

# 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  $256 \times 256$ . 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  $1 \times 10^{-4}$  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	<b>2.442</b>	<b>0.289</b>	<b>0.130</b>	<b>1.390</b>
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

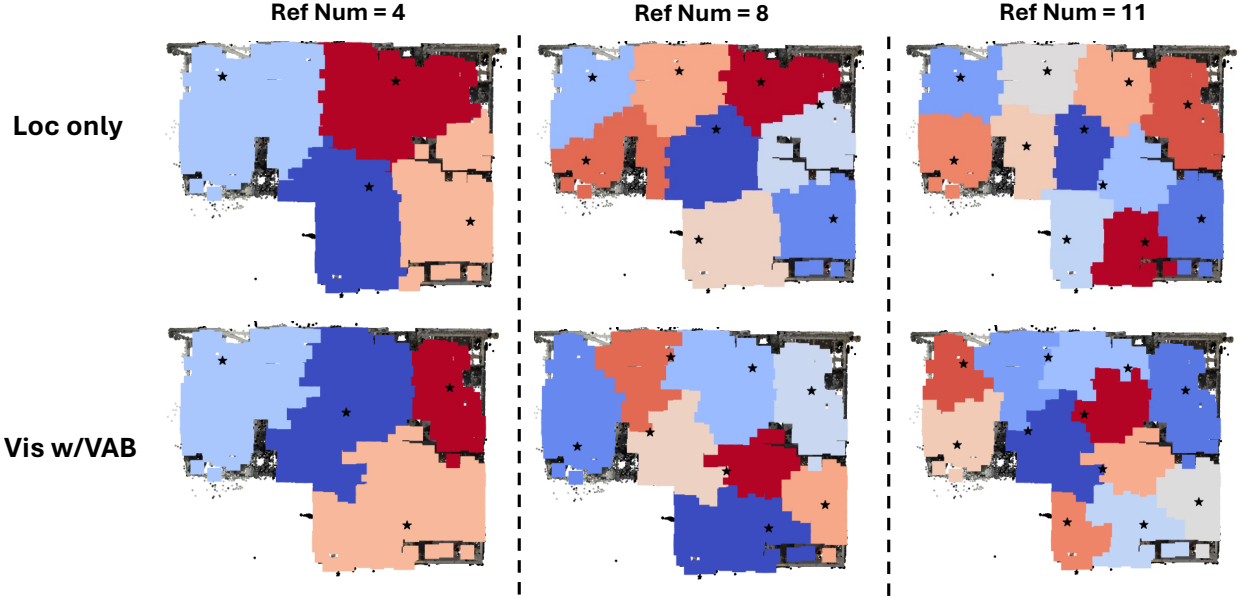


Figure 6. **Visualization of Clustering Results on N2S.** Colorized regions are different clusters. The reference location out of existed 11 references that is closest to each cluster center is marked as black star. Our sampler, *Vis w/VAB*, consistently groups locations that are free from obstacles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



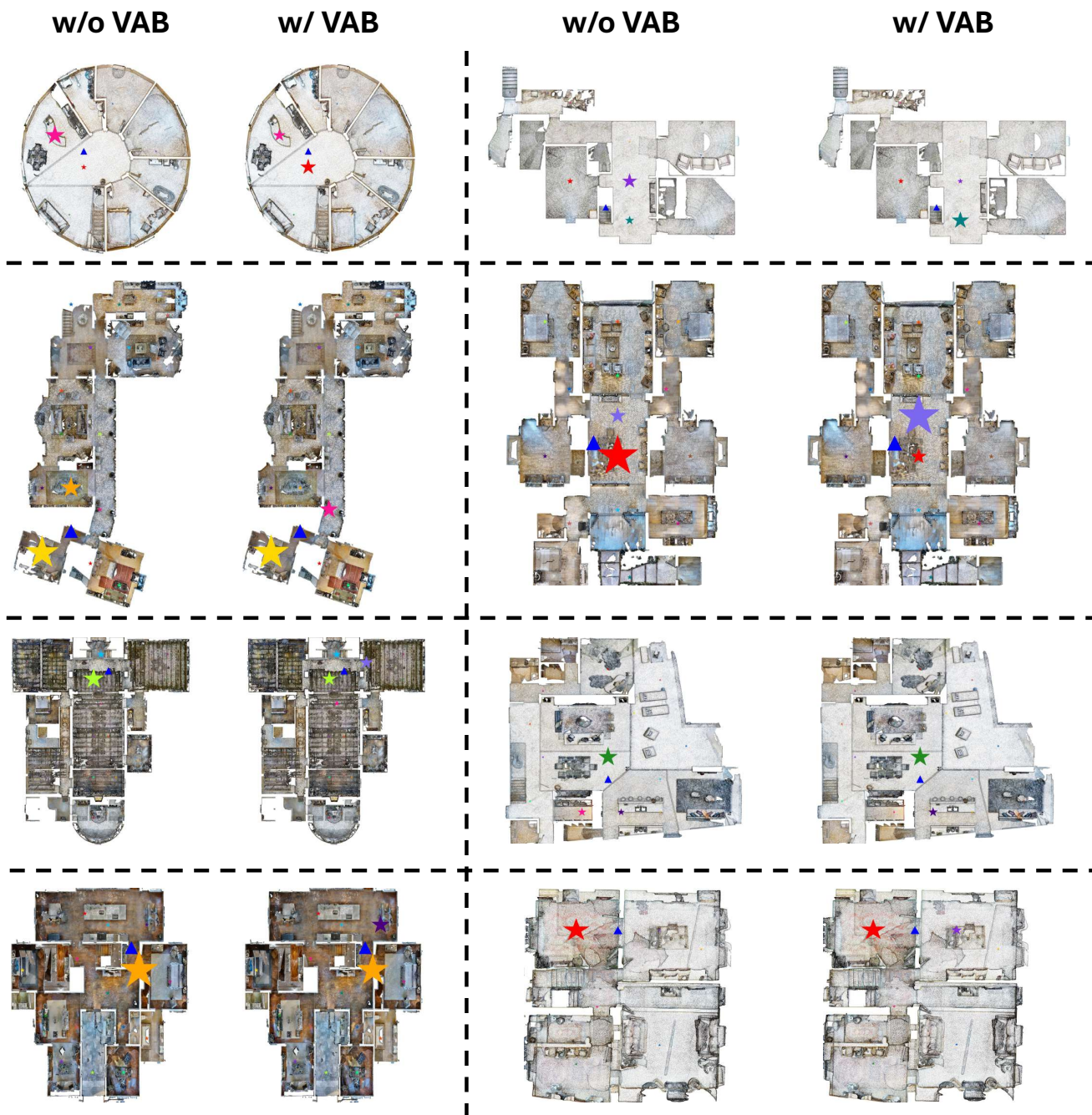


Figure 7. **Visualization of Reference Contribution Weights.** Colored stars (size proportional to weights) indicate the references and the blue triangle for the target. *w/ VAB* effectively incorporates visual cues to make the contribution weights more reasonable.

## 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.

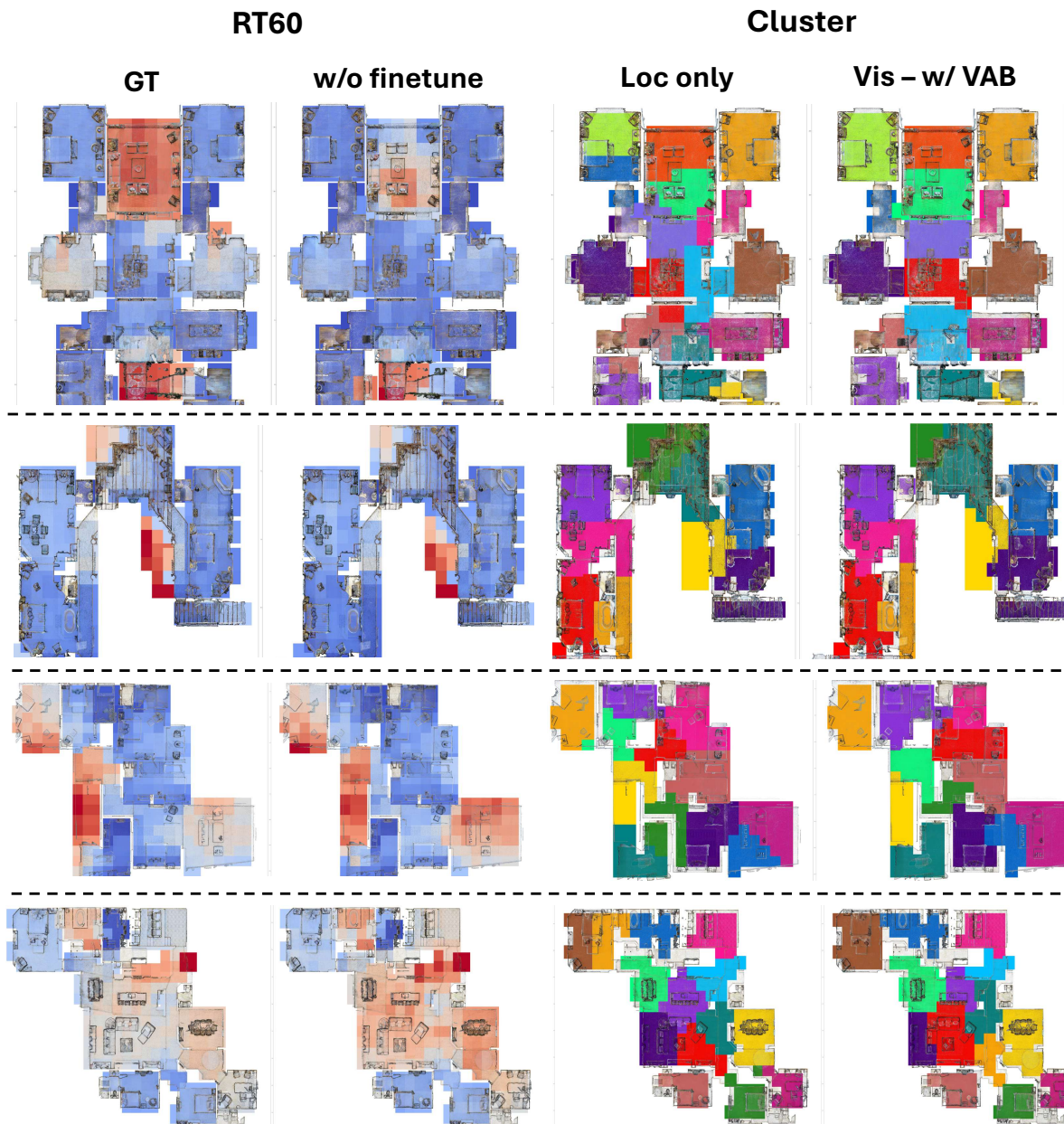


Figure 8. **Seen Scenes from Soundspace-Ambient Matterport3D Benchmark.** First two columns:  $RT_{60}$  maps, with warmer colors indicating higher values (longer energy decay). Last two columns: Cluster results comparison, with different colors marking different clusters. Our sampler, *Vis w/VAB*, provides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



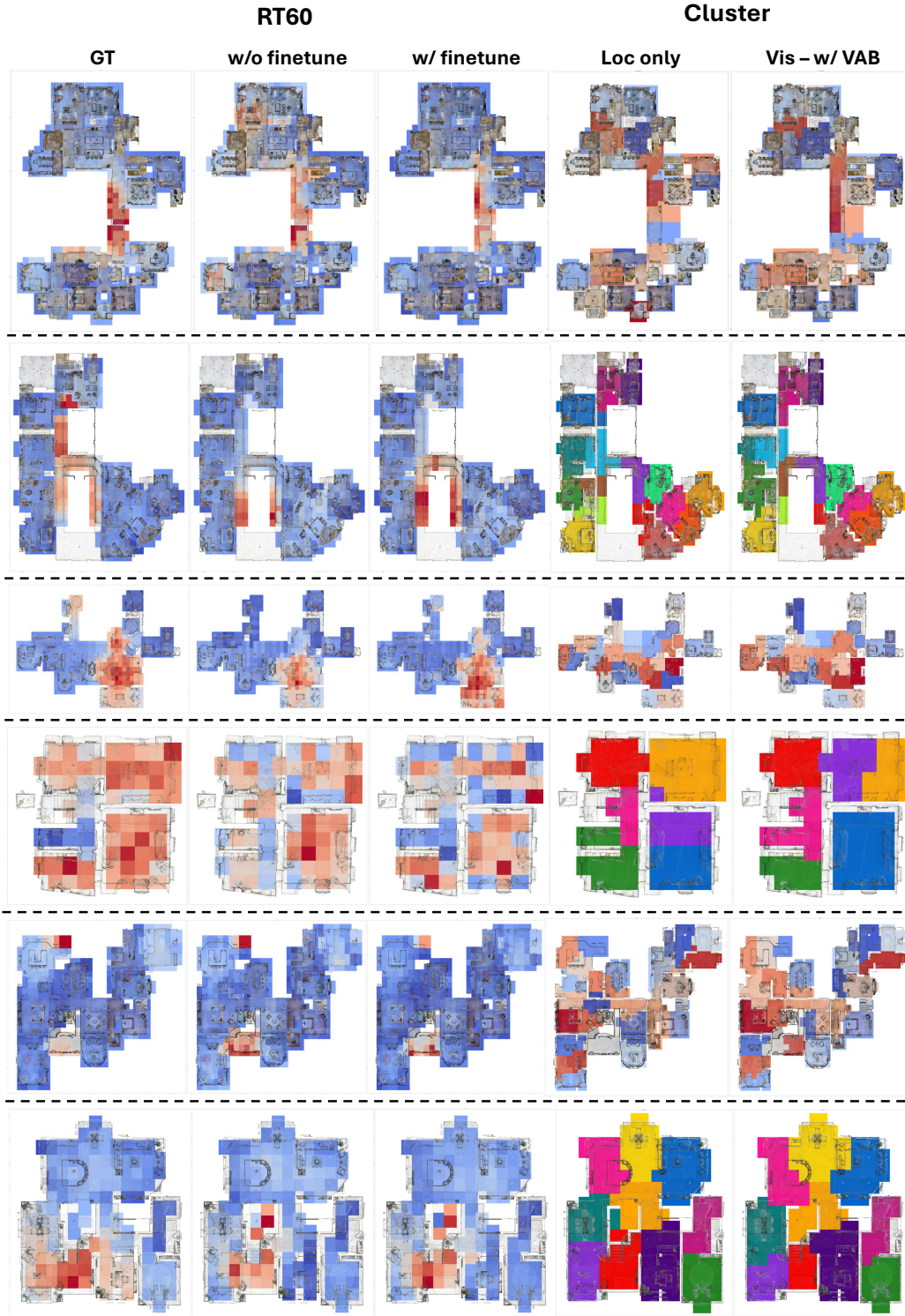


Figure 9. **Unseen Scenes from Soundspace-Ambient Matterport3D Benchmark.** First three columns:  $RT_{60}$  maps, warmer colors indicate higher values. *w/ finetune* enhances  $RT_{60}$  prediction with few-shot finetuning. Last two columns: Cluster results comparison, with colors marking clusters.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