Sonic4D: Spatial Audio Generation for Immersive 4D Scene Exploration

Siyi Xie[1*], Hanxin Zhu[1,2*†], Xinyi Chen[1], Tianyu He[3], Xin Li[1‡], Zhibo Chen[1,2‡],

[1]University of Science and Technology of China
[2]Zhongguancun Academy, Beijing, China
[3]Microsoft Research Asia
{ustc2020xsy, hanxinzhu, chenxinyi0022}@mail.ustc.edu.cn,
tianyuhe@microsoft.com , {xin.li, chenzhibo}@ustc.edu.cn

Abstract

Recent advancements in 4D generation have demonstrated its remarkable capability in synthesizing photorealistic renderings of dynamic 3D scenes. However, despite achieving impressive visual performance, almost all existing methods overlook the generation of spatial audio aligned with the corresponding 4D scenes, posing a significant limitation to truly immersive audiovisual experiences. To mitigate this issue, we propose **Sonic4D**, a novel framework that enables spatial audio generation for immersive exploration of 4D scenes. Specifically, our method is composed of three stages: 1) To capture both the dynamic visual content and raw auditory information from a monocular video, we first employ pre-trained expert models to generate the 4D scene and its corresponding monaural audio. 2) Subsequently, to transform the monaural audio into spatial audio, we localize and track the sound sources within the 4D scene, where their 3D spatial coordinates at different timestamps are estimated via a pixel-level visual grounding strategy. 3) Based on the estimated sound source locations, we further synthesize plausible spatial audio that varies across different viewpoints and timestamps using physics-based simulation. Extensive experiments have demonstrated that our proposed method generates realistic spatial audio consistent with the synthesized 4D scene in a training-free manner, significantly enhancing the immersive experience for users.

Introduction

Benefiting from large-scale data available (Soomro, Zamir, and Shah 2012; Schuhmann et al. 2021; Yu et al. 2023) and recent advancement in generative models (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020), 4D generation (*i.e.*, dynamic 3D scene generation) (Jiang et al. 2024; Ren et al. 2023; Zeng et al. 2024b; Zhu et al. 2025) has emerged as a promising direction due to its powerful capability in modeling complex spatiotemporal dynamics of real-world scenes. By enabling spatiotemporally consistent renderings from arbitrary camera viewpoints, 4D generation facilitates various downstream applications such as AR/VR (Li et al. 2024a; Fritsch and Klein 2017), robotics (Khalid et al. 2022; Hann et al. 2020), and autonomous driving (Wang et al. 2024; Min et al. 2024).

However, while existing 4D generation methods (Hann et al. 2020; Liu et al. 2025b; Zhang et al. 2024; Zeng et al. 2024a) achieve impressive visual results, they typically neglect the generation of spatial audio consistent with the corresponding 4D scene (*i.e.*, audio that varies with the listener's viewpoint and follows physical acoustic principles), limiting the overall immersive experience to a large extent.

To address this limitation, in this paper we propose Sonic4D, a novel framework that enables free-viewpoint rendering of both dynamic visual content and spatially consistent audio, thereby ensuring more immersive audiovisual exploration within the generated 4D scene. To this end, as shown in Fig. 1, we design a three-stage pipeline: **1) Dynamic Scene and Monaural Audio Generation.** To empower novel-view visual renderings and provide essential spatial priors for spatial audio generation, we first leverage a pre-trained 4D generative model to synthesize the dynamic 3D scene from a monocular video. In parallel, a video-to-audio generative model is utilized to produce monaural audio that is semantically aligned with the generated 4D scene, serving as the raw acoustic input for subsequent spatial audio rendering. **2) 3D Sound-Source Localization and Tracking.** To enable accurate physical simulation of dynamic sound propagation, we further estimate the sound source's trajectory in 3D space (*i.e.*, the 3D locations of the sound source at different timestamps). To achieve this goal, we first use a multimodal large language model (MLLM) to perform pixel-level visual grounding on the input video, identifying the 2D coordinates of the sound source in each frame. These 2D positions are then back-projected onto the dynamic point cloud reconstructed in the previous stage, yielding a sequence of 3D coordinates that represent the sound source's trajectory. **3) Physics-Driven Spatial Audio Synthesis.** Given the estimated sound source location and the generated 4D environment, a physics-based simulation using acoustic room impulse responses (RIRs) is employed to simulate plausible spatial audio, enabling more immersive experiences in complex 4D scenes.

Notably, our method adopts a modular architecture that facilitates the integration of state-of-the-art pre-trained expert models, ensuring that the framework remains extensible and future-proof, with performance continually improving

*Equal Contribution
†Project Lead
‡Corresponding Authors

as more advanced expert models become available.

The main contributions of this paper can be summarized as follows:

- We propose Sonic4D, a novel framework that achieves spatial audio generation for immersive 4D scene exploration. To the best of our knowledge, this is **the first work** that introduces spatial audio into the context of 4D generation.
- We propose a three-stage framework to achieve spatial audio generation: **1) Dynamic Scene and Monaural Audio Generation** to extract semantically aligned visual and audio priors from a single video; **2) 3D Sound-Source Localization and Tracking** to recover sound source's 3D trajectory for precise acoustic simulation; **3) Physics-Driven Spatial Audio Synthesis** to render dynamic, viewpoint-adaptive binaural audio via physics-based room impulse response simulation.
- Extensive experiments have demonstrated that our framework can effectively generate spatial audio consistent with 4D visual content, enabling much more immersive and coherent audiovisual experiences.

Related Work

4D Generation

Early advances in 4D generation have built upon techniques such as Score Distillation Sampling (SDS) (Poole et al. 2022), which leverages generative priors of pre-trained diffusion models to optimize dynamic scene representations. This paradigm laid the groundwork for methods such as Dream-in-4D (Zheng et al. 2024), which further integrated motion priors from pretrained video diffusion models into dynamic NeRF-based representations (Pumarola et al. 2021; Yan, Li, and Lee 2023). Concurrently, models like Consistent4D (Jiang et al. 2024) and 4Diffusion (Zhang et al. 2024) focused on enhancing spatiotemporal consistency, utilizing multiview interpolation, temporal alignment modules, or synchronized training strategies. Works such as Diffusion4D (Liang et al. 2024), PLA4D (Miao et al. 2024), and 4Dynamic (Yuan et al. 2024) further emphasized controllability and efficiency by leveraging pixel-aligned supervision, video-based guidance, or hybrid representations combining mesh and Gaussian structures. Recent studies such as Free4D (Liu et al. 2025b) and MVTokenFlow (Huang et al. 2025) further improve spatial-temporal coherence and visual quality without tuning large generative models.

Though achieving remarkable results, all these methods focus solely on visual content rendering, neglecting the generation of spatial audio that aligns with the dynamic scene. In contrast, in this paper we propose a novel framework for spatial audio generation that enables more coherent and immersive exploration of 4D scenes from arbitrary viewpoints.

Spatial Audio Generation

Early spatial audio generation methods (Gao and Grauman 2019; Xu et al. 2021; Leng et al. 2022) predominantly took monaural inputs and "binauralized" them via neural networks conditioned on visual cues. In recent years, the advent of powerful generative architectures has spurred end-to-end models that accept text prompts (Sun et al. 2024), images (Sun et al. 2024; Dagli et al. 2024), spatial parameters (Heydari et al. 2025; Kushwaha et al. 2025), silent videos (Kim, Yun, and Kim 2025), or full 360° panoramas (Liu et al. 2025a) to jointly learn semantic and spatial audio features. For example, SpatialSonic (Sun et al. 2024) leverages spatial-aware encoders and azimuth state matrices within a latent diffusion framework to provide fine-grained multimodal spatial guidance for stereo audio generation. ViSAGe (Kim, Yun, and Kim 2025) fuses CLIP visual embeddings with autoregressive neural audio codec modeling to generate coherent FOA directly from silent video frames. Recently, OmniAudio (Liu et al. 2025a) addressed the 360° spatial audio gap by using a dual-branch architecture to fuse panoramic and FoV streams, enabling high-fidelity FOA generation from full spherical video content.

Although these generative methods achieve impressive multimodal results, they still face two key limitations. First, data scarcity and domain gaps in end-to-end learning restrict the fidelity of spatial audio rendering. Second, even with 360° video inputs, the synthesized sound field is bound to a fixed camera pose and cannot support dynamic listener motion or free-viewpoint changes. In contrast, our method decouples semantic audio synthesis from spatial rendering and leverages RIR-based convolution to physically simulate spatial acoustics for a moving listener in dynamic scenes.

Sound Source Localization

Sound source localization aims to find the location of sound sources in an image or video frame. Initial studies (Senocak et al. 2018; Chen et al. 2021; Senocak et al. 2023; Um, Kim, and Kim 2023) addressed this problem by leveraging methods like cross-modal attention and contrastive learning to establish effective alignment between audio and visual modalities, mainly focusing on single sound source localization. Some works have taken this further to achieve multi-sound source localization (Mo and Tian 2023; Kim et al. 2024), enabling the simultaneous localization of multiple sound sources from mixed audio and visual inputs. However, these approaches focus on exploring audio–visual alignments and cannot localize sound sources without ground-truth audio. Recently, the rapid advancement of Multimodal Large Language Models (Munasinghe et al. 2024; Li et al. 2024b) (MLLMs) has made it possible to perform audio-visual grounding with strong generalization capabilities. In this work, we leverage the powerful prior knowledge of pretrained MLLMs to directly perform pixel-level visual grounding on the video input for sound source localization, followed by back-projection into 3D space to achieve accurate 3D sound source positioning and further spatial audio simulation.

Methods

Given a monocular input video $\boldsymbol{V}^s = \{\boldsymbol{I}_i^s\}_{i=1}^n \in \mathbb{R}^{n\times 3\times H\times W}$, our goal is to generate both novel-view videos and corresponding spatial audio given arbitrary target camera trajectory $\boldsymbol{T}^r = \{\boldsymbol{T}_i^r\}_{i=1}^n \in \mathbb{R}^{n\times 4\times 4}$, where each ma-

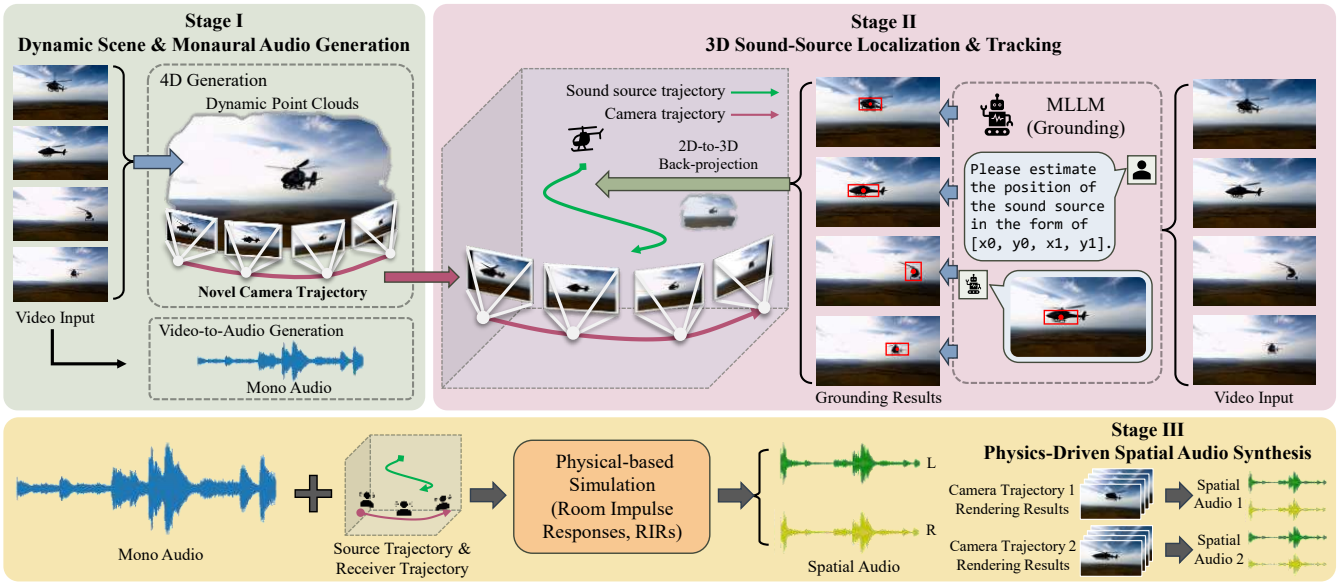

Figure 1: Pipeline of Sonic4D. Our method is composed of three stages: 1) Dynamic Scene and Monaural Audio Generation: extracting semantically aligned visual scenes and audio priors from monocular videos; 2) 3D Sound-Source Localization and Tracking: estimating the sound source's trajectory in 3D space for physically accurate sound propagation modeling; 3) Physics-Driven Spatial Audio Synthesis: leveraging a physics-based room impulse response simulation to realize spatial audio simulation.

trix $\boldsymbol{T}_i^r$ denotes the homogeneous camera pose at time i. As shown in Fig. 1, the overall pipeline is divided into three stages: 1) Dynamic scene and monaural audio generation; 2) 3D sound-source localization and tracking; 3) Physics-driven spatial audio synthesis. Each stage is discussed in detail in the following subsections.

Stage I: Dynamic Scene and Monaural Audio Generation

To obtain a novel-view video $\boldsymbol{V}^r$ while simultaneously capturing dynamic spatial priors for spatial audio simulation, a model that can produce high-fidelity, free-viewpoint renderings of dynamic scenes with strong spatiotemporal consistency is required. Motivated by recent progress in 4D generation, we adopt TrajectoryCrafter (YU et al. 2025), a pretrained video generative model capable of synthesizing high-fidelity videos along arbitrary camera trajectories from a single-view input video, as our 4D content generator.

Specifically, we first apply a monocular depth estimator (Hu et al. 2025) to estimate the depth maps $\boldsymbol{D}^s = \{\boldsymbol{D}_i^s\}_{i=1}^n \in \mathbb{R}^{n\times h\times w}$ of the input video $\boldsymbol{V}^s$. Subsequently, we lift $\boldsymbol{V}^s$ into a set of dynamic point clouds $\boldsymbol{P} = \{\boldsymbol{P}_i\}_{i=1}^n$ using inverse perspective projection $\boldsymbol{\Phi}^{-1}$, which is formulated as follows:

$$\boldsymbol{P}_i = \Phi^{-1}([\boldsymbol{I}_i^s, \boldsymbol{D}_i^s], \boldsymbol{K}), \tag{1}$$

where $\boldsymbol{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix. Using the estimated dynamic point clouds, we synthesize novel-view renderings $\boldsymbol{I}^r = \{\boldsymbol{I}_i^r\}_{i=1}^n$ by projecting $\boldsymbol{P}$ onto the target camera trajectory $\boldsymbol{T}^r = \{\boldsymbol{T}_i^r\}_{i=1}^n \in \mathbb{R}^{n\times4\times4}$ using the following equation:

$$\boldsymbol{I}_i^r = \Phi\left(\boldsymbol{T}_i^r \cdot \boldsymbol{P}_i, \boldsymbol{K}\right), \tag{2}$$

where $\boldsymbol{\Phi}$ denotes the projection operation. The rendered results $\boldsymbol{I}^r$ are then used as conditioning inputs to the video diffusion model of TrajectoryCrafter, guiding the generation of high-fidelity novel-view videos.

Next, we prepare the raw monaural audio signal required for subsequent physics-based spatial audio simulation. To this end, we adopt MMAudio (Cheng et al. 2024), a transformer-based video-to-audio synthesis model, to generate monaural audio $\boldsymbol{A}^m \in \mathbb{R}^{1\times T}$ that is temporally and semantically aligned with the input video $\boldsymbol{V}^s$. By decomposing the task of immersive 4D scene exploration into two sub-tasks—*i.e.*, 4D scene generation and spatial audio synthesis—and leveraging pretrained expert models to extract the required visual and acoustic information, we can achieve greater flexibility and continually improve performance as upstream models evolve.

Stage II: 3D Sound Source Localization and Tracking

With the dynamic point cloud $\boldsymbol{P}$ and its associated monaural audio $\boldsymbol{A}^m$, we proceed to localize and track the sound source in 3D space to facilitate physics-based spatial audio generation. However, directly tracking the sound source in a 4D environment remains highly challenging. Therefore, we first localize the sound source in the 2D pixel space and then back-project the estimated 2D coordinates into 3D space, thereby modeling the sound source trajectory within the generated dynamic scene.

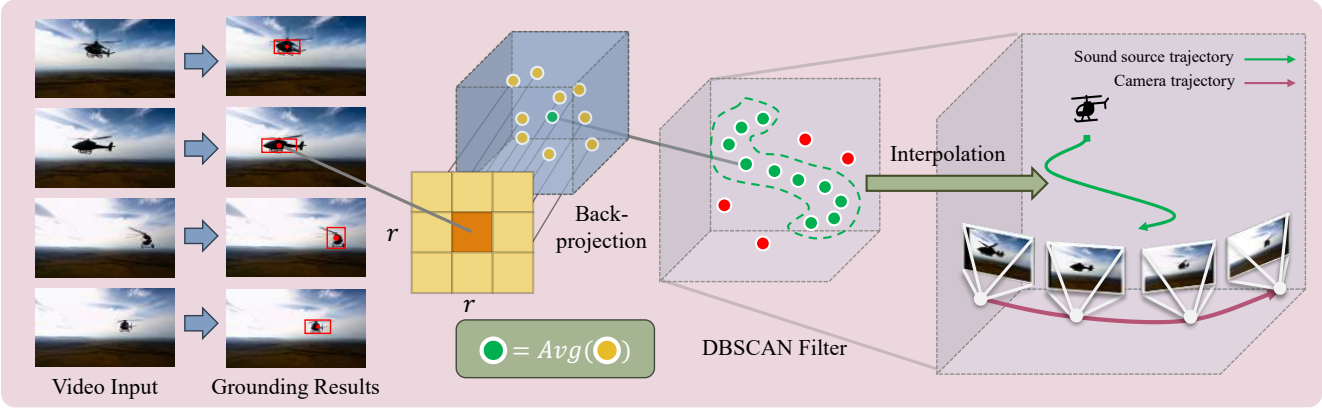

Figure 2: Illustration of Stage II: We localize the sound source in each frame using GroundingGPT (Li et al. 2024b), back-project the 2D grounding results to 3D via depth, and apply DBSCAN (Ester et al. 1996) to obtain a smooth trajectory.

To achieve this goal, existing 2D sound-source localization methods (Um, Kim, and Kim 2023; Senocak et al. 2023; Park, Senocak, and Chung 2024) typically require ground-truth audio inputs and are often tailored to specific scenarios and input sizes. This dependence on "true" audio runs counter to Sonic4D's very generation-centric design. In contrast, given the rich commonsense knowledge and reasoning abilities of multimodal large language models (MLLMs), we propose to leverage GroundingGPT (Li et al. 2024b), a highly generalizable MLLM focused on grounding tasks and trained on a variety of audio–visual datasets for sound source localization, as our 2D sound source localization component.

Concretely, as shown in Fig. 1, we craft a textual prompt and feed the source video $\boldsymbol{V}^s$ frame by frame together with the prompt into GroundingGPT (Li et al. 2024b). Assuming the grounding result for the i-th frame $\boldsymbol{I}_i^s$ is represented by a bounding box $[x_0, y_0, x_1, y_1]$, we define the sound source pixel coordinates as the center of this bounding box, *i.e.*:

$$(u_i, v_i) = \left(\left\lfloor \frac{x_0+x_1}{2} \times W \right\rfloor, \left\lfloor \frac{y_0+y_1}{2} \times H \right\rfloor \right). \quad (3)$$

where W and H are the width and height of $\boldsymbol{V}^s$. Let $d_i = \boldsymbol{D}(u_i, v_i)$ denote the depth at pixel (u_i, v_i). Then the 3D point of the sound source in camera coordinate system corresponding to frame i is given by

$$\boldsymbol{X}_i = \pi^{-1}\big((u_i, v_i), d_i\big). \quad (4)$$

Here, π^{-1} denotes the back-projection operator that maps a pixel coordinate and its depth into a 3D point. Considering the sporadic errors introduced by monocular depth estimation at individual pixels, we compute the 3D coordinates of an $r \times r$ pixel patch surrounding (u_i, v_i) and average them to obtain the sound-source's 3D location as shown in Figure 2:

$$\bar{\boldsymbol{X}}_i = \frac{1}{|\mathcal{P}^r(u_i, v_i)|} \sum_{(u,v)\in\mathcal{P}^r(u_i,v_i)} \pi^{-1}\big((u, v), \boldsymbol{D}(u, v)\big). \quad (5)$$

where $\mathcal{P}^r(u_i, v_i)$ denotes the set of all pixels (u, v) such that $|u - u_i| \le (r-1)/2$ and $|v - v_i| \le (r-1)/2$. Nonetheless, due to uncertainty in the bounding-box predictions, the back-projected trajectory $\{\bar{\mathbf{X}}_i\}$ may still exhibit significant fluctuations. To mitigate this, we apply DBSCAN (Ester et al. 1996) (Density-Based Spatial Clustering of Applications with Noise), a clustering-based outlier detection algorithm that groups dense regions while marking sparse points as noise, to filter out spurious estimates. Points marked as noise are then temporally filled in via linear interpolation between the nearest non–noise neighbours to produce a smooth trajectory $\mathbf{Traj}_{src} = \{\boldsymbol{X}_i\}_{i=1}^n$.

In addition to the sound source's trajectory, the receiver trajectory is also required for acoustic simulation. We use the user-specified camera trajectory as the receiver trajectory. For each frame i, the camera pose is defined as

$$\boldsymbol{T}_i^r = \begin{bmatrix} \boldsymbol{R}_i^r & \boldsymbol{t}_i^r \\ \mathbf{0}^\top & 1 \end{bmatrix}, \quad (6)$$

where $\boldsymbol{R}_i^r \in \mathbb{R}^{3\times3}$ is the rotation matrix and $\boldsymbol{t}_i^r \in \mathbb{R}^3$ is the translation vector. The receiver's 3D position is therefore $\boldsymbol{r}_i = \boldsymbol{t}_i^r$,and the full receiver trajectory is $\mathbf{Traj}_{rsv} = \{\boldsymbol{r}_i\}_{i=1}^n$.

Stage III: Physics-Driven Spatial Audio Synthesis

Given the sound source's trajectory $\mathbf{Traj}_{src}$, the receiver trajectory $\mathbf{Traj}_{rsv}$, and the monaural audio signal $\boldsymbol{A}^m$, we perform a physics-based spatial audio simulation.

A central element of our simulation is the computation of Room Impulse Responses (RIRs) via the Image Source Method (ISM) (Allen and Berkley 1979). Building on the Image-Source Method for static RIRs, we apply gpuRIR (Diaz-Guerra, Miguel, and Beltran 2021) to simulate dynamic sources and receivers by segmenting the trajectories into short intervals over which they are assumed stationary. Concretely, let the total signal length be N samples and split it into M segments at sample indices

$$0 = n_0 < n_1 < \cdots < n_M = N,$$

For each segment $i \in [1, M]$, let $\mathbf{x}_i^s = \mathbf{Traj}_{src}[i]$ denote the position of the sound source and $\mathbf{x}_i^r = \mathbf{Traj}_{rsv}[i]$ denote the position of the receiver. We simulate binaural microphone placement by offsetting from $\mathbf{x}_i^r$ in the direction parallel to the camera's rendering plane, yielding the left and right microphone positions $\mathbf{x}_{i,L}^r$ and $\mathbf{x}_{i,R}^r$. We compute the room impulse responses at the left and right microphones using the Image Source Method (ISM):

$$h_{i,L}(\tau) = \mathrm{ISM}\big(\mathbf{x}_i^s,\ \mathbf{x}_{i,L}^r\big), \tag{7}$$

$$h_{i,R}(\tau) = \mathrm{ISM}\big(\mathbf{x}_i^s,\ \mathbf{x}_{i,R}^r\big). \tag{8}$$

Let the mono audio block for segment i be

$$b_i[k] = \boldsymbol{A}^m[n_i + k], \quad 0 \le k < n_{i+1} - n_i. \tag{9}$$

Then the left and right binaural signals are obtained by block-wise convolution:

$$y_L[n] = \sum_{i=0}^{M-1} \big(b_i * h_{i,L}\big)[\,n - n_i\,], \tag{10}$$

$$y_R[n] = \sum_{i=0}^{M-1} \big(b_i * h_{i,R}\big)[\,n - n_i\,], \tag{11}$$

where $h_{i,L}$ and $h_{i,R}$ are the left- and right-ear impulse responses in h_i. Putting it all together, the final binaural signal is

$$\boldsymbol{A}^b = \big(y_L[n],\, y_R[n]\big). \tag{12}$$

To ensure the output audio is in a valid range for playback, we normalize the signal to the range $[-1, 1]$:

$$\boldsymbol{A}^b \leftarrow \frac{\boldsymbol{A}^b}{\max\left(|\boldsymbol{A}^b|\right)}. \tag{13}$$

Compared with learning-based methods for spatial audio generation, our physics-based audio rendering ensures physical plausibility by explicitly modeling sound propagation using room impulse responses, resulting in more realistic spatial cues such as directionality and reverberation.

By decomposing the task into three modular stages, our framework achieves high flexibility and extensibility. Each stage leverages either pretrained models or physical priors, allowing Sonic4D to naturally scale with future advances. This decoupled design further enables plug-and-play upgrades, making Sonic4D applicable to a wide range of immersive content creation scenarios.

Experiments

Experimental Settings

Since there is no existing benchmark for spatial audio in 4D scene exploration, we use two complementary evaluations.

SELD Evaluation. Inspired by (Shimada et al. 2024), We adopt a pretrained Sound Event Localization and Detection(SELD) model (Diaz-Guerra et al. 2024), which extracts a Multi-ACCDOA representation from 5s of video+stereo audio (Shimada et al. 2022). Following (Diaz-Guerra et al. 2024), we evaluate on a curated subset of STARSS23 (Shimada et al. 2023), selecting perspective-view clips of musical instruments to avoid mosaic artifacts that may interfere with generative models.

We compare our method with the state-of-the-art ViSAGe model (Kim, Yun, and Kim 2025). As ViSAGe requires additional source location inputs (elevation θ and azimuth ϕ), we use the global mean elevation from its training dataset (YT-Ambigen), and for azimuth, we utilize the ground-truth annotations provided in STARSS23. Following the approach in (Shimada et al. 2024), we convert the ViSAGe-generated FOA audio into stereo format by first rotating the audio according to the azimuth, and then applying a simple transformation: $left = W + Y$ and $right = W - Y$, where W is the omnidirectional component and Y is the first-order horizontal (left-right) component of the FOA audio.

To quantify spatial audio performance, we employ the joint localization and detection metrics from the DCASE SELD challenge (Diaz-Guerra et al. 2024; Shimada et al. 2024). Specifically, we report:

- **Location-aware F-score at 20° ($\mathbf{F}_{20^\circ}$)**, which counts a detection as correct only if the class matches, the DoA angular error is within 20°, and the relative distance error is less than 1.0.
- **Directions of Arrival Angular Error (Δ_ϕ)**, the mean azimuth angular error. (Due to the SELD model, we only measure azimuth here.)
- **Distance Error (Δ_d)** and **Relative Distance Error (Δ_{rd})**, measuring absolute and normalized distance estimation errors.

These metrics together capture both detection and localization quality in our spatial audio evaluations.

User Study. We choose MMAudio (Cheng et al. 2024) as the baseline method, whose generated audio is monaural and contains no spatialization. To determine whether our generated spatial audio conveys a sense of space, we paired the same viewpoint-specific video with either the spatialized stereo audio or the non-spatialized mono audio for comparison. Participants were first asked, for each video pair (identical visuals but different audio), to choose which audio track felt more spatial. Next, they rated both audio tracks on two criteria from 1 to 5:

- **Spatial Localization Accuracy:** Measures how faithfully the interaural level and time-difference cues recreate the true 3D source position. Participants rate from 1 ("no sense of source direction") to 5 ("precise and stable perception of source location").
- **Audio–Visual Spatial Alignment:** Assesses how consistently the evolving audio cues (e.g. changes in loudness or binaural disparity) track the on-screen motion of the subject or camera. Participants rate from 1 ("audio movement conflicts with visuals") to 5 ("audio shifts perfectly match visual motion").

To more comprehensively demonstrate our model's capabilities, we categorized the videos based on custom camera trajectories into **static** and **dynamic** viewpoints. Static viewpoints indicate using the original camera view unchanged or only modest transformations of that original view, with the viewpoint remaining fixed throughout; Dynamic viewpoints include translational camera movements, arc-shaped camera trajectories, and push-in/pull-back motions.

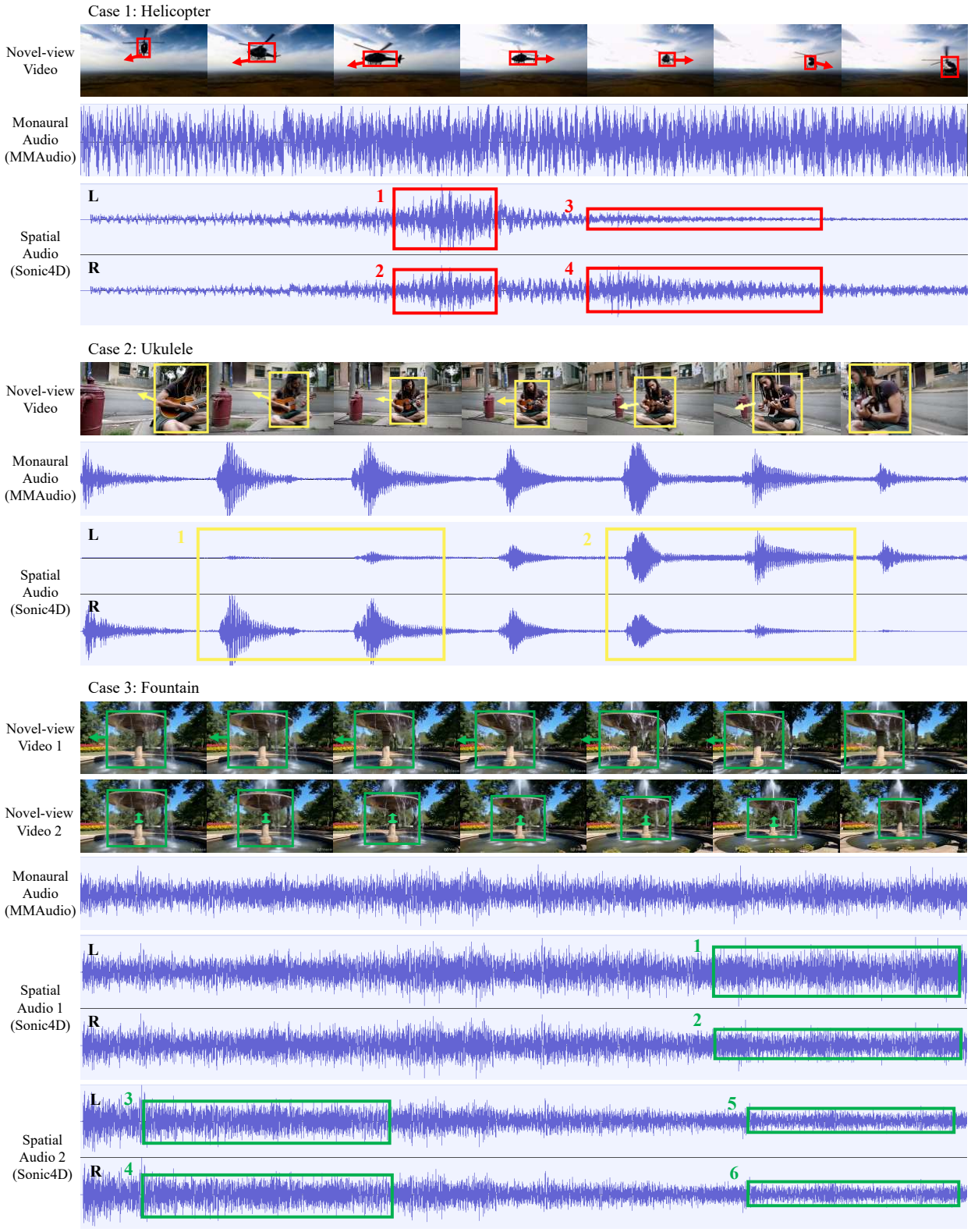

Figure 3: Qualitative results across different scenarios. We present comparisons of the spatial audio generated by Sonic4D conditioning on various camera trajectories, including static camera viewpoints, camera circling around the subject, rightward panning, and pulling out. These examples demonstrate the temporal and spatial alignment between the generated spatial audio and the motion of the visual subject.

Quantitative Comparisons

SELD Evaluation. Table 1 summarizes the SELD evaluation results on our curated STARSS23 subset. Our method attains a relatively higher F-score (**4.0%**) than ViSAGe, while reducing DOA angular error to **18.6**° and distance errors to **75.20** (absolute) and **0.44** (relative). These results demonstrate that Sonic4D produces more accurate spatial audio cues than purely learning-based methods, highlighting its ability to capture spatial information in a zero-shot setting.

Method	$\mathbf{F}_{20^\circ} \uparrow$	$\Delta_\phi(^\circ) \downarrow$	$\Delta_d \downarrow$	$\Delta_{rd} \downarrow$
Ground truth	6.5%	7.2	23.67	0.15
ViSAGe	0.4%	31.5	101.18	0.49
Sonic4D(*ours*)	**4.0%**	**18.6**	**75.20**	**0.44**

Table 1: SELD evaluation results on the curated STARSS23 subset for Sonic4D vs. ViSAGe (Kim, Yun, and Kim 2025).

User Study. We present the results of user study evaluation comparing the spatial audio rendered by our method (Sonic4D) against the non-spatialized mono audio from MMAudio (Cheng et al. 2024). As shown in Table 2, Sonic4D achieved a strong overall preference rate of **89.03%**, indicating that the spatial audio generated by Sonic4D exhibits a significant perceptual difference in spatiality compared to the original mono audio, thus demonstrating the effectiveness of our model. In terms of Mean Opinion Scores (MOS), our method outperforms the baseline in both evaluation dimensions. For Spatial Localization Accuracy, Sonic4D achieved a mean score of **4.013**, compared to MMAudio (Cheng et al. 2024)'s **2.322**. This suggests that participants were clearly able to perceive accurate and stable sound source positions when listening to our spatialized audio. A similar advantage is observed for Audio–Visual Spatial Alignment, where Sonic4D scored **3.977** versus **2.418** for MMAudio (Cheng et al. 2024), indicating a much stronger alignment between auditory motion cues and on-screen visual dynamics.

Metric	**All**	**Static**	**Dynamic**
Preference (%)			
MMAudio	10.97%	12.41%	9.90%
Sonic4D (*ours*)	**89.03%**	**87.59%**	**90.10%**
MOS-SLA			
MMAudio	2.322	2.357	2.296
Sonic4D (*ours*)	**4.013**	**3.986**	**4.033**
MOS-AVSA			
MMAudio	2.418	2.493	2.363
Sonic4D (*ours*)	**3.977**	**3.980**	**3.975**

Table 2: Subjective evaluation results for Sonic4D vs. MMAudio (Cheng et al. 2024).

Qualitative Comparisons

As shown in Fig. 3, we present additional qualitative comparisons with the baseline method. We predefined three types of camera trajectories: static viewpoint (case 1), camera circling around the subject (case 2), rightward panning and pulling out (case 3). For each case, we show the rendered video from the novel viewpoint, annotated with the motion direction of the visual subject within the frame. We then present the original waveform along with the stereo spatial audio waveform generated by Sonic4D. We highly recommend using headphones to listen to the specific examples provided in our supplementary material.

In Case 1, as the helicopter moves left then right, the spatial audio waveform reflects this motion: early segments (**Regions 1** and **2**) show the left channel dominant, whereas later segments (**Regions 3** and **4**) show the right channel surpassing the left corresponding to the helicopter's trajectory, aligning well with the expected spatial perception of the moving object. In case 2, the camera starts on the right side of a ukulele player and orbits to the left. The left-right disparity in **Regions 1** and **2** (Yellow), as well as the temporal progression within those regions, closely mirrors the trajectory, indicating strong spatial tracking in the generated audio. Case 3 demonstrates that the same source video can yield notably different spatial audio depending on the camera motion, validating the view-dependence of our model. In video 1, as the camera pans rightward, the fountain becomes closer to the left ear, leading to increased amplitude in the left channel (**Regions 1, 2**, Green). In video 2, the camera gradually pulls out, producing a fading splash sound that matches the increasing distance (**Regions 3, 4, 5, 6**, Green).

These qualitative results demonstrate that the spatial audio generated by Sonic4D aligns closely with the visual cues, reflecting both spatial position and dynamic changes in the scene.

Conclusions

In this paper, we present Sonic4D, a novel paradigm that enables spatial audio generation for immersive 4D scene exploration. Specifically, our method consists of three stages: **1) Dynamic Scene and Monaural Audio Generation:** we first employ pre-trained expert models to generate the 4D scene and its corresponding monaural audio, aiming to provide sufficient spatial and acoustic priors for subsequent spatial audio simulation. **2) 3D Sound-Source Localization and Tracking:** we further propose a pixel-level visual grounding strategy to estimate the sound source's trajectory in 3D space, facilitating spatial alignment between audio and visual content in dynamic scenes. **3) Physics-Driven Spatial Audio Synthesis:** Based on the estimated sound source's trajectory, we synthesize plausible, viewpoint-adaptive spatial audio using physics-based room impulse response (RIR) simulation. Extensive experiments demonstrate that our proposed method generates realistic spatial audio that is consistent with the synthesized 4D scene in a training-free manner, significantly enhancing the user's immersive experience.

Acknowledgements

This work was supported in part by NSFC under Grant 62371434, 62021001, 623B2098 and ZGCA Project-C20250302.

References

Allen, J. B.; and Berkley, D. A. 1979. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4): 943–950.

Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2021. Localizing visual sounds the hard way. In *CVPR*, 16867–16876.

Cheng, H. K.; Ishii, M.; Hayakawa, A.; Shibuya, T.; Schwing, A.; and Mitsufuji, Y. 2024. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*.

Dagli, R.; Prakash, S.; Wu, R.; and Khosravani, H. 2024. See-2-sound: Zero-shot spatial environment-to-spatial sound. *arXiv preprint arXiv:2406.06612*.

Diaz-Guerra, D.; Miguel, A.; and Beltran, J. R. 2021. gpuRIR: A python library for room impulse response simulation with GPU acceleration. *Multimedia Tools and Applications*, 80(4): 5653–5671.

Diaz-Guerra, D.; Politis, A.; Sudarsanam, P.; Shimada, K.; Krause, D. A.; Uchida, K.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Shibuya, T.; Mitsufuji, Y.; and Virtanen, T. 2024. Baseline Models and Evaluation of Sound Event Localization and Detection with Distance Estimation in DCASE2024 Challenge. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2024 Workshop (DCASE2024)*, 41–45. Tokyo, Japan.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Fritsch, D.; and Klein, M. 2017. 3D and 4D modeling for AR and VR app developments. In *2017 23rd International Conference on Virtual System & Multimedia (VSMM)*, 1–8. IEEE.

Gao, R.; and Grauman, K. 2019. 2.5 d visual sound. In *CVPR*, 324–333.

Hann, S. Y.; Cui, H.; Nowicki, M.; and Zhang, L. G. 2020. 4D printing soft robotics for biomedical applications. *Additive Manufacturing*, 36: 101567.

Heydari, M.; Souden, M.; Conejo, B.; and Atkins, J. 2025. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, W.; Gao, X.; Li, X.; Zhao, S.; Cun, X.; Zhang, Y.; Quan, L.; and Shan, Y. 2025. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2005–2015.

Huang, H.; Liu, Y.; Zheng, G.; Wang, J.; Dou, Z.; and Yang, S. 2025. MVTokenFlow: High-quality 4D Content Generation using Multiview Token Flow. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jiang, Y.; Zhang, L.; Gao, J.; Hu, W.; and Yao, Y. 2024. Consistent4D: Consistent 360° Dynamic Object Generation from Monocular Video. In *The Twelfth International Conference on Learning Representations*.

Khalid, M. Y.; Arif, Z. U.; Ahmed, W.; Umer, R.; Zolfagharian, A.; and Bodaghi, M. 2022. 4D printing: Technological developments in robotics applications. *Sensors and Actuators A: Physical*, 343: 113670.

Kim, D.; Um, S. J.; Lee, S.; and Kim, J. U. 2024. Learning to visually localize sound sources from mixtures without prior source knowledge. In *CVPR*, 26467–26476.

Kim, J.; Yun, H.; and Kim, G. 2025. ViSAGe: Video-to-Spatial Audio Generation. In *The Thirteenth International Conference on Learning Representations*.

Kushwaha, S. S.; Ma, J.; Thomas, M. R.; Tian, Y.; and Bruni, A. 2025. Diff-SAGe: End-to-End Spatial Audio Generation Using Diffusion Models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Leng, Y.; Chen, Z.; Guo, J.; Liu, H.; Chen, J.; Tan, X.; Mandic, D.; He, L.; Li, X.; Qin, T.; et al. 2022. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *NeurIPS*, 35: 23689–23700.

Li, R.; Pan, P.; Yang, B.; Xu, D.; Zhou, S.; Zhang, X.; Li, Z.; Kadambi, A.; Wang, Z.; Tu, Z.; et al. 2024a. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*.

Li, Z.; Xu, Q.; Zhang, D.; Song, H.; Cai, Y.; Qi, Q.; Zhou, R.; Pan, J.; Li, Z.; Vu, V. T.; et al. 2024b. Groundinggpt: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*.

Liang, H.; Yin, Y.; Xu, D.; Liang, H.; Wang, Z.; Plataniotis, K. N.; Zhao, Y.; and Wei, Y. 2024. Diffusion4D: Fast Spatial-temporal Consistent 4D Generation via Video Diffusion Models. *arXiv preprint arXiv:2405.16645*.

Liu, H.; Luo, T.; Jiang, Q.; Luo, K.; Sun, P.; Wan, J.; Huang, R.; Chen, Q.; Wang, W.; Li, X.; et al. 2025a. OmniAudio: Generating Spatial Audio from 360-Degree Video. *arXiv preprint arXiv:2504.14906*.

Liu, T.; Huang, Z.; Chen, Z.; Wang, G.; Hu, S.; Shen, l.; Sun, H.; Cao, Z.; Li, W.; and Liu, Z. 2025b. Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency. *arXiv preprint arXiv:2503.20785*.

Miao, Q.; Quan, J.; Li, K.; and Luo, Y. 2024. Pla4d: Pixel-level alignments for text-to-4d gaussian splatting. *arXiv preprint arXiv:2405.19957*.

Min, C.; Zhao, D.; Xiao, L.; Zhao, J.; Xu, X.; Zhu, Z.; Jin, L.; Li, J.; Guo, Y.; Xing, J.; et al. 2024. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. In *CVPR*, 15522–15533.

Mo, S.; and Tian, Y. 2023. Audio-visual grouping network for sound localization from mixtures. In *CVPR*, 10565–10574.

Munasinghe, S.; Gani, H.; Zhu, W.; Cao, J.; Xing, E.; Khan, F. S.; and Khan, S. 2024. VideoGLaMM: A Large Multimodal Model for Pixel-Level Visual Grounding in Videos. *arXiv preprint arXiv:2411.04923*.

Park, S.; Senocak, A.; and Chung, J. S. 2024. Can clip help sound source localization? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5711–5720.

Poole, B.; Jain, A.; Barron, J. T.; Mildenhall, B.; Abbeel, P.; and Srinivasan, P. 2022. DreamFusion: Text-to-3D using 2D diffusion. In *NeurIPS*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 10318–10327.

Ren, J.; Pan, L.; Tang, J.; Zhang, C.; Cao, A.; Zeng, G.; and Liu, Z. 2023. DreamGaussian4D: Generative 4D Gaussian Splatting. *arXiv preprint arXiv:2312.17142*.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Senocak, A.; Oh, T.-H.; Kim, J.; Yang, M.-H.; and Kweon, I. S. 2018. Learning to localize sound source in visual scenes. In *CVPR*, 4358–4366.

Senocak, A.; Ryu, H.; Kim, J.; Oh, T.-H.; Pfister, H.; and Chung, J. S. 2023. Sound source localization is all about cross-modal alignment. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7777–7787.

Shimada, K.; Koyama, Y.; Takahashi, S.; Takahashi, N.; Tsunoo, E.; and Mitsufuji, Y. 2022. Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 316–320. IEEE.

Shimada, K.; Politis, A.; Sudarsanam, P.; Krause, D. A.; Uchida, K.; Adavanne, S.; Hakala, A.; Koyama, Y.; Takahashi, N.; Takahashi, S.; et al. 2023. STARSS23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36: 72931–72957.

Shimada, K.; Simon, C.; Shibuya, T.; Takahashi, S.; and Mitsufuji, Y. 2024. SAVGBench: Benchmarking Spatially Aligned Audio-Video Generation. *arXiv preprint arXiv:2412.13462*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Sun, P.; Cheng, S.; Li, X.; Ye, Z.; Liu, H.; Zhang, H.; Xue, W.; and Guo, Y. 2024. Both Ears Wide Open: Towards Language-Driven Spatial Audio Generation. *arXiv preprint arXiv:2410.10676*.

Um, S. J.; Kim, D.; and Kim, J. U. 2023. Audio-visual spatial integration and recursive attention for robust sound source localization. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3507–3516.

Wang, L.; Zheng, W.; Ren, Y.; Jiang, H.; Cui, Z.; Yu, H.; and Lu, J. 2024. Occsora: 4d occupancy generation models as world simulators for autonomous driving. *arXiv preprint arXiv:2405.20337*.

Xu, X.; Zhou, H.; Liu, Z.; Dai, B.; Wang, X.; and Lin, D. 2021. Visually informed binaural audio generation without binaural audios. In *CVPR*, 15485–15494.

Yan, Z.; Li, C.; and Lee, G. H. 2023. Nerf-ds: Neural radiance fields for dynamic specular objects. In *CVPR*, 8285–8295.

Yu, J.; Zhu, H.; Jiang, L.; Loy, C. C.; Cai, W.; and Wu, W. 2023. Celebv-text: A large-scale facial text-video dataset. In *CVPR*, 14805–14814.

YU, M.; Hu, W.; Xing, J.; and Shan, Y. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. *arXiv preprint arXiv:2503.05638*.

Yuan, Y.-J.; Kobbelt, L.; Liu, J.; Zhang, Y.; Wan, P.; Lai, Y.-K.; and Gao, L. 2024. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv preprint arXiv:2407.12684*.

Zeng, B.; Yang, L.; Li, S.; Liu, J.; Zhang, Z.; Tian, J.; Zhu, K.; Guo, Y.; Wang, F.-Y.; Xu, M.; Ermon, S.; and Zhang, W. 2024a. Trans4D: Realistic Geometry-Aware Transition for Compositional Text-to-4D Synthesis. *arXiv preprint arXiv:2410.07155*.

Zeng, Y.; Jiang, Y.; Zhu, S.; Lu, Y.; Lin, Y.; Zhu, H.; Hu, W.; Cao, X.; and Yao, Y. 2024b. Stag4d: Spatial-temporal anchored generative 4d gaussians. In *European Conference on Computer Vision*, 163–179. Springer.

Zhang, H.; Chen, X.; Wang, Y.; Liu, X.; Wang, Y.; and Qiao, Y. 2024. 4Diffusion: Multi-view Video Diffusion Model for 4D Generation. *arXiv preprint arXiv:2405.20674*.

Zheng, Y.; Li, X.; Nagano, K.; Liu, S.; Hilliges, O.; and Mello, S. D. 2024. A Unified Approach for Text- and Image-guided 4D Scene Generation. In *CVPR*.

Zhu, H.; He, T.; Yu, X.; Guo, J.; Chen, Z.; and Bian, J. 2025. AR4D: Autoregressive 4D Generation from Monocular Videos. *arXiv preprint arXiv:2501.01722*.