plogsties. Mpeg-h audio-the new standard for universal spatial/3d audio coding. *Journal of the Audio Engineering Society*, 2014. 8

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 8

[13] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024. 2

[14] Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi Lan, Manyuan Zhang, Gang Yu, Dingwen Zhang, et al. Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction. *arXiv preprint arXiv:2510.22706*, 2025. 2

[15] 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, 5, 6, 8

[16] 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. 6

[17] Shiguang Liu and Dinesh Manocha. Sound synthesis, propagation, and rendering: a survey. *arXiv preprint arXiv:2011.05538*, 2020. 1

[18] 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, 6, 8

[19] Rémi Mignot, Gilles Chardon, and Laurent Daudet. Low frequency interpolation of room impulse responses using compressed sensing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):205–216, 2013. 8

[20] 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. 1, 8

[21] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31, 2018. 2, 8

[22] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. Ir-gan: Room impulse response generator for far-field speech recognition. *arXiv preprint arXiv:2010.13219*, 2020. 8

[23] 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. 1, 8

[24] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3209–3213. IEEE, 2022. 8

[25] You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 2

[26] Christian J Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *2021*

IEEE workshop on applications of signal processing to audio and acoustics (WASPAA), pages 221–225. IEEE, 2021. 1, 8

[27] 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, 6, 8

[28] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 2

[29] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 5

[30] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024. 2

[31] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2021. 8

[32] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *European Conference on Computer Vision*, pages 52–69. Springer, 2020. 2, 8

[33] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2