

SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding

Mingfei Chen^{1,2*}

Israel D. Gebru²

Ishwarya Ananthabhotla³

Christian Richardt²

Dejan Markovic²

Jake Sandakly²

Steven Krenn²

Todd Keebler²

Eli Shlizerman¹

Alexander Richard²

¹University of Washington

²Codec Avatars Lab, Pittsburgh, Meta

³Reality Labs Research, Meta

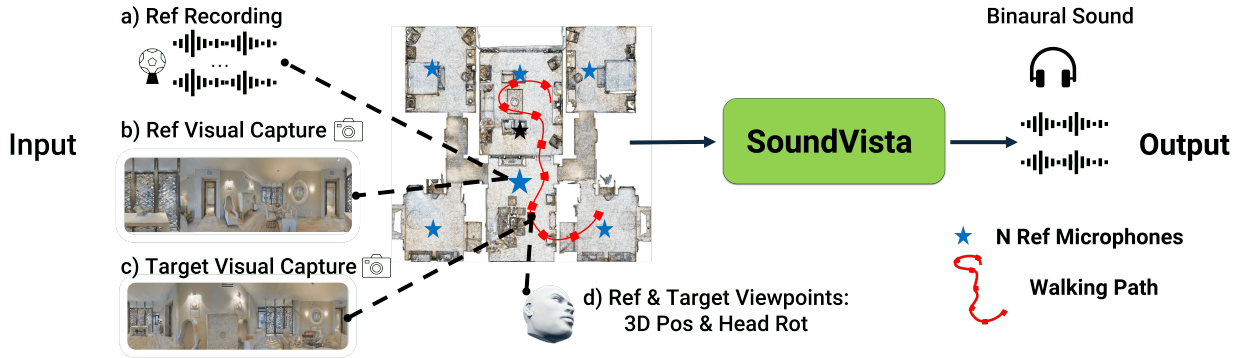


Figure 1. **SoundVista**: a novel method that synthesizes binaural ambient sound for arbitrary scenes from novel viewpoints. Our method leverages pre-acquired audio recordings and visual data captured from sparsely distributed reference points and synthesizes binaural audio consistent with the target 3D position and pose.

Abstract

We introduce *SoundVista*, a method to generate the ambient sound of an arbitrary scene at novel viewpoints. Given a pre-acquired recording of the scene from sparsely distributed microphones, *SoundVista* can synthesize the sound of that scene from an unseen target viewpoint. The method learns the underlying acoustic transfer function that relates the signals acquired at the distributed microphones to the signal at the target viewpoint, using a limited number of known recordings. Unlike existing works, our method does not require constraints or prior knowledge of sound source details. Moreover, our method efficiently adapts to diverse room layouts, reference microphone configurations and unseen environments. To enable this, we introduce a visual-acoustic binding module that learns visual embeddings linked with local acoustic properties from panoramic RGB and depth data. We first leverage these embeddings to optimize the placement of reference microphones in any given scene. During synthesis, we leverage multiple embeddings extracted from reference locations to get adaptive weights for their contribution, conditioned on target viewpoint. We benchmark the task

on both publicly available data and real-world settings. We demonstrate significant improvements over existing methods.

1. Introduction

Recent advances in 3D reconstruction and Novel-View Synthesis (NVS) have significantly enhanced our ability to create photorealistic visual models of real-world scenes [17, 34, 48, 49]. These developments have paved the way for applications in 3D virtual tours, experience recreation, and spatial media. However, the audio counterpart – Novel-View Acoustic Synthesis (NVAS) [1, 6] – has been under-explored compared to its visual counterpart and has not received the rigorous attention required to generate acoustic virtual scenes that match their visual surroundings.

To address this gap, we introduce *SoundVista*: a novel method that creates a truly immersive acoustic experience. *SoundVista* can generate realistic and spatially accurate binaural audio at novel viewpoints in arbitrary scenes given a sparse set of sample recordings from different viewpoint.

Unlike novel-view synthesis for visual 3D scenes, which are mostly static, NVAS faces significant challenges due to the dynamic nature of real-world acoustic environments. Ambient sound – the overall acoustic state describing all sounds

* Work done during an internship at Meta.

in a scene [55] – can encompass multiple time-varying, non-stationary sound sources without limitations on the type and number of sound sources. In an ideal scenario, with complete information about the ambient sound (*i.e.* clean signals of individual sound sources and their locations over time), and a fully reconstructed 3D visual model, one could synthesize the ambient sound at any location using standard acoustic renderers with room impulse responses (RIRs) [1, 4, 9, 19, 43]. In real-world scenes, however, we lack detailed information about the sound sources that comprise the ambient sound. Accurately determining their clean source content and emitter positions is a challenging problem.

NVAS also faces challenges in balancing data sampling costs and synthesis quality. Current methods in the acoustic community use grids of microphones, and techniques ranging from interpolation [28, 29] to complex signal processing [2, 50] to synthesize audio at novel locations. However, these approaches do not scale well to large spaces, and efficient sampling and data utilization remain open questions.

Recent deep learning based NVAS methods [5, 6, 8, 21, 25] utilize multimodal data, including visual inputs, to transfer sparse reference sounds to binaural sound at target viewpoints. These methods often simplify the task by focusing on re-synthesizing primary sounds such as speech and music, with a limited number of sound sources (typically 1–2). They often ignore other sounds that contribute to the natural characteristics of the ambient sound scene. While feasible with few references, they struggle to adapt to diverse, large scenes with complex layout and acoustic environment. Furthermore, the lack of an optimal reference location sampling strategy hinders adaptation, leading to a loss of acoustic details and making it difficult to synthesize high-quality sound that has accurate binaural effects for target viewpoints.

In this paper, we propose a novel approach that avoids the challenge of obtaining granular information about individual sound sources and addresses some of the shortcomings of existing NVAS techniques. Our method relies on sparsely distributed “reference microphones” to capture acoustic snapshots, taking their recorded audio signals as “reference recordings”. This allows us to get a holistic representation of the acoustic environment from references recordings. Given many examples of reference recordings, and sounds recorded at known locations within the scene, we formulate NVAS as learning the transformation from reference recordings to sound recorded at target viewpoints. At inference time, given any pre-acquired reference recordings of the scene as input and a query for an arbitrary target position, we expect the model to output binaural audio that is acoustically consistent with the query viewpoint and the content from reference recordings.

To address the limitations of existing NVAS techniques in reference microphone placement, we develop a multi-modal approach for optimal reference location sampling. Our ap-

proach can adapt to diverse, large complex scenes using a Visual-Acoustic Binding (VAB) module. VAB learns acoustically relevant features (VAB embeddings) by pretraining an encoder to align visual features from panoramic RGB-D captures with acoustic features from echo responses. Using VAB embeddings from numerous candidate reference locations, we employ a sampler that automatically identifies representative spots by finding spatial regions with similar embeddings. These spatial regions represent areas with similar acoustic properties and are usually free from obstacles that significantly hinder sound propagation. This approach optimizes reference microphone placement within a limited reference budget which enhances the overall performance.

Furthermore, to manage varying numbers of references – which typically depend on scene size and budget – and their unequal contributions to the final audio, we introduce an adaptive reference integration module. This module models sequences via a transformer, using VAB features to reweight reference inputs and viewpoints. These reweighted inputs then serve as conditions for the final binaural audio renderer. The resulting conditions are both scene-adapted and content-invariant, enabling effective handling of various reference microphone configurations and audio content.

To demonstrate the effectiveness of our approach, we benchmark the proposed model in both real-world setting and in a challenging simulated dataset derived from Matterport3D [3] scenes with SoundSpaces [4]. SoundVista outperforms existing methods [1, 5, 6, 8, 21] on scenes with multiple varieties of sound sources, handling varying numbers up to ten sound sources.

2. Related Work

Acoustic Scene Synthesis. Traditional methods focus on estimating the room impulse response (RIR) to recreate spatial audio. This is achieved by convolving the sound from each emitter with the RIR corresponding to the emitter-listener pair and summing the results [9, 20, 23, 25, 35–37, 39, 43]. These RIR-based techniques often require detailed source information, such as a clean signal of each source and its precise location, which are typically unknown in real-world scenarios, making them challenging to implement. Alternative approaches use scene-descriptor images for direct spatial sound synthesis based on reference sound input, such as visual acoustic matching [5, 7] or visual-guided audio spatialization [12, 13, 51, 57]. However, these methods may be inaccurate when viewpoints change continuously.

Novel-view Acoustic Synthesis. Novel-view acoustic synthesis techniques address limitations in acoustic scene synthesis by learning transfer functions from reference sounds to target viewpoints [6, 8, 21, 38, 46]. Methods such as *Few-shotRIR* [25] and *BEE* [8] use multiple references but lack optimal reference sampling strategies, relying on heuristic settings such as a single close reference [6, 21, 38, 46],

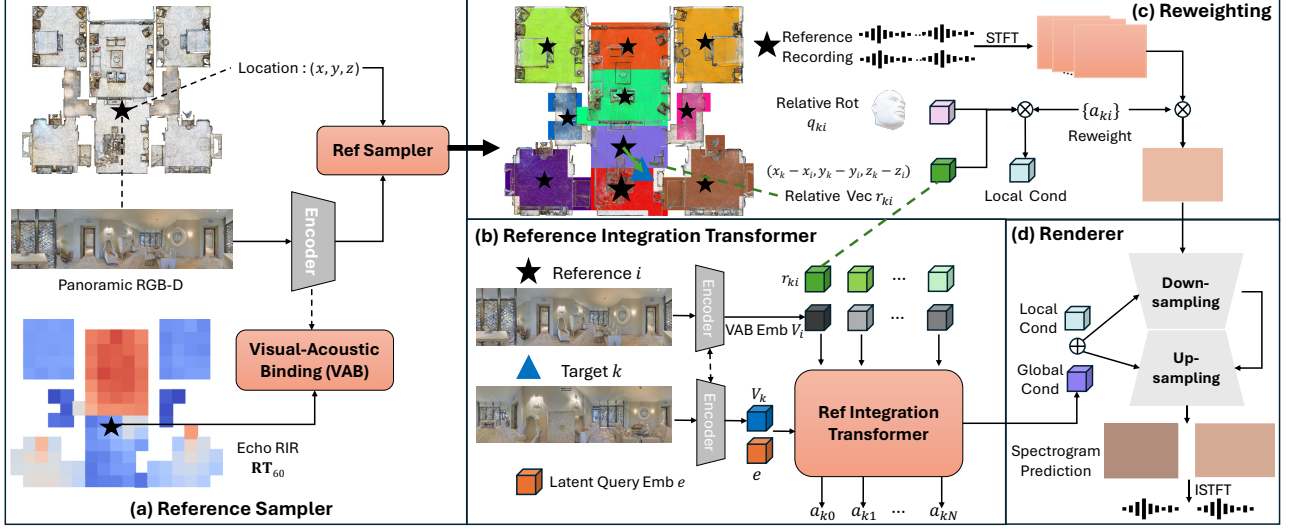


Figure 2. **Details of the SoundVista Pipeline:** (a) The reference location sampler selects optimal reference locations leveraging embeddings from visual-acoustic binding (VAB). (b) The reference integration transformer uses VAB embeddings to derive contribution weights for each reference. (c) Reweighting by contribution weights adjusts and integrates reference recording channels and pose conditioning for precise sound synthesis. (d) The spatial audio renderer converts reweighted channels and conditions to binaural sound at the target viewpoint.

random sampling [25], or fixed references [8], which limit adaptation to diverse scene layouts. In contrast, our method optimizes reference location sampling via visual-acoustic binding and derives adaptive reference contribution weights, enhancing sound synthesis accuracy and generalization.

Acoustic-Visual Learning. Techniques such as CLIP [11, 22, 24, 31, 52] have successfully aligned visual and linguistic inputs to learn matched deep semantic representations. These advances led to recent works to leverage complementary nature of audio and visual data for acoustic related tasks. For example, AV-RIR [37] binds RIR with panoramic images to improve the late reverberation modeling of RIR. Advances in multi-modal visual understanding have facilitated sound source localization [15, 18, 27, 44], audio-visual speech enhancement and source separation [10, 26, 30, 45, 54, 56], as well as acoustic scene reconstruction [5, 6, 8, 25, 37, 41]. These works underscore the strong link between visual and acoustic modalities. Inspired by these approaches, we developed a visual encoder that aligns features from RGB-D visual data with acoustic features from echo responses. The alignment enables us to infer acoustic properties using visual data only. As a result, our method does not require knowing acoustic parameters or RIRs during inference. It is also easily adaptable to various scene and acoustic environments.

3. Method

3.1. Problem Formulation

Given N pairs of co-located microphones and cameras, we aim to strategically place them in the acoustic scene. Let $\mathbf{L}_i = (x_i, y_i, z_i) \in \mathbb{R}^3$ denote the location of the devices and $\theta_i \in \mathbb{R}^2$ denote their orientations. The microphones cap-

ture reference audio recordings \mathbf{A}_i , and the cameras capture panoramic RGB-D views \mathbf{V}_i . These “reference captures” encapsulate the ambient sound and visual properties, providing the basis for reconstructing binaural ambient sound $\tilde{\mathbf{A}}_k$ for a target listener at arbitrary location $\mathbf{L}_k = (x_k, y_k, z_k) \in \mathbb{R}^3$ with orientation $\theta_k \in \mathbb{R}^2$. The reference recording can be of any type, but we use ambisonic microphones as they are ideal for capturing sound fields from multiple directions and simplify binaural signal extraction. In contrast to existing methods [1], we avoid enumerating sound sources. Instead, we learn a transfer function \mathcal{F} from reference captures to the target as:

$$\mathcal{F} : \left(\left\{ \mathbf{A}_i, g_i, r_{ki}, \theta_i \right\}_{i=1}^N, g_k, \theta_k \right) \mapsto \tilde{\mathbf{A}}_k, \quad (1)$$

$$g_i = \phi(\mathbf{V}_i), g_k = \phi(\mathbf{V}_k), \text{ and } r_{ki} = \mathbf{L}_k - \mathbf{L}_i,$$

where g_i and g_k are Visual-Acoustic Binding (VAB) embeddings extracted with a pre-trained visual encoder $\phi(\cdot)$, designed to align visual with acoustic representations. r_{ki} is relative vector introduced to avoid overfitting on absolute locations and to handle diverse scene coordinate systems.

We parameterize \mathcal{F} using neural networks that incorporate a transformer network and spatial audio renderer module. The transformer generates an adaptive mask to weight the contribution of each reference recording based on the target viewpoint and the VAB embeddings associated with reference viewpoints. We call this the “Reference Integration Transformer”. The spatial audio renderer then processes the reweighted audio channels along with conditional information from the target view features to produce the final binaural sound. We provide an overview of these modules in Figure 2, with details explained in the following sections.

3.2. Visual-Acoustic Binding (VAB)

Obtaining acoustic properties of real-world scenes is challenging [9]. We propose to leverage RGB-D data, which can provide rich information linked to acoustic properties of a scene. For example, depth data provides information about obstacles and room geometry, while pixel-level textures reveal material differences and detailed obstacle information.

The goal is to infer the acoustic properties of a scene using only the visual data. To achieve this, we first collected extensive paired panoramic RGB-D and acoustic echo responses data by navigating walkable areas in the SoundSpaces simulator [4]. We then trained a neural network, which we refer as “Visual-Acoustic Binding (VAB) module”, to predict the acoustic representations from the visual features.

Acoustic Representation. In acoustics, an impulse response is a function of time and the positions of the emitter and the listener [47]. It describes how sound propagates through a medium and interacts with the environment [32, 33]. To simplify data collection and infer local acoustic characteristic, we focus on echo impulse responses [25], where the emitter and target locations are the same. We extracted reverberation time (aka RT_{60} parameter) from echo response to use as acoustic representation. RT_{60} measures how long it takes for the acoustic energy to decay by 60dB and can reveal information about room geometry, obstacles and reflection. Figure 2(a) shows an example RT_{60} map of a scene. Significant value changes in RT_{60} map can indicate major obstacles or surface variations, highlighting key acoustic regions.

Visual Representation. Following prior works [6, 37], we use panoramic RGB-D captures as inputs and extract visual features from ResNet-18 [14]. We train the VAB module to predict RT_{60} . Therefore, effective binding visual representation with acoustic representation.

3.3. Reference Location Sampler

While having more microphones and cameras is ideal for this task, in practice resources are often limited. To maximize performance and make the best use of available resources, we propose a strategic approach for placement of reference microphones throughout a given scene.

We argue that ideal microphone placements align with the acoustic partitions of the scene—areas with unique acoustic properties and free from obstacles such as walls. To identify these partitions, we capture panoramic RGB-D images from all candidate locations. One can use novel-view synthesis [49] to render these images; without needing actual photographs at each spot. We then extract VAB embeddings from each capture. VAB embeddings which correspond to scene acoustic parameters serve as strong cues for identifying acoustic partitions, such as distinguishing regions separated by obstacles. To enhance reliability, we combine the extracted VAB embeddings with positional information,

allowing them to complement each other. Using these embeddings, we perform data clustering of the candidate locations and take the center of each cluster as a reference location. This is illustrated in Figure 2(a).

3.4. Reference Integration Transformer

We want our model to work effectively across diverse scenes. This requires the model to adapt to varying numbers of reference audio and visual inputs. Logically, larger scenes would benefit from more microphones and cameras, while smaller spaces can operate efficiently with fewer resources. Moreover, since the task (Section 3.1) is to transfer reference recordings to target sounds at specific viewpoints, distant microphones, or those in non-adjacent acoustic partitions usually contribute significantly less to the process than closer ones. However, weighting based solely on distance is insufficient due to potential obstacles such as walls and objects.

To address these issues, we propose a transformer network. We treat each reference input as part of a sequence, allowing us to manage varying numbers of reference inputs. To derive the weights for their unequal contributions to the final audio, we exploit VAB embeddings extracted from visual captures at the reference and target viewpoints.

We formed the queries input by combining the VAB embedding at the target location g_k , with a learnable latent query embedding e . The query is initialized from a normal distribution and optimized during training, to adjust the contribution of reference microphones based on their relative location vector to the target location. Let $g_k^e = [g_k || e] \in \mathbb{R}^C$ to denote the queries. The keys and values inputs are created by combining the VAB embedding of each reference with its relative vector r_{ki} . Let $g_i^r = [g_i || r_{ki}] \in \mathbb{R}^C$ to denote the keys and values of i^{th} reference for the attention operation. The attention weight a_{ki} can be calculated as follows:

$$a_{ki} = \frac{g_k^e \cdot g_i^{r\top}}{\sqrt{C}}, \quad i = \{1, 2, \dots, N\}. \quad (2)$$

The attention weight a_{ki} indicates contribution of the reference i to the target k normalized by the Softmax function over N references. Higher weights indicate greater influence.

3.5. Spatial Audio Renderer

The Spatial Audio Renderer takes the reference recordings and conditional information to generate binaural sound at the target viewpoint.

Reweighting Reference Recordings. We apply the Short-Time Fourier Transform (STFT) to the reference recordings to extract their spectrograms. Then, we reweight the channels of each reference recording using the attention weights a_{ki} from Section 3.4 to integrate them, encoding the result into a latent space as the input for the renderer.

Reweighting Condition Inputs. The binaural sound is derived based on the target’s position and pose. To capture the

different aspects of the binaural effects, we decouple the conditioning into global and local components. This accounts for the cause of spatialization effects at a given target point, which may primarily vary from the distance to the sound source or remain invariant to head orientation. The global condition c_g determines how the target viewpoint’s relative position to reference locations influences the binaural sound. The local condition c_l accounts more for the effects of head orientation. They are defined as follows:

$$c_g = \sum_{i=1}^N a_{ki} \cdot \psi_1(g_i \parallel r_{ki}), \quad (3)$$

$$c_l = \left(\sum_{i=1}^N a_{ki} \cdot \psi_2(r_{ki} \parallel q_{ki}) \right) + \sigma(\theta_k). \quad (4)$$

Here q_{ki} is the relative orientation quaternion for target k relative to reference i ; ψ_1 and ψ_2 are MLP layers projecting concatenated inputs to a latent condition embedding; and $\sigma(\theta_k)$ denotes the target’s rotational features from θ_k .

Renderer. The Audio Renderer network is designed as a stacked U-Net [40, 42], depicted in Figure 2(d). It features down-sampling and up-sampling blocks that incorporate input conditions at each layer. To preserve quality and content details from the reference input, skip connections are employed at multiple resolutions. After processing the integrated reference recordings, we apply the inverse STFT to convert the output spectrogram into a binaural waveform.

Training from scratch in complex environments with challenging audio content can hinder the renderer to adapt to head orientation while producing high-quality audio. To address this, we pretrain the binauralization capability of our renderer by fixing the target location and varying head rotations and audio sources. After pre-training, the renderer effectively understands binauralization across different orientations, enhancing the accuracy of spatial effect modeling.

3.6. Loss

We design the loss function as a weighted combination of three components, each regulating different aspects of the model: 1) *Waveform Loss* measures the mean squared error between the predicted and target waveforms, ensuring precision of the predicted waveform amplitude and phases. 2) *Binaural Interaural Level Difference (ILD) Loss* [16] focuses on the energy difference between binaural audio channels to ensure accurate spatial effect prediction. 3) *Multi-resolution Spectrogram Magnitude Loss* [53] ensures that the predicted audio matches the target spectrogram magnitude across multiple resolutions. It includes two terms: the first compares log-scaled magnitudes, and the second, the Scaled Spectrogram Magnitude Loss, calculates magnitude distance over the target scale, addressing high variance in audio magnitude. We found that resolution with large FFT and small hop sizes can benefit modeling ambient noise, such as air vibrations.

4. Experiments

4.1. Benchmarks and Metrics

N2S Benchmark. Captured in a real room of 8.5 by 6 meters dimensions with two internal rooms, the N2S benchmark utilizes microphone arrays with 32 microphones for 32-channel recordings. Visual captures are rendered by the NeRF-based 3D NVS model, VR-NeRF [49], after it has been trained on visual data collected by navigating the room. 11 static microphone arrays are uniformly distributed to provide reference ambient recordings, while 6 mobile arrays cover 557 locations with 15 seconds of recording per location. Motion capture system, OptiTrack, records 6DoF tracking data for both reference and target microphones. 35% of these locations with 20 orientations are selected for training.

Soundspace-Ambient Matterport3D. We build this benchmark based on Matterport3D scenes, comprising 39 complex training scenes and 23 unseen scenes. It is challenging as each scene can include 2 to 10 sound sources, selected from 128 different sounds, from fan noise to speech. Reference locations are spaced every 5 meters on average, with simulated second-order ambisonics RIRs to render reference sounds by Soundspaces [4]. 85% of walkable locations are used for training with azimuth angles of 0, 90, 180, and 270.

Metrics. We evaluate sound synthesis performance using four key metrics (**the lower, the better**): 1) L1 distance of STFT spectrogram (**STFT**) [13] for left and right channels; 2) Magnitude Spectrogram Distance (**MAG**) [13], measuring the closeness of reconstructed audio to the groundtruth; 3) Energy Envelope Error (**ENV**) [13], assessing the Euclidean distance between energy envelopes of groundtruth and predicted audio channels; 4) Left-Right Energy Ratio Error (**LRE**) [6], evaluating binaural effect accuracy by calculating energy ratio difference between left and right channels.

4.2. Comparison with Baselines

We compare our method with the following baseline approaches: (1) **Nearest GT**: Binaural sound from the nearest reference microphone aligned with the target orientation. (2) **Interp GT**: Employs binaural sound at the target orientation from the four nearest reference microphones, interpolating based on distance to the target location. (3) **AV-NeRF** [21]: A NeRF-based system that synthesizes binaural audio from a given camera pose and RGB-D renderings. We adapted it to use recordings, poses, and visual context from references while preserving its core model components, enabling fair comparisons in dynamic scenes. (4) **Few-shotRIR** [25]: A transformer-based method that infers RIRs from a sparse set of observed images and echoes. Adapted to replace the reference and target impulse responses with the ambient sounds at the corresponding viewpoints. (5) **VAM** [5]: Matches the acoustics of input audio with a target image. (6) **BEE** [8]:

Table 1. **Results Comparison on Soundspace-Ambient Benchmark:** Average metrics for 10,189 samples in 39 seen scenes with novel target locations and sources, and 6,534 samples in 23 unseen scenes. Sampling strategies: *location* only, *vis+location* (our sampler), and *w/o VAB* (ours without VAB pretraining). *SoundVista* Ref Num *k*: only references with top-*k* contribution weights are used for fair comparison.

Method	Sampling	Ref Num	Seen Scenes				Unseen Scenes			
			STFT ↓	MAG ↓	ENV ↓	LRE ↓	STFT ↓	MAG ↓	ENV ↓	LRE ↓
nearest GT	location	1	4.448	0.351	0.154	1.596	4.034	0.353	0.155	1.617
nearest GT	w/o VAB	1	4.414	0.341	0.151	1.576	4.680	0.347	0.153	1.537
nearest GT	vis+location	1	3.916	0.336	0.146	1.572	3.835	0.344	0.151	1.557
interp GT	location	4	3.922	0.326	0.144	1.584	3.410	0.327	0.145	1.587
interp GT	w/o VAB	4	3.660	0.319	0.142	1.570	3.766	0.320	0.142	1.531
interp GT	vis+location	4	3.179	0.313	0.137	1.559	3.415	0.321	0.141	1.531
AV-NeRF [21]	vis+location	1	9.424	0.426	0.195	1.922	9.321	0.428	0.196	1.979
VAM [5]	vis+location	1	5.224	0.420	0.178	1.902	4.936	0.436	0.182	1.977
ViGAS [6]	vis+location	1	3.740	0.361	0.154	2.040	3.438	0.371	0.157	2.051
SoundVista (Ours)	vis+location	1	2.526	0.291	0.132	1.408	2.676	0.309	0.140	1.386
BEE [8]	vis+location	4	4.098	0.365	0.162	2.083	5.635	0.396	0.178	2.131
SoundVista (Ours)	vis+location	4	2.444	0.289	0.130	1.390	2.517	0.305	0.137	1.371
Few-shotRIR [25]	vis+location	all	5.937	0.459	0.213	1.892	5.457	0.471	0.215	1.960
SoundVista w/o vis	location	all	3.228	0.306	0.141	1.425	2.890	0.312	0.142	1.439
SoundVista (Ours)	vis+location	all	2.442	0.289	0.130	1.390	2.514	0.305	0.137	1.372

Table 2. **Testing Results Comparison on N2S Benchmark.** In this real-world scene, *SoundVista* with visual modality largely boosts the performance accuracy, especially on the binaural effect (LRE).

Method	STFT ↓	MAG ↓	ENV ↓	LRE ↓
Nearest DSP	2.420	0.212	0.136	1.447
Interp DSP	1.659	0.203	0.142	1.383
AV-NeRF [21]	2.194	0.187	0.119	0.840
Few-shotRIR [25]	1.765	0.199	0.134	0.909
VAM [5]	1.972	0.190	0.119	0.916
BEE [8]	1.471	0.200	0.141	0.995
ViGAS [6]	1.201	0.185	0.119	0.873
SoundVista w/o vis	1.242	0.185	0.118	0.894
SoundVista (Ours)	1.073	0.177	0.113	0.776

A generalizable rendering pipeline that reconstructs the binaural audio at an arbitrary listener location using inputs from sparse reference audio-visual samples in the scene. (7) **ViGAS** [6]: Transforms sound to the target viewpoint given the observed audio and visual captures at the reference viewpoint. Specifically, for *AV-NeRF*, *VAM*, and *ViGAS*, which need a single reference, we use the nearest reference microphone. *BEE* is adapted to use the nearest 4 microphones. Deep-learning baselines requiring visual inputs utilize panoramic RGB-D images and the same visual encoder as our model for consistency.

Our results in Table 1 and Table 2 demonstrate that our method, *SoundVista*, consistently surpasses baselines across diverse novel scenes and real scenarios. Despite the challenge of obtaining binaural groundtruth (GT) for arbitrary target orientations, *SoundVista* significantly reduces errors across all metrics compared with *nearest GT* and *interp GT*. On the challenging Soundspace-Ambient benchmark, with diverse sound sources and complex layouts, most deep-learning baselines underperform compared to non-learning methods due to the need for robust conditioning models.

For a fair comparison, we evaluate *SoundVista* using the top 1 and 4 reference microphones by contribution weight, respectively. Compared to learning-based methods with 1 reference, *SoundVista* outperforms the best baseline *ViGAS* by 32.5% (STFT), 19.4% (MAG), 14.3% (ENV), and 31% (LRE), demonstrating the strength of our conditioned audio renderer. The nearest microphone may miss critical audio content, while using the top 4 references improves performance, matching that of all microphones since most references are not required due to their sparse distribution. Since contribution weights are independent of audio content, fewer than 4 references can be selected per target location by pre-computing contribution weights once per scene.

While *Few-shotRIR* and *BEE* use audio content to integrate references, the diverse sound distribution makes this challenging. During testing, content often falls outside the training distribution, degrading performance. Thus, more references aren’t always better; for example, *Few-shotRIR* did not gain an increase in synthesis accuracy and it was even reduced compared to baselines using only one reference.

The deep-learning baselines perform better on the N2S benchmark, which features a single static scene. Compared to these deep-learning baselines, our *SoundVista* still significantly outperforms on accuracy of energy (ENV) and binaural effect (LRE) by 5% and 7.6%, respectively.

4.3. Qualitative Results Comparison

Figure 3 visualizes our results for qualitative comparison with several competitive sound synthesis baselines: *ViGAS*, *BEE*, and *Few-shotRIR*. The first row displays loudness maps, while the second row shows generated binaural waveforms with corresponding LRE error displayed in the third row. *ViGAS* relies on the nearest microphone, leading to discrete loudness maps in complex layouts and unreliable results

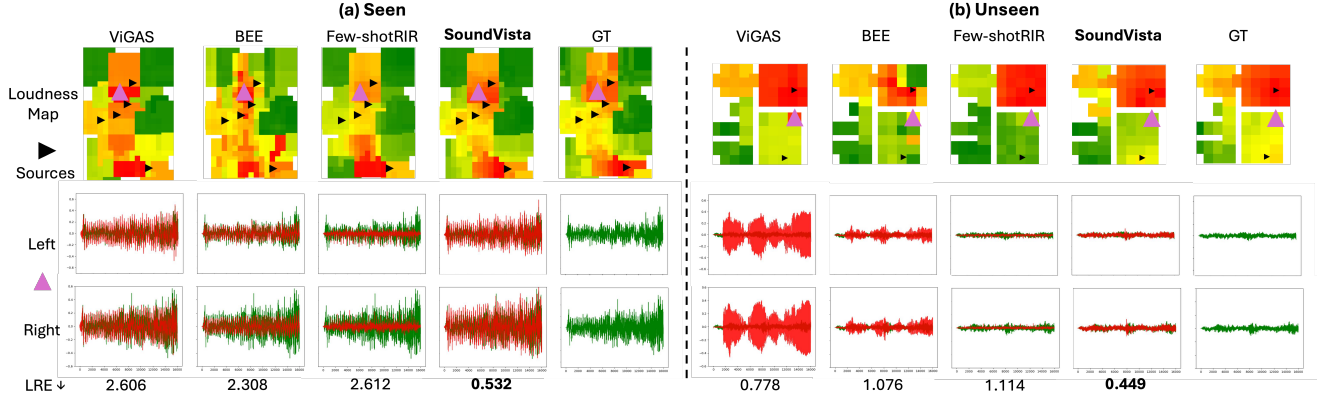


Figure 3. **Comparison of Qualitative Results:** First row: Loudness Map (black triangle: sources; purple triangle: target viewpoint). Second row: Reconstructed binaural waveform at target viewpoint (red: prediction; green: GT). Third row: LRE (lower for better binaural effect).

when obstacles are present, as shown in the unseen results. Additionally, *ViGAS* struggles with accurate binaural effects shown from the LRE results. *BEE* produces discontinuous loudness maps, deviating from the ground truth. Both *BEE* and *Few-shotRIR* have difficulty generating high-fidelity waveforms close to the groundtruth. Their reliance on incorporating audio content into the renderer’s conditioning can bias the renderer leading to confusion and reduced quality when encountering content outside the training distribution. This is particularly common in our complex task setting with diverse audio sources. In contrast, *SoundVista*, closely matches the GT in loudness maps and binaural waveforms for both seen and unseen scenes, excelling in high-quality novel-view ambient sound synthesis.

4.4. Ablation Study

We conducted ablation studies on the Soundspace-Ambient benchmark to verify the contribution of key components.

Visual-Acoustic Binding (VAB). In the design of VAB, we optimized the integration of visual features that correlate with acoustic parameters. Table 3 shows the accuracy of RT_{60} value predictions in novel scenes without finetuning (*w/o finetune*) and with few-shot finetuning at reference locations (*w/ finetune*). We tested various modality inputs: *location* only, panoramic *rgb*, *depth*, and *rgb+depth*. Depth alone significantly enhances binding, reducing error by over 50% compared to use *location* only. With finetuning, *rgb+depth* achieves superior results. Given the ability to capture references in scenes, we select *rgb+depth* as our visual input.

Reference Sampler via VAB. The Reference Sampler via Visual-Acoustic Binding (VAB) (*vis+location*) significantly enhances the accuracy of non-learning baselines such as *nearest GT* and *interp GT*, as shown in Table 1. This method improves predictions of sound magnitude (MAG, ENV) and spectrogram phases (STFT). Without pretrained VAB (*w/o VAB*), these benefits diminish, especially in seen scenes. Results in Table 1, 2, and 4 further highlight the critical role of

Table 3. **RT_{60} Prediction Results on Matterport3D.** *w/ finetune* involves few-shot finetuning on unseen scenes. *err*: distance between predicted and GT RT_{60} , *error ratio*: scaled *err* over GT.

Modality	w/o finetune		w/ finetune	
	err ↓	err ratio ↓	err ↓	err ratio ↓
location	0.170	0.411	0.067	0.199
rgb	0.117	0.236	0.044	0.126
depth	0.082	0.173	0.038	0.109
rgb+depth	0.087	0.185	0.036	0.103

Table 4. **Ablations on the SoundSpace-Ambient benchmark.**

Method	STFT ↓	MAG ↓	ENV ↓	LRE ↓
full model	2.442	0.289	0.130	1.390
w/o vis	3.228	0.306	0.141	1.425
w/o reweight	2.908	0.295	0.134	1.391
w/o render_pretrain	2.582	0.310	0.135	1.877
w/o decoup_cond	3.017	0.300	0.138	1.433
full loss	2.442	0.289	0.130	1.390
w/o ILD_loss	2.462	0.292	0.131	1.420
w/o mag_scale	2.424	0.301	0.136	1.422
w/o wav_loss	3.422	0.291	0.143	1.369

visual input, as its absence increases errors across all metrics, underscoring the importance of VAB with visual information for adapting to diverse scene layouts.

Figure 4 (left) illustrates the STFT error curve relative to reference density. Our model is trained with a reference density of 1.0, and tested on novel Soundspace-Ambient scenes with varying densities. Compared to *Ours - w/o vis* and *interp - w/o vis* (*interp GT* with location-only sampler), *Ours - w/ vis* outperforms others, with the performance gap widening as density decreases, demonstrating robustness of ours equipped with VAB vs. reference density variance.

Reference Integration. To evaluate the contribution of the adaptive reference integration, we conducted experiments with the *w/o reweight* variant, which uses only the nearest reference microphone along with its corresponding visual and pose information to calculate the condition for the audio renderer. As shown in Table 4, the results of *w/o reweight* compared to full *SoundVista* model reveal a noticeable degra-

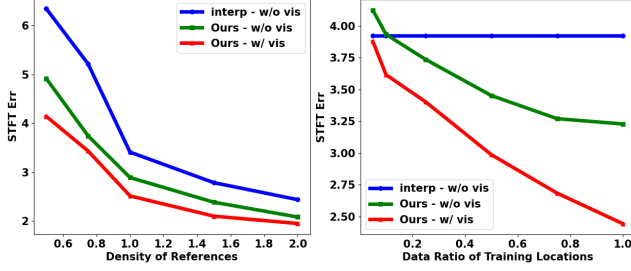


Figure 4. **STFT Error Curve** for different reference densities (left) and training location data ratios (right).

dation in performance, particularly in the STFT error, which increases by 19%. This indicates that the adaptive Reference Integration Transformer and reweighting modules effectively integrate references to enhance sound synthesis accuracy, as the nearest microphone may miss critical content required for the target sound synthesis.

Conditioned Spatial Audio Renderer. In the *w/o decoupled_cond* variant, we replaced the decoupled global and local conditions from reweighting in Section 3.5 with only the global condition plus target rotational features. This absence of local conditions significantly increases STFT error by 23.5%, MAG error by 3.8%, and ENV error by 6.2%.

As detailed in Section 3.5, we pretrain the renderer with binauralization capability to help the model learn the complex task of orientation-sensitive novel view sound synthesis. The *w/o render_pretrain* variant shows a 35% increase in LRE error, underscoring the importance of renderer pretraining for achieving a more accurate binaural effect.

Loss. The ablations in Table 4 highlight the contributions of different loss components. The Waveform Loss (*wav_loss*) constrains the energy envelope and regulates phase learning. Removing it significantly reduces STFT and ENV accuracy, increasing the STFT error by 29%. The Scaled Spectrogram Magnitude Loss (*mag_scale*) calculates magnitude loss over the target magnitude scale, addressing high variance in audio content and mitigating scale differences. Removing *mag_scale* will increase MAG error largely.

Robustness w.r.t. Training Data Ratio. To assess the impact of training location data ratio on performance, we trained two variants, *Ours - w/o vis* and *Ours - w/ vis*, using different proportions of training locations on the Soundspace-Ambient benchmark. Figure 4 (right) shows that STFT error increases as the data ratio decreases. When the data ratio is below 10%, *Ours - w/o vis* underperforms compared to the non-learning baseline *interp - w/o vis*, while *Ours - w/ vis* maintains superior performance. Additionally, *Ours - w/ vis* benefits more from increased training locations, with STFT error decreasing rapidly as the ratio nears 100%, widening the performance gap with *Ours - w/o vis* and *interp - w/o vis*.

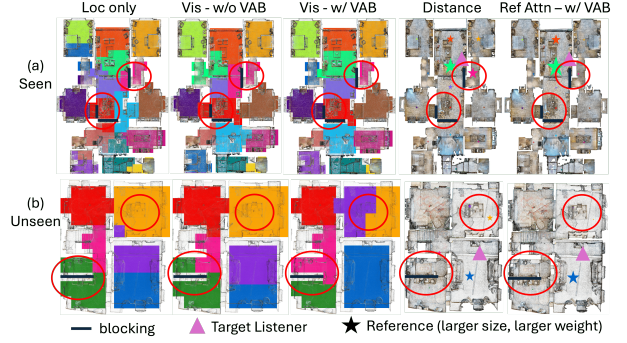


Figure 5. **Visualization Analysis.** First three columns: Clustering results for *loc only*, *Vis - w/o VAB* (ours without VAB pretraining), and *Vis - w/ VAB* (ours). Colors indicate cluster regions. VAB helps clustering to align better with room partitions. Last two columns: Attention weights visualization for references, with colored stars (size proportional to weights) and a purple triangle for the target.

4.5. Visualization Analysis

Figure 5 visualizes the clustering results of Reference Sampler via VAB, comparing *loc only*, *Vis - w/o VAB*, and *Vis - w/ VAB*. Different colors represent cluster regions. The Reference Sampler *Vis - w/ VAB* accurately segments clusters aligning with actual room partitions, proving to be more reliable. In contrast, other methods may incorrectly cluster non-adjacent rooms together which cannot share the same acoustic environment, highlighted by red-circles.

After identifying reference locations, we visualize the attention weights from the Reference Integration Transformer in Figure 5 (column 4-5 from left). Colored stars denote sampled references, sized by contribution weight, with the target marked by a purple triangle. Unlike *Distance*-only weights, which may incorrectly prioritize closer references despite obstacles, our model, *Ref Attn - w/ VAB*, trained with the adaptive Reference Integration Transformer, effectively highlights reliable references in similar acoustic environments.

5. Conclusion

We introduce *SoundVista*, a novel system designed to synthesize ambient sound from arbitrary scenes at novel view-points. Our approach introduces a visual-acoustic binding module to effectively learn visual embeddings linked with local acoustic properties. This enables our system to optimize the placement of reference microphones and derive adaptive weights for each microphone’s contribution, conditioned on the target viewpoint and visual captures, thereby, facilitating the final conditioned sound synthesis. *SoundVista* adapts to diverse room layouts, microphone configurations, and unseen environments, rendering high-quality binaural ambient sound without requiring prior constraints or detailed knowledge of sound sources. Our results, validated on both publicly available data and real-world settings, demonstrate state-of-the-art sound synthesis accuracy and generalization.

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. 1, 2, 3
- [2] Federico Borra, Israel Dejene Gebru, and Dejan Markovic. Soundfield reconstruction in reverberant environments using higher-order microphones and impulse response measurements. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 281–285. IEEE, 2019. 2
- [3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2
- [4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicens 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. 2, 4, 5
- [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, 3, 5, 6
- [6] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. *arXiv preprint arXiv:2301.08730*, 2023. 1, 2, 3, 4, 5, 6
- [7] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. In *ICASSP*, 2023. 2
- [8] Mingfei Chen, Kun Su, and Eli Shlizerman. Be everywhere-hear everything (bee): Audio scene reconstruction by sparse audio-visual samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7853–7862, 2023. 2, 3, 5, 6
- [9] 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. 2, 4
- [10] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 3
- [11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3
- [12] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019. 2
- [13] 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. 2, 5
- [14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [15] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *Advances in Neural Information Processing Systems*, 33: 10077–10087, 2020. 3
- [16] Christopher A. Ick, Gordon Wichern, Yoshiki Masuyama, François Germain, and Jonathan Le Roux. Spatially-aware losses for enhanced neural acoustic fields. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024. 5
- [17] Hyeonjoong Jang, Andreas Meuleman, Dahyun Kang, Donggun Kim, Christian Richardt, and Min H Kim. Egocentric scene reconstruction from an omnidirectional video. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022. 1
- [18] Hao Jiang, Calvin Murdock, and Vamsi Krishna Ithapu. Egocentric deep multi-channel audio-visual active speaker localization. *arXiv preprint arXiv:2201.01928*, 2022. 3
- [19] Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, 2021. 2
- [20] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Neural acoustic context field: Rendering realistic room impulse response with neural fields. *ArXiv*, 2023. 2
- [21] 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. 2, 5, 6
- [22] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard De Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 3
- [23] Andrew Luo, Yilun Du, Michael J Tarr, Joshua B Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *arXiv preprint arXiv:2021.08860*, 2021. 2
- [24] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 3
- [25] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics, 2022. 2, 3, 4, 5, 6
- [26] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021. 3

- [27] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10565–10574, 2023. 3
- [28] Bartłomiej Mróz, Marek Kabaciński, Tomasz Ciotucha, Andrzej Rumiński, and Tomasz Żernicki. Production of six-degrees-of-freedom (6dof) navigable audio using 30 ambisonic microphones. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–5. IEEE, 2021. 2
- [29] Orhun Olgun, Ege Erdem, and Hüseyin Hacıhabiboğlu. Sound field interpolation via sparse plane wave decomposition for 6dof immersive audio. In *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–10. IEEE, 2023. 2
- [30] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [32] Nikunj Raghuvanshi and John Snyder. Parametric wave field coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 33(4):1–11, 2014. 4
- [33] Nikunj Raghuvanshi and John Snyder. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 4
- [34] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 1
- [35] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. Irigan: Room impulse response generator for far-field speech recognition. *arXiv preprint arXiv:2010.13219*, 2020. 2
- [36] Anton Ratnarajah, Shi-Xiong Zhang, Meng Yu, Zhenyu Tang, Dinesh Manocha, and Dong Yu. Fast-rir: Fast neural diffuse room impulse response generator. *arXiv preprint arXiv:2110.04057*, 2021.
- [37] Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. Av-rir: Audio-visual room impulse response estimation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27154–27165, 2023. 2, 3, 4
- [38] Alexander Richard, Dejan Markovic, Israel D. Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*, 2021. 2
- [39] Alexander Richard, Peter Dodds, and Vamsi Krishna Ithapu. Deep impulse responses: Estimating and parameterizing filters with deep networks. *arXiv preprint arXiv:2202.03416*, 2022. 2
- [40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 5
- [41] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 286–295, 2021. 3
- [42] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 5
- [43] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022. 2
- [44] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3927–3935, 2021. 3
- [45] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, 2021. 3
- [46] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *Arxiv*, 2016. 2
- [47] Tor Erik Vigran. *Building acoustics*. CRC Press, 2014. 4
- [48] Xiuchao Wu, Jiamin Xu, Zihan Zhu, Hujun Bao, Qixing Huang, James Tompkin, and Weiwei Xu. Scalable neural indoor scene rendering. *ACM transactions on graphics*, 41(4), 2022. 1
- [49] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Proceedings*, 2023. 1, 4, 5
- [50] Shaoheng Xu, Jihui Aimee Zhang, Thushara D. Abhayapala, Amy Bastine, Wei-Ting Lai, and Prasanga N. Samarasinghe. Sparse sound field representation using complex orthogonal matching pursuit. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1336–1340, 2024. 2
- [51] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *CVPR*, 2021. 2
- [52] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 3

- [53] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020. [5](#)
- [54] Yuxin Ye, Wenming Yang, and Yapeng Tian. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5508–5519, 2024. [3](#)
- [55] Zechen Zhang, Nikunj Raghuvanshi, John Snyder, and Steve Marschner. Ambient sound propagation. *ACM Transactions on Graphics (TOG)*, 37(6):1–10, 2018. [2](#)
- [56] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 283–292, 2019. [3](#)
- [57] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer, 2020. [2](#)

SoundVista: Novel-View Ambient Sound Synthesis via Visual-Acoustic Binding

Supplementary Material

6. Demo Examples

Please see the attached videos in the supplementary `demo_videos` folder. We included videos from Matterport3D scene and our real-world scene (N2S). For the best experience, please **turn on your audio and use headphones**.

6.1. Real Scene: N2S Demo

This demo contains videos from a real-world scene. The scene is captured using 11 reference microphones, their spatial distribution is shown in Figure 6 (*Ref Num* = 11). **Unlike simulated scenes, the real scene presents challenges with diverse natural sounds, including diffuse machine noise and air conditioner vibrations, which are difficult to identify and localize in 3D.** Using reference sounds as input for the Novel-View Ambient Sound Synthesis task proves more effective than attempting to localize and separate sources to render with Room Impulse Responses (RIRs).

In the demo video (`0_0_n2s_soundvista.mp4`) of *SoundVista*, three dominant sound sources are clearly identifiable: a TV playing water and bird sounds, a black speaker in the corner playing music, and an air conditioner producing diffuse noise throughout the scene. The sound changes noticeably when entering a small, noisy room with considerable reverberation. As the listener continuously moves in the scene, our model was able to reconstruct these sounds without requiring source counting, localization and RIR data.

6.2. Soundspace-Ambient Matterport3D Demo

Videos prefixed with `1_x` are results from Matterport3D scenes. We show results from 10 different rooms that are part of the Soundspace-Ambient benchmark. In `1_0_mp3d_source_explain.mp4`, we outline the setup, which includes 17 reference points (green stars) and 5 sound sources (blue triangles) distributed throughout the scene. The sources produce various sounds, such as running shower water, engine noise, fireplace crackling, a phone ring, and birds chirping.

In the videos, the listener (target); shown as a red circle ● navigates between rooms throughout the scene. The binaural sound adapts naturally to both viewing orientation and source distance. Though sound transitions remain mostly smooth, crossing between rooms can create more sudden changes because of physical barriers.

6.3. Comparison with Baselines

We compare our results with two baselines: *DSP*, a traditional signal processing approach that interpolates binaural sounds from the four nearest reference points using target

orientation and distance, and *ViGAS*, a recently proposed learning-based method. *SoundVista* produces better results compared to the baseline methods. Specifically, *SoundVista* smoothly adapts binaural effects to view orientations.

For example, in the N2S scene, when navigating the TV region (video `0_1_n2s_comparison_tvclip.mp4`) and turning around, *DSP* and *ViGAS* fail to properly track the TV sound as it moves from left to front to right. Moreover, *DSP*'s simple interpolation of nearby reference points proves inadequate for handling obstacle effects, resulting in inaccurate sound magnitudes. *ViGAS* introduces sound distortions, especially with bass-heavy music and engine noise, and produces unexpected abrupt changes in sound magnitude. *SoundVista*, in contrast, delivers consistently high-quality (undistorted), smooth, and continuous audio. Similar examples are also demonstrated in comparison videos around N2S speaker (video `0_2_n2s_comparison_speakerclip.mp4`) and in the Soundspace-Ambient Matterport3D scene (video `1_1_mp3d_comparison.mp4`).

7. Implementation Details

This section details the implementation of each *SoundVista* module.

7.1. Visual Acoustic Binding (VAB)

For training, we partition the panoramic image into four RGB-D views, each of size $224 \times 224 \times 4$, and use ResNet-18 as the visual encoder to extract an embedding of dimension 256 for each view. These representations are concatenated as the VAB embedding g with a dimension of 1024.

7.2. Reference Location Sampler

To determine reference locations within a scene, we first calculate the number of reference points needed by dividing the walkable region's range by a standard distance of 8 meters. With this allocated budget, we then sample locations by clustering all potential walkable reference points.

For each location, we extract the visual representation g using the pretrained VAB visual encoder. We expand the 3-dimensional location to match the 1024-dimensional g using sinusoidal encoding and concatenate these representations. We then use K-means clustering to group the candidate locations based on the combined embeddings.

Due to the complexity of Matterport3D scenes with multiple floors, we cluster locations floor by floor. We group walkable locations by height, rounding to the nearest meter. After removing groups with fewer than three locations, we al-

locate the budget proportionally based on each group’s size. This ensures at least one location per group, with groups arranged from smallest to largest to maintain strict budget control.

After combining all clustering results, we select the walkable location nearest to each floor group’s cluster center as the sampled reference location.

7.3. Reference Integration Transformer

We deploy a three-layer cross-attention Transformer for reference integration, which features four heads and a dropout ratio of 0.1. The model has a dimension C of 256 and a feedforward hidden dimension of 512. We use a latent query embedding e with a dimension of 128. This is concatenated with the projected VAB embedding, which also has a size of 128, to form the queries. The relative vector is encoded using positional encoder with sine-cosine functions, utilizing a frequency number of 10, and is projected to a vector with a dimension of 128.

7.4. Reweighting

The dimensions of both the local and global conditions are 256. Specifically, for the local condition, we use sine-cosine functions to embed the rotation quaternion, similar to the approach used in positional encoding.

7.5. Spatial Audio Renderer

We utilize the Short-Time Fourier Transform (STFT) to convert waveform audio into the time-frequency domain. The FFT size, window length, and hop length are set to 510, 510, and 128, respectively, and a Hanning window is applied. We chunk the input waveform into segments of length 32641 to form a spectrogram of size 256×256 . The renderer consists of a U-Net structure with six downsampling layers and six upsampling layers. The conditions are multiplied to combine with the audio content within the condition layers.

7.6. Loss and Training

To balance the loss values, we assign coefficient weights to each of the three loss components: Waveform Loss, Binaural Interaural Level Difference (ILD) Loss, and Multi-resolution Spectrogram Magnitude Loss, with weights of 20, 0.025, and 1.0, respectively. We employ the Adam optimizer for optimization, using an exponentially decaying learning rate starting from 1×10^{-4} over 60 epochs. The batch size is set to 16 for the Soundspace-Ambient benchmark and 24 for the N2S benchmark. Each batch consists of various training samples from the same scene to optimize memory usage when calculating reference VAB embeddings for reference integration.

Method	STFT ↓	MAG ↓	ENV ↓	LRE ↓
w/ VAB	2.442	0.289	0.130	1.390
w/o VAB	2.580	0.295	0.134	1.403

Table 5. Ablations for VAB in Reference Integration Transformer.

8. VAB for Reference Integration

In this section, we study the effectiveness of using VAB embeddings for the Reference Integration Transformer. We implement a variant that excludes the VAB embeddings from the transformer (*w/o VAB*) to compare with *SoundVista* with VAB in the transformer (*w/ VAB*). We report the ablations results on the Soundspace-Ambient benchmark in Table 5 and visualize examples of the reference contribution weights in Figure 7. Compared with *w/o VAB*, *w/ VAB* effectively incorporates visual cues to make the contribution weights more reasonable.

9. Extrapolation Performance Analysis

In our work, the reference microphones are sparsely placed (over 5 meters apart), the edge regions of the rooms typically fall outside the convex hull formed by these microphones. Due to limited in-room data, we cannot track poses or GT sound far beyond the room to evaluate the extrapolation performance. In Figure 3, we show the loudness heatmaps for two scenes; while the errors are larger in the edge regions, the results remain reliable. Furthermore, Table 1 demonstrates that using the top selected reference microphone achieves accuracy comparable to using multiple microphones. These findings show SoundVista’s ability to extend beyond simple interpolation.

10. Acoustic Parameter Learning

We train the acoustic parameter (\mathbf{RT}_{60}) learning model on walkable locations from 39 “seen scenes” of Matterport3D in the Soundspace-Ambient benchmark. An MLP is employed to predict the \mathbf{RT}_{60} value from the VAB embedding g , using L1 loss for supervision. For testing on unseen scenes, we directly use the pretrained visual encoder without fine-tuning for the Novel-View Ambient Sound Synthesis task.

For the *w/ finetune* setting, we aim to study how our acoustic parameter predictor adapts to novel scenes through few-shot learning by finetuning the pretrained prediction model on each of the 23 unseen scenes. Specifically, we uniformly sample the reference locations given the reference budget, maintaining the same average distance as our reference location sampler, but using uniform sampling only. We obtain the \mathbf{RT}_{60} value as ground truth to supervise the prediction at these locations, which constitutes few-shot fine-tuning on sparsely sampled references. After training per scene, we test the prediction on all walkable locations for each scene

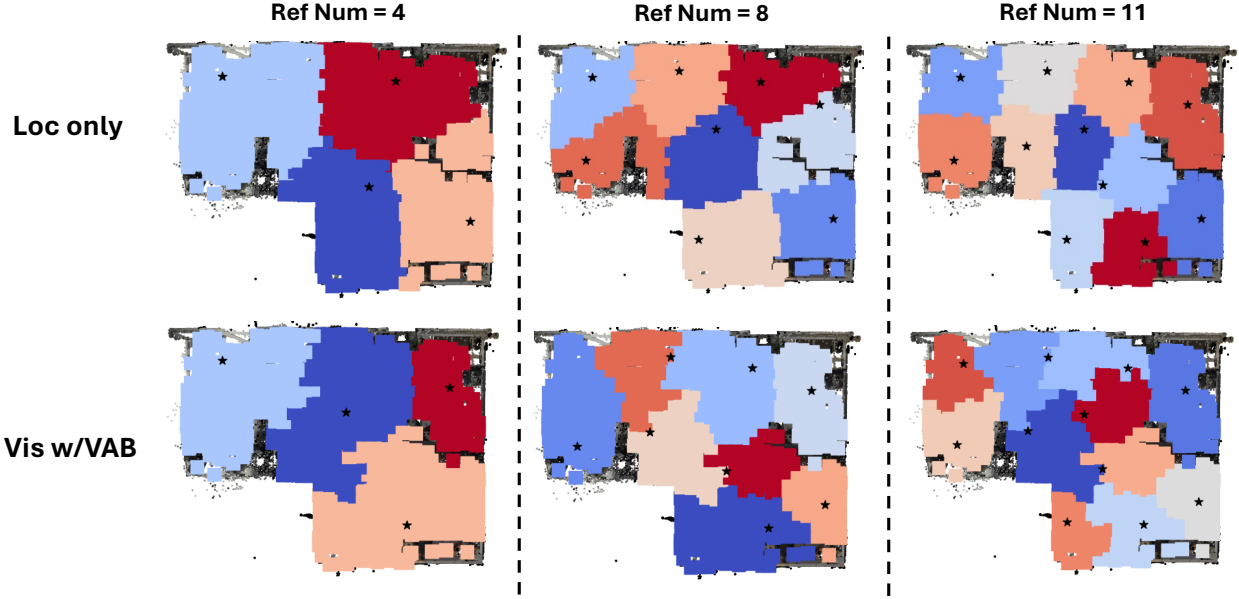


Figure 6. **Visualization of Clustering Results on N2S.** Colorized regions are different clusters. The reference location out of existed 11 references that is closest to each cluster center is marked as black star. Our sampler, *Vis w/VAB*, consistently groups locations that are free from obstacles more effectively, demonstrating reliability of VAB from simulated to real scenarios.

separately and average the RT_{60} prediction metrics to report accuracy for *w/ finetune*.

Figure 8 and Figure 9 illustrate examples of groundtruth and RT_{60} predictions for both seen and unseen scenes, respectively. The groundtruth RT_{60} map shows that RT_{60} values tend to be consistent within a room and are higher in larger spaces without many obstacles, such as open rooms or hallways. This is because sound takes longer to decay in these areas due to fewer reflections or diffusion on surfaces. The RT_{60} map is typically discontinuous in regions blocked by obstacles like walls or closed doors.

In scenes seen during training, our predictions closely match the ground truth. For unseen scenes, while the predicted values may deviate in some regions, they can still effectively distinguish different RT_{60} areas, accounting for walls and other obstacles that block sound propagation. By applying few-shot finetuning (*w/ finetune*) to correct deviated values, our prediction accuracy can improve significantly.

11. More Visualization Analysis

In this section, we present additional visualizations of our clustering results using VAB.

11.1. Sim2Real Clustering on N2S

To evaluate the simulate-to-real (sim2real) capability of VAB, which is trained on simulated data from SoundSpace, we deploy the pretrained visual encoder in a real N2S room. We cluster the walkable locations using the Reference Sampler

(see Section 7.2) to obtain clusters.

In Figure 6, we visualize the clusters with different reference numbers (*Ref Num* = 4, 8, and 11), coloring each cluster differently. We compare two samplers: *Loc only* and our sampler, *Vis w/VAB*. Since the 11 reference locations are already fixed in the real room, we mark the existing reference location closest to the cluster center with black stars, rather than selecting the walkable location nearest to the center.

Figure 6 shows that *Loc only* is more likely to incorrectly cluster locations with obstacles in between, especially with fewer reference numbers (4 and 8 compared to 11), making it less effective at identifying obstacles. In contrast, our sampler, *Vis w/VAB*, consistently groups locations that are free from obstacles more effectively, even without any training or supervision in the real scene. This demonstrates the reliability of adapting VAB from simulated to real scenarios.

11.2. Clustering via VAB

We show more visualization examples of clustering results via VAB in Figure 8 and Figure 9, covering both seen and unseen scenes, respectively. In both figures, the last two columns display scene clusters in different colors. Our sampler, *Vis w/VAB*, produces cluster segment maps that closely align with RT_{60} segments, which effectively highlights obstacles affecting sound propagation. *SoundVista* achieves this by binding visual and acoustic representation through the VAB module, enabling *Vis w/VAB* to identify acoustic regions and key obstacles more effectively than *Loc only*, resulting in more reliable clustering outcomes.

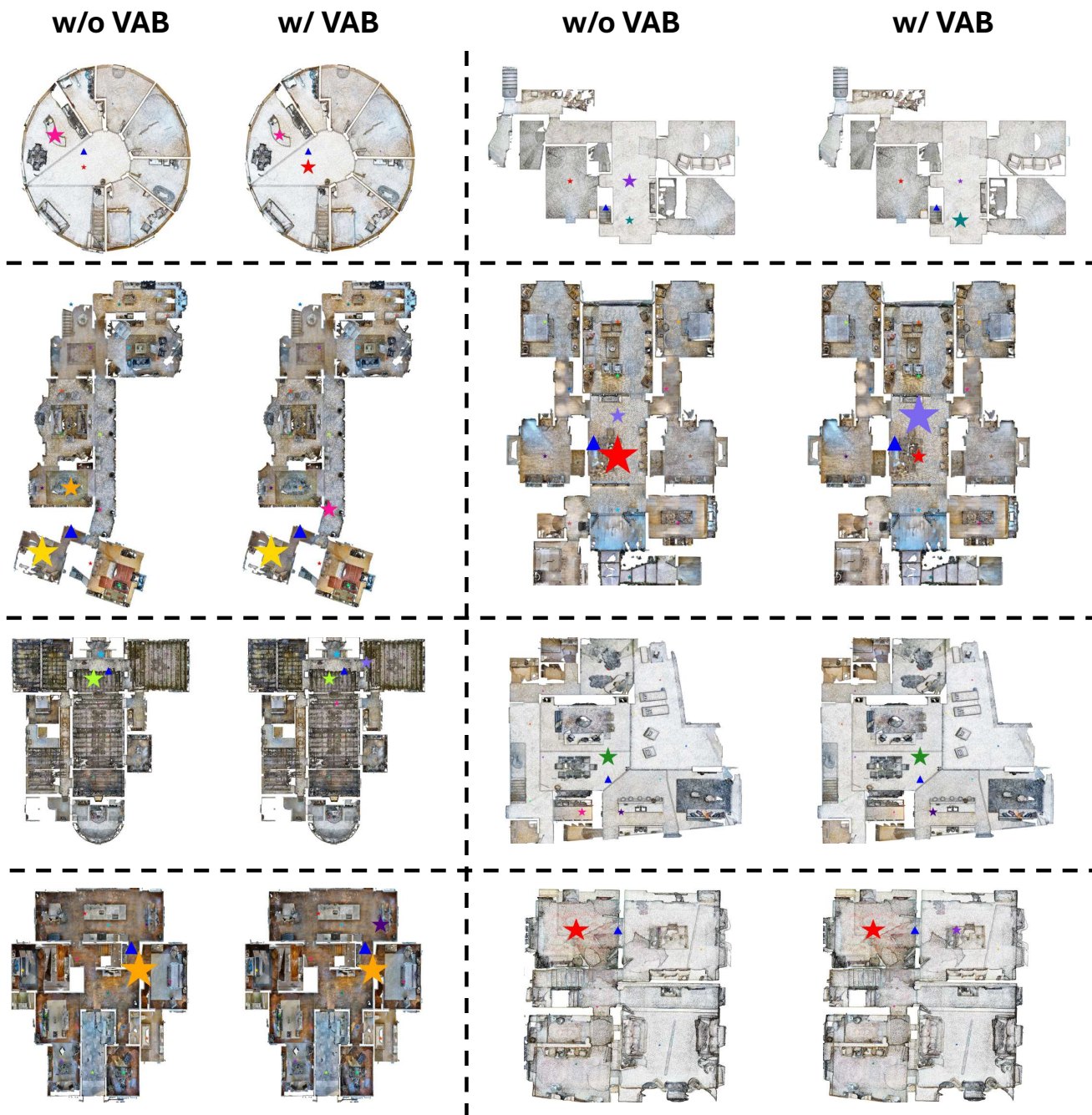


Figure 7. **Visualization of Reference Contribution Weights.** Colored stars (size proportional to weights) indicate the references and the blue triangle for the target. *w/ VAB* effectively incorporates visual cues to make the contribution weights more reasonable.

12. More Details for N2S Real Dataset

We intentionally partitioned a real room space to create distinct acoustic zones for our N2S benchmark (Section 4.1). A sound-absorbing divider separates the larger room, while the smaller concrete-walled room is more reverberant than the sound-treated main room. The top view of the geometry of

the room is shown as Figure 6. The dataset includes ambient noise from a refrigerator, coffee machine, air vents, and fans; which are challenging to isolate and measure. These add to significant acoustic complexity, although the dataset includes a single scene.

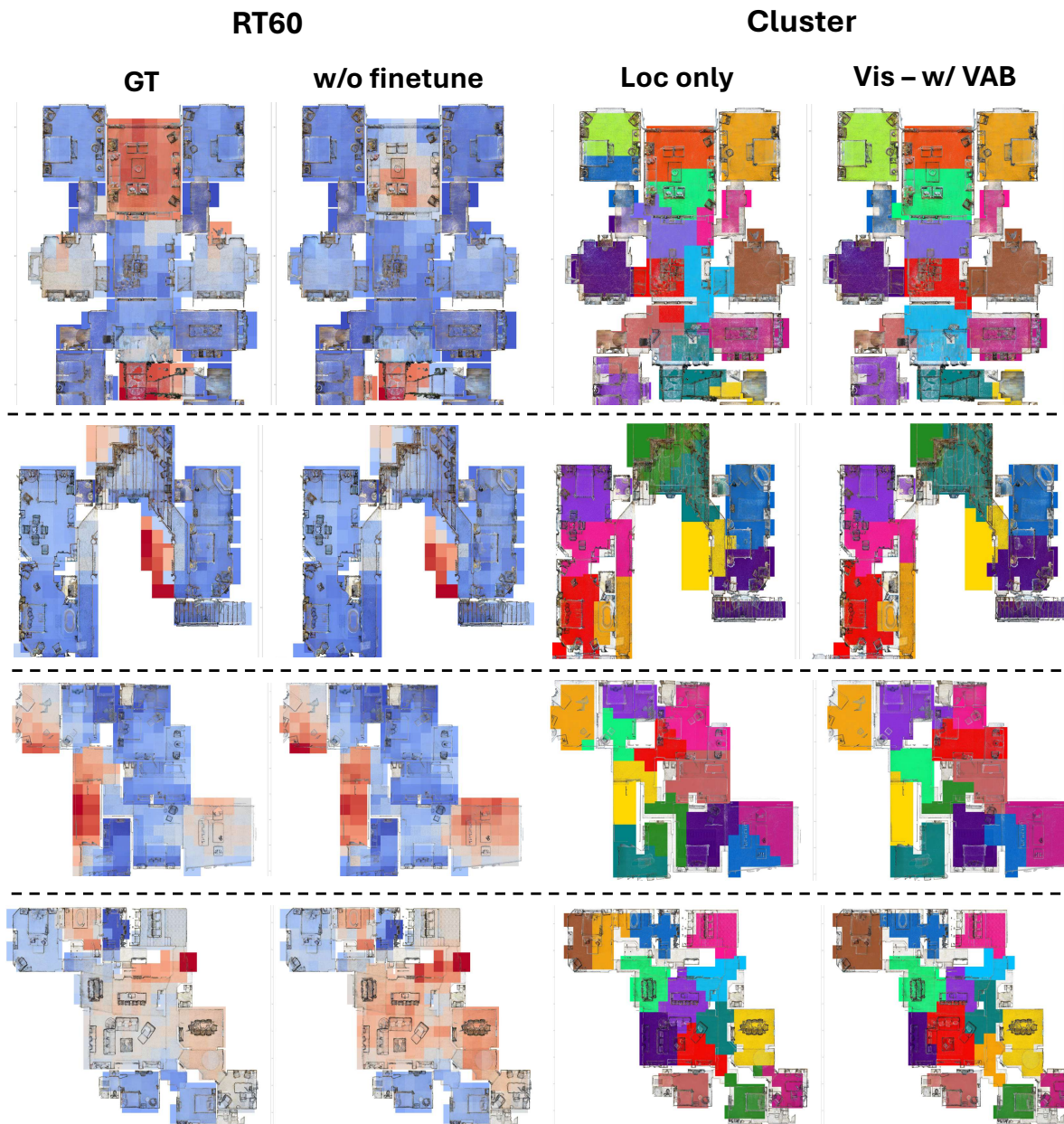


Figure 8. **Seen Scenes from Soundspace-Ambient Matterport3D Benchmark.** First two columns: RT_{60} maps, with warmer colors indicating higher values (longer energy decay). Last two columns: Cluster results comparison, with different colors marking different clusters. Our sampler, *Vis w/VAB*, provides more reliable clusters and the cluster segments better alignment with the RT_{60} map.

13. Limitations

Our method relies on reference recordings, requiring a microphone setup and data collection. These processes can be integrated with existing camera setups for NVS tasks. Additionally, the reliability of our reference sampler may decrease in regions with extremely complex scene layouts. This could be mitigated by incorporating more representative 3D visual descriptions to enhance the VAB module.

14. Broader Impact

Our pipeline can produce audio recordings that mimic real recordings from a specific room. However, this capability can lead to the creation of deceptive and misleading media. It is worth noting that, our model doesn't generate new content; instead, it primarily adapts the pre-recorded audio to sound as if it were captured from the target positions.

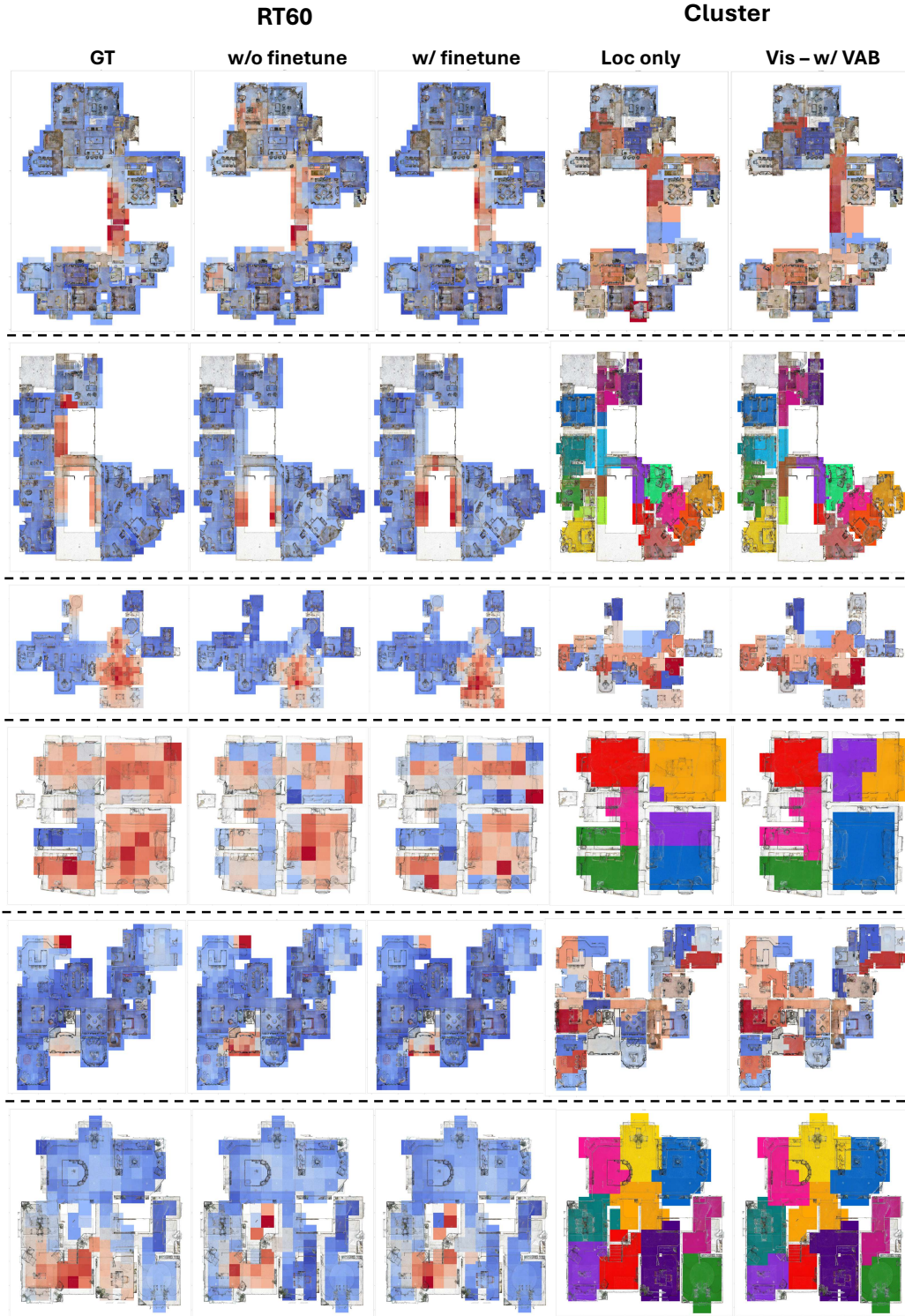


Figure 9. **Unseen Scenes from Soundspace-Ambient Matterport3D Benchmark.** First three columns: RT_{60} maps, warmer colors indicate higher values. *w/ finetune* enhances RT_{60} prediction with few-shot finetuning. Last two columns: Cluster results comparison, with colors marking clusters. Our sampler, *Vis w/VAB*, provides more reliable clusters and the cluster segments better align with the RT_{60} map.