ENABLING IMMERSIVE AUDIO-VISUAL EXPERIENCE FROM ANY VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Most videos capture only a narrow field of view and provide no spatial audio, limiting the sense of immersion they can provide. Recent video generation models can expand perspective videos into panoramic ones, but do not provide the corresponding spatial soundscape. Without spatially consistent audio, these expanded visual worlds remain incomplete. This paper presents OmniDream, a training-free framework that transforms a silent monocular video into an immersive audiovisual experience, where viewers can freely look around while sounds remain spatially aligned with the visual scene. At the core of OmniDream is an object-centric audio representation that disentangles each sound source's intrinsic audio content from its scene-dependent acoustic effects, enabling independent audio generation, physics-based simulation of propagation effects, and flexible spatial audio rendering. Experiments show improved audio-visual alignment, spatial correctness, and perceptual immersiveness over baselines. Examples are available on Hugging Face.



Figure 1: OmniDream transforms a silent perspective video into a 360° video with synchronized spatial audio. As users turn their head and inspect different parts of the scene, both the visual perspective and perceived sound change consistently with the environment.

1 INTRODUCTION

Most videos today offer a passive viewing experience. Viewers are restricted to the recorded viewing direction, while their soundtracks are often absent or non-spatial. As a result, users remain observers of a scene rather than participants within it. In contrast, immersive media such as 360° video with spatial audio allows users to turn their heads and explore different parts of the environment while sounds remain naturally and spatially anchored to the scene. For instance, as the user turns toward a waterfall, the rushing water grows louder and the pitch gets higher; as a train passes by, its sound moves across the scene, with directional cues reflecting its changing position. Together, these visual and auditory cues create a substantially stronger sense of presence and realism than ordinary videos.

Despite growing support for immersive playback (e.g., 360° video and spatial audio on YouTube), immersive content remains scarce (Yang et al., 2025; Tan et al., 2024). This is because high-quality immersive media typically requires specialized capture equipment, including panoramic cameras and microphone arrays, along with carefully calibrated production pipelines (Morgado et al., 2018; Liu et al., 2025). In contrast, ordinary perspective videos are abundant on the internet and increasingly easy to create with video generation models (Wan, 2025; Kong et al., 2024). This raises a question: *can we transform an ordinary video into an immersive audio-visual experience?*

Existing methods address separate components of immersive audio-visual generation. Video-to-audio models synthesize sound for perspective videos (Chen et al., 2025b; Cheng et al., 2025; Liu et al., 2024; Luo et al., 2023; Tian et al., 2025; Wang et al., 2025), while spatial audio methods additionally

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

provide directional cues (Karchkhadze et al., 2025; Kim et al., 2025; Liu et al., 2025; Xie et al., 2026; Zhang et al., 2025a; Zhao et al., 2025); neither expands the visual scene beyond the input's field of view. Conversely, perspective-to-360° video generation expands the visual environment but leaves the audio unaddressed (Tan et al., 2024; Fang et al., 2025; Luo et al., 2025; Li et al., 2026).

Bringing these capabilities together requires an audio representation that connects individual sound sources to the expanded visual scene. This is challenging because an immersive soundscape must respond to changes in the listener's orientation while keeping sound sources anchored to the scene. A pre-rendered binaural mixture cannot adapt on its own to arbitrary listener orientations, while first-order ambisonics (FOA) supports rotation but limits directional resolution through a first-order spherical harmonic approximation (Ben-Hur et al., 2021; Frank et al., 2015). Additionally, generating directly into these formats ties the audio representation to the chosen playback format.

We address this challenge through *object-centric audio generation and rendering*: we generate a dry audio track for each sounding object, simulate its source-to-listener impulse response (IR), and render its acoustic contribution before combining the sources into the final soundscape. This IR captures propagation delays, attenuation, and reverberation determined by source location, scene geometry, and materials (Luo et al., 2022; Wang et al., 2024; Lan et al., 2024). This formulation offers multiple advantages: (i) **object-level audio-visual grounding**, generating each dry track from its corresponding visual object and associating it with an explicit source trajectory in the 3D scene reconstructed from video, accounting for object motion and changes in camera viewpoint; (ii) **physics-grounded acoustic propagation modeling**, computing each source's IR through acoustic simulation in this reconstructed 3D scene using the source trajectories and estimated material properties; and (iii) **steerable and flexible spatial rendering**, retaining individual sources for rendering at different listening orientations and into different output formats, including binaural audio and ambisonics (Frank et al., 2015).

Building on this formulation, we present OmniDream, a training-free framework that transforms a silent perspective video into a panoramic video with spatial audio. OmniDream improves overall spatial correlation between audio and visual content by about 52% relative to the strongest baselines. In a study with 20 participants, OmniDream improves spatial correctness ratings by about 34% and overall immersiveness ratings by about 28% relative to the strongest baseline. Ablations show that object-centric tracking and full acoustic rendering contribute to spatial correspondence. Against measured IRs, the zero-shot renderer has a multi-resolution STFT error of 0.58.

In summary, our contributions are as follows:

- We present the first framework for immersive audio-visual generation that transforms a silent perspective video into a 360° video with spatially and temporally aligned audio.
- We introduce a novel object-centric formulation of audio generation and rendering that combines sound synthesis with physically simulated propagation, enabling scene-grounded, steerable audio.
- We extensively evaluate our system through objective metrics and a user study, demonstrating improvements over baseline methods. Robustness analyses and ablation studies further assess its reliability and the contribution of each component.

2 Related Work

Tab. 1 compares existing methods in terms of their visual and audio outputs, audiovisual synchronization, physics-aware rendering, and support for separate object audio tracks.

Video Outpainting and Generation. Video outpainting aims to extend the spatial extent of a video beyond its original field of view while maintaining temporal consistency and visual coherence. Early work adapted image outpainting techniques to the video domain through propagation-based or generative inpainting strategies (Huang et al., 2016; Dehan et al., 2022), and more recent approaches leverage diffusion-based video generation models to produce high-quality extrapolations with improved temporal stability(Chen et al., 2025a; Yu et al., 2025). A related line of research tackles the more ambitious goal of generating full 360° panoramic video from a narrow perspective input (Tan et al., 2024; Fang et al., 2025; Luo et al., 2025; Li et al., 2026; Wu et al., 2026). These methods demonstrate impressive visual results, yet the generated contents carry no information about audio . In our work, the expanded panoramic video serves not only as the final visual output but also as the geometric and semantic basis for immersive audio generation.

Video-to-Audio Generation. Video-to-audio generation studies how to synthesize plausible soundtracks from visual input. Recent methods have shown a strong ability to generate semantically

Method	Input	Visual Output	Audio Output	Audio-Visual Synchronization	Physics-Aware Rendering	Multi-Track Audio
MMAudio (Cheng et al., 2025)	Video	Video	Mono	✓	✗	✗
Sonic4D (Xie et al., 2026)	Video	Video	Stereo	✓	✗	✗
StereoFoley (Karchkhadze et al., 2025)	Video	Video	Stereo	✓	✗	✗
ViSAGe (Kim et al., 2025)	Video	Video	FOA	✓	✗	✗
OmniAudio (Liu et al., 2025)	360° video	360° video	FOA	✓	✗	✗
See2Sound (Dagli et al., 2025)	Image	Image	5.1 Sound	✗	✗	✓
AVObject (Li et al., 2025)	Image	Image	Mono	✗	✗	✓
SonoWorld (Jin et al., 2026)	Image	3D static scene	FOA	✗	✗	✓
Argus (Luo et al., 2025)	Video	360° video	No Audio	✗	✗	✗
CubeComposer (Li et al., 2026)	Video	360° video	No Audio	✗	✗	✗
OmniDream	Video	360° video	FOA	✓	✓	✓

Table 1: Comparison of immersive media generation methods. First-Order Ambisonics (FOA) is a spatial audio format with direction-dependent sound perception. *Physics-Aware Rendering* indicates whether the method models geometry-based acoustic propagation and environmental effects. *Multi-Track Audio* indicates whether the method decomposes the scene into individual sounding objects.

meaningful and temporally aligned audio from silent videos (Chen et al., 2025b; Cheng et al., 2025; Liu et al., 2024; Luo et al., 2023; Tian et al., 2025; Wang et al., 2025). However, most of these approaches focus on non-spatial audio and operate only within the original camera view. As a result, they do not model where sounds originate in space or how the surrounding environment shapes what is heard. In addition, existing works also explore generating audio from images (Jin et al., 2026; Dagli et al., 2025; Li et al., 2025). However, these models lack video-audio synchronization, and the visual experience remains static. In contrast, our goal is to lift a silent video into an immersive experience with temporally and spatially aligned visual and audio contents.

Spatial Audio Generation. Spatial audio methods generate directional sound in binaural (Zhang et al., 2025a; Zhao et al., 2025), stereo (Sun et al., 2024), or FOA formats (Heydari et al., 2025). These methods condition on perspective video (Gao & Grauman, 2019; Li et al., 2024; Zhang et al., 2025a; Kim et al., 2025; Karchkhadze et al., 2025; Xie et al., 2026; Zhao et al., 2025; Zhang et al., 2025b) or panoramic video, as in OmniAudio (Liu et al., 2025). Pre-rendered binaural audio cannot adapt on its own to head rotation, while FOA supports rotation with limited directional resolution due to its first-order spherical harmonic approximation (Ben-Hur et al., 2021; Frank et al., 2015). Our formulation retains individual sources and their acoustic responses for rendering into different output formats. Existing methods rely on the supplied visual scene; we expand a silent perspective video into a panoramic scene and simulate sound propagation through its reconstructed geometry.

3 METHOD

Given a silent monocular perspective video, OmniDream generates a temporally consistent panoramic video $\mathbf{P}$ together with first-order ambisonic (FOA) audio $\mathbf{A}$ that matches with the video. Fig. 2 illustrates the full pipeline, which comprises three components. **Immersive Visual Expansion** (§3.1) expands the narrow field of view into a full panoramic scene and reconstructs the surrounding 3D geometry. **Audio Content Generation** (§3.2) identifies individual sounding objects in the expanded scene and generates a separate dry audio track for each of them. **Physics-Grounded Acoustic Rendering** (§3.3) renders each dry sound into final wet sound using impulse responses simulated from the recovered scene geometry. This decomposition grounds spatialization in scene structure and enables interpretable, controllable immersive synthesis.

3.1 IMMERSIVE VISUAL EXPANSION

This component expands the current perspective view and constructs a geometrically consistent panoramic scene. The resulting scene serves three roles: the visual output delivered to the viewer, the visual context for per-object audio generation, and the geometric foundation for acoustic simulation.

Panoramic video expansion. We use an off-the-shelf model (Li et al., 2026) to generate a panoramic video $\mathbf{P} = \{p_t\}_{t=1}^{T}$ from $\mathbf{V}$. The downstream audio pipeline operates on the panoramic video and recovered scene representation rather than a specific visual model, making OmniDream compatible with different visual frontends that provide these outputs.

Depth estimation for mesh reconstruction. To recover the 3D layout of the surrounding environment, we apply a panoramic video depth estimation model (Lin et al., 2025) to each frame of $\mathbf{P}$, producing a per-frame depth map $\{d_t\}_{t=1}^{T}$. In our pipeline, we treat these per-frame predictions as relative depth and use a Vision Language Model (VLM) (Comanici et al., 2025) to estimate a metric scale for the scene. To extract the mesh from the depth images, we back-project d_t into a point cloud $\mathcal{C}_t$ for each frame t, and triangulate it by connecting neighboring pixels in the equirectangular

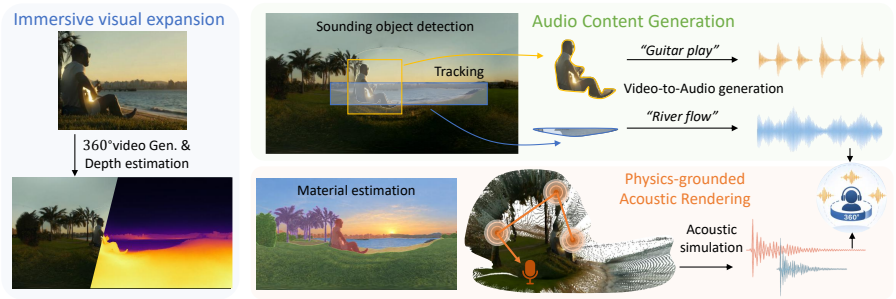

Figure 2: **OmniDream pipeline**. We first expand the narrow perspective video to 360° video and reconstruct the scene geometry (Sec. 3.1). Based on this, we analyze the sounding object in the scene and track their positions through the video and generate per-object dry sound (Sec. 3.2). Finally, we estimate the acoustic properties on the extracted scene mesh and perform ray tracing to simulate the acoustic effect for each object independently and render the final spatial audio (Sec. 3.3).

grid. To avoid bridging foreground and background across depth discontinuities, we discard any triangle whose vertex depths differ by more than a threshold τ, yielding an open triangle mesh $\mathcal{M}_t$ that represents only visible geometry. In addition, since the scene evolves through time, we construct a separate mesh per frame, resulting in a sequence of geometry $\{\mathcal{M}_t\}_{t=1}^{T}$ for accurate simulations. Please refer to App. C for more qualitative results on depth estimation.

3.2 Audio Content Generation

Given the object-centric audio representation, this component generates the audio content of each sounding object. We identify and track sounding regions in the panoramic video, then synthesize a separate dry audio track conditioned on each object's visual events and semantic description.

Sounding object proposal. We use a VLM (Comanici et al., 2025) to analyze the panoramic video $\mathbf{P}$ and propose candidate sounding objects. We sample representative keyframes and prompt the VLM to identify objects that are likely to produce sound, describing each with a semantic audio description c_k (e.g. "waterfall falling", "dog barking", "volcano is erupting lava to the air", "a monster roaring") and a spatial bounding box b_k in the panoramic image plane. Formally, the VLM produces a set of K proposals $\mathcal{B} = \{(b_k, c_k)\}_{k=1}^{K}$, where c_k is the audio description of the k-th sounding object.

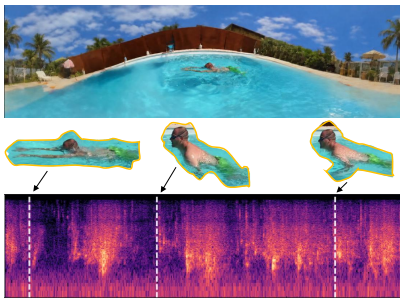

Figure 3: Object tube keyframes and the STFT spectrogram of the generated audio for a swimmer, showing accurate audio-visual timing alignment.

Sounding object tracking and segmentation. Given the bounding box proposals in $\mathcal{B}$, we use a video segmentation model (SAM2 (Ravi et al., 2024)) to propagate each proposal into a dense spatiotemporal segmentation mask. SAM2 takes each bounding box b_k on a reference frame and tracks the corresponding object region through the full video, yielding a sequence of masks $\{m_k^t\}_{t=1}^{T}$ for each object k. To better capture the interaction between a sounding object and its surroundings (e.g. water splashing against rocks, or footsteps on a surface), we expand each mask before extracting the object tube. We apply morphological dilation with a disk-shaped structuring element of radius r to each binary mask m_k^t, yielding an expanded mask $\tilde{m}_k^t = m_k^t \oplus \mathcal{D}_r$, where $\oplus$ denotes the dilation operator and $\mathcal{D}_r$ is a disk of radius r. We extract an object tube $\mathbf{O}_k = \{\pi_k^t (p_t \odot \tilde{m}_k^t)\}_{t=1}^{T}$, so that each tube captures the object together with its interactions with the environment.

Per-object audio synthesis. Each object tube $\mathbf{O}_k$ and its corresponding semantic audio description c_k is passed to a video-to-audio generation model (Cheng et al., 2025) to produce a monaural audio s_k. By conditioning on isolated object tubes rather than the full panoramic frame, audio generation is guided separately to each source, reducing interference from other visual content (Fig. 3).

Reverberation screening. Rendering assumes each source is dry sound, i.e., carrying no or little acoustic effect, since convolving wet sound with an IR would create extra acoustic effect. MMAudio output is monaural and unpanned but not guaranteed anechoic, so we screen each s_k with the pretrained blind RT60 estimator released with Visual Acoustic Matching (Chen et al., 2022) and regenerate any source exceeding a reverberation time of 0.15 s. We regenerate rather than dereverber-

ate, as dereverberation would inject artifacts into a signal that is subsequently convolved; Sec. 4.3 quantifies the residual reverberation.

3.3 Physics-Grounded Acoustic Rendering

With per-object audio signals and the reconstructed scene geometry in hand, the final step is to render each source into spatial audio by simulating how sound propagates through the physical environment.

Acoustic property assignment. In addition to geometry, acoustic simulation requires material properties. To assign these properties, we first apply SAM2 to each frame p_t of $\mathbf{P}$ in semantic segmentation mode, which segments the scene into S material regions $\{\mathcal{R}_s^t\}_{s=1}^S$. We then prompt a proprietary VLM (Comanici et al., 2025) again with these segmented regions and query it for acoustic properties including reflection coefficient α_s and scattering coefficient σ_s for each material class s at six octave bands. Since the material categories of static surfaces are stable across frames, the VLM is queried once per unique material label and the resulting (α_s, σ_s) pairs are cached and re-used. We assign these parameters to the corresponding faces of each per-frame mesh $\mathcal{M}_t$, producing the material-annotated geometry used for acoustic ray tracing (Fig. 4). See Sec. F for additional details.



Figure 4: An example of scene geometry and material segmentation. Left is the raw mesh and right is material-annotated mesh, where colors indicate surface regions with different acoustic properties assigned by the VLM.

Acoustic effect simulation. For each frame t, we lift the centroid of each tracked sounding object using the depth map d_t to estimate its 3D source position $\mathbf{x}_k^t$. AcoustiX (Lan et al., 2024) uses these positions and the material-annotated mesh to simulate a per-frame impulse response for each object. Keeping source content separate from spatial encoding allows the same tracks to be rendered in different formats, including higher-order ambisonics, without regenerating the audio. We use FOA for compatibility with immersive playback platforms and players, and for direct comparison with FOA baselines. FOA format encodes a sound field using four channels, including an omnidirectional channel W and three channels X, Y, Z aligned to the Cartesian axes. For a source at position $\mathbf{x}_k^t$ relative to the listener, the FOA encoding weights are given by the zeroth-order and first-order spherical harmonics evaluated at the source direction $\hat{\mathbf{x}}_k^t = \mathbf{x}_k^t / \|\mathbf{x}_k^t\|$: $\boldsymbol{\gamma}(\hat{\mathbf{x}}_k^t) = \left[1,\, \hat{x}_k^t,\, \hat{y}_k^t,\, \hat{z}_k^t\right]^\top$. AcoustiX traces a set of rays from $\mathbf{x}_k^t$ through $\mathcal{M}_t$, accumulating specular reflections weighted by $(1-\alpha_s)$ and diffuse scattering weighted by σ_s at each surface interaction. Each arriving wavefront at the listener is weighted by $\boldsymbol{\gamma}$ according to its angle of incidence, producing a per-frame four-channel FOA impulse response $\mathbf{h}_k^t \in \mathbb{R}^{4\times L}$:

$$\mathbf{h}_k^t[\ell] = \sum_r g_{k,r}^t \, \boldsymbol{\gamma}(\hat{\mathbf{d}}_{k,r}^t) \, \delta[\ell - \ell_{k,r}^t], \tag{1}$$

where ℓ indexes discrete delay samples, the sum is over all simulated ray paths r, $g_{k,r}^t$ is the path gain incorporating geometric and propagation absorption at frame t, $\hat{\mathbf{d}}_{k,r}^t$ is the direction of arrival, and $\ell_{k,r}^t$ is the arrival delay of path r in samples. The simulation runs for each sounding object, with the listener placed at the panoramic camera position.

Spatially extended sources. Our default renderer approximates each sounding object as a point source located at the centroid of its region. Sources such as crowds, waterfalls, and ocean waves instead occupy an extended region and should not collapse to a single direction. Let Ω_k^t denote the 3D source region obtained by lifting the tracked object mask with depth. We sample M_k emitter locations $\{\mathbf{x}_{k,m}^t\}_{m=1}^{M_k}$ uniformly with respect to surface area over Ω_k^t and simulate a per-emitter FOA IR $\mathbf{h}_{k,m}^t$ using Eq. 1. These responses form an effective regional IR:

$$\bar{\mathbf{h}}_k^t[\ell] = \sum_{m=1}^{M_k} w_{k,m} \big(\mathbf{h}_{k,m}^t * q_{k,m}\big)[\ell], \qquad \sum_{m=1}^{M_k} w_{k,m}^2 = 1, \tag{2}$$

where $q_{k,m}$ is an all-pass filter that preserves the magnitude spectrum while altering the phase, ensuring that signals from the sampled emitters do not combine coherently. The weight $w_{k,m}$ controls each emitter's contribution; for uniform sampling, we use $w_{k,m} = 1/\sqrt{M_k}$ to preserve the source energy. This superposition distributes energy across multiple directions to create a diffuse sound field. In the following, $\tilde{\mathbf{h}}_k^t$ denotes $\mathbf{h}_k^t$ for a point source and $\bar{\mathbf{h}}_k^t$ for a spatially extended source.

Method	AV Alignment		Spatial Correctness					
			Overall		1 fps		5 fps	
	IB↑	DS↓	CC↑	AUC↑	CC↑	AUC↑	CC↑	AUC↑
MMAudio	0.25	0.89	0.11	0.55	0.11	0.52	0.11	0.52
MMAudio (Spatial)	0.24	0.93	0.25	0.61	0.31	0.59	0.20	0.56
See2Sound	0.06	1.08	0.05	0.35	0.09	0.37	0.10	0.38
ViSAGe	0.15	1.00	0.33	0.60	0.35	0.59	0.22	0.61
OmniAudio	0.13	1.04	0.22	0.58	0.22	0.54	0.21	0.54
OmniDream (CubeComposer)	**0.27**	**0.87**	**0.50**	**0.72**	**0.49**	**0.70**	**0.46**	**0.68**
OmniDream (Argus)	0.26	0.89	0.48	0.71	0.46	0.67	0.43	0.63

Table 2: Evaluation of video-to-spatial-audio generation on ImmerseSet.

Spatial audio rendering and mixing. Rendering and mixing operate on the impulse responses of the chosen output format. Here, each source signal $s_k \in \mathbb{R}^{1\times N}$ is rendered using its effective FOA impulse response $\tilde{\mathbf{h}}_k^t$. Since convolution is linear, the per-emitter responses in Eq. 2 can be aggregated before applying the source signal. Both point and regional sources are therefore rendered similarly:

$$\mathbf{a}_k^t[n] = \sum_{\ell=0}^{L-1} \tilde{\mathbf{h}}_k^t[\ell]\, s_k[n-\ell]\ , \tag{3}$$

where n indexes discrete time samples within the frame. Each channel of $\mathbf{a}_k^t$ carries the corresponding FOA component (W, X, Y, Z), encoding source direction and the reverberation induced by the instantaneous scene geometry $\mathcal{M}_t$. The final FOA mixture at frame t is obtained by summing across all K sources: $\mathbf{A}^t[n] = \sum_{k=1}^{K} \mathbf{a}_k^t[n]$. The full FOA track $\mathbf{A} = \{\mathbf{A}^t\}_{t=1}^{T}$ is assembled by concatenating across frames with cross-fade effect to ensure smooth audio transition between frames. This yields an immersive four-channel audio track in which each sounding object contributes energy from its physically correct direction. The final output is $(\mathbf{P},\mathbf{A})$: a temporally consistent 360° panoramic video and a synchronized and physics-grounded FOA audio track.

4 Experiments

4.1 Experimental Setup

Evaluation Dataset. We evaluate on two test sets:

- **ImmerseSet**: We curate 65 in-the-wild monocular perspective videos, spanning indoor environments, outdoor natural settings, and urban scenes, each 8 seconds long without audio. 50 clips come from the Pexels stock-video library and 15 from current text-to-video models; we exclude clips with severe camera motion, jitter, or blur, which degrade outpainting and depth estimation upstream. Since no ground-truth panoramic video or reference audio is available, we assess semantic, temporal and spatial alignment between the generated audio and visual contents.
- **YT360**: A dataset of 360° videos with FOA audio (Morgado et al., 2020); we use a held-out subset of 1000 videos. It has reference recordings so we can measure audio similarity metrics. Since 360° video is provided, this dataset directly assesses video-to-FOA audio generation.

Baselines. Since no existing method performs the full perspective-to-immersive task, each baseline addresses a subset of it: **MMAudio** (Cheng et al., 2025), a video-to-audio model producing mono audio that we replicate to four channels as pseudo-FOA; **MMAudio (Spatial)**, a stronger directional variant we construct by running MMAudio on four directional views of the panorama and encoding each track into FOA at its view direction; **See2Sound** (Dagli et al., 2025), image-to-spatial-audio; **ViSAGe** (Kim et al., 2025), perspective-video-to-FOA; and **OmniAudio** (Liu et al., 2025), panoramic-video-to-FOA. App. G details the input each baseline receives.

Evaluation Metrics. Spatial metrics (CC, AUC) are adopted from ViSAGe (Kim et al., 2025) and measure the spatial and temporal correspondence between a generated audio energy map and a reference spatial map. We report CC/AUC over the full clip and at 1 fps and 5 fps; the harder 5 fps setting tests whether the audio tracks spatial change rather than merely placing energy in the right region. For audio-visual alignment, the ImageBind score measures semantic similarity in a shared embedding space, while DeSync estimates the temporal offset between the generated audio and video using Synchformer (Iashin et al., 2024). Audio similarity metrics (KL, FD) are standard reference-based measures of audio distribution, computed against ground-truth FOA recordings and therefore applied only on the YT360 evaluation set. Please find more details in App. H.

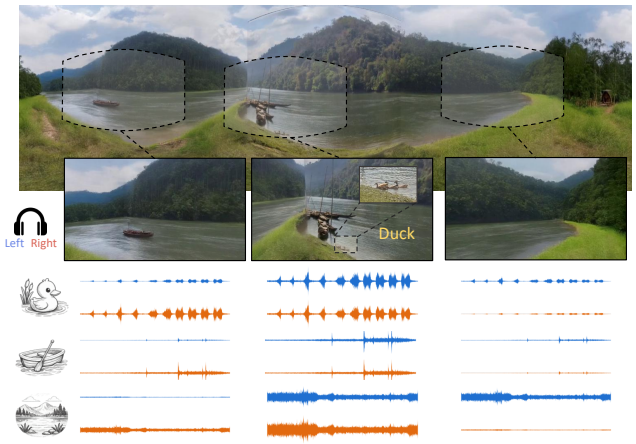

Figure 5: Example of generated visual scene with audios. When users look at different perspective views, the spatial audio will also change for different sounding object tracks.

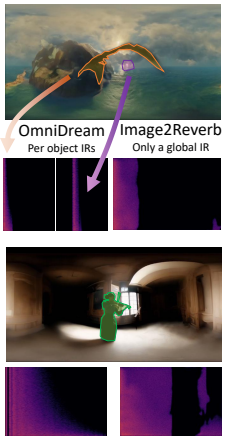

Figure 6: Comparison of rendered impulse responses.

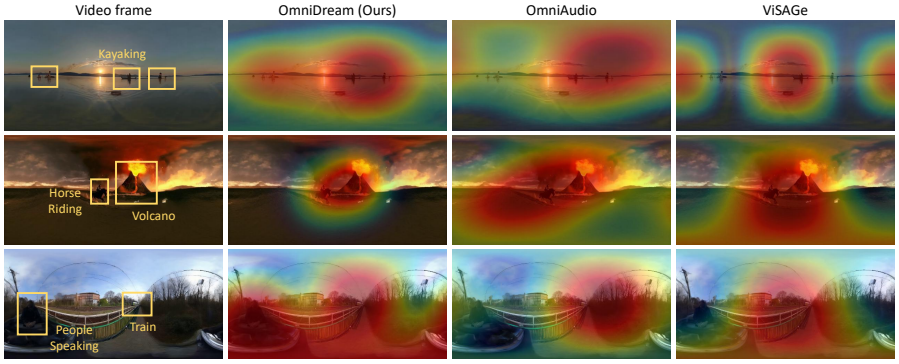

Figure 7: Spatial audio energy maps overlaid on equirectangular panoramic frames. Warmer colors indicate higher directional energy. OmniDream concentrates energy on visible sounding objects, while baseline models produce incorrect spatial distribution.

4.2 Main Results

Quantitative Results. Tab. 2 and 3 report results on ImmerseSet and YT360 respectively. On ImmerseSet, OmniDream outperforms all baselines across every metric. Spatializing MMAudio improves spatial correspondence, but coarse directional partitioning remains less effective than object-level decomposition and geometry-aware propagation. ViSAGe is limited by its fixed perspective view, while OmniAudio performs best among the baselines but still trails OmniDream, showing that panoramic input alone is insufficient without explicit scene grounding. Fig. 5 visualizes one generated scene. As the user looks toward different perspective views, each sounding object produces a different spatial sound effect, and final audio is the sum of all object tracks. On YT360, OmniDream achieves the best performance across all AV alignment and spatial correctness metrics. OmniAudio obtains lower KL and FD, since its training corpus overlaps with YT360, although evaluation is conducted on a held-out split. Nevertheless, OmniDream remains competitive in audio similarity while consistently outperforming OmniAudio in audiovisual alignment and spatial correctness.

Spatial Audio Energy Maps. We visualize the spatial energy distribution in Fig. 7. Each cell in the energy map reflects the directional audio intensity decoded from the ambisonic signal. Compared with the baseline models, OmniDream concentrates audio energy around visually identifiable sounding objects: horse riding and a volcanic eruption in the second row, and a person paddling a kayak in the row. This confirms that our object-centric decomposition and physics-grounded simulation correctly align generated audio with the panoramic video spatially.

Acoustic Rendering. OmniDream simulates unique IR for each sounding object based on the recovered acoustic environments (Fig. 6). Concert halls (on the bottom) produce long reverberant tails with early reflections, while open outdoor scenes (top) yield short and dry responses since there are fewer acoustic reflections in the environments. In the outdoor scene, the difference in arrival times between the dragon and the lightning stems from their relative distances to the camera. Since the

	AV Alignment		Spatial Correctness						Audio Similarity	
			Overall		1 fps		5 fps			
Method	IB↑	DS↓	CC↑	AUC↑	CC↑	AUC↑	CC↑	AUC↑	KL↓	FD↓
MMAudio	0.23	0.76	0.17	0.56	0.23	0.57	0.27	0.58	2.09	8.59
MMAudio (Spatial)	0.22	0.79	0.41	0.62	0.44	0.64	0.41	0.61	2.07	8.60
See2Sound	0.05	1.31	0.01	0.44	0.02	0.42	0.02	0.42	2.96	18.53
ViSAGe	0.14	1.21	0.38	0.61	0.36	0.64	0.33	0.63	2.05	11.28
OmniAudio	0.15	1.22	0.54	0.78	0.50	0.76	0.46	0.74	**1.43**	**5.50**
OmniDream	**0.24**	**0.73**	**0.58**	**0.84**	**0.67**	**0.87**	**0.64**	**0.81**	1.56	6.85

Table 3: Evaluation of video-to-spatial-audio generation on the YT360 dataset.

	AV Alignment		Spatial Correctness					
			Overall		1 fps		5 fps	
Variant	IB↑	DS↓	CC↑	AUC↑	CC↑	AUC↑	CC↑	AUC↑
w/o object tracking	0.25	0.89	0.26	0.52	0.22	0.47	0.19	0.44
w/o mask dilation	0.23	1.03	0.40	0.71	0.44	0.65	0.27	0.57
Spatialization (replacing acoustic rendering):								
azimuth panning	0.21	0.96	0.48	0.71	0.42	0.64	0.37	0.57
+ distance attenuation	0.22	0.93	0.49	0.70	0.45	0.65	0.41	0.63
+ propagation delay	0.22	0.94	0.49	0.71	0.45	0.65	0.41	0.62
OmniDream	**0.27**	**0.87**	**0.50**	**0.72**	**0.49**	**0.70**	**0.46**	**0.68**

Table 4: Ablation studies on the ImmerseSet dataset.

lightning is much farther away, its direct sound is significantly attenuated and arrives with a delay in the rendered IR. In contrast, Image2Reverb (Singh et al., 2021) that directly conditions on the whole image can only generate one global IR with clear artifacts. We additionally evaluate the accuracy of our acoustic rendering against measured real IRs in App. A.

4.3 Ablation and Analysis

Component Ablations. We ablate key design choices in Tab. 4. Ablation Variants include:

- **Object-centric audio generation**: Instead of generating audio conditioned on per-object masked visual streams, we directly feed the full 360° panoramic video to the audio generation model. We then apply spatial acoustic effects to the dominant sounding object in the scene.
- **Mask dilation**: We use binary segmentation masks without morphological dilation, so each object tube captures only the object itself. This tests whether incorporating local environmental context improves audio generation.
- **Spatialization**: We replace the acoustic rendering with simple acoustic spatialization method including azimuth panning, with distance attenuation, and with propagation delay.

Removing object-centric tracking causes the largest drop in spatial correctness, confirming that per-source decomposition is the primary driver of spatial accuracy; AV alignment is comparatively less affected since a global track can still loosely match dominant scene semantics. Removing mask dilation produces the sharpest decline in AV alignment, indicating that local environmental context around each source is critical for temporal coupling between sound events and their visual triggers. For the spatialization effect, adding distance attenuation or propagation delay to azimuth panning yields only a marginal gain. Only the full renderer improves every metric, with its widest margins at the finer temporal resolution and in semantic alignment. The benefit of physics-grounded rendering is therefore from scene-dependent acoustic simulation.

Source Audio Dryness. Physics rendering assumes the per-object signals carry no or little acoustic effect. Applying the same blind reverberation estimator to the generated sources, 90.16% receive a predicted reverberation time below 0.15 s, with a mean of 0.08 s. The generated object audio is therefore mostly dry, leaving little pre-existing acoustic effect for the simulated IR to conflict with. The few sources scoring above the threshold are almost all repetitive sounds, such as crowds and ocean waves, which already contains complicated acoustic effect. We therefore use a 0.15 s reverberation-time threshold to screen sources before physics-based acoustic rendering.

Effect of the Visual Frontend. To evaluate the modularity of OmniDream, we replace CubeComposer with Argus as the panoramic video frontend while keeping other components fixed. As shown in

Perturbation	IB↑	DS↓	CC↑	AUC↑
Material	0.26	0.90	0.44	0.69
Scene scale	0.27	0.88	0.45	0.67
Source location	0.26	0.89	0.37	0.62
OmniDream	**0.27**	**0.87**	**0.5**	**0.72**

Table 5: Robustness analysis.

Video expansion (8s video)	
CubeComposer	1680s
Argus	173s
Depth estimation	44s
Material segmentation	28s
Object tracking	88s
Audio generation	197s
Acoustic rendering	55s

Table 6: Runtime.

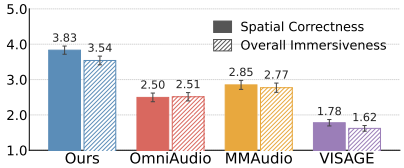

Figure 8: Human evaluation.

Tab. 2, CubeComposer performs consistently better across all end-to-end metrics. CubeComposer also achieves the highest PSNR and SSIM and the lowest FVD in the standalone panoramic video comparison (Tab. 9). Together with the qualitative examples in App. C, these results support our choice of CubeComposer as the default visual frontend.

Robustness to Scene Estimation Errors. Physics-grounded rendering relies on material, scale, and source-location estimates obtained from the input video. We therefore evaluate whether the complete system remains effective when these estimates are imperfect. We independently perturb material reflection coefficients with Gaussian noise ($\sigma = 0.12$), rescale the scene by $\pm 20\%$, and perturb each source location with Gaussian noise ($\sigma = 0.2$ m per axis). Tab. 5 reports the resulting end-to-end performance on ImmerseSet. The system is robust to substantial errors in material and scale estimation. Material perturbations retain a moderate level of performance degradation. Source-location noise has the largest effect. This behavior is consistent with the evaluation: material and scale errors alter acoustic character, whereas localization errors directly move rendered energy away from the visual source direction. Overall, the gains from physics-grounded rendering persist under plausible upstream errors rather than depending on perfect scene recovery.

Runtime. Tab. 6 reports per-stage costs on 8-second, 10 fps clips with 3.4 sources on average, measured on an NVIDIA L40 GPU. Panoramic expansion takes about 1680 s with CubeComposer, compared with 173 s with Argus. This is a substantial runtime cost for the higher visual quality and stronger end-to-end results of CubeComposer (Tabs. 9 and 2). Among the remaining stages, per-object audio generation takes 197 s and can be parallelized across sources. Panoramic expansion is therefore the current speed bottleneck. Because OmniDream has a replaceable visual frontend, faster and more capable panoramic video models can be adopted as they emerge without changing the audio generation or acoustic rendering stages.

User Study. We conduct a human evaluation with 20 participants to assess perceptual quality along two dimensions: *spatial correctness* and *overall immersiveness*. Participants listen to FOA audio decoded to binaural stereo via headphones while watching the corresponding 360° video, and rate each output on a 1–5 MOS scale. As shown in Fig. 8, OmniDream achieves the highest scores on both dimensions, with a spatial correctness MOS of 3.83 compared to 2.85 for MMAudio and 2.50 for OmniAudio. The gap is largest against ViSAGe (1.78), which struggles under the full 360° setting as it was designed for perspective video. On overall immersiveness, OmniDream (3.54) again leads over MMAudio (2.77) and OmniAudio (2.51), confirming that physics-grounded spatialization translates to a perceptually stronger sense of presence. Please refer to App. E for more details on user study.

5 Discussion

Limitation and Future Work. Currently, OmniDream does not support fully occluded sources, which lack the visual evidence our generation stage requires. OmniDream supports head rotation but not full free-viewpoint navigation. With further advances in 4D visual generation, future systems could support free viewer repositioning through novel-view synthesis across the full dynamic scene. Looking ahead, we see these threads converging on a unified paradigm for immersive media generation, in which visual expansion, geometry, and spatial audio are produced jointly rather than stage by stage, absorbing existing mono soundtracks as conditioning, supporting streaming generation, and enabling end-to-end optimization once paired supervision exists for immersive media generation. We regard our explicit decomposition as both a foundation and a baseline for that formulation.

Conclusion. We presented OmniDream, a framework that transforms a common silent video into an immersive 360° audio-visual experience. Rather than treating immersive media generation as purely a visual or audio synthesis problem, OmniDream jointly models panoramic scene reconstruction, object-centric sound generation, and physics-grounded acoustic propagation. More broadly, our work highlights a new direction for generative media: transforming passive videos into explorable audio-visual worlds. We hope OmniDream motivates future research at the intersection of multimodal generation, world model, 3D scene understanding, and interactive media.

AI USE STATEMENT

ChatGPT was used to assist with language editing and manuscript preparation. All suggestions were reviewed by the authors, who take full responsibility for the paper's content. The proposed framework itself composes pretrained generative models, including CubeComposer for panoramic video generation, MMAudio for per-object audio generation, and the Gemini API for VLM-based sounding-object proposals and acoustic-property estimation. Their roles are described explicitly in the method and implementation details.

ETHICS STATEMENT

Our user study involved only viewing immersive videos and listening to their generated audio and did not require IRB review under the applicable institutional guidelines. Participants were recruited on campus, received no compensation, were informed of the study procedure, and could withdraw at any time. We collected no personal or demographic information. All datasets were used under their applicable licenses. A potential societal risk is the misuse of generated immersive media to depict synthetic events as real; outputs should therefore be clearly disclosed as synthetic when presented outside research settings.

REPRODUCIBILITY STATEMENT

We document the model configurations, prompts, preprocessing, hyperparameters, hardware, random seeds, runtime, and evaluation protocols in the paper and appendix. Upon acceptance, we will release the implementation, evaluation scripts, prompts and configurations, generated results, ImmerseSet resources, and interactive demo, together with instructions for obtaining the required pretrained models.

REFERENCES

Zamir Ben-Hur, David Alon, Or Berebi, Ravish Mehra, and Boaz Rafaely. Binaural reproduction based on bilateral ambisonics. In Brian F. G. Katz and Piotr Majdak (eds.), *Advances in Fundamental and Applied Research on Spatial Audio*. IntechOpen, 2021. doi: 10.5772/intechopen.100402.

Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18858–18868, 2022.

Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Infinite-canvas: Higher-resolution video outpainting with extensive content generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2150–2158, 2025a.

Ziyang Chen, Israel D Gebru, Christian Richardt, Anurag Kumar, William Laney, Andrew Owens, and Alexander Richard. Real acoustic fields: An audio-visual room acoustics dataset and benchmark. *arXiv preprint arXiv:2403.18821*, 2024.

Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18770–18781, 2025b.

Ho Kei Cheng. AV-Benchmark: Evaluation toolkit for audio-visual generation. `https://github.com/hkchengrex/av-benchmark`, 2024.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. SEE-2-SOUND: Zero-shot spatial environment-to-spatial sound. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pp. 1–2, 2025.

Loïc Dehan, Wiebe Van Ranst, Patrick Vandewalle, and Toon Goedemé. Complete and temporally consistent video outpainting. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 686–694, 2022. doi: 10.1109/CVPRW56347.2022.00084.

Zixun Fang, Kai Zhu, Zhiheng Liu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Viewpoint: Panoramic video generation with pretrained diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=EKGJl7IHRw.

Matthias Frank, Franz Zotter, and Alois Sontacchi. Producing 3D audio in ambisonics. In *Audio Engineering Society 57th International Conference: The Future of Audio Entertainment Technology—Cinema, Television and the Internet*, 2015. URL https://ambisonics.iem.at/Members/zotter/2015_ambisonics_aes_hollywood.pdf.

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

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Temporally coherent completion of dynamic video. *ACM Trans. Graph.*, 35(6), December 2016. ISSN 0730-0301. doi: 10.1145/2980179.2982398. URL https://doi.org/10.1145/2980179.2982398.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

Derong Jin, Xiyi Chen, Ming C Lin, and Ruohan Gao. SonoWorld: From one image to a 3D audio-visual scene. *arXiv preprint arXiv:2603.28757*, 2026.

Tornike Karchkhadze, Kuan-Lin Chen, Mojtaba Heydari, Robert Henzel, Alessandro Toso, Mehrez Souden, and Joshua Atkins. StereoFoley: Object-aware stereo audio generation from video. *arXiv preprint arXiv:2509.18272*, 2025.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. ViSAGe: Video-to-spatial audio generation. *arXiv preprint arXiv:2506.12199*, 2025.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YCKuXkw6UL.

Lingen Li, Guangzhi Wang, Xiaoyu Li, Zhaoyang Zhang, Qi Dou, Jinwei Gu, Tianfan Xue, and Ying Shan. Cubecomposer: Spatio-temporal autoregressive 4k 360° video generation from perspective video. *arXiv preprint arXiv:2603.04291*, 2026.

Tingle Li, Baihe Huang, Xiaobin Zhuang, Dongya Jia, Jiawei Chen, Yuping Wang, Zhuo Chen, Gopala Anumanchipalli, and Yuxuan Wang. Sounding that object: Interactive object-aware image to audio generation. *arXiv preprint arXiv:2506.04214*, 2025.

Zhaojian Li, Bin Zhao, and Yuan Yuan. Cyclic learning for binaural audio generation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26669–26678, 2024.

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. *arXiv preprint arXiv:2512.16913*, 2025.

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. *arXiv preprint arXiv:2504.14906*, 2025.

Xiulong Liu, Kun Su, and Eli Shlizerman. Tell what you hear from what you see-video to audio generation through text. *Advances in Neural Information Processing Systems*, 37:101337–101366, 2024.

Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177, 2022.

Rundong Luo, Matthew Wallingford, Ali Farhadi, Noah Snavely, and Wei-Chiu Ma. Beyond the frame: Generating 360 panoramic videos from perspective videos. *arXiv e-prints*, pp. arXiv–2504, 2025.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 48855–48876, 2023.

Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *arXiv preprint arXiv:1809.02587*, 2018.

Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. Image2reverb: Cross-modal reverb impulse response synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 286–295, 2021.

Kun Su, Mingfei Chen, and Eli Shlizerman. Inras: Implicit neural representation for audio scenes. *Advances in Neural Information Processing Systems*, 35:8144–8158, 2022.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*, 2024.

Jing Tan, Shuai Yang, Tong Wu, Jingwen He, Yuwei Guo, Ziwei Liu, and Dahua Lin. Imagine360: Immersive 360 video generation from perspective anchor. *arXiv preprint arXiv:2412.03552*, 2024.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: A unified framework for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*, 2025.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, Di Zhang, and Kun Gai. AudioGen-Omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. *arXiv preprint arXiv:2508.00733*, 2025.

Mason Long Wang, Ryosuke Sawata, Samuel Clarke, Ruohan Gao, Shangzhe Wu, and Jiajun Wu. Hearing anything anywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11790–11799, 2024.

Ziyi Wu, Daniel Watson, Andrea Tagliasacchi, David J. Fleet, Marcus A. Brubaker, and Saurabh Saxena. 360Anything: Geometry-free lifting of images and videos to 360°. *arXiv*, 2026.

Yifei Xia, Shuchen Weng, Siqi Yang, Jingqi Liu, Chengxuan Zhu, Minggui Teng, Zijian Jia, Han Jiang, and Boxin Shi. PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. *arXiv preprint arXiv:2505.22016*, 2025.

Siyi Xie, Hanxin Zhu, Xinyi Chen, Tianyu He, Xin Li, and Zhibo Chen. Sonic4D: Spatial audio generation for immersive 4D scene exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11087–11095, 2026.

Zhengxian Yang et al. Imvid: Immersive volumetric videos for enhanced vr engagement. *arXiv preprint arXiv:2503.14359*, 2025.

Zhongrui Yu, Martina Megaro-Boldini, Robert W. Sumner, and Abdelaziz Djelouah. Unboxed: Geometrically and temporally consistent video outpainting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7309–7319, 2025. doi: 10.1109/CVPR52734.2025.00685.

Mengchen Zhang, Qi Chen, Tong Wu, Zihan Liu, and Dahua Lin. Visaudio: End-to-end video-driven binaural spatial audio generation. *arXiv preprint arXiv:2512.03036*, 2025a.

Yu Zhang, Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Tao Jin, and Zhou Zhao. Isdrama: Immersive spatial drama generation through multimodal prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9618–9627, 2025b.

Lei Zhao, Rujin Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Foleyspace: Vision-aligned binaural spatial audio generation. *arXiv preprint arXiv:2508.12918*, 2025.

APPENDIX

The appendix is organized as follows. Sec. A studies the performance of physics-grounded acoustic rendering. Sec. B gives the protocol and results for the panoramic video-quality evaluation. Sec. C presents additional qualitative results on the 360° visual expansion stage and depth estimation. Sec. D details FOA rendering, cross-fade mixing, and binaural playback. Sec. E describes the user study protocol. Sec. F provides implementation details including model configurations, hyper-parameters, material acoustic property parameterization, and VLM prompts. Sec. G details how each baseline is configured. Sec. H describes the evaluation metrics in full detail.

A EVALUATION OF PHYSICS-GROUNDED ACOUSTIC RENDERING

Our physics-grounded acoustic rendering essentially estimate the impulse response purely based on the visual scene. We therefore evaluate the accuracy of this component against measured impulse responses on Real Acoustic Fields (RAF) (Chen et al., 2024) dataset.

Setup. We use the furnished room of RAF, which provides densely measured real IRs together with a scanned room mesh. Instead of using dataset's ground truth geometry, we render a panoramic image from the provided mesh and then estimate everything the simulator needs with the same modules OmniDream uses at inference: panoramic depth estimation for geometry, SAM2 segmentation with VLM querying for materials, and VLM metric-scale calibration. Only the source and receiver positions come from the dataset, and the source is transformed by our estimated scale rather than the true one, so scale error propagates exactly as it would in deployment.

Metrics. All reported quantities are errors with respect to the measured IRs: C50 in dB, EDT in seconds, RT60 as a relative error, and a multi-resolution STFT error. Lower is better.

Setting	C50	EDT	RT60	STFT
Zero-shot	4.69	0.099	31.9	0.58
One-shot	2.21	0.042	11.4	0.56
INRAS	2.45	0.071	25.1	0.52

Table 7: Accuracy of IRs estimation. Zero-shot uses visual estimates only, one-shot uses one measured IR for calibration, and INRAS is trained on 300 samples.

Perturbation	ΔC50	ΔEDT	ΔRT60	ΔSTFT
Material	0.74	0.0102	8.97%	0.02
Scene scale	0.89	0.0163	13.46%	0.01
Source location	0.54	0.0089	6.06%	0.01

Table 8: Robustness of simulated IRs to errors in estimated materials, scene scale, and source location.

Results. In the zero-shot setting used by OmniDream, the simulated IR captures the overall time-frequency structure of the measured response, with an STFT error of 0.58, but does not fully recover its late-field energy, resulting in a 31.9% RT60 relative error and a 4.69 dB C50 error. As a diagnostic, we additionally use one measured reference IR to calibrate the room-level tail. This one-shot calibration reduces the C50 error to 2.21 dB and the RT60 error to 11.4%, matching or outperforming INRAS (Su et al., 2022), which is trained on 300 samples, while the STFT error changes only slightly from 0.58 to 0.56. This result indicates that the remaining error is concentrated primarily in late reverberation rather than the propagation structure, since a single measurement could not correct incorrect geometry. One-shot calibration is not part of our pipeline and is used only for this diagnostic, as such a measurement is unavailable for arbitrary videos. We therefore claim that visually estimated geometry and materials provide a useful physical approximation rather than an exact acoustic reconstruction, and that this approximation is sufficient to improve immersive audio generation (Sec. 4.3).

Robustness to estimation errors. To test the robustness of the acoustic rendering pipeline, we perturb one visually estimated quantity at a time while holding the others fixed. Scene scale has the largest effect on the simulated response (13.46% relative RT60 change) because rescaling modifies every propagation-path length. Material noise produces a smaller change (8.97%), and source-location noise is mildest at the IR level (6.06%). These results characterize how errors in the vision-derived scene representation propagate into the acoustic simulation; Sec. 4.3 separately evaluates their effect on the complete immersive-generation system.

B Panoramic Video Generation Comparisons

Since panoramic expansion supplies both the visual output and the geometry driving acoustic rendering, we quantify it directly against the three candidate backbones.

Setup. We use 30 YT360 clips, a corpus none of the three backbones trains on. Each backbone receives an identical perspective crop of the ground-truth panorama at 80° FOV and 854×480 resolution, covering the leading 8 seconds. Because the methods generate at different native resolutions and lengths, no metric is computed on raw outputs: every output and the ground truth are resampled onto a common 1024×512 equirectangular grid and a common 27-frame window, 27 being the shortest native length so that no model is rolled out beyond its trained horizon.

Metrics. PSNR and SSIM are computed on the equirectangular frames, and FVD uses the StyleGAN-V I3D TorchScript extractor at the same resolution. The generated-region mask is the exact spherical footprint of the input frustum, computed by the pipeline's own conditioning routine rather than a rectangular latitude/longitude box, and covers 5.8% of equirectangular pixels.

Method	PSNR↑	SSIM↑	FVD↓
CubeComposer (Li et al., 2026) (ours)	**11.57**	**0.37**	**2417**
Argus (Luo et al., 2025)	11.13	0.34	2974
PanoWan (Xia et al., 2025)	8.19	0.22	5099

Table 9: Panoramic expansion quality on YT360 clips.

Results. As shown in Tab. 9, CubeComposer achieves the highest PSNR (11.57 dB) and SSIM (0.37), and the lowest FVD (2417) among the three visual backbones. Compared with Argus, it improves PSNR by 0.44 dB and SSIM by 0.03, while reducing FVD by 18.7%. The improvements over PanoWan are larger: 3.38 dB in PSNR, 0.15 in SSIM, and a 52.6% reduction in FVD. These results are consistent with the qualitative comparisons in Sec. C. Together with its stronger end-to-end performance in Tab. 2, they support our choice of CubeComposer as the default visual frontend for OmniDream.

C Additional Qualitative Results

360° **Video Generation.** We compare three perspective-to-360° video generation methods including Argus (Luo et al., 2025), CubeComposer (Li et al., 2026), and PanoWan (Xia et al., 2025), on one indoor scene and one outdoor scene (Fig. 9). CubeComposer produces the most consistent panoramic expansions, better preserving the color palette and global scene structure of the original input. Argus and PanoWan exhibit more visible inconsistencies in the generated regions, including seam and warping artifacts that can propagate into depth estimation and acoustic simulation. These comparisons, together with the quantitative results in Tab. 9, justify our choice of CubeComposer as the panoramic outpainting backbone.

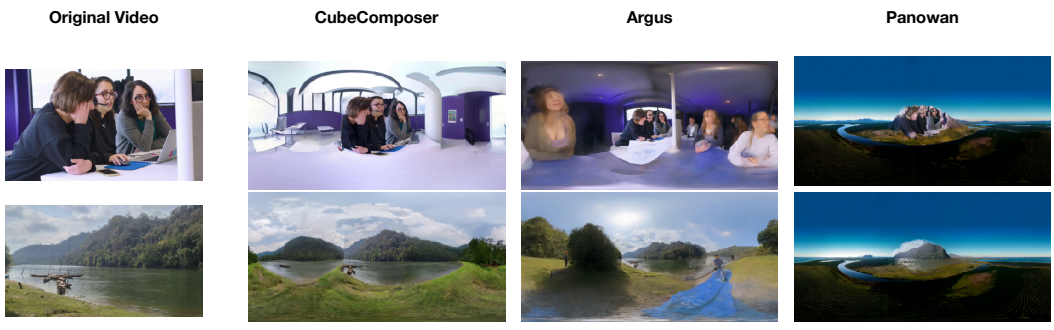

Figure 9: Comparison of perspective-to-360° video generation methods on two scenes. Each row shows the original perspective input (left) alongside the panoramic outputs of CubeComposer (Li et al., 2026), Argus (Luo et al., 2025), and PanoWan (Xia et al., 2025). CubeComposer generates the most coherent panoramic expansion, while Argus and PanoWan exhibit more visible inconsistencies in the generated regions.

Depth Estimation. We evaluate panoramic depth estimation on the generated 360° videos produced by our pipeline (Fig. 10). Despite being synthesized through panoramic outpainting, the generated scenes preserve sufficient geometric consistency for downstream depth prediction. The estimated depth maps recover the dominant room layouts, open spaces, and large structural boundaries across both indoor and outdoor environments. These geometry estimates provide foundations for our acoustic simulation module, enabling physically grounded propagation and spatial audio rendering conditioned on the generated panoramic scenes.

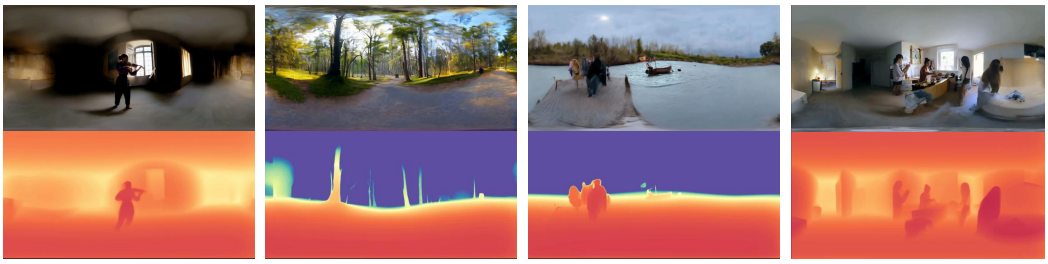

Figure 10: Example depth estimation for the generated panoramic videos.

D FOA Rendering, Cross-Fade Mixing, and Binaural Playback

Since both the scene geometry and source positions change over time, we simulate a separate four-channel FOA impulse response $\mathbf{h}_k^t \in \mathbb{R}^{4\times L}$ for each source k at every acoustic update t. In our implementation, updates occur once per second, and the IR is held fixed within each interval. Each dry source signal s_k is divided according to these update intervals and rendered with the corresponding IR as described in Eq. 3.

Cross-fade mixing. Directly concatenating rendered segments can introduce discontinuities at IR-update boundaries when the impulse response changes abruptly, for example when the source moves or the geometry changes. We therefore apply a raised-cosine cross-fade of length 300 ms at each boundary. Let C be the cross-fade length in samples, which is $C = 13{,}230$ at 44.1 kHz, and let $u \in \{0, \dots, C-1\}$ index the overlap. The complementary windows are

$$w_{\text{in}}[u] = \frac{1}{2}\left(1 - \cos\frac{\pi u}{C-1}\right), \quad w_{\text{out}}[u] = 1 - w_{\text{in}}[u]. \tag{4}$$

The two adjacent rendered segments are then combined by overlap-add:

$$\mathbf{a}_{k,\text{xf}}[b_t + u] = w_{\text{out}}[u]\,\mathbf{a}_k^t[b_t + u] + w_{\text{in}}[u]\,\mathbf{a}_k^{t+1}[b_t + u]. \tag{5}$$

where b_t is the start of the overlap between adjacent rendered segments. The windows sum to one at every overlap sample, smoothly transferring contribution between acoustic states. We apply the same scalar weights to all four FOA channels to preserve their directional relationships. After cross-fading, the rendered tracks are summed across sources to produce the final FOA mixture $\mathbf{A}$.

Binaural playback. During the user study, the viewer orientation (ψ, θ) is used to rotate the FOA field and decode it to M virtual loudspeaker directions with an ambisonic decoding matrix $\mathbf{D} \in \mathbb{R}^{M\times 4}$. Each decoded loudspeaker signal is then filtered with the corresponding left- and right-ear HRTFs:

$$\mathbf{Y}^t = \mathbf{D}(\psi, \theta)\,\mathbf{A}^t, \tag{6}$$

$$\mathbf{b}_e[n] = \sum_{m=1}^{M} \left(\mathbf{Y}_m^t * h_m^e\right)[n], \quad e \in \{L, R\}, \tag{7}$$

where h_m^L and h_m^R are the HRTFs for virtual loudspeaker direction m. This decoding runs in real time so that the perceived sound field rotates consistently with the viewer.

E User Study Protocol

Interactive viewer. To deliver the immersive experience, we implement a browser-based interactive viewer in which the participant can freely drag to rotate the viewpoint within the 360° panoramic

video. As the viewpoint changes, the FOA audio is decoded to binaural stereo following Sec. D, so the perceived sound directions update continuously with head rotation. Fig. 11 shows a screenshot of the viewer interface, including a debug overlay that displays the live decoded stereo levels (L/R RMS and peak), channel balance, and current yaw/pitch orientation.



Figure 11: Interactive viewer for the user study. Participants drag to rotate the 360° panoramic view; FOA audio is decoded to binaural stereo in real time as the viewpoint changes. The debug overlay (bottom right) shows live decoded stereo RMS/peak levels, and the current yaw/pitch orientation.

Participants and rating procedure. We recruit 20 participants to watch and listen to our immersive media. Each participant evaluates 6 clips per method (exclude See2Sound) in the ImmerseSet. Method identity is hidden and randomized throughout the study. Participants wear headphones and are instructed to freely explore the scene by dragging the view before and during rating. After experiencing each clip, participants are asked to submit their evaluation on the google form with the following rating dimensions.

Rating dimensions. Each clip is rated on a 1–5 MOS scale along two dimensions:

- **Spatial correctness**: does the audio appear to originate from the correct direction relative to the visual scene?
- **Overall immersiveness**: how strongly does the output create a sense of being present within the scene?

F Implementation Details

Model configurations. OmniDream is a training-free framework that composes the following off-the-shelf models. Panoramic outpainting uses CubeComposer (Li et al., 2026). We also alternatively use Argus (Luo et al., 2025) and PanoWan (Xia et al., 2025) for ablation studies. Panoramic depth estimation uses (Lin et al., 2025), with metric scale estimated via Gemini (Comanici et al., 2025). Sounding object proposals and acoustic property assignment both use Gemini (Comanici et al., 2025). Object tracking and segmentation use SAM2 (Ravi et al., 2024). Per-object audio generation uses MMAudio (Cheng et al., 2025). Acoustic ray tracing is performed with AcoustiX (Lan et al., 2024).

Hyperparameters. We use a mask dilation radius of $r = 200$ pixels for object tube extraction. The depth discontinuity threshold for mesh triangulation is $\tau = 2$ meter. Each per-frame acoustic simulation uses 10^6 ray paths in AcoustiX. And the sampling rate for the impulse response is set to be 44.1 kHz, which is consistent with the MMAudio (Cheng et al., 2025) output. We sample 1 keyframe per second when querying the VLM for sounding object proposals. The impulse response simulation interval is set to 1 frame per second. We simulate a new impulse response once per second and hold it constant within that interval. The cross-fade window between adjacent frames is 300 ms. The runtime measurements in Tab. 6 use an NVIDIA L40 GPU; the other inference experiments use a single NVIDIA B200 GPU.

Acoustic material properties. For each segmented surface region, the VLM estimates three frequency-dependent acoustic properties across six octave-band centre frequencies: 500 Hz, 1 kHz,

2 kHz, 4 kHz, 8 kHz, and 16 kHz. These are the reflection coefficient $\alpha_s^f \in [0, 1]$, the absorption coefficient $\beta_s^f \in [0, 1]$, and the scattering coefficient $\sigma_s^f \in [0, 1]$, where f indexes the frequency band and s indexes the material class. By design, $\alpha_s^f + \beta_s^f \approx 1$ at each frequency, as energy is either reflected or absorbed (transmission through thin surfaces is neglected). The scattering coefficient controls the fraction of reflected energy that is diffusely scattered rather than specularly reflected, which governs late reverberation diffuseness. This per-frequency parameterisation allows the ray tracer to capture frequency-dependent reverberation: hard dense surfaces such as glass ($\alpha \approx 0.97$ at 1 kHz) sustain high reflection across all bands, while porous or soft materials such as carpet ($\alpha \approx 0.10$ at 1 kHz) absorb more energy, especially at higher frequencies. As noted in Sec. 3.3, material properties are stable across frames for static surfaces, so the VLM is queried once per unique material label and the results are cached.

VLM prompts. We use three separate VLM queries during the pipeline.

Metric depth calibration. In our pipeline, we treat the panoramic model's per-frame predictions as relative depth. To recover metric scale, we query the VLM with a panoramic frame and ask it to identify a single object with a known real-world size and estimate its average distance from the camera. The returned bounding box and depth estimate are used to compute a scale factor that maps the relative depth map to metric units.

Metric Depth Calibration Prompt

You are given a 360° equirectangular panoramic image. Your task is to identify one object in the scene whose real-world size is well known (e.g. a person, a car, a door, a bicycle, a table), and estimate how far it is from the camera.

Choose the object that you are most confident about in terms of both its identity and its distance.

Return ONLY a JSON object with these fields:
`"object_name"`: (string) name of the chosen object, e.g. *"person"*, *"car"*, *"door"*
`"bbox"`: (list of 4 floats) bounding box of the object as [ymin, xmin, ymax, xmax] in 0–1000 coordinates
`"estimated_depth_m"`: (float) estimated average distance from the camera to the object centre, in metres
`"confidence"`: (float in 0–1) your confidence in the depth estimate

Notes:
– Prefer objects that are fully visible and unoccluded.
– Use typical real-world sizes as anchors: a standing adult ≈ 1.7 m tall; a standard car ≈ 4.5 m long; a door ≈ 2.0 m tall.
– If no object with a known size is visible, return `"confidence":0`.

Sounding object proposal. We sample one keyframe per second from the panoramic video and pass all frames together to the VLM, which identifies persistent sounding objects across frames and reports each one's bounding box in the most representative frame.

Sounding Object Proposal Prompt

You are given N frames sampled at 1 frame per second from a 360° equirectangular panoramic video (full sphere unwrapped).

Analyze all frames together to identify objects that plausibly produce sound throughout the video (e.g., people, musicians, animals, machinery, water, traffic).

For each identified sounding object:
– Track its presence across frames to confirm it is a genuine, persistent sound source.
– Select the frame where it is most clearly visible as the representative frame.
– Report its bounding box in that representative frame.

Return JSON with an `"objects"` list, each entry having:
`"id"` (int),
`"name"` (str),
`"prompt"` (short sound description for audio generation, e.g. *"crowd cheering"*),
`"representative_frame"` (int, 0-indexed),
`"bbox"` ([ymin, xmin, ymax, xmax] in 0–1000 coordinates),
`"confidence"` (float in 0–1).

Return at most 4 objects. Return an empty list if no sounding objects are found.

Acoustic property assignment. Given segmented material regions from SAM2, the VLM is queried once per unique material label to estimate acoustic properties.

Acoustic Property Assignment Prompt

You are an acoustic engineering expert. This image shows a single surface region extracted from a 360° equirectangular video frame. The background has been greyed out so you can focus on just this surface.

Analyse the visible material and estimate its acoustic properties at six octave-band centre frequencies: 500 Hz, 1 kHz, 2 kHz, 4 kHz, 8 kHz, and 16 kHz.

Return ONLY a JSON object with these fields:
`"material_name"`: (string) specific material name, e.g. *"polished marble"*, *"rough concrete wall"*
`"description"`: (string) brief description of the surface and why it has these acoustic properties
`"frequencies"`: (list of int) [500, 1000, 2000, 4000, 8000, 16000]
`"reflection_coefficients"`: (list of 6 floats) fraction of incident sound energy reflected at each frequency
`"absorption_coefficients"`: (list of 6 floats) fraction of incident sound energy absorbed at each frequency
`"scattering_coefficients"`: (list of 6 floats) fraction of reflected energy scattered diffusely at each frequency

Notes:
– `reflection + absorption` ≈ 1.0 at each frequency (ignore transmission through thin surfaces).
– Most soft or porous materials absorb more at higher frequencies.
– Hard dense materials (glass, concrete, marble) maintain high reflection across all bands but may show slight absorption increase above 4 kHz.
– Typical references at 1 kHz: glass ≈ 0.97 reflection; carpet ≈ 0.10 reflection; concrete ≈ 0.95 reflection.

G Baseline Configurations

No existing method maps a silent perspective video to a 360° video with FOA audio, so every baseline solves part of the task and must be given the input it was designed for.

MMAudio (Cheng et al., 2025). A state-of-the-art video-to-audio model generating single-channel audio by latent diffusion. We run it on the generated 360° video and replicate the single channel across four channels to form a pseudo-FOA signal. This carries no directional information, so its spatial scores reflect the absence of spatialization rather than a failure of the model, which was never designed to produce ambisonics.

MMAudio (Spatial). Because channel replication is a weak spatial reference, we construct a stronger variant. We partition the panorama into four directional views, run MMAudio independently on each, and encode each resulting track into FOA at its view direction. This gives MMAudio genuine directional structure while still lacking object-level decomposition and scene geometry, isolating what those two components contribute.

See2Sound (Dagli et al., 2025). An image-driven method that natively renders 5.1 surround audio from a single image. We condition it on the first frame of the generated 360° video. Being image-based, it has no access to temporal dynamics, which is reflected in its synchronization scores.

ViSAGe (Kim et al., 2025). A vision-guided model predicting directional audio from perspective video. We apply it to the original perspective input, the setting it was designed for, rather than to the panorama. It therefore reasons only about the original field of view and cannot place sources in the surrounding 360° scene.

OmniAudio (Liu et al., 2025). A panoramic-video-to-FOA model. Since it requires 360° input, we feed it the outpainted panoramic video $\mathbf{P}$ produced by our own visual expansion stage, so it receives the same visual context as OmniDream's audio stages and differs only in how spatial audio is produced.

H EVALUATION METRICS

Spatial metrics. We adopt the correlation coefficient (CC) and AUC-Judd (AUC) metrics from ViSAGe (Kim et al., 2025) to measure spatial correspondence between the generated audio and a reference spatial map. For each frame, we decode the FOA signal into a directional energy map over the sphere and compare it against the reference map. CC measures the linear correlation between the two maps, while AUC-Judd measures how well the energy map predicts high-reference locations. The reference map depends on the dataset. For **YT360**, which provides ground-truth FOA, we decode the reference energy map directly from the ambisonic audio following ViSAGe. For **ImmerseSet**, which has no ground-truth spatial audio, we construct a visually derived reference map from dense optical-flow magnitudes, aggregate it over each evaluation window, and resize and Gaussian-smooth it onto the equirectangular grid. We normalize both the predicted and reference maps before computing CC and AUC over the full clip and within 1 s and 0.2 s windows. The YT360 evaluation therefore measures agreement with reference acoustic energy, while the ImmerseSet evaluation provides a complementary measure of audio-visual spatial correspondence.

AV synchronization metrics. We compute ImageBind score (IB) and DeSync (DS) using the open-source AV-Benchmark toolkit (Cheng, 2024), which is also used by MMAudio (Cheng et al., 2025). ImageBind score measures semantic alignment between audio and video in the shared ImageBind embedding space (Girdhar et al., 2023). DeSync measures temporal misalignment as the absolute audio-video offset predicted by Synchformer (Iashin et al., 2024).

Audio similarity metrics. KL divergence and Fréchet distance (FD) are computed using the same AV-Benchmark toolkit (Cheng, 2024) between the distribution of generated audio features and ground-truth FOA recordings on the YT360 dataset. These metrics are only reported on YT360, as ImmerseSet has no ground-truth audio reference.