

SONOWORLD: From One Image to a 3D Audio-Visual Scene

Supplementary Material

Contents

1. Supplementary Video	1
2. Ablation Studies	1
3. More Qualitative Results	2
3.1. Synthetic and Real-World Scenes	2
3.2. 3D Scene Backbones	2
3.3. Panoramic Instance Merging	3
3.4. Camera and Microphone Calibration	4
4. More Details For SONOSCENE360 Dataset	4
4.1. Hardware Setup and Microphone Calibration	5
4.2. Dataset Statistics	6
5. Metric Definitions	6
5.1. Spatial Metrics	6
5.2. Semantic Metrics	7
6. More Details For SONOWORLD Framework	7
6.1. VLM Prompt	7
6.2. Gaussian-Pyramid-Based Warping	8
6.3. From Directional Field to Binaural Audio . .	9
7. Setup for One-Shot Room Acoustic Learning	10
8. Setup for Audio-Visual Spatial Source Separation	11
9. User Study Details	12
10. Discussions	12

1. Supplementary Video

The supplementary video provides a visual and auditory overview of SONOWORLD and its applications. It is organized into the following parts:

1. **Task and pipeline recap.** We first briefly recap the problem setting of generating a 3D audio-visual scene from a single image, and summarize our overall pipeline, including visual scene generation, panorama grounding, spatial audio encoding, and free-view rendering.

2. **Interactive web demo.** We then show the layout of our interactive web demo, where users can freely move the listener in 3D and rotate the head while listening to spatialized audio in real time. This section includes five screen recordings of live interactions to highlight responsiveness and stability.
3. **Baseline qualitative comparison.** Next, we present qualitative comparisons with OmniAudio [15] and MMAudio [2] on the same scenes. For each scene, we fix the camera trajectory and compare how different methods generate spatial audio.
4. **Long-trajectory visualizations.** We show longer camera trajectories rendered with our method with rendered binaural audio together with visualizations of the FOA direction of arrival (DoA) as the listener moves through the 3D scene.
5. **Extension: one-shot room acoustic learning.** We demonstrate the one-shot room acoustic learning setup, where our differentiable renderer is fit to a single source-listener FOA recording. The video shows how the learned room response generalizes to new listener positions and to new source audio played through the same scene.
6. **Extension: audio-visual spatial source separation.** Finally, we showcase audio-visual spatial source separation on a YouTube 360° video with FOA audio. Given the visual layout and the recorded spatial mixture, our method separates the spatial audio into monaural tracks, for example isolating a violin and a beatboxer from the same 360° performance.

2. Ablation Studies

Table 1 evaluates the contribution of key components in our framework (*Ours (Proprietary)* in main). First, modeling sources as finite regions rather than point emitters is important for spatial accuracy: **Ours (Point)** degrades substantially on all spatial metrics, indicating that extended source support is necessary to capture realistic directional spread. Second, the equalization module improves both spatial consistency and semantic alignment: removing it (**Ours (w/o EQ)**) yields noticeably worse Δ_{Angular} , CC,

Method	Spatial Metrics					Semantic Metrics		
	$\Delta_{\text{abs}}\theta \downarrow$	$\Delta_{\text{abs}}\varphi \downarrow$	$\Delta_{\text{Angular}} \downarrow$	CC \uparrow	AUC \uparrow	D-CLAP _R \uparrow	D-CLAP _A \uparrow	D-CLAP _T \uparrow
Ours (Point)	1.264	0.614	1.203	0.273	0.629	38.2%	0.521	0.419
Ours (w/o EQ)	0.815	0.196	0.814	0.602	0.806	58.8%	0.467	0.443
Ours (No merge)	0.800	0.359	0.843	0.586	0.807	39.7%	0.271	0.324
Ours (All merge)	1.032	0.331	1.012	0.445	0.719	45.6%	0.461	0.329
Ours (Boundary perturb)	0.760	0.228	0.791	0.627	0.824	69.1%	0.483	0.456
Ours (Depth perturb)	0.732	0.248	0.781	0.628	0.821	64.7%	0.484	0.450
Ours (Full)	0.672	0.216	0.728	0.658	0.838	67.6%	0.480	0.457

Table 1. **Ablation studies on SONOSCENE360.** We analyze the effect of source representation, equalization, grounding/merging strategy, and robustness to imperfect visual inputs. **Ours (Point)** replaces region-based sources with point sources. **Ours (w/o EQ)** removes the equalization module. **Ours (No merge)** treats tile-level masks independently, while **Ours (All merge)** merges all masks from the same category. **Ours (Boundary perturb)** and **Ours (Depth perturb)** evaluate robustness to noisy mask boundaries and depth estimates, respectively. Results show that each component contributes to the final performance, with our full model achieving the strongest overall spatial and semantic quality.

AUC, and D-CLAP_T, showing that explicit per-source amplitude correction is important for matching the rendered scene geometry with the generated audio content. Third, our SAM2-based mask voting strategy is critical for robust visual grounding. Both **Ours (No merge)**, which treats tile masks independently, and **Ours (All merge)**, which merges all masks of the same category, underperform the full system by a clear margin, confirming that our voting-and-merging design yields more reliable source extents and locations (See Figure 5 for qualitative comparisons). Finally, we test robustness to imperfect geometry and segmentation by perturbing mask boundaries and depth maps. Both **Ours (Boundary perturb)** and **Ours (Depth perturb)** incur only minor performance drops relative to the full model, indicating that our method is stable under realistic noise in visual grounding and 3D reconstruction. Overall, the full system achieves the best balance of spatial and semantic quality.

3. More Qualitative Results

3.1. Synthetic and Real-World Scenes

In Fig. 2, we visualize SONOWORLD on two synthetic scenes generated by a text-to-image diffusion model. From a single input image (top row), our method reconstructs a 3D Gaussian scene and predicts a spatial audio field that can be queried along arbitrary camera trajectories. We show several rendered views along the trajectory together with the corresponding first order ambisonic (FOA) and second order ambisonic (SOA) spherical energy maps. In the *Garden* scene, the energy clearly concentrates around the waterfall and two streams as the camera moves, while in the *Riverside Market* scene the energy follows the visually salient market area, illustrating that our model can localize multiple extended sources in synthetic environments.

Figure 3 provides additional qualitative comparisons on



Figure 1. The sound energy by second order ambisonics (SOA) in real scenes.

real scenes from SONOSCENE360. For each environment, we show a rendered panorama of the reconstructed 3D scene, the reference spherical energy map computed from the recorded FOA, and the energy maps from our method and three adapted baselines (OmniAudio [15], SEE-2-SOUND (S2S) [4], and ViSAGE [12]). Our predictions most closely match the reference both in the dominant direction of arrival and in the spread of energy around extended sources (e.g., water in *Fountain* and *Stream*, room ambience in *Kitchen*), while some baselines either over-smooth the field or collapse it to overly concentrated blobs. Please refer to the *supplementary video* for the accompanying audio. In Fig. 1, we show the in real scenes, SOA also yields sharper, more concentrated energy maps around the sounding objects.

3.2. 3D Scene Backbones

We next analyze the impact of different 3D scene backbones on the visual quality of the reconstructed environments. Figure 4 compares HunyuanWorld 1.0 [10] (with mesh output) and Marble [13] (with mesh or 3DGS, the renderings are from 3DGS) when conditioned on the same input panorama. For each real scene, we show the input view (left) and one novel view generated by HunyuanWorld and Marble.

HunyuanWorld often produces visually rich global structure but can introduce noticeable distortions and over-smoothing in nearby geometry, which is undesirable for precise audio anchoring. Marble, in contrast, yields sharper details and more faithful local geometry with fewer distortions around important objects (e.g., kitchen counters, river

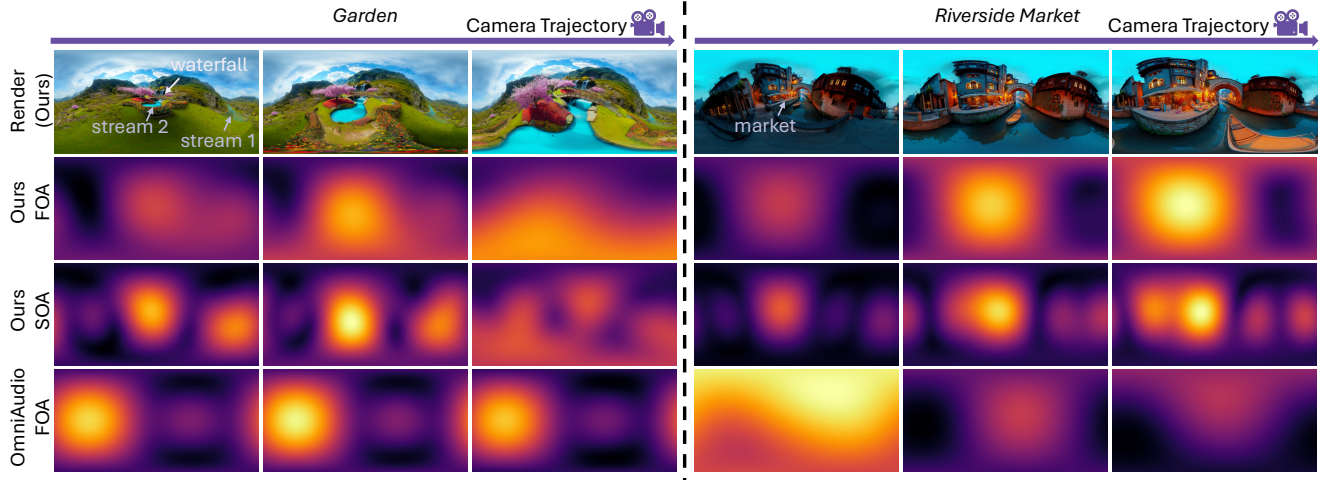


Figure 2. **Qualitative results on synthetic scenes.** From a single diffusion-generated image, SONOWORLD reconstructs a 3D Gaussian scene and predicts a spatial audio field that supports free-viewpoint exploration. We show two synthetic scenes (*Garden* and *Riverside Market*), sample views along a camera trajectory (top row), and the corresponding FOA / SOA spherical energy maps for our method and OmniAudio. The energy smoothly tracks visually grounded sources such as the waterfall, streams, and market stalls.

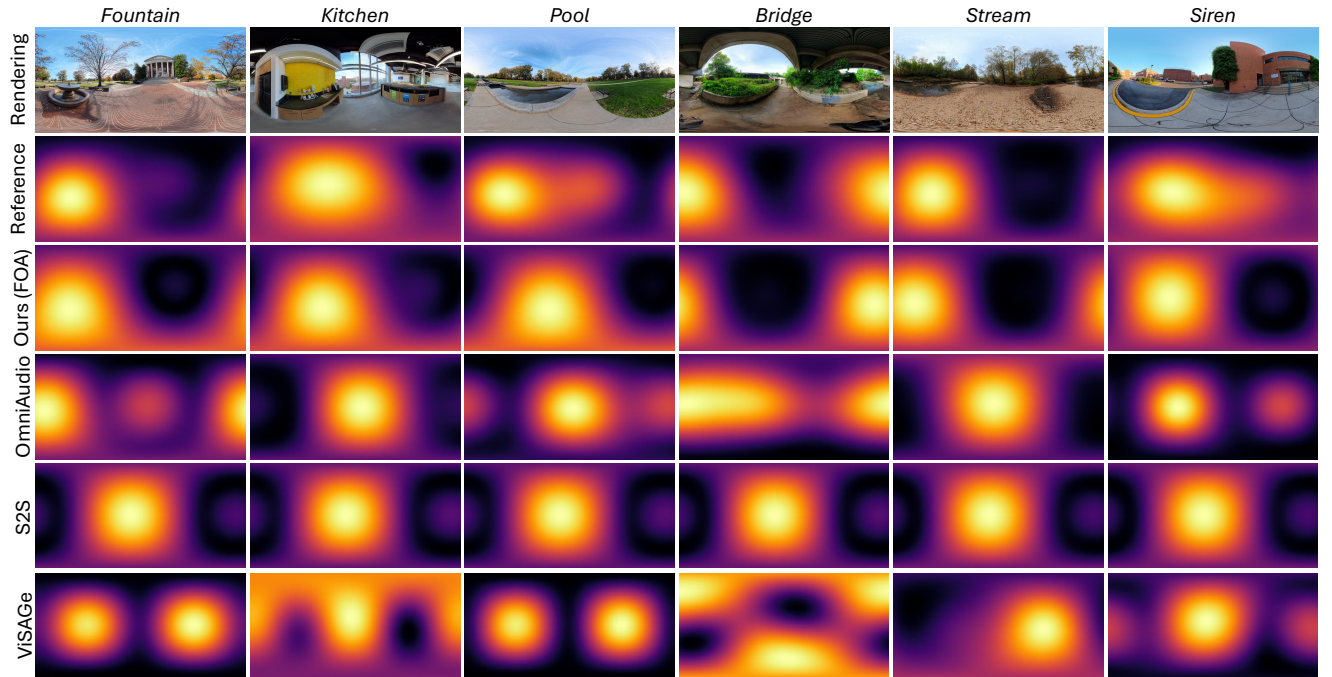


Figure 3. **Qualitative comparison on real scenes.** Additional qualitative results on six real environments from SONOSCENE360. For each scene, we show the rendered panorama from our 3D Gaussian reconstruction, the reference spherical energy map derived from the recorded FOA (second row), and the predicted energy maps from our method (FOA), OmniAudio, SEE-2-SOUND (S2S), and ViSAGE. Our method produces spatial patterns that best align with the reference both in azimuth and elevation and in the spatial extent of the energy, especially around extended water and ambient sources. Please refer to the *supplementary video* for audio.

banks, and buildings), while still enabling efficient real-time rendering. These observations support our choice of Marble as the default backbone in the main experiments.

3.3. Panoramic Instance Merging

Our panoramic grounding pipeline combines class-agnostic proposals (from SAM2-style segmentation) with open-



Figure 4. **Comparison of 3D scene backbones.** Given an input 360° panorama (left column), we compare novel views generated by HunyuanWorld 1.0 and Marble for several real scenes (*Kitchen*, *Stream*, *Siren*). HunyuanWorld produces globally coherent but sometimes distorted geometry (e.g., warped floors and façades), whereas Marble yields sharper, more faithful reconstructions that preserve local spatial layout, which is crucial for accurate spatial audio anchoring and free-viewpoint navigation.

vocabulary semantic masks to derive sound source instances. A key design choice is how to merge or split the underlying regions into semantic instances. Figure 5 visualizes this ablation on three scenes (*Siren*, *Pool*, *Train*).

The *All-Merge* variant merges all proposals belonging to the same category into a single instance, which oversmooths extended structures and loses important spatial variation (e.g., the long hedge in *Siren* and the water surface in *Pool*). The *No-Merge* variant treats each proposal as a separate instance, often leading to excessive fragmentation (dozens of instances for a single physical object), which complicates downstream spatial audio allocation. Our voting-based strategy (*Vote (Ours)*) aggregates proposals using semantic agreement while retaining a small number of coherent instances that better match human perception of sound-emitting regions.

Moreover, *Vote (Ours)* is more robust to failures introduced by splitting views. In *Siren* and *Pool*, certain regions are poorly captured in individual perspective tiles, e.g. the tops of the bushes or the middle of the pool. These missing areas create artifacts for both *All-Merge* and *No-Merge*, while our voting scheme recovers them by relying on SAM2’s global panoramic proposals, which better preserve the overall scene structure. Finally, our merging strategy is agnostic to the particular 360° panorama grounding model used, and can readily benefit from future improvements in panoramic segmentation and grounding.

3.4. Camera and Microphone Calibration

Finally, we provide additional visualizations of the camera–microphone calibration process for SONOSCENE360. As discussed in the main paper, accurate alignment between the FOA microphone and the 360° camera is critical for learning reliable audio-visual correspondences.

Figure 6 shows our annotation interface. In the left panel, annotators click on the microphone in the panorama to specify its azimuth and elevation relative to the camera. In the middle panel, they annotate the marker on the ground below the microphone. In the right panel, we visualize the estimated microphone rotation relative to the marker board. These annotations are combined with the AprilTag detections to obtain an initial estimate of the microphone pose.

Figure 7 illustrates how the calibration affects the rendered views. For each scene, we show (i) the raw camera view with the microphone visible, (ii) an “enhanced” camera view where the microphone is removed via inpainting, (iii) the rendering of the reconstructed 3D scene when the FOA reference frame is aligned using only the AprilTag pose, and (iv) the rendering after our refinement. The refined poses yield better alignment between the camera and the inferred world frame, which provides reliable foundation for evaluation on SONOSCENE360.

4. More Details For SONOSCENE360 Dataset

Remind that the input is one RGB image, and the outputs are:

- 3D Visual Scene Representation (3DGS, for ours).

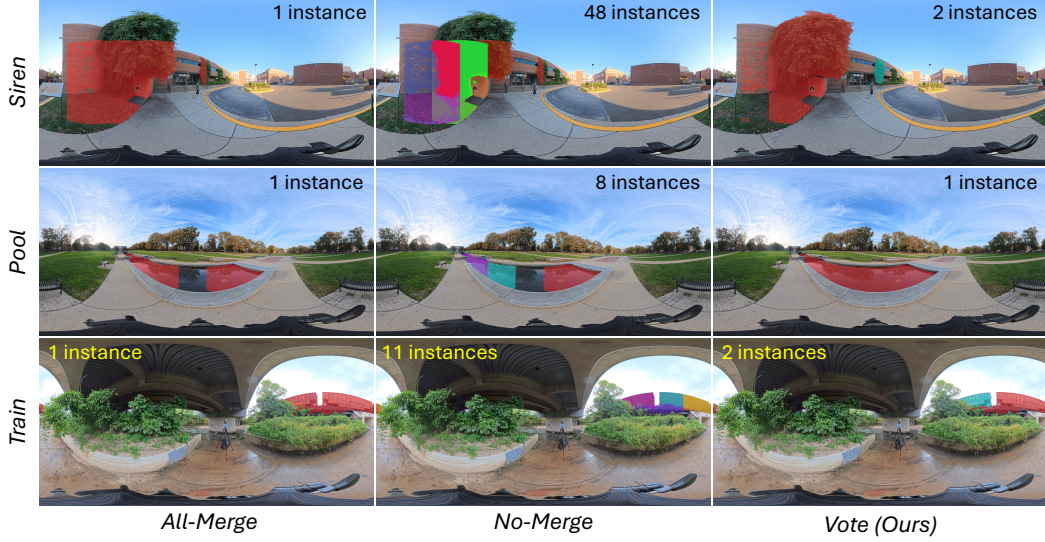


Figure 5. **Panoramic instance merging strategies.** We compare three strategies for grouping class-agnostic region proposals into semantic instances in SONOSCENE360: *All-Merge* (left), *No-Merge* (middle), and our voting-based strategy *Vote (Ours)* (right). For each scene, we overlay the resulting instances on the panorama and indicate their counts. *All-Merge* collapses large structures into a single instance, losing spatial detail; *No-Merge* over-fragments the scene into many small pieces. Our voting scheme strikes a balance, producing a small number of coherent instances that align with visually meaningful sound sources.



Figure 6. **Calibration annotation interface.** Our tool for camera–microphone calibration. Left: annotating the azimuth and elevation of the FOA microphone in the panorama. Middle: annotating azimuth and elevation for ground markers (*i.e.*, the marker right below the microphone position) that define a world reference frame. Right: visualizing the relative rotation between AprilTag marker and camera, which is used to estimate the microphone orientation in the world coordinate system.

- Spatial Audio Field **A** (Ambisonics Field in our case)

4.1. Hardware Setup and Microphone Calibration

We briefly recall the SONOSCENE360 setup and provide details about the calibration procedure (see Fig. 2 in the main paper). An Insta360 X5 camera captures 6K 360° video, and 12K 360° image for microphone calibration, and a RØDE NT-SF1 FOA microphone records ambisonics at 48 kHz. The microphone is mounted with an AprilTag [16] and ensured to be visible by the camera.

To use the FOA channels as a reference sound field aligned with the panorama, we must estimate:

- the rotation between the microphone and the camera coordinate system, and
- the absolute azimuth and elevation of the microphone axes in the world frame.

Initial extrinsic calibration with AprilTag. We attach an AprilTag board [16] to the microphone and capture a 12K image for calibration after the setup in a scene, and split the 360° image into 12 field-of-view (FOV) views. For each view, we:

1. detect the AprilTag pose in the FOV image,
2. reconstruct the board pose in the camera coordinate sys-



Figure 7. **Effect of calibration on rendered scenes.** For two real scenes (*Siren* and *River & Train*), we show the original camera view with the microphone visible, the inpainted camera view, and renderings of the reconstructed 3D scene when aligned using the AprilTag-based pose estimate vs. our refined pose. The refined calibration produces renderings whose layout and perspective are more consistent with the original capture, leading to more accurate alignment between the FOA reference frame and the visual scene.

tem, and

3. choose the one with the smallest reconstruction error as final prediction.

This yields an initial rotation quaternion $q_{\text{cam} \rightarrow \text{mic}}$ and translation offset.

Elevation and azimuth refinement. Due to inconsistent scale between reconstructed scene and real pose, we further refine the orientation using human annotation. For each scene, the annotator:

1. selects a the middle of the microphone to annotate the elevation ($\theta_{\text{mic}}, \varphi_{\text{mic}}$)
2. we stick some small markers on the ground to allow us annotate the point that is under the microphone, this will be use to query the depth and use to calculate the microphone in the generated scene, the reason we need another marker instead of using the microphone depth is that we found that the ground depth is usually more accurate/consistent than the microphone middle point.

Combine with the depth estimated at the ground, we can

4.2. Dataset Statistics

Table 2 reports scene-wise statistics for SONOSCENE360. For each scene, there might be more than one setup (e.g., *Pool-left/Pool-right*) under the scene categories described in the main paper (Fountain, Kitchen, Pool, Bridge, Stream, Siren), but we keep all subsets for evaluation.

5. Metric Definitions

5.1. Spatial Metrics

DoA from FOA intensity. Following [8, 15] for first-order ambisonics (FOA, $L = 1$), we write the four channels as

$$\mathbf{a}_1(t) = [W(t), Y(t), Z(t), X(t)]^\top, \quad (1)$$

Scene (setup)	# Mics	Clips / Mic	Total Clips
Fountain	10	2	20
Kitchen	5	2	10
Pool-left	3	2	6
Pool-right	2	2	4
Bridge-river	1	2	2
Bridge-train	1	2	2
Stream-original	5	2	10
Stream-human	5	2	10
Building-siren	1	1	1
Building-birds	1	3	3
Total	34	—	68

Table 2. **Statistics of the SONOSCENE360 dataset.** Each scene subset corresponds to one environment; we vary microphone layouts within a scene to capture diverse listening positions.

where W is the omnidirectional channel and (X, Y, Z) encode directional components in a right-handed coordinate system. We approximate the intensity vector by correlating W with the directional channels over the clip:

$$I_X = \sum_t W(t) X(t), \quad (2)$$

$$I_Y = \sum_t W(t) Y(t), \quad (3)$$

$$I_Z = \sum_t W(t) Z(t), \quad (4)$$

where t indexes audio samples or short-time frames (we use STFT frames in practice). The intensity vector $\mathbf{I} = [I_X, I_Y, I_Z]^\top$ defines a dominant direction of arrival (DoA) on the unit sphere.

We convert \mathbf{I} to azimuth θ and elevation φ :

$$\theta = \text{atan2}(I_Y, I_X), \quad (5)$$

$$\varphi = \text{atan2}(I_Z, \sqrt{I_X^2 + I_Y^2}), \quad (6)$$

with $\theta \in [-\pi, \pi]$, $\varphi \in [-\frac{\pi}{2}, \frac{\pi}{2}]$.

Given ground-truth and predicted DoAs, $(\theta_{\text{gt}}, \varphi_{\text{gt}})$ and $(\theta_{\text{pred}}, \varphi_{\text{pred}})$, we report:

1. **Azimuth error**

$$\Delta_{\text{abs}}\theta = \min(|\theta_{\text{gt}} - \theta_{\text{pred}}|, 2\pi - |\theta_{\text{gt}} - \theta_{\text{pred}}|). \quad (7)$$

2. **Elevation error**

$$\Delta_{\text{abs}}\varphi = |\varphi_{\text{gt}} - \varphi_{\text{pred}}|. \quad (8)$$

3. **Geodesic angular error.** We convert both DoAs to points on the unit sphere,

$$\mathbf{u}_{\text{gt}} = \begin{bmatrix} \cos \varphi_{\text{gt}} \cos \theta_{\text{gt}} \\ \cos \varphi_{\text{gt}} \sin \theta_{\text{gt}} \\ \sin \varphi_{\text{gt}} \end{bmatrix}, \quad \mathbf{u}_{\text{pred}} = \begin{bmatrix} \cos \varphi_{\text{pred}} \cos \theta_{\text{pred}} \\ \cos \varphi_{\text{pred}} \sin \theta_{\text{pred}} \\ \sin \varphi_{\text{pred}} \end{bmatrix}, \quad (9)$$

and compute the geodesic distance

$$\Delta_{\text{Angular}} = \arccos(\text{clip}(\mathbf{u}_{\text{gt}}^\top \mathbf{u}_{\text{pred}}, -1, 1)). \quad (10)$$

For completeness, we also report the haversine-style formulation used in the main text:

$$a = \sin^2\left(\frac{\Delta_{\text{abs}}\theta}{2}\right) + \cos \varphi_{\text{pred}} \cos \varphi_{\text{gt}} \sin^2\left(\frac{\Delta_{\text{abs}}\varphi}{2}\right), \quad (11)$$

$$\Delta_{\text{Angular}} = 2 \arctan\left(\sqrt{\frac{a}{1-a}}\right). \quad (12)$$

Spherical energy maps, CC and AUC. Following [12], given ambisonics coefficients $a_L(t)$, we render the scalar pressure at a direction (θ, φ) as

$$a(\theta, \varphi, t) = y_L(\theta, \varphi)^\top a_L(t). \quad (13)$$

We then compute a time-aggregated energy map

$$E(\theta, \varphi) = \sum_t |a(\theta, \varphi, t)|^2, \quad (14)$$

discretized on a balanced spherical grid $\Omega = \{(\theta_i, \varphi_i)\}_{i=1}^N$ (we use an equiangular grid with uniform weights).

Let $\mathbf{e}_{\text{pred}}, \mathbf{e}_{\text{gt}} \in \mathbb{R}^N$ be the flattened predicted and ground-truth energy maps after min-max normalization to $[0, 1]$. The *correlation coefficient* (CC) is

$$\text{CC} = \frac{\sum_i (\mathbf{e}_{\text{pred},i} - \bar{\mathbf{e}}_{\text{pred}})(\mathbf{e}_{\text{gt},i} - \bar{\mathbf{e}}_{\text{gt}})}{\sqrt{\sum_i (\mathbf{e}_{\text{pred},i} - \bar{\mathbf{e}}_{\text{pred}})^2} \sqrt{\sum_i (\mathbf{e}_{\text{gt},i} - \bar{\mathbf{e}}_{\text{gt}})^2}}. \quad (15)$$

To compute AUC, we treat \mathbf{e}_{gt} as a soft foreground mask by binarizing it at its median value. Using \mathbf{e}_{pred} as scores, we form the ROC curve over all thresholds and report the area under the curve (AUC), following ViSAGE [12]. Higher CC and AUC indicate closer spatial energy patterns to the reference.

5.2. Semantic Metrics

Directional CLAP. We evaluate semantic consistency by probing the ambisonics field along four canonical FOA-aligned directions:

$$\text{left: } \mathbf{u}_L = (\theta = \frac{\pi}{2}, \varphi = 0), \quad (16)$$

$$\text{right: } \mathbf{u}_R = (\theta = -\frac{\pi}{2}, \varphi = 0), \quad (17)$$

$$\text{front: } \mathbf{u}_F = (\theta = 0, \varphi = 0), \quad (18)$$

$$\text{back: } \mathbf{u}_B = (\theta = \pi, \varphi = 0). \quad (19)$$

For each direction \mathbf{u}_d , we render a monaural waveform

$$a_d(t) = y_L(\mathbf{u}_d)^\top a_L(t). \quad (20)$$

Let $f_a(\cdot)$ and $f_t(\cdot)$ be the audio and text encoders of CLAP [24]. For a caption c describing a sounding source and directional audio a_d , we define

$$s_{\text{CLAP-T}}(d, c) = \cos(f_a(a_d^{\text{pred}}), f_t(c)), \quad (21)$$

$$s_{\text{CLAP-A}}(d) = \cos(f_a(a_d^{\text{pred}}), f_a(a_d^{\text{gt}})), \quad (22)$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. We report:

- **D-CLAP_T**: averaged $s_{\text{CLAP-T}}$ between directional audio and its text caption;
- **D-CLAP_A**: averaged $s_{\text{CLAP-A}}$ between predicted and ground-truth directional audio;
- **D-CLAP_R**: for each annotation (d_{gt}, c) , we rank all four directions by $s_{\text{CLAP-T}}(d, c)$ and compute top-1 accuracy that d_{gt} is ranked highest.

All metrics are averaged over clips in SONOSCENE360.

6. More Details For SONOWORLD Framework

6.1. VLM Prompt

We query a vision-language model (VLM) with the input image I to obtain a list of sounding categories, their types (point/cluster/global), and audio prompts for text-to-audio generation. In Figure 8 we provide an example prompt used in practice.

The JSON output is parsed and converted into the category set \mathcal{C} and equalization parameters used by our spatial audio encoder.

The warping operator W_G reprojects a calibrated perspective image into an equirectangular panorama while avoiding aliasing near the poles and at large viewing angles.

TASK
You are given ONE image. Infer plausible sounds from VISIBLE things in the image.

OUTPUT
Return a JSON array with 2{8 items, ordered from loudest (peak_db closest to 0) to softest (most negative).
Output JSON ONLY|no extra text.

FOR EACH ITEM (object)

- "diffusion_prompt": <=10 simple words describing the sound (common words only).
- "grounding_label": 1{2 word VISIBLE object to ground on (e.g., river, tree, door).
 - If the sound comes from a hidden agent (bird, insect, person off-frame), map it to a VISIBLE HOST object (e.g., tree, bush, window, street).
- "peak_db": integer NEGATIVE dB for target PEAK (0 dB = full-scale, where [-1, 1] is 0 dB). Never use values > -6 dB.
- "source_type": Choose from "area", "point", and "background", area means the sound comes from an area, e.g., river, leaves, "point" means the sound comes from a point source, "background" is reserved for global background bed.

GLOBAL BACKGROUND BED (required as LAST item)

- Add ONE final item that captures the scene's background as a reusable, loopable bed:
 - e.g., "open room soft air hum", "damp cave low drip hush", "quiet library room tone", or "silence" if none is implied.
- Use "grounding_label": "global".
- Set "peak_db" soft (\approx -26...-32). If truly silent, use -120.
- Set "source_type" as "background"

SELECTION RULES

- Include only sources that are visible or strongly implied by what is visible (moving water -> water sound; swaying trees -> wind; visible vents -> HVAC).
- No voices/music unless people/speakers are visible.
- Keep words simple; avoid jargon, metaphors, and long hyphen chains.
- Use Ascii.

LEVEL GUIDE (choose by strength)

- Foreground/strong: -8...-14 dB
- Mid/medium: -14...-20 dB
- Far/quiet: -20...-26 dB
- Background bed: -26...-32 dB (or -120 for silence)

CONSTRAINTS

- Keys must be exactly: "diffusion_prompt", "grounding_label", "peak_db", "source_type".
- Use integers for "peak_db".
- "source_type" options: area, point, background
- Do not add other fields.

OUTPUT-TEMPLATE
- Output in JSON format:

```

'''
[
  {"diffusion_prompt": "<prompt>", "grounding_label": "<object>", "peak_db": <int>, "source_type": <area/point>},
  {"diffusion_prompt": "<prompt>", "grounding_label": "<host object>", "peak_db": <int>, "source_type": <area/point>},
  {"diffusion_prompt": "<generative background of the scene>", "grounding_label": "global", "peak_db": <int>}
]
'''

```

Figure 8. VLM prompt used to query the vision-language model.

Mapping from panorama to camera. For each output panorama pixel (u, v) with resolution $W_{\text{pano}} \times H_{\text{pano}}$, we convert to spherical angles

$$\theta = 2\pi \left(\frac{u + 0.5}{W_{\text{pano}}} - \frac{1}{2} \right), \quad (23)$$

$$\varphi = \pi \left(\frac{v + 0.5}{H_{\text{pano}}} - \frac{1}{2} \right), \quad (24)$$

and obtain a direction vector

$$\mathbf{d}(\theta, \varphi) = \begin{bmatrix} \cos \varphi \cos \theta \\ \cos \varphi \sin \theta \\ \sin \varphi \end{bmatrix}. \quad (25)$$

6.2. Gaussian-Pyramid-Based Warping

Given the camera extrinsics from GeoCalib [22], we transform \mathbf{d} into the camera frame and project to normalized image coordinates (x, y) using the calibrated focal length f from Eq. (4) in the main text. These are mapped to pixel coordinates (u', v') in the input image.

Gaussian pyramid and anti-aliased sampling. Fore-shortening near the panorama poles and along the vertical direction leads to non-uniform sampling: some panorama pixels correspond to large footprints in the input image. To mitigate aliasing, we construct a Gaussian pyramid

Algorithm 1 Mask Voting for Category c

```

1: Input:  $\mathbf{M}_{\text{pano}}, \mathbf{M}_{\text{OVS},c}, \tau_{\text{vote}}, \tau_{\text{IOU}}$ 
2: Output:  $\mathcal{M}_c$ 
3:  $\mathbf{M}_c \leftarrow \emptyset$ 
4: for each  $\mathcal{M}_i^{\text{pano}} \in \mathbf{M}_{\text{pano}}$  do
5:    $s_i \leftarrow \frac{\sum_p \text{PIXELSCORE}(p)}{\text{VISIBILITY}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})}$ 
6:   if  $s_i \geq \tau_{\text{vote}}$  then
7:      $\widehat{\mathcal{M}}_i \leftarrow \mathcal{M}_i^{\text{pano}} \vee \text{COMBINE}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})$ 
8:    $\mathbf{M}_c \leftarrow \mathbf{M}_c \cup \{\widehat{\mathcal{M}}_i\}$ 
9: Return  $\mathbf{M}_c$ 

```

$\{I^{(s)}\}_{s=0}^{S-1}$ from the input image I , where $I^{(0)} = I$ and

$$I^{(s+1)} = \text{downsample}_2(\text{GaussianBlur}(I^{(s)})). \quad (26)$$

For each panorama pixel, we approximate a local magnification factor ρ from the Jacobian of the equirectangular-to-camera mapping and choose a pyramid level

$$s^* = \text{clip}(\lfloor \log_2 \rho \rfloor, 0, S-1), \quad (27)$$

where $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer. We then sample $I^{(s^*)}$ at (u', v') using bilinear interpolation to obtain the warped color. This yields

$$I_{\text{warp}} = \mathcal{W}_G(I, \varphi, f), \quad (28)$$

which is used as input to the panorama outpainting model g_{outpaint} .

Voting We address the mismatch between FoV-tile-wise open-vocabulary segmentation (OVS) masks and globally consistent panoramic masks by letting SAM2 [18] proposals and X-Decoder [25] instance masks vote via Alg. 1, preserving SAM2’s global geometry while inheriting X-Decoder’s category-wise semantics.

For each SAM2 proposal $\mathcal{M}_i^{\text{pano}}$ and candidate sounding category c proposed by GPT-5 [17], Alg. 1 computes a vote score s_i by aggregating per-pixel confidences from overlapping open-vocabulary masks $\mathbf{M}_{\text{OVS},c}$. $\text{PIXELSCORE}(p)$ assigns to each pixel p in $\mathcal{M}_i^{\text{pano}}$ the maximum confidence over all instance masks in $\mathbf{M}_{\text{OVS},c}$ whose IoU with $\mathcal{M}_i^{\text{pano}}$ exceeds τ_{IOU} . $\text{VISIBILITY}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})$ counts the pixels in $\mathcal{M}_i^{\text{pano}}$ that receive at least one such vote and normalizes the sum of $\text{PIXELSCORE}(p)$ to obtain s_i . If $s_i \geq \tau_{\text{vote}}$, $\text{COMBINE}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})$ takes the union of all contributing instance masks (those with $\text{IoU} > \tau_{\text{IOU}}$) with $\mathcal{M}_i^{\text{pano}}$ to produce the refined mask $\widehat{\mathcal{M}}_i$, which is then added to \mathbf{M}_c , the final set of retained and refined panoramic instance masks for category c .

Point Downsampling Given a semantic mask \mathcal{M}_i for object i in the equirectangular panorama and its associated 3D point set $\mathcal{P}_{\text{raw},i} = \{\mathbf{x}_j\}_{j=1}^{N_i}$, where each point has elevation e_j , depth d_j , surface normal \mathbf{n}_j , and view direction \mathbf{v}_j , we compute per-point importance weights to form a compact set of representatives while preserving extended geometry:

$$w_j = \frac{d_j^2 \cos e_j}{\max(|\mathbf{n}_j^\top \mathbf{v}_j|, \varepsilon)}, \quad (29)$$

where the $\cos e_j$ term compensates for equirectangular sampling density, d_j^2 implements distance-based weighing, and the normal term deprioritizes visible surfaces, and ε is a small number for numerical stability. We normalize $\pi_j = w_j / \sum_{\ell=1}^{N_i} w_\ell$ and perform weighted down-sampling to at most $N_{\text{max}} = 1000$ representatives, denote \mathcal{P}_i as the final down-sampled point cloud of object i and \mathbf{x}_{ik} as the k^{th} point in \mathcal{P}_i :

$$\mathcal{P}_i = \{\mathbf{x}_{ik}\}_{k=1}^{K_i}, \quad \mathbf{x}_{ik} \sim \mathcal{P}_{\text{raw},i} \quad (30)$$

where $\Pr(\mathbf{x}_j) = \pi_j$ and $K_i = \min(N_{\text{max}}, N_i)$.

6.3. From Directional Field to Binaural Audio

We elaborate on the HRTF-based binaural decoding used in Sec. 4.4 of the main paper. Given the directional sound field $a(\theta, \varphi, t)$ and left/right head-related impulse responses (HRIRs) $h_{\text{left}}(\theta, \varphi, \tau)$ and $h_{\text{right}}(\theta, \varphi, \tau)$, the binaural signals can be written as

$$b_{\text{left}}(t) = \sum_{\theta, \varphi} \sum_{\tau} (h_{\text{left}}^{\text{left}}(\theta, \varphi, \tau) \cdot a(\theta, \varphi, t - \tau)), \quad (31)$$

$$b_{\text{right}}(t) = \sum_{\theta, \varphi} \sum_{\tau} (h_{\text{right}}^{\text{right}}(\theta, \varphi, \tau) \cdot a(\theta, \varphi, t - \tau)), \quad (32)$$

where the sum is over a discrete sampling of the sphere.

Using the ambisonics expansion (Eq. (1) in main)

$$a(\theta, \varphi, t) = \sum_{\ell, m} Y_\ell^m(\theta, \varphi) a_{\ell, m}(t), \quad (33)$$

we can precompute ambisonics-domain HRIRs:

$$h_{\ell, m}^{\text{left/right}}(\tau) = \sum_{\theta, \varphi} w(\theta, \varphi) h^{\text{left/right}}(\theta, \varphi, \tau) Y_\ell^m(\theta, \varphi), \quad (34)$$

where $w(\theta, \varphi)$ are weights for spherical integration. Substituting into the binaural equations yields

$$\begin{bmatrix} b_{\text{left}}(t) \\ b_{\text{right}}(t) \end{bmatrix} = \sum_{\ell=0}^L \sum_{m=-\ell}^{\ell} \begin{bmatrix} h_{\ell, m}^{\text{left}} * a_{\ell, m} \\ h_{\ell, m}^{\text{right}} * a_{\ell, m} \end{bmatrix} (t), \quad (35)$$

which matches Eq. (11) in the main paper after stacking channels into vectors. In practice, we precompute $h_{\ell, m}^{\text{left}}$ and $h_{\ell, m}^{\text{right}}$ by a regular HRTF set (e.g., SADIE-II [1] dataset).

7. Setup for One-Shot Room Acoustic Learning

We expand on Sec. 5.4 of the main paper. Here, the goal is to fit the acoustic parameters of our differentiable renderer so that the predicted ambisonics match a *single* first-order ambisonics (FOA) recording at one microphone pose.

Task formulation. Let $\tilde{a}_L(t)$ denote the ground-truth FOA signal at a fixed pose $\tilde{\mathbf{p}}$, and $a_{\text{src}}(t)$ be the monaural dry source audio. Let $\mathbf{A}(\mathbf{p}, t; \theta)$ be our renderer with learnable parameters θ , including:

- (i) the attenuation constant α controlling geometric decay,
- (ii) the average frequency-dependent reflection response $R[f]$ of the room surfaces,
- (iii) per-source equalization coefficients s (gain and tilt),
- (iv) and the predicted RT60 \hat{T}_{60} (in seconds), parameterized as a base RT60, by ρ and a frequency slope, by γ .

The rendered FOA sound field at listener pose \mathbf{p} is

$$\mathbf{A}(\mathbf{p}, t; \theta) = [\text{RIR}_L(\mathbf{p}, \cdot; \theta) * a_{\text{src}}(\cdot)](t), \quad (36)$$

where $\text{RIR}_L(\mathbf{p}, t; \theta) \in \mathbb{R}^{(L+1)^2}$ is the ambisonic room impulse response that models the transfer function between the source and \mathbf{p} . Following [11, 23], we decompose it into early reflections and late diffuse reverberation:

$$\text{RIR}_L(\mathbf{p}, t; \theta) = \text{Blend}[\text{RIR}_L^{\text{early}}(\mathbf{p}, t; \theta), \text{RIR}_L^{\text{late}}(\mathbf{p}, t; \theta)]. \quad (37)$$

The early part is modeled as a sum over geometric paths $p \in \mathbb{P}$ (direct path and a sparse set of early reflections from beam tracing [6, 7, 11, 14, 21]):

$$\text{RIR}_L^{\text{early}}(\mathbf{p}, t; \theta) \quad (38)$$

$$= \frac{se^{-\alpha t}}{c_{\text{sound}}\tau_p} \sum_{p \in \mathbb{P}} \mathbf{y}_L(\mathbf{d}_p) \mathcal{F}_{\min}^{-1} \left\{ R[f]^{|p|} \right\} (t - \tau_p), \quad (39)$$

where $\mathbf{y}_L(\mathbf{d}_p)$ is the ambisonic encoding of the path ending direction \mathbf{d}_p , $|p|$ is the number of reflections along path p , τ_p is the path delay, and \mathcal{F}_{\min}^{-1} denotes min-phase transform. The reflection response $R[f]$ and equalization s control the spectral characteristics of early reflections, while α governs distance-dependent attenuation modeling air absorption.

Late reverberation parameterization from RT60. The late part $\text{RIR}_L^{\text{late}}$ captures the dense, diffuse tail beyond the early reflection window. We approximate it using a stochastic, frequency-dependent exponential decay that is fully determined by a small number of RT60 parameters.

In our implementation, the late tail is first synthesized as a mono signal $r^{\text{late}}(t; \theta)$ and then mapped to ambisonics

under a diffuse-field assumption (i.e., equal energy in all directions). For each band b we construct an exponentially decaying envelope

$$e_b(t) = \exp\left(-\frac{\ln(1000)}{\hat{T}_{60}(f_b)} t\right), \quad (40)$$

where $\ln(1000)$ corresponds to a 60 dB decay (i.e., the definition of RT60). The band-wise late reverberation signals are then

$$r_b(t; \theta) = e_b(t) \tilde{n}_b(t), \quad (41)$$

and we sum across bands to obtain a full-band late tail

$$r^{\text{late}}(t; \theta) = \sigma(g) \sum_{b=1}^B r_b(t; \theta), \quad (42)$$

where g is a learnable gain parameter and $\sigma(\cdot)$ is a sigmoid to keep the overall late tail level bounded and stable. Finally, we normalize r^{late} to have unit peak magnitude during training.

From mono tail to ambisonics. Under a diffuse-field assumption, late reverberation is approximately isotropic. We therefore lift the mono late tail $r^{\text{late}}(t)$ to ambisonics by:

$$\text{RIR}_L^{\text{late}}(\mathbf{p}, t; \theta) = r^{\text{late}}(t; \theta) \mathbf{1}, \quad (43)$$

where $\mathbf{1} \in \mathbb{R}^{(L+1)^2}$ broadcast r^{late} to all channels. This gives a spatially smooth, low-variance late tail that complements the directional early reflections.

Early/late blending. We model the full room impulse response as a smooth combination of a deterministic early part and a stochastic late tail. Intuitively, the early reflections (direct path and a few specular bounces) encode precise geometric information, while the late reverberation behaves more like a diffuse sound field. To avoid audible discontinuities between these two regimes, we introduce a time-dependent blend:

$$\text{RIR}_L(\mathbf{p}, t; \theta) \quad (44)$$

$$= w_{\text{early}}(t) \text{RIR}_L^{\text{early}}(\mathbf{p}, t; \theta) + w_{\text{late}}(t) \text{RIR}_L^{\text{late}}(\mathbf{p}, t; \theta), \quad (45)$$

where $w_{\text{early}}(t)$ and $w_{\text{late}}(t)$ are scalar envelopes that satisfy

$$w_{\text{early}}(t) \approx 1, w_{\text{late}}(t) \approx 0 \quad \text{at very early times,}$$

and

$$w_{\text{early}}(t) \approx 0, w_{\text{late}}(t) \approx 1 \quad \text{well into the late tail.}$$

We choose a physically motivated early/late cutoff time T_e (proportional to the window used to define early reflections) and construct a short cross-fade region around T_e .

Before this region, $w_{\text{early}}(t)$ stays close to one and $w_{\text{late}}(t)$ stays close to zero; after the cutoff, the roles are reversed. In the transition interval, we use a smooth cosine-shaped cross-fade so that both envelopes vary continuously and the total energy does not exhibit sharp jumps.

Finally, the overall level of the late tail is normalized relative to the early part: we set the initial amplitude of $\text{RIR}_L^{\text{late}}$ such that its peak is comparable to the peak energy of $\text{RIR}_L^{\text{early}}$ per ambisonics channel. This ensures a perceptually continuous decay from the last prominent early reflection into the diffuse reverberant tail, while still allowing the late component to adapt its decay rate and spectral color through the RT60-based parameterization described above.

One-shot fitting objective. Given the dry source $a_{\text{src}}(t)$ and measured FOA $\tilde{a}_L(t)$ at pose $\tilde{\mathbf{p}}$, we optimize:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{MAG}} \quad (46)$$

And evaluated on \mathcal{L}_{MAG} , \mathcal{L}_{ENV} and Δ_{Angular} where:

$$\mathcal{L}_{\text{MAG}} = \left\| \log |S(A(\tilde{\mathbf{p}}, t; \theta))| - \log |S(\tilde{a}_L(t))| \right\|_1, \quad (47)$$

$$\mathcal{L}_{\text{ENV}} = \left\| \text{Env}(A(\tilde{\mathbf{p}}, t; \theta)) - \text{Env}(\tilde{a}_L(t)) \right\|_2^2, \quad (48)$$

$S(\cdot)$ is the STFT, $\text{Env}(\cdot)$ computes the Hilbert-envelope per channel, and Δ_{Angular} is the geodesic angular error defined earlier. We optimize θ with Adam for a small number of iterations (one-shot setting), while keeping the 3D geometry and semantic anchors fixed.

8. Setup for Audio-Visual Spatial Source Separation

We also treat our renderer as a differentiable spatialization module for audio-visual source separation.

Mixture model. Given:

- visually localized sources with 3D anchors, and
 - a mixture FOA recording $\tilde{a}_L(t)$ at pose $\tilde{\mathbf{p}}$,
- we seek per-source monaural signals $\{s_i(t)\}_{i \in \mathcal{O}}$ such that

$$\sum_{i \in \mathcal{O}} \mathbf{A}_i(\tilde{\mathbf{p}}, t; s_i) \approx \tilde{a}_L(t), \quad (49)$$

where \mathbf{A}_i denotes the contribution of source i under our encoder in main Sec. 4.3.

Optimization objective. We parameterize s_i either directly as learnable waveforms constrained by audio priors, or as latent codes for a text-to-audio prior that we decode at each iteration. The loss combines reconstruction and spatial regularization:

$$\mathcal{L}_{\text{sep}} = \mathcal{L}_{\text{MAG}}(\sum_i \mathbf{A}_i, \tilde{a}_L), \quad (50)$$

where \mathcal{L}_{MAG} is as above. In practice, we implement separation via diffusion posterior sampling [3] over the per-source latents, guided by \mathcal{L}_{sep} .

Source Separation by Diffusion Posterior Sampling [3]

In our final model, we adopt a generative approach: each s_i is sampled from a pretrained text-to-audio diffusion prior conditioned on the visual and textual description of source i , and the renderer acts as a differentiable observation model that ties all sources together through the FOA mixture.

Let $x_t^{(i)}$ denote the noisy latent of source i at reverse-diffusion time step t , and let $p_{\text{prior}}(x_t^{(i)})$ be the corresponding prior distribution given by the pretrained diffusion model. The posterior over latents given the observed mixture is

$$p_{\text{post}}(\{x_t^{(i)}\}_{i \in \mathcal{O}} | \tilde{a}_L) \propto \left[\prod_{i \in \mathcal{O}} p_{\text{prior}}(x_t^{(i)}) \right] \exp(-\lambda \mathcal{L}_{\text{sep}}), \quad (51)$$

where λ controls the strength of the guidance.

During sampling, we approximate the posterior score for each source j as

$$\nabla \log p_{\text{post}}(x_t^{(j)}) \quad (52)$$

$$\approx \nabla \log p_{\text{prior}}(x_t^{(j)}) - \lambda \nabla_{x_t^{(j)}} \mathcal{L}_{\text{sep}}(\{\mathbf{A}_i(\tilde{\mathbf{p}}, t; x_t^{(i)})\}, \tilde{a}_L(t)), \quad (53)$$

where the second term backpropagates the separation loss through the renderer and the text-to-audio decoder into the latent of source j . This *posterior-guided* score replaces the unconditional prior score in the reverse diffusion update, yielding a *diffusion posterior sampler* that steers each source towards waveforms that (i) remain likely under the pretrained prior and (ii) jointly reconstruct the observed FOA mixture with spatial patterns consistent with the 3D audio-visual scene.

Intuitively, the pretrained diffusion model maintains the naturalness and diversity of individual source signals, while the renderer and \mathcal{L}_{sep} enforce that their spatialized superposition matches the recorded sound field and respects the visual layout.

Initialization. We initialize all sources in a simple, physically motivated way. For each source, we first pan the mixture FOA into a directional mono signal according to Eq. (1) in the main paper, using the angle predicted by the grounding model for that source. We then apply ZeroSep [9] to enhance this directional signal and use the result as the initial waveform. Finally, starting from an intermediate diffusion time $t = 0.5$, we run diffusion posterior sampling on Stable Audio Open [5] using the DPM++ (3M) SDE solver.

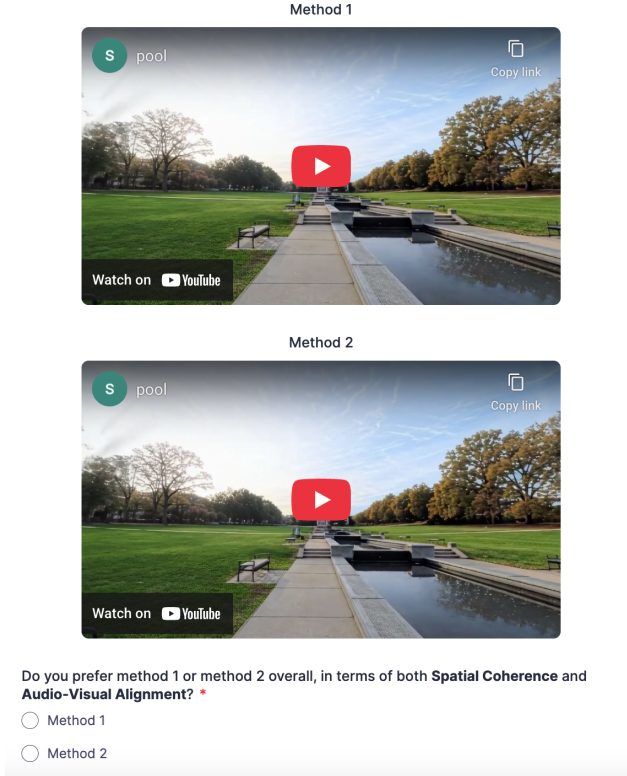


Figure 9. **User Study Interface.** An example pairwise comparison shown to participants. Each question presents two spatial-audio videos (“Method 1” and “Method 2”) for the same scene. Participants listen with headphones and select the method that provides better spatial coherence and audio–visual alignment.

9. User Study Details

We conduct an anonymized user study to evaluate the spatial audio generation quality of our method in comparison to MMAudio [2] and OmniAudio [15]. We recruit 50 participants via Prolific, selecting AI Taskers from diverse geographic regions. The study is administered through Zoho Forms. For each of the 12 test scenes, we form three pairwise comparisons: Ours vs. MMAudio, Ours vs. OmniAudio, and OmniAudio vs. MMAudio. This yields a total of 36 questions, which are presented in randomized order with the method names hidden to avoid biases. Before beginning, participants read a detailed explanation of the evaluation criteria and are instructed to judge each pair based on spatial coherence and audio–visual alignment. To ensure proper perception of binaural audio, participants must also confirm that they are wearing headphones prior to proceeding. Figure 9 shows an example question from our user study form.

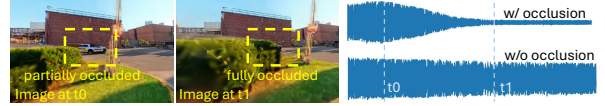


Figure 10. Physics-aware occlusion add-on: as a clustered source becomes partially and then fully blocked between t_0 and t_1 , visibility-based point reweighting yields smooth attenuation of the rendered signal.

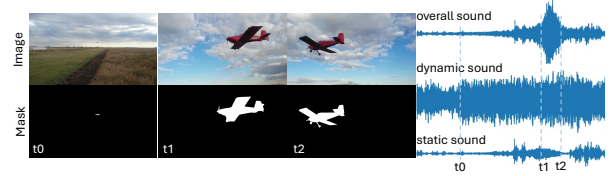


Figure 11. Dynamic source rendering on a plane fly-by clip: a time-varying 3D anchor from SAM3 and DepthAnythingV3 separates dynamic and static sound components, with the dynamic layer peaking near closest approach and attenuating with distance

10. Discussions

Outdoor-focused propagation and reverberation. Our propagation model is designed primarily for outdoor scenes, where recordings are often close to dry [19] and listeners generally have weaker expectations of strong reverberation [20]. Since MMAudio is trained on web videos that typically exhibit little room-like reverberation, its outputs for common outdoor sources are also usually near-dry. As a result, double reverberation is unlikely in our setting. We therefore focus on the central challenges of semantic coherence, accurate DoA alignment, and heterogeneous source decomposition.

Physics-aware extensions. SONOWORLD naturally accommodates lightweight geometry-driven extensions. As one example, we incorporate a simple occlusion heuristic for clustered sources based on visibility-aware point reweighting, which yields smooth attenuation as a source transitions from partially visible to fully occluded (Fig. 10).

Dynamic sources. Our renderer also naturally supports dynamic sources when a time-varying 3D anchor is available. In a newly tested plane fly-by example, SAM3 and DepthAnythingV3 recover a source trajectory that we render into dynamic and static layers. The dynamic component peaks near the point of closest approach and attenuates with distance. Looking ahead, advances in 4D visual reconstruction may enable richer image-to-4D audiovisual scene generation (Fig. 11).

References

- [1] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11):2029, 2018. [9](#)
- [2] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *ArXiv*, abs/2412.15322, 2024. [1](#), [12](#)
- [3] Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*, 2023. [11](#)
- [4] Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. See-2-sound: Zero-shot spatial environment-to-spatial sound. *ArXiv*, abs/2406.06612, 2024. [2](#)
- [5] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. [11](#)
- [6] Thomas Funkhouser, Ingrid Carlbom, Gary Elko, Gopal Pingali, Mohan Sondhi, and Jim West. A beam tracing approach to acoustic modeling for interactive virtual environments. In *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, page 21–32, New York, NY, USA, 1998. Association for Computing Machinery. [10](#)
- [7] John Kenneth Haviland and Balakrishna D. Thanedar. Monte carlo applications to acoustical field solutions. *The Journal of the Acoustical Society of America*, 54(6):1442–1448, 1973. [10](#)
- [8] Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP*, 2024. [6](#)
- [9] Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, Jing Bi, Wenqiang Liu, Nima Mesgarani, and Chenliang Xu. Zerosp: Separate anything in audio with zero training. *arXiv preprint arXiv:2505.23625*, 2025. [11](#)
- [10] Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025. [2](#)
- [11] Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. In *ICCV*, 2025. [10](#)
- [12] Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. *ICLR*, abs/2506.12199, 2025. [2](#), [7](#)
- [13] World Labs. Marble. <https://marble.worldlabs.ai/>, 2025. [2](#)
- [14] Christian Lauterbach, Anish Chandak, and Dinesh Manocha. Interactive sound rendering in complex and dynamic scenes using frustum tracing. *IEEE Transactions on Visualization and Computer Graphics*, 13:1672–1679, 2007. [10](#)
- [15] Huadai Liu, Tianyi Luo, Kaicheng Luo, Qikai Jiang, Peiwen Sun, Jialei Wang, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, et al. Omniaudio: Generating spatial audio from 360-degree video. *ICML*, 2025. [1](#), [2](#), [6](#), [12](#)
- [16] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011. [5](#)
- [17] OpenAI. Gpt-5 system card. <https://openai.com/index/gpt-5-system-card/>, 2025. [9](#)
- [18] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. [9](#)
- [19] James Traer and Josh H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48): E7856–E7865, 2016. [12](#)
- [20] James Traer, Sam V. Norman-Haignere, and Josh H. McDermott. Causal inference in environmental sound recognition. *Cognition*, 214:104627, 2021. [12](#)
- [21] Dirk van Maercke and Jacques Martin. The prediction of echograms and impulse responses within the epidaure software. *Applied Acoustics*, 38(2):93–114, 1993. [10](#)
- [22] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. [8](#)
- [23] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. [10](#)
- [24] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023. [7](#)
- [25] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, and Jianfeng Gao. Generalized decoding for pixel, image, and language. In *CVPR*, 2023. [9](#)