SONICVISTA: FROM ONE VIDEO TO A 4D AUDIO-VISUAL SCENE VIA DIFFUSION SYNCHRONIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in visual scene generation has extended beyond static 3D scenes toward the more challenging problem of generating dynamic 4D scenes from minimal input. However, sound, despite being intrinsically coupled with dynamic objects and events, remains largely unmodeled in prior work. We introduce VIDEO-TO-4DAVSCENE, a new task that turns a video into an exploraable, time-varying 4D audio-visual scene, and SONICVISTA, the first framework for this task. Given a monocular video, our method generates a 360° static background, reconstructs explicit motion trajectories of dynamic, sound-emitting foreground objects, and synthesizes a physically-grounded sound field consistent across multiple sources through dual-space audio diffusion synchronization. The resulting 4D scene supports real-time novel-view visual rendering and spatial audio rendering. Evaluation on real-world dynamic scenes and a user study demonstrates SONICVISTA generates 4D audio-visual scenes with perceptually realistic, physically grounded spatial audio and geometry-grounded audio-visual consistency.
Interactive demo: `https://sonicvista.github.io/demo.html`

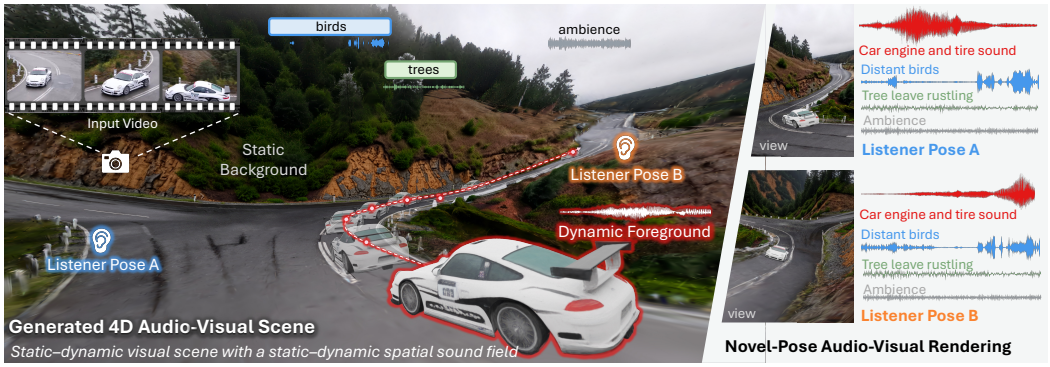


Figure 1: From a single monocular video, SONICVISTA generates an explorable 4D audio-visual scene, where you can navigate to novel viewpoints and follow arbitrary listener trajectories while hearing spatial audio synchronized with dynamic events and grounded in moving sound sources (e.g, a racing car), static sources (e.g., birds and rustling trees), and background ambience.

1 INTRODUCTION

Advances in 3D scene reconstruction and generation (Mildenhall et al., 2020; Kerbl et al., 2023; Yu et al., 2025a; 2024), and more recently in dynamic 4D scene generation (Wu et al., 2024; Ren et al., 2025; Lei et al., 2024), makes it possible to transform monocular videos into explorable, time-varying worlds with moving objects and temporally coherent geometry. Yet these worlds remain incomplete without sound, which is intrinsically coupled with dynamic objects and events: engine and wheel noise grow louder as a car approaches, sweeping past and fading into the distance; a dancer's tap shoes strike the stage in sharp, rhythmic beats synchronized with each step; and ambient sounds such as wind and crowds continuously conveys the structure and scale of the environment.

Despite its importance, limited prior work has modeled spatial audio in 3D visual scenes. Existing approaches either focus on static scenes without time-varying geometry or moving sound

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.

sources (Jin et al., 2026; Yoshida et al., 2026), or spatialize audio for dynamic videos under constrained camera and interaction settings (Xie et al., 2025). Generating fully explorable 4D audio-visual scenes therefore remains an open challenge, requiring joint modeling of dynamic geometry, moving sound sources, and spatial audio.

We introduce VIDEO-TO-4DAVSCENE, a new problem that aims to turn a monocular video into an explorable *4D audio-visual scene*: an interactive, time-varying world-space scene with dynamic sound sources that supports synchronized rendering of both visual content and spatial audio from novel viewpoints and listener trajectories. The generated scene should contain a spatially grounded sound field: ambient sound should be tied to the static environment, dynamic sounds should follow moving visual objects, and the rendered audio should remain consistent as the virtual listener moves through the reconstructed 4D audio-visual world.

Solving this problem introduces several new challenges. *First*, unlike recent 4D visual generation methods that primarily optimize visual fidelity, 4D audio-visual scene generation requires explicit scene geometry, camera motion, dynamic object trajectories, and static-dynamic structure to place and render sound sources in a physically meaningful coordinate system. *Second*, dynamic scenes contain heterogeneous acoustic components: ambient sound should align with the background, localized static sources should remain fixed in world space, and dynamic sources should follow foreground objects while obeying distance attenuation, occlusion, and motion-dependent effects. *Third*, the system must maintain consistency across coordinate spaces: input audio is observed in camera space, object sounds are represented in object-centric canonical space, and the output must be rendered for arbitrary listener trajectories in world space. These spaces must be synchronized to keep sound temporally aligned with the input video and geometrically consistent with the evolving scene.

To address these challenges, we propose SONICVISTA, a world-space framework for 4D audio-visual scene generation. Given an input video, we reconstruct an *audio-aware 4D visual scene representation*, comprising a 360° static background and dynamic foreground objects with world-space trajectories, providing geometry and motion for spatial audio rendering while supporting interative visual exploration. Building on this representation, we introduce a *diffusion-based synchronization* mechanism that jointly synthesizes a global camera-space audio signal and per-object canonical audio signals to generate a dynamic spatial sound field. Specifically, it repeatedly projects object-space sounds into camera space through acoustic rendering and maps the global mixture back to object space via inverse rendering, enforcing cross-space consistency throughout the diffusion process. This synchronization keeps the generated audio aligned with the input video while preserving persistent, re-renderable object sounds that follow dynamic sources under novel listener trajectories.

We evaluate SONICVISTA on real-world dynamic scenes with moving foreground sound sources exhibiting rigid-body, human, and general deformable motion. We also establish objective evaluation protocols for VIDEO-TO-4DAVSCENE to assess the spatial accuracy and semantic consistency of generated audio. Experiments show that our framework generates 4D audio-visual scenes with spatial audio that better aligns with scene geometry, object motion, and listener viewpoint than baseline methods. Ablation studies and a user study further confirm the effectiveness of diffusion synchronization in generating coherent spatial audio with multiple sound sources.

2 RELATED WORK

4D Visual Scene Generation. Recent 4D scene generation methods can be broadly grouped into video diffusion, dynamic Gaussian reconstruction, and dynamic foreground reconstruction. *Video diffusion* based methods (Ren et al., 2025; Xie et al., 2026; Yu et al., 2025b; Lin et al., 2026; Yang et al., 2026; Bai et al., 2025) synthesize photorealistic videos under controlled camera trajectories, but are trajectory-specific and often lack explicit geometry grounding. *Dynamic Gaussian reconstruction* methods (Lei et al., 2024; Wang et al., 2025; Shih et al., 2025) reconstruct geometry-grounded dynamic scenes from monocular videos and support real-time rendering, but are largely limited to visible regions and cannot reliably recover unseen object surfaces or a complete panoramic background. *Dynamic foreground reconstruction* methods further recover animatable 4D assets from monocular inputs, including category-agnostic deformable or articulated objects (Liao et al., 2025; Remy Sabathier, 2026; Hongyuan et al., 2026; Chen et al., 2025a; Litman et al., 2026) and human-specific avatars (Qiu et al., 2025; Hu et al., 2024; Moon et al., 2024). Our visual pipeline builds on these advances to construct a 4D visual scene representation, combining dynamic foreground objects with a static background to enable generation of a dynamic spatial sound field.

Spatial Audio Generation. Audio generation methods synthesize realistic monaural audio from text or video (Cheng et al., 2025; Shan et al., 2025; Tian et al., 2025; Chen et al., 2025b), but typically lack spatial cues. Complementary approaches spatialize existing audio: RIR-based rendering and acoustic-field methods model propagation effects (Chen et al., 2020; 2022; Luo et al., 2022; Wang et al., 2024; Jin & Gao, 2025; Lan et al., 2024; Liu et al., 2025b), while visually guided methods use scene geometry, object locations, or motion to predict binaural or spatial audio from mono input (Gao & Grauman, 2019; Garg et al., 2023; 2021; Morgado et al., 2018; Li et al., 2018; Liang et al., 2023). More recent methods directly generate ambisonics from videos (Kim et al., 2025; Liu et al., 2025a), spatialize generated audio for 4D scene exploration (Xie et al., 2025), or construct static 3D audio-visual scenes from single images (Jin et al., 2026; Yoshida et al., 2026). In contrast, SONICVISTA generates explorable dynamic 4D audio-visual scenes with moving sound sources, synchronizing spatial audio with object motion while modeling listener-dependent acoustic effects.

Diffusion Synchronization. Diffusion synchronization couples multiple diffusion trajectories at inference time to satisfy shared constraints. Primarily explored for image generation, it enables coherent large images and region-controlled layouts (Bar-Tal et al., 2023), enforces consistency across overlapping generations (Lee et al., 2023), and models correlations between coupled trajectories (Lee et al., 2025). Recent work extends these ideas to synchronize image and audio diffusion on a shared canvas (Chen et al., 2024) and extend long-form audio separation beyond the training length (Shi et al., 2025). We also apply diffusion synchronization to audio, but target spatial audio rendering. Specifically, we synchronize camera-space and object-space audio trajectories through forward and inverse acoustic rendering, allowing generated source sounds to remain consistent with the camera-space mixture while being re-renderable under source motion and novel listener trajectories. To our knowledge, this is the first use of diffusion synchronization for spatial audio generation with an explicit rendering component.

3 THE VIDEO-TO-4DAVSCENE TASK

We define VIDEO-TO-4DAVSCENE as the problem of generating an explorable 4D audio-visual scene from a single monocular video. Given an input video $V = \{I_t\}_{t=1}^{T}$, where I_t denotes the RGB frame at time t, the problem is to generate:

$$V \mapsto \{\mathcal{S}, \mathcal{A}\}, \tag{1}$$

where $\mathcal{S}$ denotes a time-varying 4D visual scene with moving objects and $\mathcal{A}$ denotes a dynamic spatial sound field associated with the scene. The output should be explorable: for any target camera/listener trajectory, it should support synchronized novel-view visual rendering and listener-dependent spatial audio rendering.

Formally, given a target trajectory $p_t = [R_t, \mathbf{x}_t] \in SE(3)$, where R_t and $\mathbf{x}_t$ denote the listener orientation and position at time t, the generated scene should render an audio-visual sequence

$$\{\mathcal{S}(p_t), \mathcal{A}(p_t)\} = \left\{\hat{I}_t^{\,p},\, \hat{a}_t^{\,p}\right\}. \tag{2}$$

Here, $\hat{I}_t^{\,p}$ is the visual observation rendered from pose p_t, and $\hat{a}_t^{\,p}$ is the synchronized spatial audio perceived by the listener at the same pose. The visual scene $\mathcal{S}$ contains both the static environment and time-varying foreground content. The spatial sound field $\mathcal{A}$ should be geometrically and temporally grounded in $\mathcal{S}$: ambient sound and static localized sound sources should remain fixed in static world space, and dynamic sound sources should move consistently with their corresponding foreground objects. Thus, the generated audio should not only match the semantic and temporal content of the input video, but also vary coherently with listener position, listener orientation, source motion, distance attenuation, and visibility.

The SONICVISTA360 Dataset. To support quantitative evaluation of this new task, we collect SONICVISTA360, a evaluation dataset of 54 clips of dynamic scenarios with moving sound sources such as cars and humans. Specifically, each scene contains an 8-second 360° video captured at 30 FPS using an Insta360 X5 camera, paired with synchronized first-order ambisonic (FOA) audio recorded at 48 kHz using a RØDE NT-SF1 microphone placed at a novel listening pose.The FOA microphone is placed at a listening pose different from the camera pose, enabling evaluation of whether a method can render spatial audio at a novel listener location rather than merely reproducing camera-view audio. Further details about SONICVISTA360 are provided in Appendix.

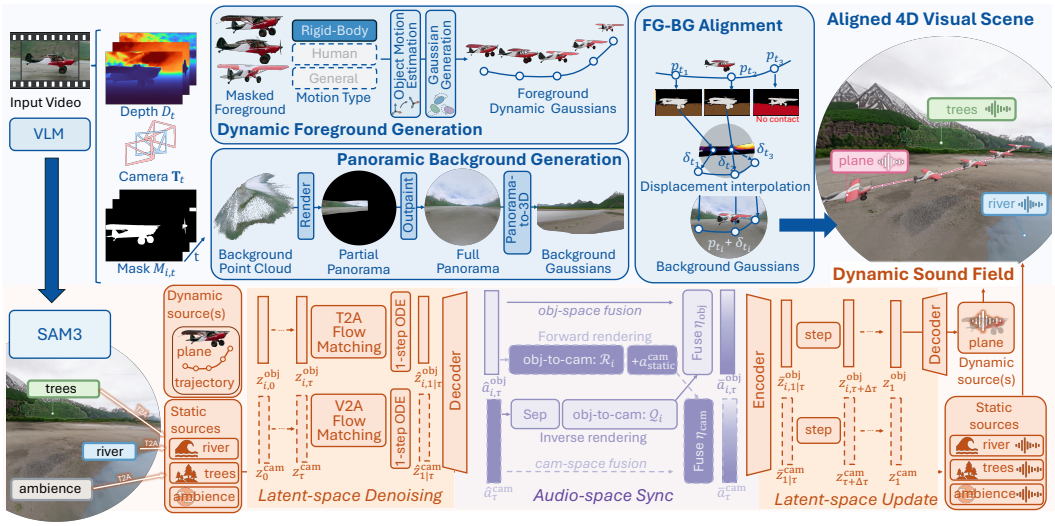


Figure 2: Given a monocular video, SONICVISTA first reconstructs a 4D visual scene (Sec. 4.2) with a static background and dynamic foreground objects. Based on this aligned visual scene representation, it then generates a dynamic spatial sound field (Sec. 4.3) by jointly predicting object-space and camera-space audio through diffusion synchronization. The resulting scene supports real-time interactive exploration (Sec. 4.4), rendering semantically and geometrically synchronized visual and spatial audio streams with moving sound sources.

Evaluation Metrics. We evaluate spatial audio generation in terms of spatial accuracy and semantic consistency. For *spatial accuracy*, we report correlation coefficient (CC) and area under the curve (AUC) between generated and reference spatial energy maps (Kim et al., 2025), together with direction-of-arrival (DoA) errors, including absolute azimuth error $\Delta_{\text{abs}}\theta$, absolute elevation error $\Delta_{\text{abs}}\phi$, and angular error Δ_{Angular} (Liu et al., 2025a). For *semantic consistency*, we compute CLAP similarity between the rendered audio and manually annotated object descriptions. Spatial metrics are evaluated at 5 Hz and averaged over time, while CLAP is computed at the clip level.

4 APPROACH

4.1 OVERVIEW

Given a monocular video $V = \{I_t\}_{t=1}^{T}$, SONICVISTA generates an explorable 4D audio-visual scene $\{\mathcal{S}, \mathcal{A}\}$. We distinguish three coordinate systems: world space is the shared, fixed reference frame for the scene; camera space is the camera's coordinate system at each frame; and object space is the local coordinate system attached to each foreground object. To support object-space audio modeling, we decompose dynamic foreground objects from the static background and explicitly model each object's trajectory, allowing its sound to follow its motion even when multiple objects move independently. All static objects are treated as part of the background. We represent the visual scene as $\mathcal{S} = (\mathcal{G}_{\text{bg}}^{\text{world}}, \{\mathcal{G}_i^{\text{world}}\}_{i=1}^{N})$, where $\mathcal{G}_{\text{bg}}^{\text{world}}$ denotes the static background and $\mathcal{G}_i^{\text{world}}$ denotes the i-th dynamic foreground object, both represented by Gaussians that support real-time rendering.

We represent the spatial sound field as

$$\mathcal{A} = \left(a^{\text{amb}}, \{a_j^{\text{bg}}\}_{j=1}^{M}, \{a_i^{\text{obj}}\}_{i=1}^{N}\right), \tag{3}$$

where a^{amb} denotes diffuse ambient audio, a_j^{bg} denotes audio from the j-th static background source, and a_i^{obj} denotes object-space audio from the i-th dynamic source. The reconstructed scene provides the source locations, trajectories, and geometry required to render these signals from arbitrary listener poses.

Figure 2 illustrates our framework. We first reconstruct a decomposed 4D visual scene in world space (Sec. 4.2): the static component provides geometry and context for ambient sound, while the dynamic component provides object-centric trajectories for moving sound sources. We then reconstruct a dynamic spatial sound field (Sec 4.3) by jointly generating object-space and camera-space

audio through diffusion synchronization, allowing source sounds to remain aligned with object motion and consistent with the rendered camera mixture. During free-viewpoint rendering (Sec. 4.4), a target listener trajectory is used to render novel-view visuals and spatialize the generated object sounds into synchronized audio coherently blended with the background sound.

4.2 4D Visual Scene Generation via Foreground-Background Decomposition

We first preprocess the input video to estimate scene geometry, camera poses, and foreground masks. Conditioned on the observed background regions and visible foreground pixels, we then generate panoramic background Gaussians and dynamic foreground Gaussians, respectively. Foreground generation supports three motion types: rigid-body, human, and general deformable motion, using a shared formulation for object motion estimation and deformable Gaussian generation. Finally, we align the generated foreground objects and background in a common world frame. The resulting 4D scene provides the geometry and object trajectories used for spatial audio generation and rendering.

Video Preprocessing. Given an input video $\{I_t\}_{t=1}^{T}$ with T frames, we use a geometry foundation model (Wang et al., 2026b) to estimate depth maps $\{D_t\}_{t=1}^{T}$ and camera poses $\{\mathbf{T}_t\}_{t=1}^{T}$, and use a segmentation model (Carion et al., 2026) to obtain foreground masks $\{M_{i,t}\}_{i=1,t=1}^{N,T}$, where N is the number of foreground objects. We further use a vision-language model (VLM) (Team et al., 2026) to infer auditory context from the video. It produces a video-level description d_{vid} capturing the overall acoustic scene, including ambient and background sounds, and an object-level description d_i for each sound-producing foreground object i, capturing its sound semantics and temporal activity. The former conditions camera-space audio generation, while the latter conditions the corresponding object-space source generation in Sec. 4.3.

Panoramic Background Generation. We generate the background from observed background regions in the input video. To consolidate these observations into a single panorama for completion, we select the frame with the largest visible background region as an anchor frame and take its coordinate system as the world frame. We then unproject background pixels from the depth maps across all frames and rasterize the aggregated points in the world frame to obtain a partial background panorama $I_{\text{pano}}^{\text{partial}}$. Next, we apply single-camera calibration (Veicht et al., 2024) to align the panorama with the gravity direction and use a panorama diffusion outpainting model (nomadoor, 2026; Labs, 2025a) to generate a complete 360° background panorama $I_{\text{pano}}^{\text{full}}$. Finally, a panorama-to-3D model (Team HY-World, 2026; Labs, 2025b) lifts it into background Gaussians $\mathcal{G}_{\text{bg}}^{\text{world}}$.

Dynamic Foreground Generation. We generate a dynamic Gaussian representation for each foreground object conditioned on its visible pixels in the input video. Dynamic object modeling remains an active area of research, with visual backbones offering different capabilities across motion types. We therefore adopt a unified formulation with motion-type-specific instantiations for three categories: *rigid-body motion*, which preserves object shape; *human motion*, which includes body articulation; and *general deformable motion*, which allows flexible shape changes without assuming a human body model. Across all three types, we model global motion through per-frame scale, rotation, and translation, while treating local deformation differently.

Generation comprises two shared components: *i) Object Motion Estimation:* For each foreground object i and frame t, we estimate a similarity transformation $\mathbf{H}_{i,t}^{\text{obj}\to\text{cam}} = (\alpha_{i,t}, \mathbf{R}_{i,t}, \mathbf{p}_{i,t}) \in \text{Sim}(3)$, where $\alpha_{i,t}$, $\mathbf{R}_{i,t}$, and $\mathbf{p}_{i,t}$ denote scale, rotation, and translation, respectively. This transformation maps the object's locally deformed geometry from object space into camera space at frame t. These parameters describe its scale, orientation, and position, while local shape changes are modeled by deformable Gaussian generation below. *ii) Deformable Gaussian Generation:* Conditioned on visible foreground pixels of object i, we generate its canonical Gaussians $\mathcal{G}_i^{\text{obj}}$ and a time-dependant deformation field $\Phi_{i,t}$. This field maps $\mathcal{G}_i^{\text{obj}}$ to their deformed configuration in object space at frame t, capturing local articulation or shape changes, and reduces to the identity for rigid objects. We apply $\Phi_{i,t}$, followed by the object-to-camera transformation $\mathbf{H}_{i,t}^{\text{obj}\to\text{cam}}$ and the camera-to-world transformation $\mathbf{T}_t^{\text{c}\to\text{w}}$, to obtain dynamic foreground Gaussians $\mathcal{G}_{i,t}^{\text{world}}$. The full formulation and motion-type-specific instantiations are provided in Appendix. The resulting foreground Gaussians and object trajectories are subsequently aligned with the panoramic background.

Foreground-Background Alignment. We align the independently generated foreground and background through global scale alignment and local displacement correction. We first estimate a background scale s in closed form by matching the generated background to the video-derived point

cloud over their common visible region. For each foreground object, we then estimate displacements $\delta_{i,t}$ from correspondences between its local contact neighborhood and the scale-aligned background. For frames without reliable contact, we interpolate the displacements from supported frames, retaining the nearest displacement outside their temporal range. We apply these corrections to the object's world-space positions, $\mathbf{p}_{i,t}^{\mathrm{world}} + \delta_{i,t}$, and its Gaussians to obtain an aligned 4D visual scene.

4.3 Dynamic Spatial Sound Field Generation via Diffusion Synchronization

The reconstructed 4D visual scene provides the geometry and trajectories of sound sources, while the video provides their semantic and temporal context. Our goal is to generate persistent source signals that can be spatially re-rendered along arbitrary listener trajectories while remaining consistent with the acoustic events observed in the input video.

To this end, we represent audio in two complementary spaces. We denote by a^{cam} a video-conditioned camera-space mixture, which captures the overall acoustic content observed by the input camera, and by a_i^{obj} the object-space signal of dynamic source $i \in \mathcal{I}_{\mathrm{dyn}}$, which can be independently re-rendered from novel listener poses. Generating these signals independently, however, does not guarantee that the object tracks explain the camera-space mixture. We therefore couple the two representations through the acoustic renderer induced by the reconstructed 4D scene.

Static Background Sound Generation. For each background sound label, we use a text-to-audio model (Shan et al., 2025) to generate a source signal conditioned on its textual description. Each signal is associated with its corresponding static source region in the reconstructed scene for spatial audio rendering.

Bidirectional Camera-Object Consistency. For each dynamic source $i \in \mathcal{I}_{\mathrm{dyn}}$, the reconstructed scene defines an acoustic rendering operator $\mathcal{R}_i$ that maps its object-space signal to its contribution along the observed camera trajectory. Rendering and summing the source tracks gives a camera-space prediction:

$$\tilde{a}^{\mathrm{cam}} = \sum_{i \in \mathcal{I}_{\mathrm{dyn}}} \mathcal{R}_i\Big(a_i^{\mathrm{obj}}\Big) + a_{\mathrm{static}}^{\mathrm{cam}}, \tag{4}$$

where $a_{\mathrm{static}}^{\mathrm{cam}}$ denotes the rendered static-source and ambient contributions. We require this rendered mixture to remain consistent with the video-conditioned prediction a^{cam}.

Conversely, the camera-space mixture provides a video-aligned cue for each dynamic source. We first isolate source i using a query-conditioned separator Sep_{q_i} and then approximately invert its acoustic propagation:

$$\tilde{a}_i^{\mathrm{obj}} = \mathcal{Q}_i\big(\mathrm{Sep}_{q_i}(a^{\mathrm{cam}})\big), \tag{5}$$

where q_i is the source-specific separation prompt and $\mathrm{Sep}(\cdot)$ is instantiated with SAM-Audio (Shi et al., 2025). Here, $\mathcal{Q}_i$ denotes an approximate inverse renderer, *i.e.*, $\mathcal{Q}_i \circ \mathcal{R}_i \approx \mathrm{Id}$. Together, Eq. 4 and Eq. 5 define the bidirectional consistency constraints $a^{\mathrm{cam}} \approx \tilde{a}^{\mathrm{cam}}$, $a_i^{\mathrm{obj}} \approx \tilde{a}_i^{\mathrm{obj}}$, for all $i \in \mathcal{I}_{\mathrm{dyn}}$. Static and ambient tracks contribute to the forward camera-space mixture but do not receive camera-to-object synchronization feedback.

Dual-Space Diffusion Synchronization. We enforce the above constraints during sampling by synchronizing one camera-space diffusion trajectory with the object-space trajectories of all dynamic sources. We use the flow-matching formulation of HY-VideoFoley (Shan et al., 2025). At each flow step, we first predict clean audio for every branch, synchronize the predictions in audio space through Eq. 4 and Eq. 5, and then map the synchronized predictions back to the latent flow.

Clean prediction. Let $\tau \in [0, 1)$ denote flow time, and let z_τ^k and v_τ^k denote the latent state and predicted velocity of branch k, where $k \in \{\mathrm{cam}\} \cup \mathcal{I}_{\mathrm{dyn}}$. We extrapolate each branch to its predicted clean endpoint and decode it with the audio decoder Dec.

$$\hat{z}_{1|\tau}^k = z_\tau^k + (1-\tau)v_\tau^k, \qquad \hat{a}_\tau^k = \mathrm{Dec}\Big(\hat{z}_{1|\tau}^k\Big). \tag{6}$$

Audio-space synchronization. We first render the object-space predictions into camera space:

$$\tilde{a}_\tau^{\mathrm{cam}} = \sum_{i \in \mathcal{I}_{\mathrm{dyn}}} \mathcal{R}_i\left(\hat{a}_{i,\tau}^{\mathrm{obj}}\right) + a_{\mathrm{static}}^{\mathrm{cam}}. \tag{7}$$

We synchronize the camera-space prediction using

$$\bar{a}_\tau^{\text{cam}} = (1-\eta_{\text{cam}})\hat{a}_\tau^{\text{cam}} + \eta_{\text{cam}}\tilde{a}_\tau^{\text{cam}}, \qquad \eta_{\text{cam}} \in [0,1]. \tag{8}$$

Conversely, the camera-space prediction provides an object-space hint $\tilde{a}_{i,\tau}^{\text{obj}} = \mathcal{Q}_i\left(\text{Sep}_{q_i}\left(\hat{a}_\tau^{\text{cam}}\right)\right)$. The synchronized object-space target is then

$$\bar{a}_{i,\tau}^{\text{obj}} = (1-\eta_{\text{obj}})\,\hat{a}_{i,\tau}^{\text{obj}} + \eta_{\text{obj}}\tilde{a}_{i,\tau}^{\text{obj}}, \qquad \eta_{\text{obj}} \in [0,1]. \tag{9}$$

Synchronized flow update. We encode each synchronized audio target into a clean latent endpoint $\bar{z}_{1|\tau}^k = \text{Enc}\left(\bar{a}_\tau^k\right)$, and replace the branch velocity by the direction toward this synchronized endpoint:

$$v_{\text{sync},\tau}^k = \frac{\bar{z}_{1|\tau}^k - z_\tau^k}{1-\tau}, \qquad \frac{\mathrm{d}z_\tau^k}{\mathrm{d}\tau} = v_{\text{sync},\tau}^k. \tag{10}$$

Repeating this procedure throughout sampling couples the video-conditioned camera trajectory with the object-conditioned source trajectories, producing source signals that remain both video-consistent and independently re-renderable.

4.4 Free-Viewpoint Rendering of Dynamic Scenes

Given a target camera/listener trajectory $p_t = [R_t, \mathbf{x}_t] \in SE(3)$, we render the visual and audio streams in the same world coordinate system. For visuals, we render the static background and dynamic foreground objects at time t using the reconstructed Gaussian splats, compositing the foreground objects according to their estimated trajectories $\mathbf{X}_i(t)$. For audio, we first render each generated object-space signal into the listener coordinate frame using the acoustic renderer, which accounts for source position, motion, attenuation, and visibility. The resulting source contributions are then encoded into ambisonics:

$$a_L(t) = \sum_{i=1}^{N} \sigma_i(t)\, a_i^{\text{obj}}(t_i^{\text{em}})\, y_L\left(R_t^\top \frac{\mathbf{s}_i(t) - \mathbf{x}_t}{\|\mathbf{s}_i(t) - \mathbf{x}_t\|}\right), \tag{11}$$

where $\mathbf{s}_i(t)$ is the world-space source position, t_i^{em} is the emission time, $\sigma_i(t)$ models distance attenuation and occlusion, and y_L denotes spherical harmonics up to order L. Finally, the ambisonics signal is decoded to binaural audio with an HRTF decoder:

$$b(t) = \sum_{\ell=0}^{L} \sum_{m=-\ell}^{\ell} h_{\ell,m} \star a_{\ell,m}(t), \tag{12}$$

where $h_{\ell,m}$ are the left/right HRIR filters for each ambisonics channel. This produces synchronized novel-view video and listener-dependent spatial audio along the target trajectory.

5 Experiments

5.1 Experiment Settings

We evaluate on our newly collected dataset, SonicVista360, with metrics introduced in Sec. 3, by rendering spatial audio at the microphone pose and comparing it with the ground-truth FOA recording. We compare against four prior methods that generate spatial audio from visual inputs:

- VisAGe (Kim et al., 2025): generates ambisonics at the input camera pose from a field-of-view (FoV) video, without supporting novel-view audio rendering. To evaluate it at the target microphone pose, we provide a FoV video rendered from our 4D visual scene at that pose.
- OmniAudio (Liu et al., 2025a): generates ambisonics at the input camera pose from a panoramic video, likewise without supporting novel-view audio rendering. We therefore provide a panoramic video rendered from our 4D visual scene at the target microphone pose.
- SonoWorld (Jin et al., 2026): generates a static 3D audio-visual scene from a single image. Since it does not model scene dynamics, we use the middle frame of the input video and otherwise follow its original setting for evaluation.

Method	Spatial Metrics					Semantic Metrics	
	$\Delta_{\text{abs}}\theta \downarrow$	$\Delta_{\text{abs}}\varphi \downarrow$	$\Delta_{\text{Angular}} \downarrow$	CC ↑	AUC ↑	CLAP_{A} ↑	CLAP_{T} ↑
VISAGE	74.1	24.5	75.9	0.189	0.615	0.078	−0.059
OMNIAUDIO	80.9	30.0	83.1	0.124	0.579	0.234	0.057
SONOWORLD	65.7	24.3	66.6	0.333	0.672	0.365	0.154
SONIC4D	54.3	17.2	55.5	0.446	0.728	0.223	0.103
Ours (T2A)	45.7	16.8	48.4	0.566	0.787	0.349	0.185
Ours (V2A)	43.2	16.9	46.1	0.591	0.800	0.527	0.261
Ours	**36.6**	16.9	**40.7**	**0.647**	**0.835**	**0.560**	**0.274**

Table 1: **Quantitative comparison on SONICVISTA360 for video-to-4D audio-visual scene generation.** Angular errors are reported in degrees.

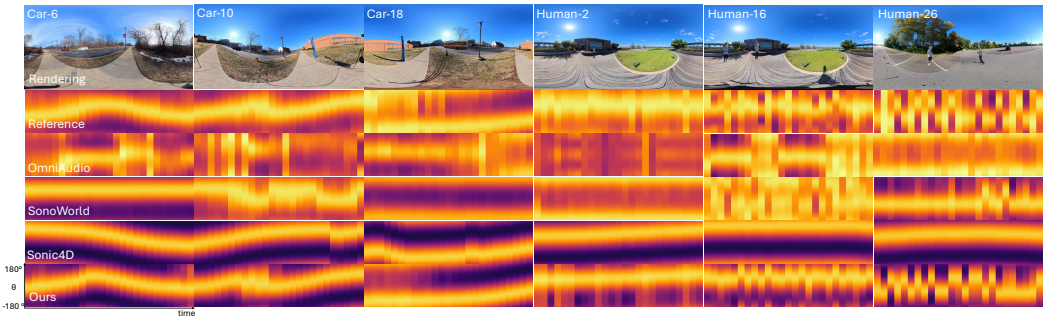


Figure 3: **Time-azimuth RMS visualization of ambisonics.** Top: mid-frame panoramas of each video rendered by our method at microphone poses in SONICVISTA360. Remaining rows compare the reference FOA recordings with baselines and ours. Each map shows the peak-normalized directional RMS response over time and azimuth after spherical-area-weighted aggregation over elevation. Our method more closely follows the reference azimuthal trajectories and temporal structure, while the baselines exhibit directional mismatches or reduced spatiotemporal variation.

- SONIC4D (Xie et al., 2025): is the closest prior method for spatializing video-conditioned monaural audio in 4D scenes. It supports novel-view rendering but assumes a single sound source. We retain its visual backbone and audio pipeline, track the dominant moving sound source, and render at the target microphone pose. For comparison with ground-truth FOA, we adapt its binaural renderer using one omnidirectional and three co-located, orthogonal figure-eight virtual microphones.

We also compare with two variants: *Ours (T2A)*, which uses text-to-audio generation for each dynamic-source track, and *Ours (V2A)*, which instead uses video-to-audio generation conditioned on the input video. Beyond our main task, we further evaluate our synchronization framework for multi-track video-to-audio generation on FAIR-Play (Gao & Grauman, 2019) in App. A.1.

5.2 QUANTITATIVE RESULTS

Quantitative Evaluation on SONICVISTA360. Table 1 shows results on VIDEO-TO-4DAVSCENE using the metrics defined in Sec. 3. Compared with prior video-to-audio generation baselines, including VISAGE (Kim et al., 2025) and OMNIAUDIO (Liu et al., 2025a), SONICVISTA achieves substantially better performance. Compared with SONIC4D (Xie et al., 2025), we also show large gains, confirming the effectiveness of our explicit object-space audio generation and dual-space synchronization mechanisms, which improves spatial audio consistency over a trajectory-based dynamic spatialization baseline.

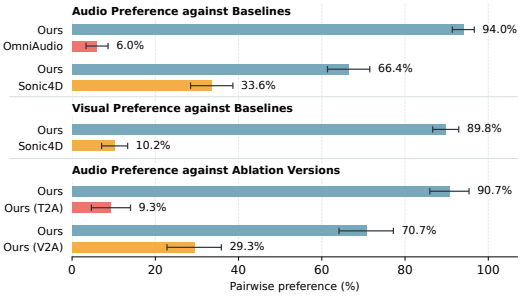


Figure 5: **Pairwise user preferences**. Bars show mean per-participant preference rates, and error bars indicate 95% confidence intervals.

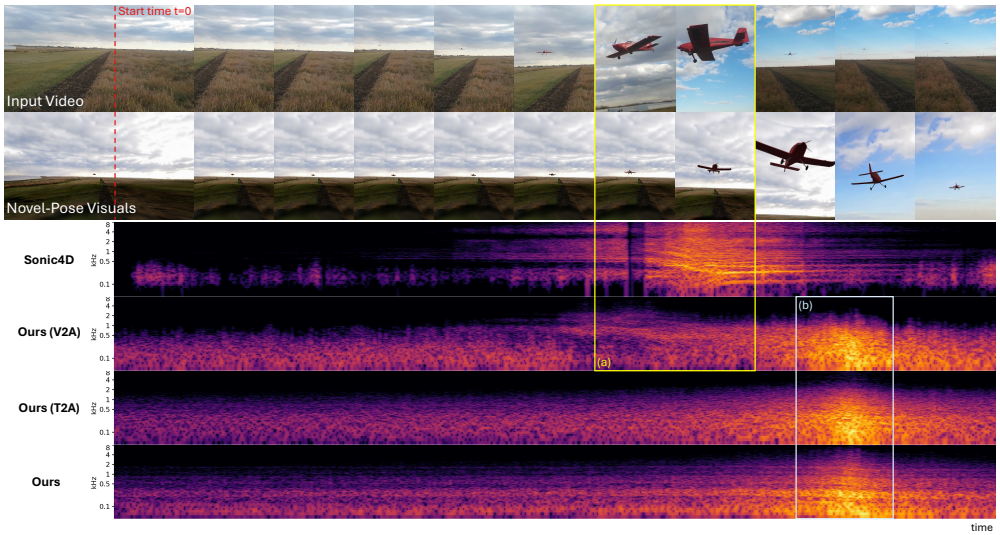


Figure 4: **STFT visualization of novel-listener spatial audio.** (a) and (b) denote the airplane's closest approach to the input camera and novel listener, respectively. While the baselines remain tied to (a), respond at both locations, or lack clear spectral structure, our full model exhibits a dominant response and clear Doppler-related variation at (b), consistent with the novel-view rendering.

User Study. We conduct a user study with 50 participants comparing our method against baselines and ablated variants against nine dynamic scenes from our dataset and web videos, covering multi-source scenarios and rigid-body, human, and general deformable motion. Participants use headphones to assess audio-visual synchronization, spatial direction and distance cues, audio quality, and consistency with a shared sound source description. Comparisons with Sonic4D also assess overall visual quality. As shown in Fig. 5, participants prefer our audio over OmniAudio and Sonic4D in 94.0% and 66.4% of comparisons, respectively, and our visuals over Sonic4D in 89.8% of comparisons. Our full method is also preferred over the T2A and V2A variants in 90.7% and 70.7% of audio comparisons, respectively. See appendix for details.

5.3 Qualitative Results

We qualitatively evaluate the spatial and temporal consistency of generated audio and its adaptation to novel listener positions. Fig. 3 compares time-azimuth RMS maps of generated ambisonics with reference FOA recordings on SonicVista360. Our method more closely follows the reference sound directions and temporal structure, while the baselines exhibit directional mismatches or reduced spatiotemporal variation. Fig. 4 further illustrates adaptation to a novel listener position in the airplane scene. Our full model produces a dominant response and Doppler-related spectral variation around the plane's closest approach to the novel listener, whereas the baselines and ablated variants retain input-view timing, respond at both locations, or lack clear spectral structure. See supplementary video and our interactive demos for more qualitative results.

6 Conclusion

We presented Video-to-4DAVScene, a new task that transforms a monocular video into an explorable, time-varying 4D audio-visual scene, and SonicVista, a training-free framework for this task. Our framework combines a 4D visual scene representation that decomposes scenes into static backgrounds and dynamic foreground objects in a unified world coordinate system, with a novel diffusion synchronization mechanism tailored for dynamic spatial sound field generation. This enables real-time interactive exploration with moving sound sources and temporally, semantically, and geometrically aligned visual and spatial audio rendering. Experiments on real-world dynamic scenes demonstrate that SonicVista outperforms a series of baselines on both objective metrics and in a user study. Future work will explicitly model object materials to support more physically accurate rendering of interaction-driven sounds in dynamic scenes.

AI USE STATEMENT

We used AI tools to refine author-written text and assist with code development for this work. The authors reviewed and revised the AI-assisted text and code and take full responsibility for the final manuscript and implementation.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

Jan Kautz Stan Birchfield Bowen Wen, Wei Yang. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020.

Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS*, 2022.

Jianqi Chen, Biao Zhang, Xiangjun Tang, and Peter Wonka. V2m4: 4d mesh animation reconstruction from a single monocular video. In *ICCV*, 2025a.

Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. *NeurIPS*, 2024.

Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. 2025b.

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 *CVPR*, 2025.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision, 2022. URL https://arxiv.org/abs/2206.04769.

Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019.

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

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

Chen Hongyuan, Chen Xingyu, Youjia Zhang, Xu Zexiang, and Chen Anpei. Motion 3-to-4: 3d motion reconstruction for 4d synthesis. *arXiv preprint arXiv:2601.14253*, 2026.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *CVPR*, 2024.

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

Derong Jin, Xiyi Chen, Ming Lin, and Ruohan Gao. Sonoworld: From one image to a 3d audio-visual scene. In *CVPR*, 2026.

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

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. In *ICLR*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025a.

World Labs. Marble. `https://marble.worldlabs.ai/`, 2025b.

Zitong Lan, Chenhao Zheng, Zhiwei Zheng, and Mingmin Zhao. Acoustic volume rendering for neural impulse response fields. *arXiv preprint arXiv:2411.06307*, 2024.

Hyunjun Lee, Hyunsoo Lee, and Sookwan Han. Syncsde: A probabilistic framework for diffusion synchronization. In *CVPR*, 2025.

Junwon Lee, Juhan Nam, and Jiyoung Lee. Hear what matters! text-conditioned selective video-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36680–36690, June 2026.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *NeurIPS*, 2023.

Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026.

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

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *ICCV*, 2025.

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

Ting-Hsuan Liao, Haowen Liu, Yiran Xu, Songwei Ge, Gengshan Yang, and Jia-Bin Huang. Pad3r: Pose-aware dynamic 3d reconstruction from casual videos. *SIGGRAPH ASIA*, 2025.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4d point clouds. In *CVPR*, 2026.

Yehonathan Litman, Xiaoxuan Ma, Manan Shah, Nicolás Ugrinovic, Kris Kitani, Fernando De la Torre, and Shubham Tulsiani. Lift4d: Harmonizing single-view 3d estimation for 4d reconstruction in-the-wild. *arXiv preprint arXiv:2606.23688*, 2026.

Huadai Liu, Tianyi Luo, Qikai Jiang, Kaicheng Luo, Peiwen Sun, Jialei Wan, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. Omniaudio: Generating spatial audio from 360-degree video. In *ICML*, 2025a.

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*, 2025b.

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

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar. In *ECCV*, 2024.

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

nomadoor. FLUX.2 Klein 9B 360 ERP Outpaint LoRA. `https://huggingface.co/nomadoor/flux-2-klein-9B-360-erp-outpaint-lora`, 2026. Hugging Face model repository.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019.

Lingteng Qiu, Peihao Li, Heyuan Li, Qi Zuo, Xiaodong Gu, Yuan Dong, Weihao Yuan, Rui Peng, Siyu Zhu, Xiaoguang Han, Guanying Chen, and Zilong Dong. Lhm++: An efficient large human reconstruction model for pose-free images to 3d. *arXiv preprint arXiv:2503.10625*, 2025.

Niloy Mitra Tom Monnier Remy Sabathier, David Novotny. Actionmesh: Animated 3d mesh generation with temporal 3d diffusion. In *CVPR*, 2026.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, 2025.

Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation, 2025.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. Sam audio: Segment anything in audio. 2025.

Meng-Li Shih, Ying-Huan Chen, Yu-Lun Liu, and Brian Curless. Prior-enhanced gaussian splatting for dynamic scene reconstruction from casual video. In *SIGGRAPH Asia*, 2025.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. 2025. URL `https://arxiv.org/abs/2511.16624`.

Team HY-World. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026. doi: 10.48550/arXiv.2604.14268. URL `https://arxiv.org/abs/2604.14268`.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*, 2025.

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

Helin Wang, Bowen Shi, Andros Tjandra, John Hoffman, Yi-Chiao Wu, Apoorv Vyas, Najim Dehak, Ann Lee, and Wei-Ning Hsu. Sam audio judge: A unified multimodal framework for perceptual evaluation of audio separation, 2026a. URL https://arxiv.org/abs/2601.19702.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

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

Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *ICCV*, 2025.

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.

Mingyang Xie, Numair Khan, Tianfu Wang, Naina Dhingra, Seonghyeon Nam, Haitao Yang, Zhuo Hui, Christopher Metzler, Andrea Vedaldi, Hamed Pirsiavash, et al. Lavr: Scene latent conditioned generative video trajectory re-rendering using large 4d reconstruction models. *arXiv preprint arXiv:2601.14674*, 2026.

Siyi Xie, Hanxin Zhu, Tianyu He, Xin Li, and Zhibo Chen. Sonic4d: Spatial audio generation for immersive 4d scene exploration. *arXiv preprint arXiv:2506.15759*, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation, 2023. URL https://arxiv.org/abs/2309.16429.

Masaki Yoshida, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Scene2sound: Auditory-grounded soundscape generation for 3d gaussian worlds. *arXiv preprint arXiv:2608.00463*, 2026.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *CVPR*, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *ICCV*, 2025b.

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.

A Additional Quantitative Results

A.1 Multi-track Video-to-Audio Generation.

Our diffusion synchronization framework is not specific to 4D spatial audio. Without a 4D scene representation or acoustic rendering, it can jointly generate a video-conditioned global track and multiple source-specific tracks. We demonstrate this capability on FAIR-Play (Gao & Grauman, 2019), using a random subset of 120 videos containing two visible performers playing musical instruments, singing, or both. For each video, the task is to generate two performer-specific tracks together with their global mixture. We condition the global branch on the full video and each source branch on the corresponding SAM3-masked video. We otherwise use the same synchronization parameters as in our 4D setting, while setting $\mathcal{R}_i = \mathcal{Q}_i = \mathrm{Id}$ to remove spatial rendering.

Table 2: Multi-track V2A on FAIR-Play. Values are means with 95% confidence intervals.

Method	SAM Judge ↑	CLAP ↑	AV-Align ↑
SAM3 → V2A	3.769 [3.625, 3.910]	0.326 [0.314, 0.337]	0.393 [0.370, 0.416]
V2A → SAM Audio	3.297 [3.105, 3.485]	0.166 [0.143, 0.188]	0.293 [0.269, 0.318]
SelVA	3.661 [3.514, 3.804]	0.314 [0.301, 0.328]	**0.444** [0.424, 0.464]
Ours	**4.177** [4.051, 4.296]	**0.372** [0.356, 0.388]	0.418 [0.397, 0.438]

We compare against three baselines. SAM3→V2A first isolates each performer visually and generates its audio independently; V2A→SAM-Audio first generates a global audio track and then separates it into source-specific tracks using SAM Audio (Shi et al., 2025); and SelVA (Lee et al., 2026) directly generates audio for a text-specified visual source. We evaluate source-specific track quality using SAM Audio Judge (Wang et al., 2026a), a reference-free perceptual metric for audio separation, semantic consistency using CLAP (Elizalde et al., 2022), and audio-visual temporal correspondence using AV-Align (Yariv et al., 2023) with a 100 ms tolerance.

As shown in Table 2, our method achieves the highest SAM-Judge and CLAP scores, indicating stronger source-specific perceptual quality and semantic alignment. It also improves AV-Align over both straightforward V2A–separation compositions, while SelVA obtains the highest AV-Align score. Together, these results show that diffusion synchronization extends beyond spatial rendering to jointly generate coherent, source-specific audio tracks from video.

B Motion-Type-Specific Instantiations for Dynamic Foreground Generation

Our dynamic foreground generation follows the same two components for all motion types: *Object Motion Estimation*, which estimates the global similarity transformation $\mathbf{H}_{i,t}^{\mathrm{obj}\to\mathrm{cam}}$ from object space to camera space at each frame, and *Deformable Gaussian Generation*, which produces canonical Gaussians $\mathcal{G}_i^{\mathrm{obj}}$ and their time-dependent deformation $\Phi_{i,t}$. We instantiate these components with different backbones for rigid-body, human, and general deformable motion, while retaining the shared formulation:

$$\mathcal{G}_{i,t}^{\mathrm{world}} = \mathbf{T}_t^{\mathrm{c}\to\mathrm{w}} \left[\left(\mathbf{H}_{i,t}^{\mathrm{obj}\to\mathrm{cam}} \circ \Phi_{i,t} \right) \left(\mathcal{G}_i^{\mathrm{obj}} \right) \right]. \tag{13}$$

Below, we describe how each motion type instantiates the two components. All outputs are subsequently aligned with the panoramic background through the foreground-background alignment described in the main paper.

B.1 Rigid-Body Motion

Object Motion Estimation. We use MV-SAM3D (Li et al., 2026) to jointly reconstruct a proxy mesh and canonical Gaussians from multiple frames of the input video. Using this mesh, we apply FoundationPose (Bowen Wen, 2024) to estimate object rotation and translation from the observed RGB-D sequence via pose registration or tracking.

We select reliable anchor views and perform pose registration at each anchor. These anchor poses then initialize tracking. Between consecutive anchors, we track forward from the earlier anchor and backward from the later anchor, averaging the pose estimates where both directions provide

valid results and retaining the available estimate where only one direction succeeds. Before the first anchor and after the last anchor, we perform backward and forward tracking, respectively.

We skip direct pose estimation in frames where the object is heavily occluded, too small, outside the image, or severely truncated by the image boundary. Tracking chains terminate upon encountering such unreliable observations. We complete the remaining trajectory by interpolating between reliable poses and extrapolating beyond their temporal range. Translation is initialized by linear interpolation and constant-velocity extrapolation, while rotation is interpolated using SLERP within the reliable range and held at the nearest endpoint outside it. Where usable masks remain, we further use their centroids and apparent areas to initialize extrapolated positions, followed by refinement under silhouette-containment and temporal smoothness constraints. Together with reconstruction-derived scale estimates, the completed poses specify $\mathbf{H}_{i,t}^{\mathrm{obj}\to\mathrm{cam}}$.

Deformable Gaussian Generation. We use the canonical Gaussians $\mathcal{G}_i^{\mathrm{obj}}$ jointly produced with the proxy mesh by MV-SAM3D. Reconstruction is conditioned on foreground images, masks, and depth-derived pointmaps from selected keyframes, excluding severely occluded or truncated observations. Since rigid objects preserve their shape, we set $\Phi_{i,t} = \mathrm{Id}$ and reuse the same canonical Gaussians throughout the sequence.

B.2 Human Motion

Object Motion Estimation. We use GEM (Li et al., 2025) to recover SMPL-X (Pavlakos et al., 2019) parameters from the foreground video, including global body motion and articulated poses. The global body pose provides the rotation and translation needed for $\mathbf{H}_{i,t}^{\mathrm{obj}\to\mathrm{cam}}$. We reconcile the human reconstruction with the scene scale using the ratio between predicted scene depth and rendered human depth over foreground pixels. The camera poses from video preprocessing then map the recovered motion into the common world frame.

Deformable Gaussian Generation. We use LHM++ (Qiu et al., 2025) to reconstruct an animatable Gaussian avatar from reference images of the person. The avatar provides $\mathcal{G}_i^{\mathrm{obj}}$, and skeletal skinning driven by the estimated SMPL-X pose sequence instantiates $\Phi_{i,t}$. This captures body articulation while maintaining a shared appearance representation across frames. In our implementation, the posed Gaussians are stored relative to the SMPL-X pelvis, with the pelvis trajectory retained separately, allowing subsequent alignment to adjust the global trajectory without changing the recovered articulation.

B.3 General Deformable Motion

Object Motion Estimation. We use SAM3D (Team et al., 2025) to reconstruct a proxy mesh for each frame, conditioned on the foreground image, mask, and depth-derived pointmap. As in the rigid-body instantiation, we use FoundationPose (Bowen Wen, 2024) to estimate object rotation and translation via pose registration and tracking. Here, however, we use a separately reconstructed proxy mesh for each frame instead of a single fixed mesh, allowing motion estimation to account for the object's changing shape. Together with reconstruction-derived scale estimates, the resulting poses specify $\mathbf{H}_{i,t}^{\mathrm{obj}\to\mathrm{cam}}$.

Deformable Gaussian Generation. We initialize the canonical Gaussians $\mathcal{G}_i^{\mathrm{obj}}$ from the SAM3D reconstruction of a reference frame and use Lift4D (Litman et al., 2026) to fit the time-dependent deformation field $\Phi_{i,t}$ to the per-frame reconstructions and observed foreground images. The deformation field is parameterized by learned control nodes, supporting non-human articulation and flexible shape changes without assuming a human body model. We combine the deformed Gaussians with the estimated object-to-camera transformations using the same formulation as the other two instantiations.

C Details on Foreground-Background Alignment

Foreground reconstruction and panorama-to-3D generation produce geometry independently. Although their coordinate orientations are established by camera estimation and panorama calibration,

their scales and local surfaces may disagree. We address these discrepancies through a scene-level background scale followed by time-varying foreground displacement corrections.

C.1 Global Scale Alignment

We render the video-derived background point cloud and the generated background Gaussians into the same panoramic view. Let $d_{\text{init}}(u)$ and $d_{\text{bg}}(u)$ denote their radial depths at panorama pixel u, respectively. We retain pixels with valid depths in both renderings and positive observation confidence $w_u = C_{\text{obs}}(u)$. To suppress inconsistent correspondences, we reject outliers in $\log(d_{\text{init}}(u)/d_{\text{bg}}(u))$ using the weighted median and median absolute deviation. Let $\mathcal{U}$ denote the retained pixel set.

We estimate a single background scale shared by all foreground objects:

$$s = \arg\min_{\tilde{s}>0} \sum_{u\in\mathcal{U}} w_u \left(\tilde{s}\, d_{\text{bg}}(u) - d_{\text{init}}(u)\right)^2 = \frac{\sum_{u\in\mathcal{U}} w_u\, d_{\text{bg}}(u) d_{\text{init}}(u)}{\sum_{u\in\mathcal{U}} w_u\, d_{\text{bg}}(u)^2}. \tag{14}$$

Equivalently, defining $\mathbf{X}_u = d_{\text{bg}}(u)\mathbf{r}(u)$ and $\mathbf{Y}_u = d_{\text{init}}(u)\mathbf{r}(u)$ along the shared unit panorama ray $\mathbf{r}(u)$, the solution is

$$s = \frac{\sum_{u\in\mathcal{U}} w_u\, \mathbf{X}_u^\top \mathbf{Y}_u}{\sum_{u\in\mathcal{U}} w_u\, \|\mathbf{X}_u\|_2^2}. \tag{15}$$

We apply this scale to the generated background geometry and transform it into the common world frame using the known panorama-to-world transformation.

C.2 Local Displacement Estimation

Global scale alignment does not resolve local geometric differences between the two reconstructions. For each object i and frame t, we collect background points in a neighborhood of its foreground mask $M_{i,t}$. We exclude foreground pixels and use depth and surface-normal discontinuities to restrict the neighborhood to connected geometry. For supported objects, such as a car on a road, this neighborhood captures the nearby supporting surface.

We project these points into the reference panorama and retain correspondences that are visible and depth-consistent with the initialized background, with valid geometry in the generated background. For each retained correspondence p, let $\mathbf{x}_{i,t,p}$ be the video-derived point and $\mathbf{y}_{i,t,p}$ its corresponding point on the scale-aligned background, both in world space. Their displacement is $\Delta\mathbf{x}_{i,t,p} = \mathbf{y}_{i,t,p} - \mathbf{x}_{i,t,p}$.

After rejecting displacement outliers, we estimate the local translation by a weighted average:

$$\widehat{\delta}_{i,t} = \frac{\sum_{p\in\Omega_{i,t}} w_{i,t,p}\, \Delta\mathbf{x}_{i,t,p}}{\sum_{p\in\Omega_{i,t}} w_{i,t,p}}, \tag{16}$$

where $\Omega_{i,t}$ is the retained support set and $w_{i,t,p}$ combines depth confidence with background rendering opacity. We accept a displacement estimate only when its support is sufficient and its residual displacements are consistent.

C.3 Displacement Interpolation

Frames without reliable local support, such as an airborne interval, do not provide direct displacement constraints. We propagate the accepted estimates through a temporally regularized correction field. Let $\mathcal{T}_i$ denote the supported frames and $c_{i,t}$ the confidence of each accepted estimate. We solve

$$\min_{\{\delta_{i,t}\}_{t=1}^{T}} \sum_{t\in\mathcal{T}_i} c_{i,t}\, \rho\Big(\|\delta_{i,t} - \widehat{\delta}_{i,t}\|_2\Big) + \eta \sum_{t=1}^{T-1} \|\delta_{i,t+1} - \delta_{i,t}\|_2^2, \tag{17}$$

where ρ is a Huber penalty and η controls temporal smoothness. We solve this objective using iteratively reweighted least squares.

This optimization smooths the supported displacement estimates and linearly interpolates the resulting offsets across intervals without reliable contact. Outside the supported temporal range, the

correction remains equal to the nearest supported offset. If no reliable support is available throughout the sequence, we set $\delta_{i,t} = \mathbf{0}$ and retain the original trajectory.

Finally, we update each object's world-space position as

$$\widetilde{\mathbf{p}}_{i,t}^{\text{world}} = \mathbf{p}_{i,t}^{\text{world}} + \delta_{i,t}, \tag{18}$$

and apply the same translation to its Gaussians, preserving object rotations and local deformations. Together with the scale-aligned background, these corrected foreground objects form the aligned 4D visual scene.

D Ablation on Visual Components

We examine two components of visual scene generation: the displacement correction in foreground-background alignment and the panorama-to-3D backbone used for background generation. Fig. 6 provides qualitative comparisons.

Foreground-background displacement correction. We compare our full alignment, which combines global scale alignment with local displacement corrections, against a variant that retains only scale alignment. Global scale alone cannot resolve local geometric discrepancies between the independently reconstructed foreground and background. As shown in Fig. 6, the scale-only variant places much of the car below the road surface. Adding displacement corrections substantially reduces this interpenetration and improves object-ground contact. This comparison illustrates the importance of accounting for local geometry in addition to global scale.

Panorama-to-3D backbone. We compare HYWorld 2.0 (Team HY-World, 2026) and Marble (Labs, 2025b) for lifting the completed background panorama into 3D Gaussians. Both produce usable background reconstructions, demonstrating that our framework accommodates different panorama-to-3D backbones. In our qualitative comparisons, however, HYWorld 2.0 exhibits more visible artifacts, including distorted vegetation and spurious geometry, while Marble produces cleaner renderings. For reproducibility, we use HYWorld 2.0 for all quantitative evaluations. For improved visual quality, we use the proprietary Marble model in our qualitative comparisons and demos.

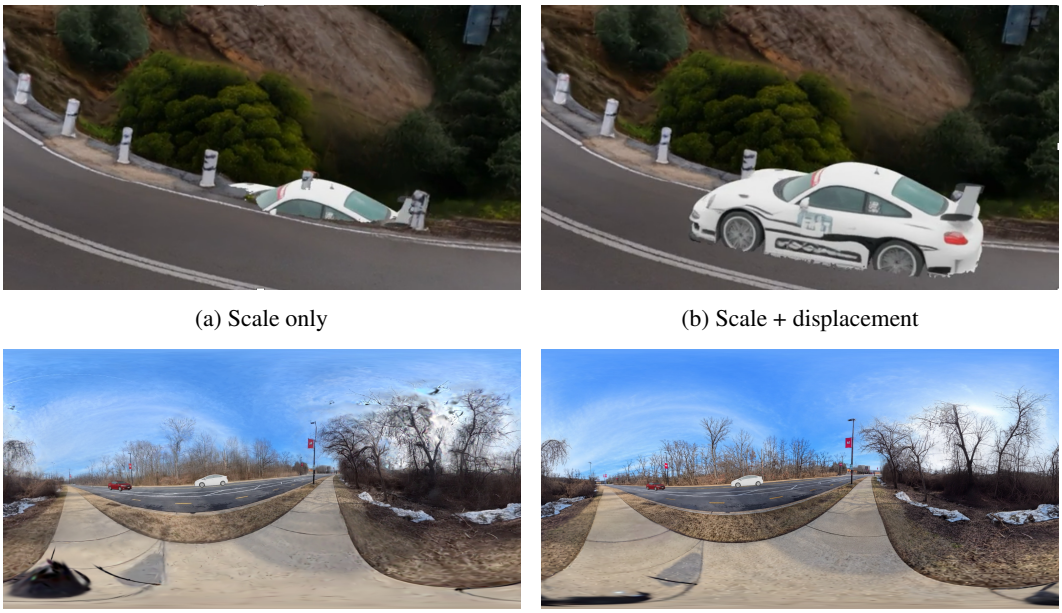
E Details on Our SonicVista360 Dataset

We collect SonicVista360 to evaluate spatial audio generation at listener positions distinct from the input camera viewpoints. The dataset contains 54 dynamic scenes: 25 vehicle scenes and 29 human scenes, covering rigid-body motion and human articulation. Each scene includes a panoramic video and synchronized first-order ambisonic (FOA) audio recorded at a separate listening position. The dataset includes both single-source and multi-source scenarios, with 13 scenes containing two dynamic foreground sound sources.

Vehicle scenes. We capture 25 vehicle scenes across three locations, featuring police vehicles, pickup trucks, sedans, and buses. These scenes cover different vehicle appearances and acoustic characteristics, with vehicles moving through the scene along different trajectories. Three scenes contain two vehicles, requiring the generated audio to account for multiple moving sources and their respective spatial cues. The remaining 22 scenes contain a single foreground vehicle.

Human scenes. We capture 29 human scenes across three locations. Activities include walking, jogging, jumping, foot stomping, clapping, burpees, and clapping while walking. These activities produce footsteps, hand claps, and ground-contact sounds with distinct temporal patterns, enabling evaluation of both spatial audio and synchronization with visible actions. Ten scenes contain two people performing either the same activity or different activities, such as walking and jogging, or clapping and jogging. The remaining 19 scenes contain one person. Some recordings additionally include static background sounds, such as flowing water, providing mixtures of dynamic foreground events and persistent background audio.

Together, these scenes cover diverse source motions, acoustic events, and overlapping sound sources. The spatial separation between the input camera and the recording microphone enables evaluation of generated audio from a novel listening position.



(a) Scale only

(b) Scale + displacement

(c) HYWorld 2.0

(d) Marble

Figure 6: Qualitative ablation on visual components. **Top:** Ablation of displacement correction in foreground-background alignment. Both variants include global scale alignment; adding local displacement corrections reduces foreground interpenetration with the ground. **Bottom:** Both panorama-to-3D backbones produce usable scenes, but Marble yields fewer visible reconstruction artifacts in the illustrated example.

We follow the annotation and capture-device setup protocols of SONOSCENE360 (Jin et al., 2026).