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

Derong Jin[*] Xiyi Chen[*] Ming C. Lin Ruohan Gao
University of Maryland, College Park

Abstract

Tremendous progress in visual scene generation now turns a single image into an explorable 3D world, yet immersion remains incomplete without sound. We introduce Image2AVScene*, the task of generating a 3D audio-visual scene from a single image, and present* SonoWorld*, the first framework to tackle this challenge. From one image, our pipeline outpaints a 360° panorama, lifts it into a navigable 3D scene, places language-guided sound anchors, and renders ambisonics for point, areal, and ambient sources, yielding spatial audio aligned with scene geometry and semantics. Quantitative evaluations on a newly curated real-world dataset and a controlled user study confirm the effectiveness of our approach. Beyond free-viewpoint audio-visual rendering, we also demonstrate applications to one-shot acoustic learning and audio-visual spatial source separation. Project website:* `https://humathe.github.io/sonoworld/`

1. Introduction

The past few years have seen rapid progress in visual scene generation [57, 85, 87, 94]. Building on recent advances in 3D scene generation [37, 44, 57, 64, 84, 86], today's systems can generate photorealistic 3D worlds from a single 2D image. For example, in Fig. 1, one photo of a garden becomes an explorable 3D scene: you step onto a wooden bridge spanning a turquoise stream, peer beneath the arch as water tumbles through, and shift viewpoints to inspect the tiered waterfalls and overhanging cherry blossoms. Such capabilities enable compelling applications in VR/AR, content creation, and robotics. Yet, these models share a striking limitation: they produce *silent* worlds you can walk through, but not *listen* to.

Immersion in the real world is inherently multisensory. In the same scene, sound is crucial for perceiving and understanding space: the waterfalls should thunder from upstream and swell as you approach; birds should chirp and leaves rustle from the canopy; insects should buzz near the flowerbeds and shift with head turns. Without these semantically meaningful sounds, together with the directional and distance cues, the world may look convincing yet remains perceptually incomplete.

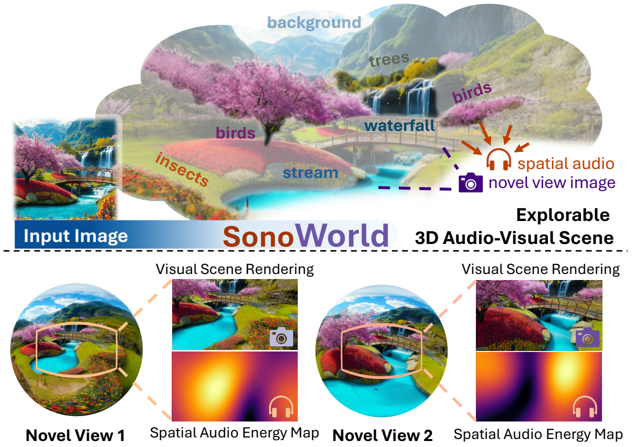


Figure 1. From one image, SonoWorld generates an explorable 3D audio-visual scene, where you can navigate to novel views and locations, while listening to spatial audio aligned with scene semantics and the 3D locations of heterogeneous sound sources.

To realize this vision, we introduce Image2AVScene, a new task that aims to generate an explorable 3D audio-visual scene from a single RGB image. Given one image of a scene, the goal is to generate (i) a navigable 3D visual scene and (ii) a spatial sound field that is semantically and geometrically aligned with the visual content. This enables users not only to look around the generated 3D scene, but also *listen* from any location and viewpoint.

Jointly generating a coherent 3D scene and its spatial sound field from a single image is challenging. *First,* unlike traditional audio synthesis [7, 9, 43, 51, 62] that often targets isolated objects or events, scene-level audio generation must compose heterogeneous source types and scales: point sources (*e.g.*, a chirping bird), areal sources (*e.g.*, a flowing river), and ambient soundscapes (*e.g.*, forest insects and wind). Each behaves differently over time and distance, and must remain coherent as the listener moves. *Second,* the system requires holistic *semantic* scene understanding to infer what is likely to make sound, how it sounds, and how loud all from the visual context: waterfalls roar and fluctuate, birds call intermittently, insects form a high-frequency

[*] Equal contribution.

bed near flowers and trees—while silent objects (*e.g.*, a wooden bridge) should remain silent. *Third,* all generated sounds must be grounded to plausible 3D locations and spatial extents inferred from the image, and rendered with perceptually realistic *spatial* effect (*e.g.*, direction of arrival and distance-dependent attenuation).

We propose SONOWORLD, a *training-free* framework to systematically address these challenges. Given a single image as input, we first reproject it to an equirectangular panorama with elevation correction and outpaint a complete 360° view. We then lift this panorama into a navigable 3D representation (3D Gaussian splats [40]) to obtain dense, view-consistent geometry. To bridge vision and sound, we perform 360° semantic grounding: a language-guided set of sounding categories seeds open-vocabulary instance discovery on tiled perspective views; these instances are reconciled with panoramic mask proposals and back-projected into 3D to produce instance/region anchors for sound. Finally, a spatial audio encoder converts these anchors into ambisonics coefficients, supporting point-like and areal sources as well as ambient sound, yielding geometry-aware spatialization consistent with the scene's 3D structure and semantics.

To facilitate evaluation of this new task, we contribute an evaluation dataset of 68 clips that contains both panorama videos and ambisonics recordings across six diverse real-world scenes, and we define a suite of metrics assessing both the semantic fidelity and spatial accuracy of the generation outputs. Across these metrics and a user study, our method consistently outperforms a series of baselines that generate spatial audio from visual input. Beyond free-viewpoint audio-visual rendering and exploration in the generated 3D audio-visual scene, we also show evidence that our framework extends to two additional 3D audio-visual learning tasks: one-shot room acoustic learning and audio-visual spatial source separation.

In summary, our key contributions are:

- We introduce IMAGE2AVSCENE, a novel task to generate an interactive 3D visual scene together with a spatial sound field that is semantically and geometrically grounded to the visual context, along with the first effective framework, SONOWORLD, to tackle this task.
- We collect SONOSCENE360, an evaluation dataset with well-calibrated 360° video and ambisonic audio, and define a suite of semantic and spatial metrics to comprehensively assess generation quality.
- Our method outperforms strong baselines across all metrics and in a perceptual user study on visually-guided spatial audio generation. We further apply our framework to one-shot room acoustic learning and audio-visual spatial source separation, highlighting its potential to extend to other audio-visual learning tasks in 3D.

2. Related Work

3D Scene Generation. Recent single-image 3D scene generation methods have coalesced into three categories: iterative, video diffusion, and panoramic, each advancing scale, coherence, and controllability in complementary ways. *Iterative* methods [4, 12, 86, 87, 89] grow scenes by alternating diffusion-based outpainting with 3D lifting and optimization (often via Gaussian Splatting [40]), typically guided by geometry cues such as monocular depth and light trajectory planning with text prompts. They enable scalable outpainting and interactive refinement but can accumulate drift over long trajectories and often require post-hoc geometric cleanup. *Video diffusion* methods [17, 25, 28, 35, 64, 65, 79, 88] leverage temporally coherent generators plus cached geometry such as point clouds to improve cross-view alignment, provide precise camera control, and extend from static to dynamic scenes. Their strengths include reduced flicker and controllable trajectories, while challenges include computational cost and long-range consistency when stitching extended worlds. *Panoramic* methods [37, 82, 84, 92, 95] first outpaint a coherent 360° panorama (typically equirectangular with upright and field-of-view constraints) and then lift it into 3D via depth alignment and Gaussian optimization, achieving full-environment coverage and seamless horizon continuity. Our approach follows this panoramic paradigm but is distinguished by jointly modeling spatial audio alongside visuals. We adopt a panoramic representation because it captures the full 360° field of view and provides a unified, scene-level coordinate frame for ambisonics rendering.

Spatial Audio Generation. A large body of audio generation work produces high-quality monoaural audio from either text or video inputs [7–9, 30, 43, 51, 53, 62, 96]. To spatialize such audio, prior methods either explicitly model room acoustics by predicting room impulse responses (RIRs) for rendering spatial audio [39, 45, 55, 56, 68, 78], or directly perform mono to spatial conversion conditioned on visual scene structure from videos to localize sources and synthesize spatial channels [21, 26, 27, 47, 49, 59].

Most related to our work are recent methods that directly synthesizes spatial audio, either from text [32, 69] or images/videos [14, 42, 54]. Concurrent work Sonic4D [81] further couples spatial audio generation with 4D dynamic scenes, but it is limited to single objects, narrow views, and offline processing. In contrast, our method generates a full 3D audio-visual scene from a single image input, producing scene-consistent FOA tightly anchored to the visual context and supporting free, real-time navigation with physically grounded point, areal, and ambient sources across 360° environments, including off-screen emitters.

Localizing Sounds in Visual Scenes. Sound localization aims to identify the pixels or regions corresponding to sound sources in images or videos. Early approaches model audio-visual mutual information [18, 31] or use canonical correlation analysis [41]. Later deep methods exploit audio-visual correspondence with varying levels of supervision [1, 2, 6, 33, 58, 66, 67, 70, 93]. Recently, Audio-Visual Large Language Models [10, 48, 91] have significantly improved the generalizability of audio-visual grounding. We also localize sound sources in visual scenes; however, unlike prior work above that treats localization as the end goal, we localize and ground potential sound sources in the generated 3D panoramic scene using foundational Vision-Language Models (VLMs) [97] to enable subsequent spatial audio generation at the corresponding 3D locations.

Audio-Visual Source Separation. Prior work on audio-visual source separation has leveraged visual cues to guide separation of speech [13, 15, 23], musical instruments [20, 22, 83, 90], and general sound sources in-the-wild [24, 73, 74]. Differently, our work tackles spatial audio generation for all sound sources grounded in a generated 3D visual scene, rather than separating an existing sound mixture. We further demonstrate an application to separate spatial sound sources in 3D via diffusion posterior sampling [11].

3. The IMAGE2AVSCENE Task

This section reviews ambisonic spatial audio (Sec. 3.1), formalizes IMAGE2AVSCENE (Sec. 3.2), introduces SONOSCENE360, our curated evaluation dataset (Sec. 3.3), and defines the evaluation metrics (Sec. 3.4).

3.1. Background on Ambisonics

Ambisonics represents the sound field around *a point* as a weighted sum of spherical harmonics. Let $\mathbf{u}(\theta, \varphi)$ denote a unit direction with azimuth $\theta \in [-\pi, \pi]$ and elevation $\varphi \in [-\pi/2, \pi/2]$, and let $a(\theta, \varphi, t)$ denote the directional sound field. Ambisonics expand this directional function on the sphere using real spherical harmonics $Y_\ell^m(\theta, \varphi)$:

$$a(\theta,\varphi,t) \approx \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} Y_\ell^m(\theta,\varphi) a_{\ell,m}(t) = \mathbf{y}_L(\theta,\varphi)^\top \mathbf{a}_L(t), \tag{1}$$

where L is the ambisonic order, $\mathbf{y}_L$ stacks the $(L+1)^2$ basis functions, and $\mathbf{a}_L(t)$ are the corresponding ambisonic coefficients (channels of sound). Given coefficients $\mathbf{a}_L$, a virtual microphone at direction $\mathbf{u}$ is obtained by $a_\mathbf{u} = \mathbf{y}_L(\mathbf{u})^\top \mathbf{a}_L$. For a single point source $a_{\text{src}}(t)$ at $\mathbf{u}$ with distance d, the ambisonic coefficients at $[0, 0, 0]$ are:

$$\mathbf{a}_L^{\text{single}}(t) = \sigma(d) a_{\text{src}}(t) \mathbf{y}_L(\mathbf{u}), \tag{2}$$

where $\sigma(d)$ is the attenuation and decay. Listener head rotation $R \in SO(3)$ enables *3-DoF* rendering from a single ambisonics capture at a fixed point. Translation is not natively represented by a single-point capture. However, if ambisonics can be *encoded at arbitrary listener locations* (as in our model), composing head rotations with listener positions enables full *6-DoF* exploration in the sound field.

A common choice for ambisonics is first-order ambisonics (FOA; $L = 1$), which has four channels corresponding to the zeroth- and first-order harmonics. FOA provides a first-order approximation of the sound field on the sphere and is widely supported in capture and playback pipelines. We use FOA in most experiments, but our rendering framework is *order-agnostic* and supports arbitrary ambisonic orders L, trading channel count $(L+1)^2$ for spatial accuracy.

3.2. Task Formulation of IMAGE2AVSCENE

Given a single input image I, our objective is to develop a framework $\mathcal{G}$ that jointly generates a visual representation $\mathbf{V}$ and the corresponding spatial audio field $\mathbf{A}$, forming an interactive 3D audio-visual scene:

$$\mathcal{G} : I \to \{\mathbf{V}, \mathbf{A}\}. \tag{3}$$

Here, $\mathbf{V}$ encapsulates the geometric and appearance structure of the scene, while $\mathbf{A}$ defines a spatially coherent sound field that aligns with $\mathbf{V}$ both geometrically and semantically. For an observer at pose $\mathbf{p}$, the scene can be rendered into an image $\mathbf{V}(\mathbf{p})$, and the corresponding spatial audio signal is given by $\mathbf{A}(\mathbf{p}, t)$.

Our Instantiation. We parameterize the visual scene $\mathbf{V}$ as 3D Gaussian Splats [40] (Sec. 4.1), and represent the auditory scene $\mathbf{A}$ as a point-cloud-based representation that supports ambisonics rendering (Sec. 4.3): given a listener pose $\mathbf{p}$, it synthesizes the ambisonics $\mathbf{a}_L(t) = \mathbf{A}(\mathbf{p}, t) \in \mathbb{R}^{(L+1)^2}$ at arbitrary order L.

3.3. The SONOSCENE360 Dataset

Quantitative evaluation of IMAGE2AVSCENE requires paired 3D visual scenes with spatial audio at listener positions offset from the camera pose—conditions not met by existing datasets. Therefore, with the data collection setup shown in Fig. 2, we collect SONOSCENE360, a well-curated evaluation dataset for this task that contains 68 clips of synchronized 360° video captured by Insta360 X5 and FOA audio recorded by RØDE NT-SF1 from six real-world scenes (*Fountain*, *Kitchen*, *Pool*, *Bridge*, *Stream*, *Siren*). See Fig. 6 for scene visualizations. For audible sources within each scene, we provide the semantic labels (*e.g.*, *fountain*, *microwave*, *bushes*), a brief description of the sound, and a coarse direction relative to the microphone, (*e.g.*, source at *left/right/front/back*). These annotations enable both spatial and semantic evaluation described in Sec 3.4. See Supp. for dataset and calibration details.

3.4. Evaluation Metrics

We evaluate the quality of generated spatial audio with a suite of metrics along two complementary axes: i) *spatial coherence*—whether the spatial effect of sound sources in

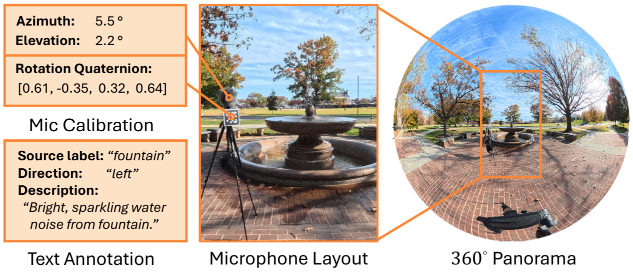


Figure 2. Illustration of real-world audio-visual scene data collection and curation for SONOSCENE360.

the 3D scene is accurate and realistic, and ii) *semantic alignment*—whether sounds originate from the correct visually-depicted objects/regions. Below we outline each metric; see Supp. for detailed metric definitions.

Spatial Metrics. We compare the predicted spatial audio in two terms: (i) deviation in direction of arrival (DoA), and (ii) correlation of sphereical energy patterns. After estimating DoA in azimuth θ and elevation φ, we report the absolute azimuth error $\Delta_{\text{abs}}\theta$, absolute elevation error $\Delta_{\text{abs}}\varphi$, and the geodesic angular error Δ_{Angular}, following prior works [32, 54]. To handle multi-source scenes where a single DoA is insufficient, we compute two saliency metrics on the spherical energy map: correlation coefficient (CC) and area under curve (AUC), following [42].

Semantic Metrics. To assess semantic consistency, we form directional monaural renderings by placing a virtual microphone oriented toward direction $\mathbf{u} \in \mathbb{S}^2$. Specifically, we render along four FOA-aligned principal directions: *left*, *right*, *front*, and *back*, which cover the majority of sound source placements. Using semantic labels of sound sources, we introduce two CLAP-based similarities [80] on these directional audios: D-CLAP$_\text{T}$ (audio-text) between directional audio and the semantic label, and D-CLAP$_\text{A}$ (audio-audio) between predicted and reference directional audio. Beyond similarity, we report a directional CLAP-based R-Precision, namely D-CLAP$_\text{R}$, inspired by text-to-3D evaluation [38]. For annotations of the form (direction, label), we render monaural audio toward the four principal directions, rank directions by CLAP-T score for the label, and compute R-Precision (with $R = 1$), *i.e.*, the top-1 accuracy that the correct direction receives the highest similarity score.

4. Approach

This section presents SONOWORLD, our *training-free* pipeline to tackle the IMAGE2AVSCENE task (illustrated in Fig. 3): Sec. 4.1 details how we construct a 360° panorama and lift it into a 3D Gaussian scene $\mathbf{V}$; Sec. 4.2 describes 360° semantic grounding to localize the sounding entities in 3D; Sec. 4.3 explains how we design the ambisonics encoder $\mathbf{A}$, which produces spatial sound fields that are geometrically aligned and semantically consistent with 3D visual scene $\mathbf{V}$; and Sec. 4.4 demonstrates how we render binaural audio from $\mathbf{A}$ at any pose.

4.1. Panorama-Based Visual Scene Generation

We adopt a panoramic representation for single-image scene generation because it naturally captures the full 360° field of view (FoV) and unifies holistic audio-visual scene synthesis. It contains the following three steps: 1) camera calibration of the input image; 2) panoramic outpainting from the corrected image; and 3) final 3D reconstruction from the outpainted panorama, detailed below.

Single-Image Camera Calibration. Prior single-image panorama outpainting methods [36, 37, 82] implicitly assume a level camera and place the input view along the equator of the panorama, which can cause vertical misalignment and distortion when the input image is captured with an upward or downward tilt. To address this and prepare for the outpainting step, we first use GeoCalib [76], a single-image calibration network that jointly infers gravity direction and camera FoV through learned geometric optimization, to obtain the camera elevation and FoV as:

$$(\varphi, f) = \text{Calib}(I). \tag{4}$$

Image to Panorama Outpainting. We then reproject the input perspective image I into an equirectangular panorama using a warping operator $\mathcal{W}_G$ that rectifies the camera and performs multi-scale anti-aliased sampling based on a Gaussian pyramid. The warped image is subsequently outpainted using the panorama outpainting model g_{outpaint} from WorldGen [82]:

$$I_{\text{pano}} = g_{\text{outpaint}}\big(\mathcal{W}_G(I, \varphi, f)\big), \tag{5}$$

which completes the missing regions beyond the input image and produces a full 360° panorama. Please refer to Supp. for implementation details of $\mathcal{W}_G$.

Panorama to 3D Scene Reconstruction. Finally, we lift the completed panorama into a 3D visual scene using a panorama-to-3D reconstruction method, denoted as $\mathbf{V} = \mathcal{G}_\mathbf{V}(I_{\text{pano}})$. Multiple approaches can be used for this component [37, 44, 82, 84], including World Labs' Marble model [44], which currently offers the best visual rendering quality, and the open-source model HunyuanWorld-1.0 [37]. Both transform the panorama into a 3D scene parameterized either by 3D Gaussian splats or a textured mesh, yielding a photorealistic 3D environment that supports real-time interactive exploration.

4.2. 360° Audio-Visual Semantic Grounding

To bridge visual and audio generation from a single image I, we 1) use a vision-language model (VLM) to propose sounding categories and their acoustic attributes; 2) predict instance masks on FoV images with an open-vocabulary segmenter; 3) refine them into globally consistent panoramic segmentations that align with both semantics and 3D geometry; and 4) unproject them to localize 3D sound sources in $\mathbf{V}$.

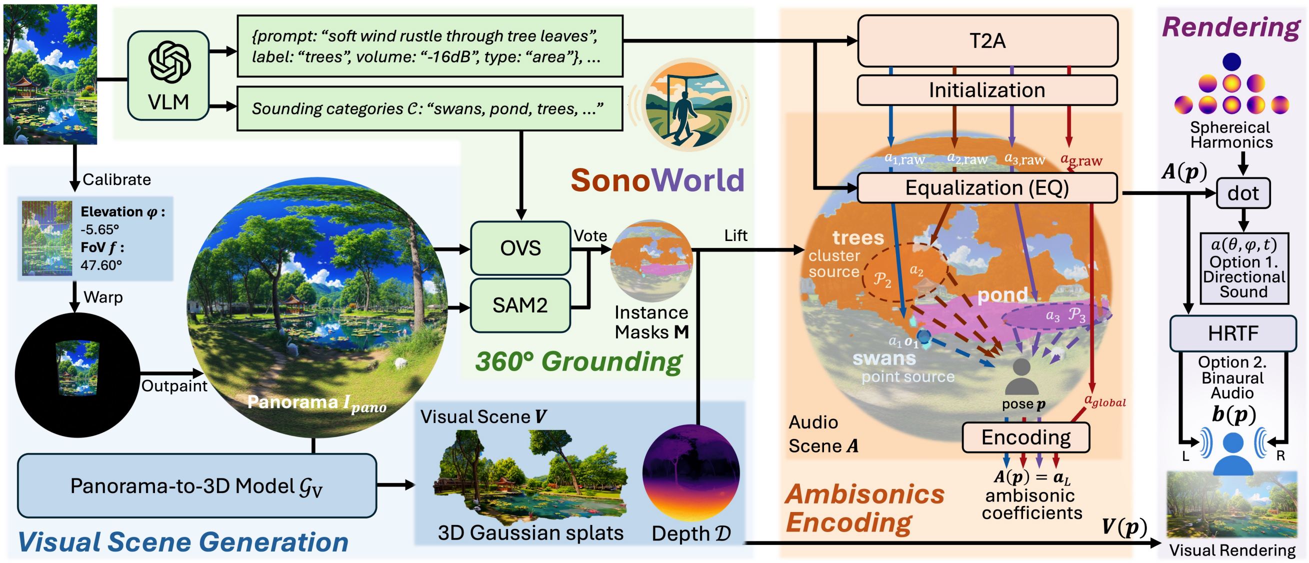


Figure 3. Given a single image I, SONOWORLD jointly generates a 3D visual scene $\mathbf{V}$ and a semantically and geometrically aligned audio scene $\mathbf{A}$. It consists of: 1) **Visual Scene Generation** (Sec. 4.1): single-image calibration and warping followed by panorama outpainting to obtain the a full 360° panorama image I_{pano}, which is further lifted into a 3D Gaussian scene via a panorama-to-3D reconstruction model $\mathcal{G}_{\text{V}}$; 2) **360° Semantic Grounding** (Sec. 4.2): a VLM extracts the categories $\mathcal{C}$ of potential sounding sources, which are used to generate panoramic instance masks $\mathbf{M}$ with the help and coordination of both an open-vocabulary segmentation model (OVS) and a class-agnostic segmentation model (SAM2 [63]); 3) **Ambisonics Encoding** (Sec. 4.3): based on the audio prompt and equalization parameters from the VLM model, a text-to-audio (T2A) model generates per-source waveforms that are equalized, and mapped to ambisonics coefficients based on the 3D locations and its source type; and 4) **Free-Viewpoint Rendering** (Sec. 4.4): the ambisonics coefficients are decoded into pose-dependent binaural audio $\mathbf{b}(\mathbf{p})$ using head related transfer function (HRTF) and synchorinzed with Gaussian rendering $\mathbf{V}(\mathbf{p})$.

Sounding Category Proposal. We first query a VLM—either the proprietary GPT-5 [61] or the open-source LLaVA-Next-34B [52]—with the input image I to obtain a set of candidate sounding categories $\mathcal{C}$ and their attributes: source-type labels (point, clustered, or ambient), text prompts for audio synthesis, and amplitude-equalization parameters. These categories then guide subsequent visual grounding.

Open-Vocabulary Segmentation. Since open-vocabulary segmentation (OVS) models are trained on perspective FoV, not panoramic, images, we split the generated panorama I_{pano} into overlapping FoV tiles and run X-Decoder [97] on each tile conditioned on $\mathcal{C}$. The resulting instance masks for each category $c \in \mathcal{C}$ are reprojected back to panoramic coordinates and grouped per category to provide semantically labeled, tile-aggregated mask predictions $\mathbf{M}_{\text{OVS},c}$.

Panoramic Mask Refinement. Tile-wise X-Decoder masks reprojected to the panorama are not globally consistent: limited vertical FoV and seams at tile boundaries can leave large regions (*e.g.*, sky or ground) incomplete and introduce broken edges. In contrast, foundational segmentation models like SAM2 [63] can accurately segment arbitrary image formats, including equirectangular panoramas, but yield only class-agnostic regions. We therefore use SAM2 to produce panorama-wide proposals $\mathbf{M}_{\text{pano}}$, and then, for each category c, let the corresponding open-vocabulary predictions $\mathbf{M}_{\text{OVS},c}$ cast confidence-weighted votes on overlapping SAM2 regions based on spatial agreement. Proposals in $\mathbf{M}_{\text{pano}}$ with strong semantic support from $\mathbf{M}_{\text{OVS},c}$ are retained and, when appropriate, slightly refined to include nearby pixels endorsed by $\mathbf{M}_{\text{OVS},c}$. For each category c, we aggregate the retained and refined proposals into a panoramic instance set $\mathbf{M}_c$, and collect the final result as the union of all category-specific masks, $\mathbf{M} = \bigcup_{c \in \mathcal{C}} \mathbf{M}_c$, preserving SAM2's global geometric consistency while inheriting X-Decoder's semantic precision.

Unprojection to 3D. Finally, using the panoramic depth $\mathcal{D}$ rendered from the generated 3D visual scene from Sec. 4.1, we unproject each final panoramic instance mask $\mathcal{M}_i \in \mathbf{M}$ into 3D to obtain its spatial locations $\mathcal{P}_i = \text{Lift}(\mathcal{M}_i, \mathcal{D})$. Collecting all instances yields $\mathcal{P}$ which specifies the 3D locations of all sounding objects within the scene.

4.3. Ambisonics Encoding

Based on the VLM prediction of source type, source content, and equalization parameters, we generate an audio scene $\mathbf{A}$ and encode ambisonics at any listener pose. We use MMAudio [9] to synthesize audio from text prompts at two levels. For each grounded sound source i, MMAudio generates a raw waveform $a_{i,\text{raw}}$ from a source-specific prompt. It also generates a scene-level ambient waveform $a_{\text{global}}(t)$ from a global prompt describing the overall background atmosphere. Then, we amplify the volume of each sound (equlaization) via the *predicted sound energy* v (in dB, 0 for the loudest). The resulting audio content of source i is:

$$a_i(t) = 10^{v_i/20} a_{i,\text{raw}}(t). \tag{6}$$

Next, we spatialize the grounded sounds a_i based on their 3D positions $\mathcal{P}_i$ and source types predicted by the VLM. Let $\mathcal{O}$ denote the set of grounded sound sources only. We partition $\mathcal{O}$ into two disjoint subsets, $\mathcal{O} = \mathcal{O}_{\text{point}} \cup \mathcal{O}_{\text{cluster}}$ corresponding to point-like and clustered sources. For a listener pose $\mathbf{p} = [\mathbf{R}, \mathbf{t}] \in SE(3)$, we model distance attenuation and air absorption by $\sigma(d) = e^{-\alpha d}/d$, where d is the distance from sound source to listener position $\mathbf{t}$. Omitting $\mathbf{p}$ and t for brevity, the ambisonic coefficient is the sum of the grounded-source contributions and the global ambient term:

$$\mathbf{A} = \mathbf{A}_{\text{grounded}} + \mathbf{A}_{\text{global}} = \mathbf{A}_{\text{point}} + \mathbf{A}_{\text{cluster}} + \mathbf{A}_{\text{global}} \tag{7}$$

- **Point sources.** We approximate source i by the centroid $\mathbf{o}_i$ of its point cluster $\mathcal{P}_i$, and, following Eq. 2, we write:

$$\mathbf{A}_{\text{point}} = \sum_{i \in \mathcal{O}_{\text{point}}} a_{i,L} \sigma(\|\mathbf{d}_i\|) \mathbf{y}_L \left(\mathbf{R}^\top \frac{\mathbf{d}_i}{\|\mathbf{d}_i\|} \right), \tag{8}$$

  where $\mathbf{d}_i = \mathbf{t} - \mathbf{o}_i$ is the relative offset vector from the listener to source i.
- **Clustered sources.** For spatially extended emitters (*e.g.*, a flowing river), we average over the point cloud $\mathcal{P}_i$ to create a diffuse field:

$$\mathbf{A}_{\text{cluster}} = \sum_{i \in \mathcal{O}_{\text{cluster}}} \frac{a_i}{|\mathcal{P}_i|} \sum_{\mathbf{o} \in \mathcal{P}_i} \sigma(\|\mathbf{d}\|) \mathbf{y}_L \left(\mathbf{R}^\top \frac{\mathbf{d}}{\|\mathbf{d}\|} \right), \tag{9}$$

  where $\mathbf{d} = \mathbf{t} - \mathbf{o}$, analogous to Eq. 8. For an area sound that surrounds the listener, the directional terms ($Y_{\ell,m}, \ell > 0$) tend to cancel each other and aggregate into the omnidirectional component, making the perceived directivity less sensitive to head rotation.
- **Global ambience.** Scene-level background (*e.g.*, wind, distant traffic) is reflected only in the omnidirectional part with no spatial variation:

$$\mathbf{A}_{\text{global}} = a_{\text{global}} \left[1,\ 0, \ldots,\ 0 \right]^\top. \tag{10}$$

The final ambisonics signal is obtained by combining grounded sources and the ambient term in Eq. 7. This encoding is fully differentiable with respect to the audio buffers a_i, which makes it applicable to other 3D audio-visual learning tasks (Sec. 5.4). Next, we render $\mathbf{A}(\mathbf{p}, t) \in \mathbb{R}^{(L+1)^2}$ into binaural audio for free-viewpoint rendering.

4.4. Free-Viewpoint Rendering

Given a camera/microphone pose $\mathbf{p}$ for both image and audio synthesis, we render synchronized visuals in images and encode spatial sound in ambisonics. The visual frame $\mathbf{V}(\mathbf{p})$ is obtained by standard 3D Gaussian Splatting rendering. The ambisonics signals $\mathbf{a}_L(t)$ are decoded to binaural output via an Head Related Transfer Function (HRTF)-based decoder. Let $h^{\text{left}}_{\ell,m}$ and $h^{\text{right}}_{\ell,m}$ denote the left/right HRIRs for ambisonics channel m, ℓ. The final binaural waveform $\mathbf{b}(\mathbf{p}) = [b_{\text{left}}, b_{\text{right}}]^\top$ is computed as:

$$\begin{bmatrix} b_{\text{left}} \\ b_{\text{right}} \end{bmatrix}(t) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \begin{bmatrix} h^{\text{left}}_{\ell,m} * a_{\ell,m} \\ h^{\text{right}}_{\ell,m} * a_{\ell,m} \end{bmatrix}(t), \tag{11}$$

where $*$ denotes convolution. Updating the camera and microphone pose $\mathbf{p}$ enables interactive navigation with view-consistent, spatialized audio. See Supp. for details.

5. Experiments

5.1. Experiment Settings

Dataset. Since no existing dataset provides ground-truth spatial audio recorded at listener positions distinct from camera viewpoints, we rely on our SonoScene360 dataset (Sec. 3.3) for quantitative evaluation. For qualitative evaluation, user studies, and demos, we additionally use online photographs and diffusion-generated images, spanning a wide range of audio-visual events (*e.g.*, plane landing, riverside market, volcano eruption) that are often difficult to capture in real-world recordings. Together, these real-world and synthetic scenes provide a complementary and comprehensive evaluation of generation quality.

Baselines. No existing approach directly addresses the Image2AVScene task we propose. We compare with prior methods that generate spatial audio from visual input (SEE-2-SOUND [14], ViSAGe [42], OmniAudio [54]), as well as MMAudio [9], a representative monaural audio generation method. For comparison, we adapt each method to our setting by feeding it the 3D visual scene rendered by SonoWorld and using it to generate spatial audio, while keeping their archetectures unchanged. Below, we describe the specific adaptation for each baseline.

- **MMAudio [9]**: It generates monaural audio conditioned on a FoV video, which we render from our reconstructed scene. We further guide it with a text prompt composed of all detected sounding categories and directly pan the synthesized audio sources using their ground-truth locations to obtain spatial audio, both to the baseline's advantage.
- **SEE-2-SOUND [14]**: It is designed for FoV image inputs and lifts audio anchors into 3D using per-region depth estimates. We render FoV images from our reconstructed 3D scene to guide its ambisonics generation.
- **ViSAGe [42]**: generates ambisonics from a FoV video, which we render from our reconstructed scene.
- **OmniAudio [54]**: generates ambisonics from a panorama video, which we render from our reconstructed scene.

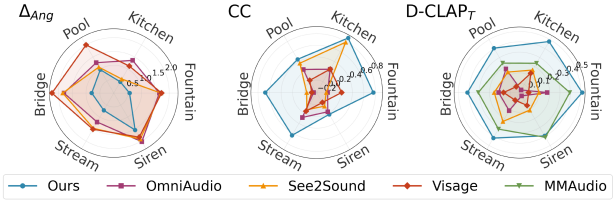


Figure 4. **Per-scene results on our SonoScene360 dataset.** Results on representative metrics show that our method consistently outperforms baselines on all scenes.

5.2. Quantative Results

Quantitative Evaluation on SonoScene360. Table 1 reports results on both spatial and semantic metrics introduced in Sec. 3.4 on SonoScene360. We report two versions of our results: **Ours (Open-source)**, a fully reproducible open-source version, where we use HunyuanWorld-1.0 [37] for panorama-to-3D scene reconstruction and LLaVA-Next-34B [52] for sound-source proposal; and **Ours (Proprietary)**, which instead uses Marble [44] and GPT-5 [61] for the best performance. Our approach consistently outperforms all baselines across every metric: compared to all spatial audio generation baselines, we reduce DOA error by 47% and improve CC and AUC by >239% and 34%. On semantic metrics, our method achieves higher D-CLAP scores, improving them by more than 39% even relative to the state-of-the-art video-to-audio baseline MMAudio, and by more than 117% over spatial-audio baselines. Note that all baselines benefit from outputs produced by proprietary models, yet our method still significantly outperforms them even in the open-source setting. In addition, we estimate the uncertainty via 95% confidence intervals computed as $1.96 \times$ SEM (standard error of the mean): $\Delta_{\text{Angular}} = 0.728 \pm 0.100$, CC$= 0.658 \pm 0.063$, and D-CLAP$_{\text{T}} = 0.457 \pm 0.014$, indicating stable gains. Fig. 4 shows per-scene results. See Supp. for ablation studies.

User Study. We conduct a user study with 50 participants across 12 scenes (6 real scenes from SonoScene360 and 6 synthetic scenes), comparing our method with MMAudio [9] and Omniaudio [54]. For each scene, we render a fixed-trajectory video from our generated visual scene and pair it with each method's spatial audio. Participants perform three pairwise comparisons (Ours vs. MMAudio, Ours vs. Omniaudio, and Omniaudio vs. MMAudio), selecting the preferred video for each pair based on both spatial coherence and audio-visual semantic alignment. Note that the visuals are identical across methods in the user study; only the spatial audio differ. We report pairwise preference rates averaged across scenes. Across both real and synthetic scenes, our method achieves the highest human preference, in line with quantitative evaluation results above.

5.3. Qualitative Results

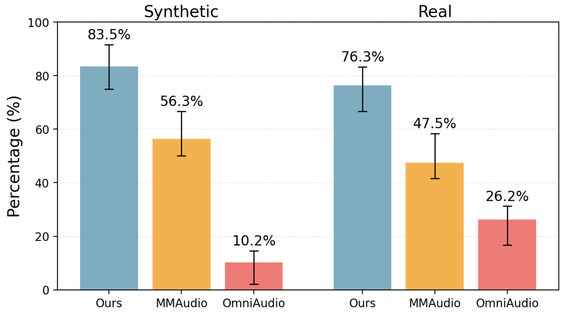


Figure 5. **User Study of Spatial Audio Quality.** Preference rates across *synthetic* and *real* scenes for our method, MMAudio [9], and OmniAudio [54]. Each bar shows the average per-user, per-scene preference (%); error bars indicate the interquartile range (25th–75th percentile) across 50 participants.

Free-Viewpoint Audio-Visual Exploration. Our framework enables interactive free-viewpoint navigation in the generated 3D audio-visual scenes, with an audio callback under 1 ms on an Apple M3 Pro laptop for the *Fountain* scene, well within the 5.3 ms latency of a 256-sample buffer at 48kHz, easily meeting real-time constraints. We further implement a public browser-based viewer in which visuals are rendered with Three.js and spatial audio is synthesized via the WebAudio API, running entirely on a standard laptop CPU. Please check our project website for details of our interactive interface and video demos across diverse scenes.

Ambisonics Energy Map Visualization. We visualize per-direction ambisonics energy maps in Fig. 6 for scenes in SonoScene360. The columns correspond to different scenes and rows to ambisonics predictions from different methods, with brighter regions indicating higher sound energy. On these real scenes, our FOA maps best match the ground truth, with high-energy lobes aligned to the correct directions and spatial extents of sounding objects, while OmniAudio yields oversmoothed or weakly correlated patterns. Please refer to Supp. for qualitative results on synthetic data with camera movements. The *Siren* scene in Fig. 6 shows a typical failure case, where the source (police car) is moving; since our method takes a static image as input, it has no cues about source motion.

5.4. Extensions and Applications

A key property of our framework is that the entire spatial audio rendering pipeline is differentiable, connecting source audio and scene parameters to the resulting spatial audio field via analytic gradients. This enables direct optimization of scene- and source-level variables from audio observations, opening up a range of 3D audio-visual learning tasks beyond scene generation. We highlight two such extensions—*one-shot room acoustic learning* and *audio-visual spatial source separation*—as examples of broader 3D audio-visual applications enabled by our formulation. Please see Supp. for the detailed setup of these two tasks.

One-Shot Room Acoustic Learning. Given a synthesized 3D scene with fixed geometry and visually localized

Method	Spatial Metrics					Semantic Metrics		
	$\Delta_{\text{abs}}\theta \downarrow$	$\Delta_{\text{abs}}\varphi \downarrow$	$\Delta_{\text{Anglular}} \downarrow$	CC ↑	AUC ↑	D-CLAP$_{\text{R}}$ ↑	D-CLAP$_{\text{A}}$ ↑	D-CLAP$_{\text{T}}$ ↑
MMAudio [9]	—	—	—	—	—	33.8%	0.345	0.322
SEE-2-SOUND [14]	1.354	0.254	1.397	0.194	0.603	22.1%	0.221	0.156
ViSAGe [42]	1.598	0.426	1.649	0.142	0.624	22.1%	0.088	0.019
OmniAudio [54]	1.508	0.317	1.449	0.148	0.588	39.7%	0.101	0.104
Ours (Open-source)	0.976	0.251	0.975	0.491	0.753	52.9%	0.464	0.413
Ours (Proprietary)	**0.672**	**0.216**	**0.728**	**0.658**	**0.838**	**67.6%**	**0.480**	**0.457**

Table 1. **Quantitative comparison on SonoScene360.** We report two versions of our method: **Ours (Open-source)**, a fully reproducible version using only open-source models, and **Ours (Proprietary)**, which leverages proprietary models for the best performance.

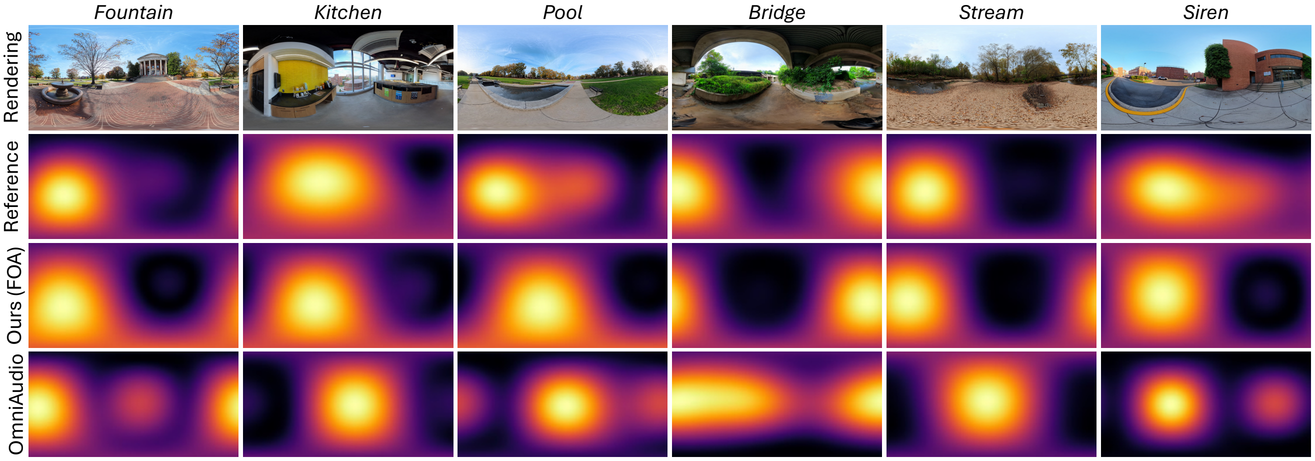


Figure 6. **Ambisonics energy map visualization on real-world scenes.** Top row: panoramic renderings from our method at novel views (microphone poses) on SonoScene360. Second row: ground-truth FOA energy maps. Remaining rows: FOA predictions from our method and OmniAudio [54]. Our maps most closely follow the ground-truth reference spatial patterns, while the baselines produce oversmoothed or misaligned energy. For more qualitative baseline comparisons, please refer to Supp.

Method	$\Delta_{\text{Angular}} \downarrow$	MAG ↓	ENV ↓
NAF [56]	1.76	3.96	3.60
INRAS [68]	1.64	5.06	5.97
AV-NeRF [49]	1.58	4.58	1.89
Ours	**0.22**	**3.46**	**1.22**

Table 2. Results on one-shot room acoustic learning.

sources, we optimize the parameters of our differentiable renderer such that the predicted ambisonics match a target FOA recording at one microphone pose. Tab. 2 shows that our method substantially outperforms neural acoustic field baselines NAF [56], INRAS [68], and AV-NeRF [49] on all metrics, indicating that our differentiable spatial audio renderer can serve as an effective and data-efficient surrogate for room acoustics, even in the challenging one-shot setting.

Audio-Visual Spatial Source Separation. We further treat our renderer as a differentiable spatialization module for audio-visual source separation. Given a mixture spatial recording and visually localized sources in 3D, we estimate per-source waveforms whose rendered sum matches the observed mixture, encouraging each source to explain energy near its corresponding visual region. This provides a spatially grounded separation objective built directly in 3D. We show that we can perform source separation in 3D on FOA recording based on the semantic understanding of the visual scene. See Supp. for demonstration.

6. Conclusion

We presented Image2AVScene—generating a 3D audio-visual scene from a single image—and SonoWorld, the first training-free framework for this task. Our method produces explorable scenes that support free-viewpoint audio-visual rendering: photorealistic novel views paired with perceptually realistic ambisonics aligned with scene geometry and sound source semantics. Quantitative evaluation on spatial audio generation and a user study both confirm the effectiveness of our proposed method. Beyond the core task, we showed preliminary extensions of our differentiable design to one-shot acoustic learning and audio-visual spatial source separation. As future work, we aim to extend our framework to more challenging 3D audio-visual learning tasks and dynamic scenes with moving sources.

Acknowledgments: This project was supported in part by a gift from Dolby Laboratorie and Barry Mersky and Capital One Endowed Professorships. We thank the anonymous reviewers for their useful feedback that helped improve this manuscript.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020. 3

[2] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 3

[3] 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. 21

[4] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *ICCV*, 2023. 2

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 23

[6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 3

[7] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. Deep cross-modal audio-visual generation. In *on Thematic Workshops of ACM Multimedia*, 2017. 1, 2

[8] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *CVPR*, 2025.

[9] 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, 2, 5, 6, 7, 8, 13, 22

[10] Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Jun Chen, Mohamed Elhoseiny, Ruohan Gao, and Dinesh Manocha. Meerkat: Audio-visual large language model for grounding in space and time. In *ECCV*, 2024. 3

[11] 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. 3, 22

[12] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *ArXiv*, abs/2311.13384, 2023. 2

[13] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. In *INTERSPEECH*, 2020. 3

[14] 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, 6, 8, 14

[15] 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. In *SIGGRAPH*, 2018. 3

[16] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open, 2024. 22

[17] Hao Feng, Zhi Zuo, Jia-hui Pan, Ka-hei Hui, Yihua Shao, Qi Dou, Wei Xie, and Zhengzhe Liu. Wonderverse: Extendable 3d scene generation with video generative models. *arXiv preprint arXiv:2503.09160*, 2025. 2

[18] John W Fisher III, Trevor Darrell, William T Freeman, and Paul A Viola. Learning joint statistical models for audio-visual fusion and segregation. In *NeurIPS*, 2001. 3

[19] 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. 21

[20] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 3

[21] Ruohan Gao and Kristen Grauman. 2.5d visual sound. *CVPR*, 2018. 2

[22] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 3

[23] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 3

[24] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 3

[25] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *NeurIPS*, 2024. 2

[26] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *BMVC*, 2021. 2

[27] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Visually-guided audio spatialization in video with geometry-aware multi-task learning. In *IJCV*, 2023. 2

[28] Junlin Hao, Peiheng Wang, Haoyang Wang, Xinggong Zhang, and Zongming Guo. Gaussvideodreamer: 3d scene generation with video diffusion and inconsistency-aware gaussian splatting. *arXiv preprint arXiv:2504.10001*, 2025. 2

[29] 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. 21

[30] Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Mmdisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation. In *ICLR*, 2025. 2

[31] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *NeurIPS*, 2000. 3

[32] Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP*, 2024. 2, 4, 18

[33] 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. *NeurIPS*, 2020. 3

[34] Chao Huang, Yuesheng Ma, Junxuan Huang, Susan Liang, Yunlong Tang, Jing Bi, Wenqiang Liu, Nima Mesgarani, and Chenliang Xu. Zerosep: Separate anything in audio with zero training. *arXiv preprint arXiv:2505.23625*, 2025. 22

[35] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson WH Lau, Wangmeng Zuo, and Chunchao Guo. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *arXiv preprint arXiv:2506.04225*, 2025. 2

[36] Yukun Huang, Jiwen Yu, Yanning Zhou, Jianan Wang, Xintao Wang, Pengfei Wan, and Xihui Liu. Omnix: From unified panoramic generation and perception to graphics-ready 3d scenes. *arXiv preprint arXiv:2510.26800*, 2025. 4

[37] Team HunyuanWorld. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. *arXiv preprint*, 2025. 1, 2, 4, 7, 14

[38] Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022. 4

[39] Derong Jin and Ruohan Gao. Differentiable room acoustic rendering with multi-view vision priors. In *ICCV*, 2025. 2, 21

[40] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 2, 3

[41] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *CVPR*, 2005. 3

[42] Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. *ICLR*, abs/2506.12199, 2025. 2, 4, 6, 8, 14, 18

[43] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *ICLR*, 2023. 1, 2

[44] World Labs. Marble. `https://marble.worldlabs.ai/`, 2025. 1, 4, 7, 14

[45] Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *NeurIPS*, 2024. 2

[46] 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. 21

[47] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng. Scene-aware audio for 360 videos. *TOG*, 2018. 2

[48] Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *ACL*, 2024. 3

[49] Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *NeurIPS*, 2023. 2, 8

[50] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 23

[51] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *ICML*, 2023. 1, 2

[52] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 5, 7

[53] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *TASLP*, 2024. 2

[54] 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. 2, 4, 6, 7, 8, 13, 14, 18, 22

[55] Xiulong Liu, Anurag Kumar, Paul Calamia, Sebastià V. Amengual Garí, Calvin Murdock, Ishwarya Ananthabhotla, Philip Robinson, Eli Shlizerman, Vamsi Krishna Ithapu, and Ruohan Gao. Hearing anywhere in any environment. In *CVPR*, 2025. 2

[56] Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *NeurIPS*, 2022. 2, 8

[57] YU Mark, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *ArXiv*, abs/2503.05638, 2025. 1

[58] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *NeurIPS*, 2022. 3

[59] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *NeurIPS*, 2018. 2

[60] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011. 17

[61] OpenAI. Gpt-5 system card. `https://openai.com/index/gpt-5-system-card/`, 2025. 5, 7, 19

[62] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016. 1, 2

[63] 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. 5, 19

[64] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Muller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed

world-consistent video generation with precise camera control. *ArXiv*, abs/2503.03751, 2025. 1, 2

[65] Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. In *SIGGRAPH Asia*, 2025. 2

[66] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 3

[67] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *ICCV*, 2023. 3

[68] Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *NeurIPS*, 2022. 2, 8

[69] Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yi-Ting Guo. Both ears wide open: Towards language-driven spatial audio generation. *ArXiv*, abs/2410.10676, 2024. 2

[70] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 3

[71] 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. 23

[72] James Traer, Sam V. Norman-Haignere, and Josh H. McDermott. Causal inference in environmental sound recognition. *Cognition*, 214:104627, 2021. 23

[73] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *ICLR*, 2021. 3

[74] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *ECCV*, 2022. 3

[75] Dirk van Maercke and Jacques Martin. The prediction of echograms and impulse responses within the epidaure software. *Applied Acoustics*, 38(2):93–114, 1993. 21

[76] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. 4, 19

[77] Mason Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 21

[78] Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *CVPR*, 2024. 2

[79] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. CAT4D: Create Anything in 4D with Multi-View Video Diffusion Models. *arXiv:2411.18613*, 2024. 2

[80] 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. 4, 19

[81] Siyi Xie, Hanxin Zhu, Tianyu He, Xin Li, and Zhibo Chen. Sonic4d: Spatial audio generation for immersive 4d scene exploration. *ArXiv*, abs/2506.15759, 2025. 2

[82] Ziyang Xie. Worldgen: Generate any 3d scene in seconds. https://github.com/ZiYang-xie/WorldGen, 2025. 2, 4

[83] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *ICCV*, 2019. 3

[84] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. *SIGGRAPH*, 2025. 1, 2, 4

[85] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1

[86] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T. Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, and Charles Herrmann. Wonderjourney: Going from anywhere to everywhere. *CVPR*, 2023. 1, 2

[87] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, 2024. 1, 2

[88] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 2

[89] Qihang Zhang, Yinghao Xu, Chaoyang Wang, Hsin-Ying Lee, Gordon Wetzstein, Bolei Zhou, and Ceyuan Yang. 3ditscene: Editing any scene via language-guided disentangled gaussian splatting. *arXiv preprint arXiv:2405.18424*, 2024. 2

[90] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh H. McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 3

[91] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 3

[92] Haiyang Zhou, Xinhua Cheng, Wangbo Yu, Yonghong Tian, and Li Yuan. Holodreamer: Holistic 3d panoramic world generation from text descriptions. *arXiv preprint arXiv:2407.15187*, 2024. 2

[93] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, 2022. 3

[94] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint*, 2025. 1

[95] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang

Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *ECCV*, 2024. 2

[96] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, 2018. 2

[97] 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. 3, 5, 19

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

Supplementary Material

Contents

A. 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 [54] and MMAudio [9] 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.

B. Ablation Studies

Table 3 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, AUC, and D-CLAP$_{\text{T}}$, 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 11 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.

C. More Qualitative Results

C.1. Synthetic and Real-World Scenes

In Fig. 8, 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.

	Spatial Metrics					Semantic Metrics		
Method	$\Delta_{\text{abs}}\theta \downarrow$	$\Delta_{\text{abs}}\varphi \downarrow$	$\Delta_{\text{Anglular}} \downarrow$	CC ↑	AUC ↑	D-CLAP$_{\text{R}}$ ↑	D-CLAP$_{\text{A}}$ ↑	D-CLAP$_{\text{T}}$ ↑
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 3. **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.

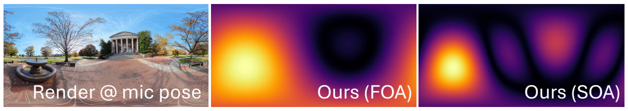


Figure 7. The sound energy by second order ambisonics (SOA) in real scenes.

Figure 9 provides additional qualitative comparisons on 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 [54], SEE-2-SOUND (S2S) [14], and ViSAGe [42]). 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. 7, we show the in real scenes, SOA also yields sharper, more concentrated energy maps around the sounding objects.

C.2. 3D Scene Backbones

We next analyze the impact of different 3D scene backbones on the visual quality of the reconstructed environments. Figure 10 compares HunyuanWorld 1.0 [37] (with mesh output) and Marble [44] (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 oversmoothing 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 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.

C.3. Panoramic Instance Merging

Our panoramic grounding pipeline combines class-agnostic proposals (from SAM2-style segmentation) with open-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 11 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.

C.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 12 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.

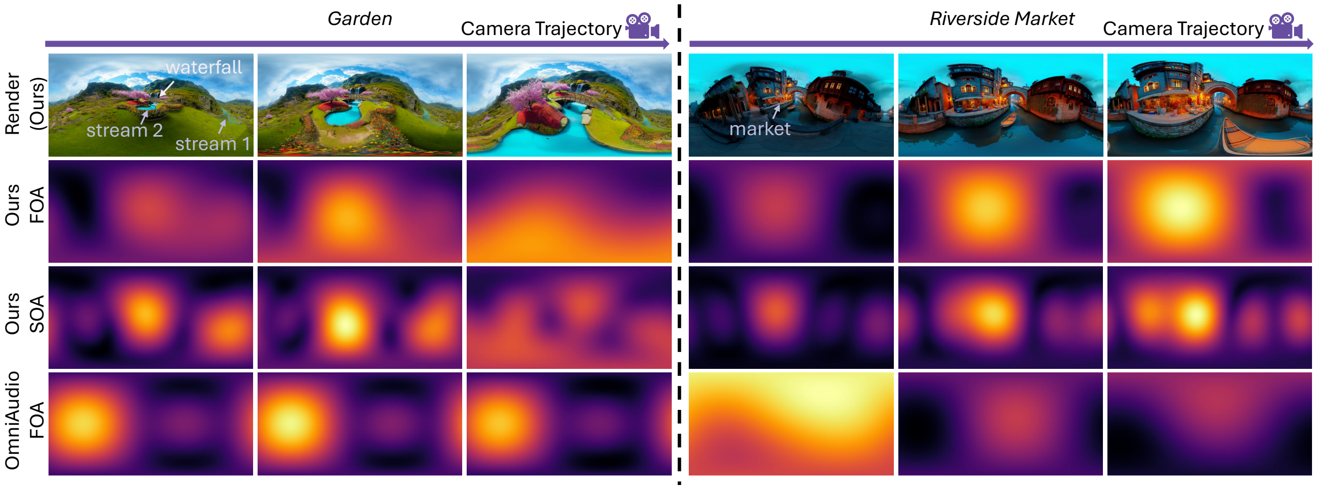


Figure 8. **Qualitative results on synthetic scenes.** From a single diffusion-generated image, SonoWorld reconstructs a 3D Gaussian scene and predicts a spatial audio field that supports free-viewpoint exploration. We show two synthetic scenes (*Garden* and *Riverside Market*), sample views along a 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.

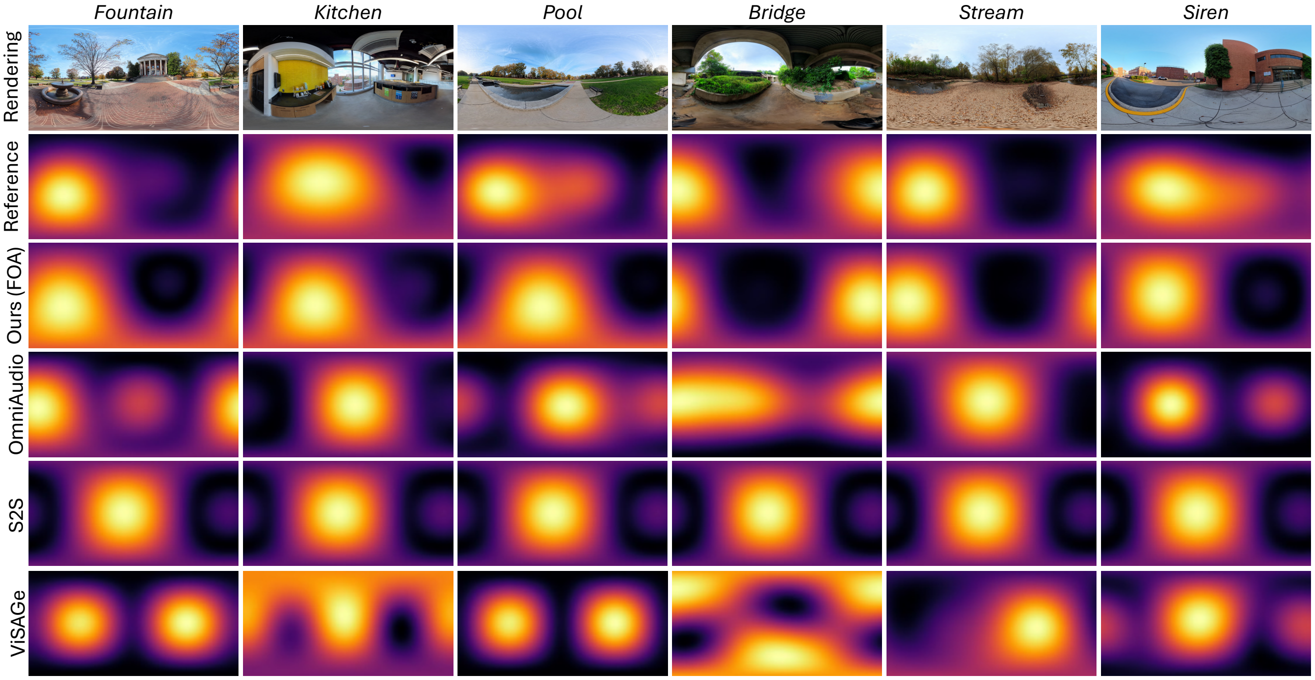


Figure 9. **Qualitative comparison on real scenes.** Additional qualitative results on six real environments from SonoScene360. For each scene, we show the rendered panorama from our 3D Gaussian reconstruction, the reference spherical energy map derived from the recorded FOA (second row), and the predicted 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.

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 13 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 re-

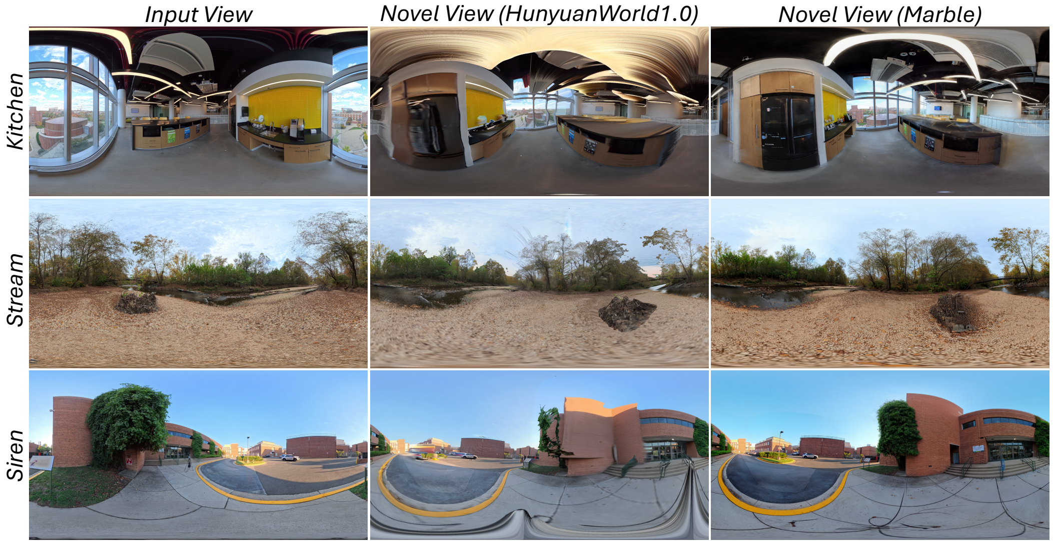


Figure 10. **Comparison of 3D scene backbones.** Given an input 360° panorama (left column), we compare novel views generated by HunyuanWorld 1.0 and Marble for several real scenes (*Kitchen*, *Stream*, *Siren*). HunyuanWorld produces globally coherent but sometimes distorted geometry (e.g., warped floors 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.

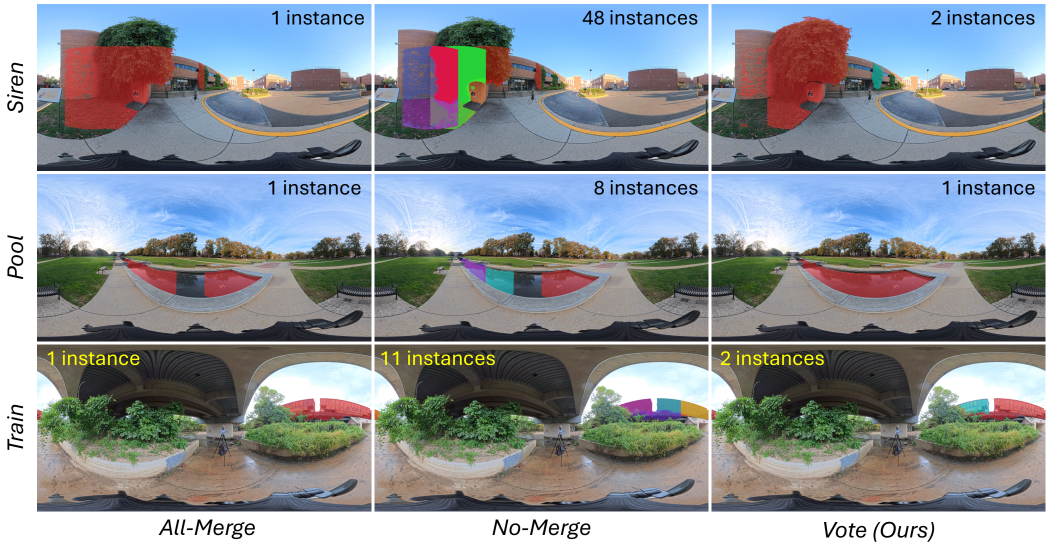


Figure 11. **Panoramic instance merging strategies.** We compare three strategies for grouping class-agnostic region proposals into semantic instances in SonoScene360: *All-Merge* (left), *No-Merge* (middle), and our voting-based strategy *Vote (Ours)* (right). For each scene, we overlay the resulting 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.

constructed 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.

D. More Details For SonoScene360 Dataset

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

- 3D Visual Scene Representation (3DGS, for ours).
- Spatial Audio Field $\mathbf{A}$ (Ambisonics Field in our case)

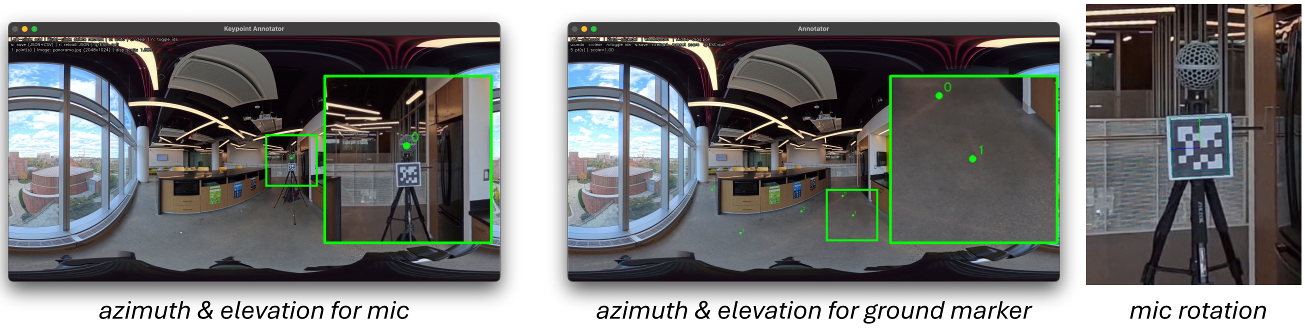

Keypoint Annotator
Annotator
azimuth & elevation for mic
azimuth & elevation for ground marker
mic rotation

Figure 12. **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.

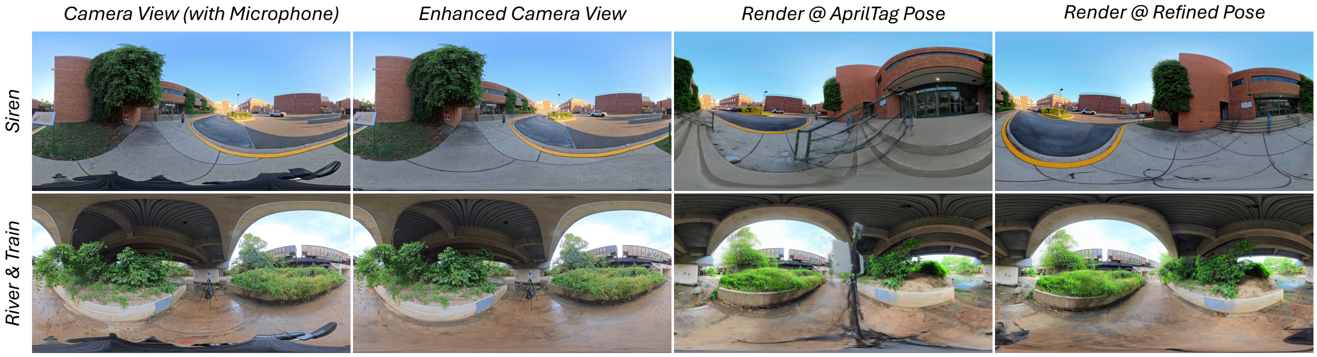


Figure 13. **Effect of calibration on rendered scenes.** For two real scenes (*Siren* and *River & Train*), we show the original camera view with the microphone visible, the inpainted camera view, and renderings of the reconstructed 3D scene when aligned using the AprilTag-based pose estimate vs. our 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.

D.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 [60] 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 [60] 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 system, 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 scele and real pose, we further refine the orientation using human annotation. For each scene, the annotator:

1. selects a the middle of the micorphone 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 istead 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

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

D.2. Dataset Statistics

Table 4 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.

E. Metric Definitions

E.1. Spatial Metrics

DoA from FOA intensity. Following [32, 54] for first-order ambisonics (FOA, $L = 1$), we write the four channels as

$$\mathbf{a}_1(t) = [W(t), Y(t), Z(t), X(t)]^\top, \tag{12}$$

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:

$$I_X = \sum_t W(t)\, X(t), \tag{13}$$

$$I_Y = \sum_t W(t)\, Y(t), \tag{14}$$

$$I_Z = \sum_t W(t)\, Z(t), \tag{15}$$

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:

$$\theta = \operatorname{atan2}(I_Y, I_X), \tag{16}$$

$$\varphi = \operatorname{atan2}\left(I_Z,\ \sqrt{I_X^2 + I_Y^2}\right), \tag{17}$$

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**

$$\Delta_{\text{abs}}\theta = \min\left(|\theta_{\text{gt}} - \theta_{\text{pred}}|, 2\pi - |\theta_{\text{gt}} - \theta_{\text{pred}}|\right). \tag{18}$$

2. **Elevation error**

$$\Delta_{\text{abs}}\varphi = \left|\varphi_{\text{gt}} - \varphi_{\text{pred}}\right|. \tag{19}$$

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

$$\mathbf{u}_{\text{gt}} = \begin{bmatrix} \cos\varphi_{\text{gt}} \cos\theta_{\text{gt}} \\ \cos\varphi_{\text{gt}} \sin\theta_{\text{gt}} \\ \sin\varphi_{\text{gt}} \end{bmatrix}, \quad \mathbf{u}_{\text{pred}} = \begin{bmatrix} \cos\varphi_{\text{pred}} \cos\theta_{\text{pred}} \\ \cos\varphi_{\text{pred}} \sin\theta_{\text{pred}} \\ \sin\varphi_{\text{pred}} \end{bmatrix}, \tag{20}$$

and compute the geodesic distance

$$\Delta_{\text{Angular}} = \arccos\left(\operatorname{clip}(\mathbf{u}_{\text{gt}}^\top \mathbf{u}_{\text{pred}}, -1, 1)\right). \tag{21}$$

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

$$a = \sin^2\left(\frac{\Delta_{\text{abs}}\theta}{2}\right) + \cos\varphi_{\text{pred}} \cos\varphi_{\text{gt}} \sin^2\left(\frac{\Delta_{\text{abs}}\varphi}{2}\right), \tag{22}$$

$$\Delta_{\text{Angular}} = 2\arctan\left(\sqrt{\frac{a}{1-a}}\right). \tag{23}$$

Spherical energy maps, CC and AUC. Following [42], given ambisonics coefficients $a_L(t)$, we render the scalar pressure at a direction (θ, φ) as

$$a(\theta, \varphi, t) = y_L(\theta, \varphi)^\top a_L(t). \tag{24}$$

We then compute a time-aggregated energy map

$$E(\theta, \varphi) = \sum_t |a(\theta, \varphi, t)|^2, \tag{25}$$

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

$$\text{CC} = \frac{\sum_i (\mathbf{e}_{\text{pred},i} - \bar{e}_{\text{pred}})(\mathbf{e}_{\text{gt},i} - \bar{e}_{\text{gt}})}{\sqrt{\sum_i (\mathbf{e}_{\text{pred},i} - \bar{e}_{\text{pred}})^2} \sqrt{\sum_i (\mathbf{e}_{\text{gt},i} - \bar{e}_{\text{gt}})^2}}. \tag{26}$$

To compute AUC, we treat $\mathbf{e}_{\text{gt}}$ as a soft foreground mask by binarizing it at its median value. Using $\mathbf{e}_{\text{pred}}$ as scores, we form the ROC curve over all thresholds and report the area under the curve (AUC), following ViSAGe [42]. Higher CC and AUC indicate closer spatial energy patterns to the reference.

E.2. Semantic Metrics

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

$$\text{left:} \quad \mathbf{u}_{\text{L}} = (\theta = \tfrac{\pi}{2}, \varphi = 0), \tag{27}$$

$$\text{right:} \quad \mathbf{u}_{\text{R}} = (\theta = -\tfrac{\pi}{2}, \varphi = 0), \tag{28}$$

$$\text{front:} \quad \mathbf{u}_{\text{F}} = (\theta = 0, \varphi = 0), \tag{29}$$

$$\text{back:} \quad \mathbf{u}_{\text{B}} = (\theta = \pi, \varphi = 0). \tag{30}$$

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

$$a_d(t) = y_L(\mathbf{u}_d)^\top a_L(t). \tag{31}$$

Let $f_a(\cdot)$ and $f_t(\cdot)$ be the audio and text encoders of CLAP [80]. For a caption c describing a sounding source and directional audio a_d, we define

$$s_{\mathrm{CLAP-T}}(d, c) = \cos\left(f_a(a_d^{\mathrm{pred}}), f_t(c)\right), \tag{32}$$

$$s_{\mathrm{CLAP-A}}(d) = \cos\left(f_a(a_d^{\mathrm{pred}}), f_a(a_d^{\mathrm{gt}})\right), \tag{33}$$

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

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

All metrics are averaged over clips in SONOSCENE360.

F. More Details For SONOWORLD Framework

F.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 14 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.

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

$$\theta = 2\pi\left(\frac{u + 0.5}{W_{\text{pano}}} - \frac{1}{2}\right), \tag{34}$$

$$\varphi = \pi\left(\frac{v + 0.5}{H_{\text{pano}}} - \frac{1}{2}\right), \tag{35}$$

and obtain a direction vector

$$\mathbf{d}(\theta, \varphi) = \begin{bmatrix} \cos\varphi\cos\theta \\ \cos\varphi\sin\theta \\ \sin\varphi \end{bmatrix}. \tag{36}$$

F.2. Gaussian-Pyramid-Based Warping

Given the camera extrinsics from GeoCalib [76], 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.

Algorithm 1 Mask Voting for Category c

```
Input: $\mathbf{M}_{\text{pano}}$, $\mathbf{M}_{\text{OVS},c}$, $\tau_{\text{vote}}$, $\tau_{\text{IOU}}$
Output: $\mathcal{M}_c$
$\mathbf{M}_c \leftarrow \emptyset$
for each $\mathcal{M}_i^{\text{pano}} \in \mathbf{M}_{\text{pano}}$ do
    $s_i \leftarrow \frac{\sum_p \textsc{PixelScore}(p)}{\textsc{Visibility}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})}$
    if $s_i \geq \tau_{\text{vote}}$ then
        $\widehat{\mathcal{M}}_i \leftarrow \mathcal{M}_i^{\text{pano}} \vee \textsc{Combine}(\mathcal{M}_i^{\text{pano}}, \mathcal{M}_{\text{OVS},c})$
        $\mathbf{M}_c \leftarrow \mathbf{M}_c \cup \{\widehat{\mathcal{M}}_i\}$
Return $\mathbf{M}_c$
```

Gaussian pyramid and anti-aliased sampling. Foreshortening 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 $\{I^{(s)}\}_{s=0}^{S-1}$ from the input image I, where $I^{(0)} = I$ and

$$I^{(s+1)} = \text{downsample}_2(\text{GaussianBlur}(I^{(s)})). \tag{37}$$

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

$$s^\star = \text{clip}(\lfloor \log_2 \rho \rceil, 0, S - 1), \tag{38}$$

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

$$I_{\text{warp}} = \mathcal{W}_G(I, \varphi, f), \tag{39}$$

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 [63] proposals and X-Decoder [97] 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 [61], Alg. 1 computes a vote score s_i by aggregating per-pixel confidences from overlapping open-vocabulary masks $\mathbf{M}_{\text{OVS},c}$. $\textsc{PixelScore}(p)$ 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}. $\textsc{Visibility}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})$ counts the pixels in $\mathcal{M}_i^{\text{pano}}$ that receive at least one such vote and normalizes the sum of $\textsc{PixelScore}(p)$ to obtain s_i. If $s_i \geq \tau_{\text{vote}}$, $\textsc{Combine}(\mathcal{M}_i^{\text{pano}}, \mathbf{M}_{\text{OVS},c})$ takes the union of all contributing instance masks (those with 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,

~~~~
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 14. VLM prompt used to query the vision-language model.

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:

$$w_j = \frac{d_j^2 \cos e_j}{\max\left(|\mathbf{n}_j^\top \mathbf{v}_j|, \varepsilon\right)}\,, \tag{40}$$

where the $\cos e_j$ term compensates for equirectangular sampling density, d_j^2 implements distance-based weigthing, 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 $N_{\max} = 1000$ 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 $\mathcal{P}_i$:

$$\mathcal{P}_i = \{\mathbf{x}_{ik}\}_{k=1}^{K_i}, \qquad \mathbf{x}_{ik} \sim \mathcal{P}_{\text{raw},i} \tag{41}$$

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

F.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

$$b_{\text{left}}(t) = \sum_{\theta,\varphi} \sum_{\tau} \left(h^{\text{left}}(\theta, \varphi, \tau) \cdot a(\theta, \varphi, t - \tau)\right), \tag{42}$$

$$b_{\text{right}}(t) = \sum_{\theta,\varphi} \sum_{\tau} \left(h^{\text{right}}(\theta, \varphi, \tau) \cdot a(\theta, \varphi, t - \tau)\right), \tag{43}$$

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

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

$$a(\theta,\varphi,t) = \sum_{\ell,m} Y_\ell^m(\theta,\varphi)\, a_{\ell,m}(t), \tag{44}$$

we can precompute ambisonics-domain HRIRs:

$$h_{\ell,m}^{\text{left/right}}(\tau) = \sum_{\theta,\varphi} w(\theta,\varphi)\, h^{\text{left/right}}(\theta,\varphi,\tau)\, Y_\ell^m(\theta,\varphi), \tag{45}$$

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

$$\begin{bmatrix} b_{\text{left}}(t) \\ b_{\text{right}}(t) \end{bmatrix} = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} \begin{bmatrix} h_{\ell,m}^{\text{left}} * a_{\ell,m} \\ h_{\ell,m}^{\text{right}} * a_{\ell,m} \end{bmatrix}(t), \tag{46}$$

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 [3] dataset).

G. 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

$$\mathbf{A}(\mathbf{p},t;\theta) = \big[\mathrm{RIR}_L(\mathbf{p},\cdot;\theta) * a_{\text{src}}(\cdot)\big](t), \tag{47}$$

where $\mathrm{RIR}_L(\mathbf{p},t;\theta) \in \mathbb{R}^{(L+1)^2}$ is the ambisonic room impulse response that models the transfer function between the source and $\mathbf{p}$. Following [39, 77], we decompose it into early reflections and late diffuse reverberation:

$$\mathrm{RIR}_L(\mathbf{p},t;\theta) = \mathrm{Blend}[\mathrm{RIR}_L^{\text{early}}(\mathbf{p},t;\theta), \mathrm{RIR}_L^{\text{late}}(\mathbf{p},t;\theta)]. \tag{48}$$

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 [19, 29, 39, 46, 75]):

$$\begin{aligned} &\mathrm{RIR}_L^{\text{early}}(\mathbf{p},t;\theta) && (49)\\ =&\frac{se^{-\alpha t}}{c_{\text{sound}}\tau_p} \sum_{p\in\mathbb{P}} \mathbf{y}_L(\mathbf{d}_p)\, \mathcal{F}_{\text{min}}^{-1}\Big\{R[f]^{|p|}\Big\}(t-\tau_p), && (50) \end{aligned}$$

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 $\mathcal{F}_{\text{min}}^{-1}$ 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 aborption.

Late reverberation parameterization from RT60. The late part $\mathrm{RIR}_L^{\text{late}}$ 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

$$e_b(t) = \exp\left(-\frac{\ln(1000)}{\hat{T}_{60}(f_b)}\,t\right), \tag{51}$$

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

$$r_b(t;\theta) = e_b(t)\,\tilde{n}_b(t), \tag{52}$$

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

$$r^{\text{late}}(t;\theta) = \sigma(g) \sum_{b=1}^{B} r_b(t;\theta), \tag{53}$$

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:

$$\mathrm{RIR}_L^{\text{late}}(\mathbf{p},t;\theta) = r^{\text{late}}(t;\theta)\,\mathbf{1}, \tag{54}$$

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:

$$\begin{aligned} &\mathrm{RIR}_L(\mathbf{p},t;\theta) && (55)\\ =&w_{\text{early}}(t)\,\mathrm{RIR}_L^{\text{early}}(\mathbf{p},t;\theta) + w_{\text{late}}(t)\,\mathrm{RIR}_L^{\text{late}}(\mathbf{p},t;\theta), && (56) \end{aligned}$$

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 $\mathrm{RIR}_L^{\text{late}}$ such that its peak is comparable to the peak energy of $\mathrm{RIR}_L^{\text{early}}$ 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:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{MAG}} \tag{57}$$

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

$$\mathcal{L}_{\text{MAG}} = \big\| \log |S(A(\tilde{p}, t; \theta))| - \log |S(\tilde{a}_L(t))| \big\|_1, \tag{58}$$

$$\mathcal{L}_{\text{ENV}} = \big\| \mathrm{Env}(A(\tilde{p}, t; \theta)) - \mathrm{Env}(\tilde{a}_L(t)) \big\|_2^2, \tag{59}$$

$S(\cdot)$ is the STFT, $\mathrm{Env}(\cdot)$ 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.

H. 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 $\tilde{p}$,

we seek per-source monaural signals $\{s_i(t)\}_{i\in\mathcal{O}}$ such that

$$\sum_{i\in\mathcal{O}} \mathbf{A}_i(\tilde{p}, t; s_i) \approx \tilde{a}_L(t), \tag{60}$$

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:

$$\mathcal{L}_{\text{sep}} = \mathcal{L}_{\text{MAG}}(\textstyle\sum_i \mathbf{A}_i, \tilde{a}_L), \tag{61}$$

where $\mathcal{L}_{\text{MAG}}$ is as above. In practice, we implement separation via diffusion posterior sampling [11] over the per-source latents, guided by $\mathcal{L}_{\text{sep}}$.

Source Separation by Diffusion Posterior Sampling [11] 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

$$p_{\text{post}}\big(\{x_t^{(i)}\}_{i\in\mathcal{O}} \mid \tilde{a}_L\big) \propto \left[\prod_{i\in\mathcal{O}} p_{\text{prior}}(x_t^{(i)})\right] \exp\big(-\lambda\,\mathcal{L}_{\text{sep}}\big), \tag{62}$$

where λ controls the strength of the guidance.

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

$$\begin{aligned} &\nabla \log p_{\text{post}}\big(x_t^{(j)}\big) && (63)\\ \approx &\nabla \log p_{\text{prior}}\big(x_t^{(j)}\big) - \lambda\, \nabla_{x_t^{(j)}} \mathcal{L}_{\text{sep}}\Big(\{\mathbf{A}_i(\tilde{\mathbf{p}}, t; x_t^{(i)})\}, \tilde{a}_L(t)\Big), && (64)\end{aligned}$$

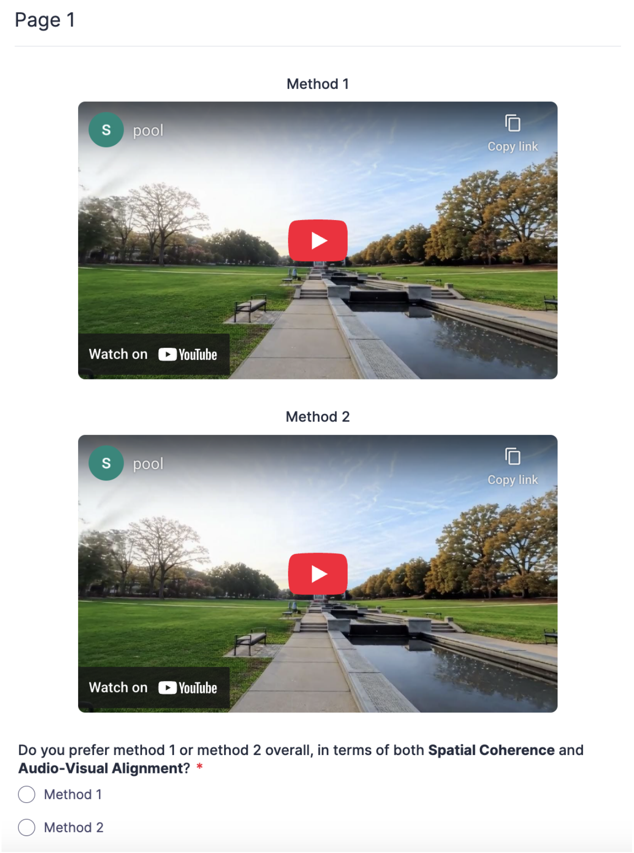
where the second term backpropagates the separation loss through the renderer and the text-to-audio decoder into the latent of source j. This *posterior-guided* score replaces the unconditional prior score in the reverse diffusion update, yielding a *diffusion posterior sampler* that steers each source 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 $\mathcal{L}_{\text{sep}}$ 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 [34] 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 [16] using the `DPM++(3M) SDE` solver.

I. User Study Details

We conduct an anonymized user study to evaluate the spatial audio generation quality of our method in comparison to MMAudio [9] and OmniAudio [54]. 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



Page 1

Method 1

Method 2

Do you prefer method 1 or method 2 overall, in terms of both **Spatial Coherence** and **Audio-Visual Alignment?** *

○ Method 1

○ Method 2

Figure 15. **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.

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 15 shows an example question from our user study form.

J. Discussions

Outdoor-focused propagation and reverberation. Our propagation model is designed primarily for outdoor scenes, where recordings are often close to dry [71] and listeners generally have weaker expectations of strong reverberation [72]. 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. 16).

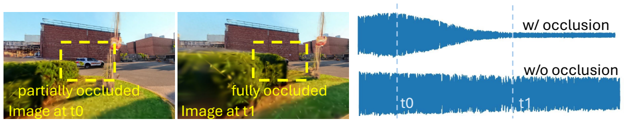


Figure 16. Physics-aware occlusion add-on: as a clustered source becomes partially and then fully blocked between t_0 and t_1, visibility-based point reweighting yields smooth attenuation of the rendered signal.

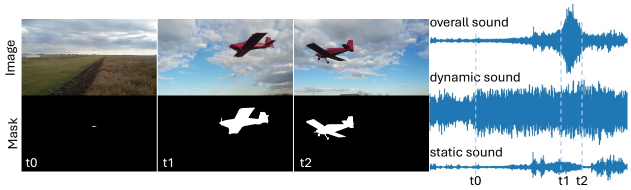


Figure 17. Dynamic source rendering on a plane fly-by clip: a time-varying 3D anchor from SAM3 [5] and Depth Anything 3 [50] separates dynamic and static sound components, with the dynamic layer peaking near closest approach and attenuating with distance

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 [5] and Depth Anything 3 [50] 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 audio-visual scene generation (Fig. 17).