PhysWave: Physics-Guided Latent Diffusion Models for Controllable Spatial Audio Generation

Anonymous ACL submission

Abstract

Text-to-spatial audio generation, such as text-to-First-Order Ambisonics (FOA), provides a convenient way to create spatial audio for billion-dollar gaming and film industries. However, existing text-to-FOA methods are largely data-driven and may produce audio that violates acoustic relations between source direction and distance. They also separate descriptive and parametric control, forcing users to trade usability for precision. In this paper, we present *PhysWave*, a physics-guided latent diffusion model for controllable text-to-FOA generation. PhysWave unifies natural-language and trajectory control through a shared waypoint-caption representation, and augments diffusion training with two differentiable acoustic priors: spherical-harmonic direction consistency and inverse-square distance consistency. To support dynamic spatial generation, we further construct a 300K-clip FOA dataset with diverse sound categories and source trajectories. Extensive experiment results show that the proposed priors help PhysWave generate spatially consistent FOA audio while maintaining competitive audio quality. Further analyses show that these physics priors improve spatial consistency during training and can also be used as inference-time guidance for training-free spatial refinement. Demos are available at https://physwave.github.io/physwave-demo/.

1 Introduction

Spatial audio generation is an essential part of the billion-dollar gaming and film industries, especially for immersive virtual reality (VR)/augmented reality (AR) gaming and 3D/4D movies, which require audio to convey a three-dimensional sound field. First-Order Ambisonics (FOA) (Zotter and Frank, 2019; Gerzon, 1985; Malham and Myatt, 1995) provides a practical format for this purpose. Unlike channel-based formats tied to fixed playback layout, FOA represents the sound field with four spherical-harmonic channels and can be decoded to headphones, loudspeaker arrays, and other playback systems. This device-agnostic property makes FOA a natural format for spatial audio generation. However, creating FOA audio remains challenging, which typically requires specialized tools, manual scene design, and spatial audio expertise. Those requirements limit FOA's use by non-expert users. That motivates text-conditioned FOA generation, where users can create spatial audio from natural-language descriptions.

Text-to-audio generation has made strong progress in producing audio that faithfully matches text prompts. For example, monaural text-to-audio methods (Liu et al., 2023, 2024; Huang et al., 2023b; Majumder et al., 2024) mainly focused on audio quality and semantic alignment, but they did not model spatial structure. Stereo audio generation (Sun et al., 2024; Feng et al., 2025; Zhao et al., 2026) introduced spatial cues such as interaural time difference (ITD) and interaural level difference (ILD), providing a sense of left-right localization. However, two-channel audio is limited in representing a full three-dimensional sound field. Recent work has therefore moved toward FOA generation. ImmerseDiffusion (Heydari et al., 2025) studied text-conditioned FOA synthesis for static sound scenes, and SonicMotion (Templin et al., 2025) extended FOA generation to moving sound sources. Despite this progress, current text-to-FOA methods still have two major limitations.

Limitation 1: Physical consistency cannot be guaranteed. FOA audio follows known acoustic relations. For a source arriving from azimuth θ and elevation ϕ, the FOA directional channels should match the first-order spherical-harmonic encoding. For a source at distance r, the received energy should vary with distance, which is commonly modeled by an inverse-square relation in a free field. Existing FOA generation methods

mainly learn these relations implicitly from data. As a result, a generated clip may sound plausible but still encode an incorrect source direction in its FOA channels, or produce an energy envelope that does not match the source-listener distance. Since these relations are known and differentiable, they can be leveraged as explicit physical priors during training and sampling.

Limitation 2: User interfaces are separated. Current FOA generation methods often provide two separate control interfaces (Heydari et al., 2025; Templin et al., 2025). A *parametric* mode accepts numerical spatial parameters such as azimuth, elevation, distance, and motion, enabling precise control but requiring users to specify low-level spatial variables. A *descriptive* mode accepts natural-language prompts, which are easy to use but usually provide coarse control over direction, distance, and time-varying motion. This separation forces users to choose between ease of use and fine-grained spatial control. A unified interface is therefore needed to support both interaction modes within the same generation model.

To address these limitations, we propose *PhysWave*, a physics-guided latent diffusion model for text-to-FOA generation. PhysWave augments the standard diffusion denoising objective with two differentiable acoustic losses: (i) a *spherical-harmonic direction consistency* loss that aligns FOA cross-channel relations with the target source direction, and (ii) an *inverse-square distance consistency* loss that regularizes the generated energy envelope according to the source-listener distance. These losses provide explicit physical priors and guide the model to generate FOA audio that matches the requested spatial trajectory. PhysWave also introduces a unified *waypoint-caption* representation for spatial control. A user can describe a sound scene in natural language, which is parsed into a non-spatial acoustic caption and an editable waypoint trajectory, or directly provide waypoints for fine-grained control. Both input forms are processed by the same diffusion transformer (DiT)-based model. To support training and evaluation, we construct a 300K-clip dynamic FOA dataset with diverse sound categories and source trajectories. Experiments show that PhysWave improves source direction and distance consistency while maintaining competitive audio quality. Further analyses demonstrate that the proposed physics priors can also be applied as inference-time guidance for training-free spatial refinement. Our salient contributions are summarized as follows.

- We propose **PhysWave**, a novel physics-guided latent diffusion framework for controllable text-to-FOA generation.
- We introduce **differentiable acoustic priors** for spherical-harmonic direction consistency and inverse-square distance consistency, which improve spatial consistency during training and enable inference-time spatial refinement without retraining.
- We design a **waypoint-caption representation** that unifies natural-language descriptions and editable spatial trajectories within the same generation model.
- We construct a **300K-clip dynamic FOA dataset** with diverse sound categories and source trajectories, supporting both large-scale training and fine-grained evaluations.

2 Related Work

2.1 Text-to-Audio Generation

Existing audio generation methods can be grouped by output format: monaural, stereo/binaural, and FOA spatial audio. Monaural text-to-audio methods, such as AudioLDM series (Liu et al., 2023, 2024), Make-An-Audio series (Huang et al., 2023b,a), TANGO series (Ghosal et al., 2023a; Majumder et al., 2024), and Stable Audio (Evans et al., 2024), have improved audio fidelity and semantic alignment. However, their outputs do not encode source direction, distance, or motion. Stereo audio generation methods introduce spatial cues through two-channel audio. AudioSpa (Feng et al., 2025) and DualSpec (Zhao et al., 2026) study text-guided binaural generation, while SpatialSonic (Sun et al., 2024) supports controllable stereo audio generation from language and other modalities. These methods improve spatial perception, but binaural audio remains a listener-centric format and does not fully represent the three-dimensional sound field.

FOA has recently been used as a device-agnostic format for spatial audio generation. DiffSAGe (Kushwaha et al., 2025) studies FOA generation from sound category and source location. ImmerseDiffusion (Heydari et al., 2025) introduces text-conditioned FOA generation for static sound scenes, and SonicMotion (Templin et al., 2025) extends FOA generation to moving sound sources.

These works show the potential of FOA for immersive audio generation. However, existing FOA generators remain mainly data-driven. They learn spatial structure from paired training data without explicitly using acoustic relations in the training objective. They also often separate descriptive and parametric control, which limits either ease of use or precise spatial control. In contrast, PhysWave introduces differentiable acoustic priors for FOA generation and uses a unified waypoint-caption representation for both natural-language and trajectory-based control. A detailed comparison is provided in Appendix A.

2.2 Physics-Guided Generative Models

Physics-guided learning incorporates known physical relations into neural network training. Physics-Informed Neural Networks (PINNs) (Raissi et al., 2019) penalize violations of governing equations, and related ideas have been extended to generative models. For example, physics-informed diffusion models (Bastek et al., 2025) use residual losses or physical constraints to improve generation in scientific domains, including flow fields (Shu et al., 2023), temperature downscaling (Rosu et al., 2025), and infrared imagery (Mao et al., 2026). In acoustics, physics-informed learning has been studied for sound field estimation and reconstruction, often using priors based on the wave equation, Helmholtz equation, or physics-constrained kernels (Olivieri et al., 2024; Ribeiro et al., 2024; Koyama et al., 2025). These works focus on inverse reconstruction from measurement rather than text-conditioned audio generation. PhysWave differs by applying simple differentiable acoustic priors to controllable FOA generation, where spherical-harmonic direction consistency and inverse-square distance consistency directly match the spatial cues specified by source trajectories.

3 Spatial Audio Dataset and Representation

Controllable FOA generation requires paired data that specify both the acoustic event and its listener-relative motion. However, real FOA recordings with accurate source trajectories are difficult to collect at scale. We therefore construct a large synthetic dataset by spatializing captioned monaural audio along sampled source trajectories. As shown in Figure 1, our pipeline consists of three stages: monaural audio preparation, spatial trajectory sampling, and physics-based FOA simulation.

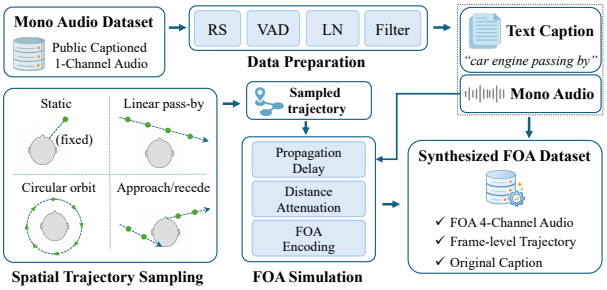

Figure 1: Dataset construction pipeline. Captioned monaural audio is paired with sampled trajectories and rendered into FOA audio through simulation.

Data preparation. We collect monaural clips from AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), and Clotho (Drossos et al., 2020). Each clip is processed in four steps. First, we apply resampling (RS) to convert all audio to 16 kHz. Second, we use voice activity detection (VAD) to select 10-second segments with sufficient active content. Third, we perform loudness normalization (LN) to reduce scale variation across sources. Finally, we filter weakly matched caption-audio pairs using CLAP similarity. More details are provided in Appendix B.

FOA representation. FOA represents a three-dimensional sound field using four spherical-harmonic channels. We denote an FOA waveform as

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

where W is the omnidirectional component, and (X, Y, Z) encode directional projections along the front–back, left–right, and up–down axes. For a point source with pressure signal $s(t)$ arriving from azimuth θ and elevation ϕ, the FOA encoding is

$$W(t) = \frac{1}{\sqrt{2}} s(t), \qquad X(t) = s(t)\cos\phi\cos\theta,$$
$$Y(t) = s(t)\cos\phi\sin\theta, \quad Z(t) = s(t)\sin\phi. \qquad (2)$$

The four channels are therefore not independent audio streams. Their cross-channel relations encode the spatial direction, which is essential for trajectory-controllable FOA generation.

Trajectory representation. We describe source motion by a listener-relative trajectory $(\theta(t), \phi(t), r(t))$, where $\theta(t)$ is the azimuth, $\phi(t)$ is the elevation, and $r(t)$ is the source-listener distance. To make trajectories compact and

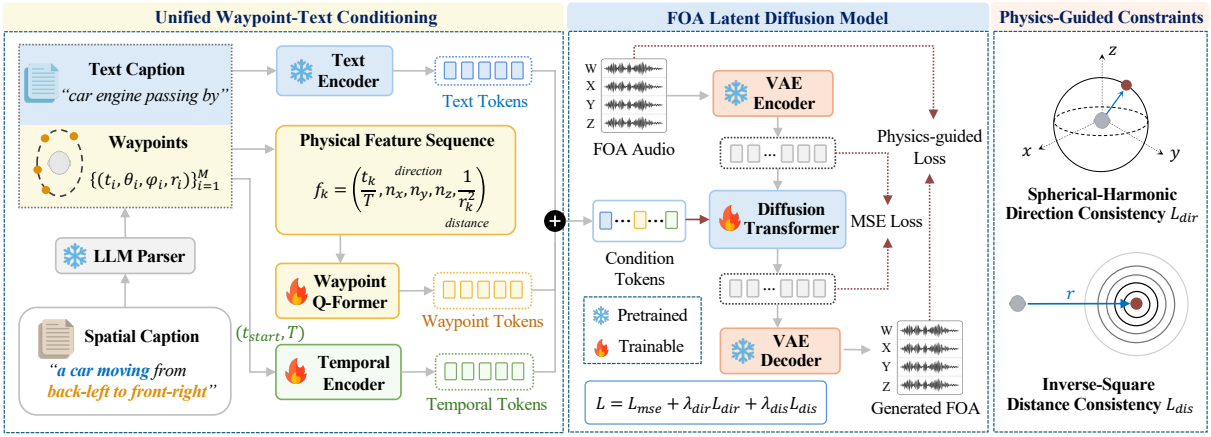

Figure 2: Overview of PhysWave. A spatial caption is parsed into an acoustic caption and a waypoint trajectory. Text, waypoint, and temporal tokens jointly condition an FOA latent diffusion model, which is trained with latent denoising loss and physics-consistency losses on the decoded FOA waveform.

editable, each trajectory is specified by a waypoint sequence

$$\mathbf{W} = \{(t_i, \theta_i, \phi_i, r_i)\}_{i=1}^{M}, \tag{3}$$

where M controls the temporal granularity. During simulation, we linearly interpolate the waypoint sequence to obtain sample-level directions and distances. As illustrated in Figure 1, we sample trajectories from four motion families: *static* sources with fixed direction and distance, *circular orbits* with changing direction and nearly constant distance, *approach/recede* trajectories with strong radial distance changes, and *linear pass-bys* with coupled direction and distance changes.

FOA simulation. Given a monaural clip $s(t)$ and a sampled trajectory $(\theta(t), \phi(t), r(t))$, we render FOA audio with an acoustic simulation pipeline. *(i) Propagation delay* resamples the source signal on the retarded-time grid $t_{\text{emit}} = t - r(t)/c$, where c is the speed of sound. This makes each output sample depend on the pressure emitted at the corresponding source distance and induces Doppler-like frequency changes under radial motion. *(ii) Distance attenuation* scales the delayed signal by $1/r(t)$ to model free-field pressure decay. *(iii) FOA encoding* projects the resulting pressure signal onto the four spherical-harmonic channels using Eq. (2). The final synthesized dataset contains triples of {FOA audio, frame-level trajectory, original caption}.

4 Physics-Guided Latent Diffusion Model

4.1 Overview

Our goal is to generate FOA audio that is both semantically aligned with a text prompt and spatially consistent with a user-specified source trajectory. As shown in Figure 2, PhysWave consists of three main components: unified waypoint-caption conditioning, an FOA latent diffusion backbone, and physics-guided training objectives.

Given a spatial caption, a large language model (LLM) parser decomposes it into two structured conditions: a text caption that describes the sound event and a waypoint sequence specifying the source motion. The text caption is encoded into text tokens. The waypoint sequence is converted into physical trajectory features and encoded into waypoint tokens. A temporal encoder further maps temporal information into temporal tokens. These token sets are concatenated as the joint condition for the diffusion model.

PhysWave follows a latent diffusion design. A four-channel FOA waveform $\mathbf{a}$ is compressed into a continuous latent $\mathbf{x}$ by an FOA variational autoencoder (VAE) encoder. A DiT then denoises the latent under the joint condition. Unlike standard latent diffusion training, which applies supervision only in latent space, we decode the predicted clean latent back to an FOA waveform during training and apply differentiable physical constraints to the decoded signal. This encourages the model to match both the acoustic caption and the target source trajectory.

4.2 Spatial Physical Priors

A source trajectory provides two physical cues that a faithful FOA signal should follow: the instantaneous direction of arrival and the distance-induced energy envelope. We formalize them as two physical priors, which are used for both waypoint condi-

tioning in Section 4.3 and physics-guided training in Section 4.5.

Spherical-harmonic direction prior. For a source arriving from azimuth $\theta(t)$ and elevation $\phi(t)$, the listener-relative unit direction is

$$\mathbf{n}(t) = \begin{bmatrix} \cos\phi(t)\cos\theta(t) \\ \cos\phi(t)\sin\theta(t) \\ \sin\phi(t) \end{bmatrix}. \tag{4}$$

We use $\mathbf{n}(t)$ as the per-frame direction target. As defined in Eq. (2), FOA represents direction through cross-channel relations between the omnidirectional channel W and the directional channels (X, Y, Z). A spatially consistent FOA signal should therefore recover the target direction from this cross-channel structure.

Inverse-square distance prior. Free-field propagation attenuates pressure approximately by $1/r(t)$, where $r(t)$ is the source-listener distance. The corresponding distance-induced energy component follows an inverse-square profile $1/r(t)^2$. Since source pressure $s(t)$ scales the FOA channels in Eq. (2), this profile provides a trajectory-conditioned attenuation prior for generated FOA audio. We use it to constrain the relative temporal energy trend of the generated signal while leaving the absolute source loudness unconstrained.

4.3 Unified Waypoint-Caption Conditioning

A spatial caption is first parsed by the frozen LLM parser into a text caption and a waypoint sequence, as illustrated in Figure 2. The text caption specifies the sound event and is encoded by a pretrained T5 encoder (Raffel et al., 2020) into *text tokens*. The waypoint sequence specifies the source motion:

$$\mathbf{W} = \{(t_i, \theta_i, \phi_i, r_i)\}_{i=1}^{M}. \tag{5}$$

To expose the physical structure of this trajectory to the model, we interpolate the waypoints into a dense per-frame feature sequence:

$$\mathbf{f}_k = \big[\underbrace{\tfrac{t_k}{T}}_{\text{time}}, \underbrace{n_x(t_k),\, n_y(t_k),\, n_z(t_k)}_{\text{direction}}, \underbrace{\tfrac{1}{r(t_k)^2}}_{\text{distance}}\big], \tag{6}$$

where T is the clip duration, (n_x, n_y, n_z) is the unit direction vector from Eq. (4), and $1/r(t_k)^2$ is the inverse-square distance profile.

A waypoint Q-Former attends a small set of learnable queries to $\{\mathbf{f}_k\}$ and produces *waypoint tokens*. In parallel, a temporal encoder embeds the clip start time and total duration (t_{start}, T) into *temporal tokens*. We concatenate the text tokens, waypoint tokens, and temporal tokens into the joint condition $\mathbf{c}$, which drives the DiT through cross-attention.

4.4 FOA Latent Diffusion Backbone

We use a four-channel FOA VAE to compress FOA waveforms into continuous latents. Because FOA channels jointly encode spatial structure, we pretrain the VAE with standard waveform reconstruction losses and the spherical-harmonic direction consistency loss $\mathcal{L}_{\text{dir}}$ in Eq. (11), applied to the VAE output. VAE training details are provided in Appendix E.1. After pretraining, VAE is frozen during diffusion training.

Given a clean latent $\mathbf{x}$, we sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and construct the noisy latent

$$\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \tag{7}$$

where α_t and σ_t follow a cosine noise schedule. The DiT predicts the velocity target

$$\mathbf{v}_t = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}, \tag{8}$$

conditioned on $\mathbf{c}$. The standard latent diffusion objective is

$$\mathcal{L}_{\text{mse}} = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t}\Big[\big\|\hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_t\big\|_2^2\Big]. \tag{9}$$

4.5 Physics-Guided Training Objective

To improve the physical consistency of generated FOA audio, we compute auxiliary losses on the decoded waveform rather than on the latent. These losses encourage the generated waveform to follow the direction and distance relations specified by the input trajectory. Given the velocity prediction $\hat{\mathbf{v}}_\theta$, we recover the predicted clean latent $\hat{\mathbf{x}}_0 = \alpha_t \mathbf{x}_t - \sigma_t \hat{\mathbf{v}}_\theta(\mathbf{x}_t, t, \mathbf{c})$ and decode it with the frozen VAE decoder $\mathcal{D}$ to obtain the predicted FOA waveform $\hat{\mathbf{a}} = \mathcal{D}(\hat{\mathbf{x}}_0)$. We then apply one consistency loss for each physical prior.

Spherical-harmonic direction consistency. For each frame k, we estimate a short-time FOA intensity vector (Adavanne et al., 2018) from the generated waveform:

$$\hat{\mathbf{I}}_k = \big[\langle \hat{W}_k \hat{X}_k \rangle,\ \langle \hat{W}_k \hat{Y}_k \rangle,\ \langle \hat{W}_k \hat{Z}_k \rangle\big]^\top, \tag{10}$$

where $\langle \cdot \rangle$ denotes averaging within frame k. The direction loss aligns this vector with the target direction $\mathbf{n}_k$:

$$\mathcal{L}_{\text{dir}} = \frac{1}{K}\sum_{k=1}^{K}\left(1 - \frac{\hat{\mathbf{I}}_k^\top \mathbf{n}_k}{\|\hat{\mathbf{I}}_k\|\,\|\mathbf{n}_k\| + \varepsilon}\right). \tag{11}$$

where K is the number of frames and ε is a small constant for numerical stability.

Inverse-square distance consistency. We measure the distance-induced energy envelope using the omnidirectional FOA channel. For each frame k, we compute the generated W-channel energy:

$$\hat{E}_k = \langle \hat{W}_k^2 \rangle. \tag{12}$$

The target energy profile is $E_k^* = 1/r_k^2$. Since absolute loudness depends on the source content, we compare only the normalized log-energy shape. Let $\widetilde{u}_k = \log(u_k + \varepsilon) - \frac{1}{K}\sum_j \log(u_j + \varepsilon)$. The distance consistency loss is

$$\mathcal{L}_{\text{dist}} = \frac{1}{K}\sum_{k=1}^{K} \left(\widetilde{\hat{E}}_k - \widetilde{E}_k^*\right)^2. \tag{13}$$

Final objective. The final objective combines latent denoising with the two physics-guided constraints:

$$\mathcal{L} = \mathcal{L}_{\text{mse}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}, \tag{14}$$

where λ_{dir} and λ_{dist} control the strength of the direction and distance constraints.

5 Experiments

5.1 Implementation Details

Dataset. We use the synthetic FOA dataset described in Section 3. Each monaural clip is paired with a sampled listener-relative source trajectory and rendered into four-channel FOA audio. Appendix B provides further details.

Model. We use Qwen3.5-4B (Qwen Team, 2026) as the LLM parser to convert spatial captions into structured waypoint-caption conditions. Our generative backbone follows Stable Audio Open (Evans et al., 2025): we adapt its continuous DAC-based VAE from stereo to four-channel FOA WXYZ input/output, and use its 24-layer DiT denoiser with hidden size 768 and 12 attention heads. Text captions are encoded by a pretrained T5-base encoder (Raffel et al., 2020) with a maximum length of 128 tokens. Waypoints are encoded by a lightweight Q-Former-style encoder into 16 learnable query tokens, and temporal scalars are encoded with scalar embedding layers. Further details are provided in Appendix E.

5.2 Evaluation Metrics

We evaluate generated FOA audio from two aspects: audio quality and spatial fidelity.

Audio quality. Following standard text-to-audio evaluation protocols, we report CLAP score, Fréchet Distance (FD), CLAP-based Fréchet Audio Distance (FAD_{CLAP}), Inception Score (IS), and Kullback–Leibler divergence (KL). These metrics measure semantic alignment and perceptual quality of generated audio. Since they are mainly designed for single-channel audio, we compute them on the omnidirectional W channel of the FOA signal.

Spatial fidelity. We evaluate spatial quality in terms of direction consistency and distance consistency. For direction consistency, we estimate the direction of arrival (DoA) at each frame from the FOA intensity vector $(\langle WX \rangle, \langle WY \rangle, \langle WZ \rangle)$ and convert it to azimuth $\hat{\theta}$ and elevation $\hat{\phi}$. We then compute the spherical angular error (Heydari et al., 2025) between the estimated and ground-truth directions using the haversine formula. For distance consistency, we evaluate whether the generated signal follows the distance attenuation profile specified by the ground-truth trajectory. Specifically, we compute the per-frame W-channel energy and compare it with the target inverse-square profile $1/r_k^2$. Since absolute loudness depends on the audio content, we convert both sequences to log scale and mean-center them before comparison. To reduce the effect of silent or low-energy frames, we apply an energy gate on the generated W-channel and compute the metrics over active frames. We report RMS error (InvSqErr, in dB), which measures profile mismatch, and Pearson correlation (InvSqCorr), which measures temporal agreement. Lower InvSqErr and higher InvSqCorr indicate better distance-consistent attenuation. Full metric definitions are provided in Appendix F.

5.3 Main Results

Spatial fidelity. We compare PhysWave with two representative FOA generation methods: ImmerseDiffusion (Heydari et al., 2025), which focuses on static-source generation, and SonicMotion (Templin et al., 2025), which supports moving-source generation under circular motion. Since these methods are not publicly released, we cannot directly re-evaluate them on our dataset. We therefore report their published results as reference numbers under their original evaluation settings. As shown in Table 2, PhysWave achieves a moving-source angular error of 4.65° under waypoint-based trajectory control. In comparison, the published moving-source angular errors

Variant	Audio quality					Spatial fidelity			
	FD ↓	FAD_{CLAP} ↓	CLAP ↑	KL ↓	IS ↑	Static (°) ↓	Moving (°) ↓	InvSq Err. (dB) ↓	InvSq Corr. ↑
Trajectory text	23.13	0.20	0.32	1.74	8.05	7.33	17.62	4.59	0.62
PhysWave (w/o physics)	21.92	0.20	0.33	1.69	8.22	2.05	6.50	5.15	0.48
PhysWave (w/o $\mathcal{L}_{\text{dist}}$)	21.79	0.20	0.33	1.68	8.15	1.75	4.81	4.79	0.59
PhysWave (w/o $\mathcal{L}_{\text{dir}}$)	21.34	0.20	0.33	1.67	8.25	2.47	6.60	**3.78**	**0.74**
PhysWave	**21.22**	0.21	**0.33**	**1.66**	**8.31**	**1.73**	**4.65**	4.06	0.73

Table 1: Ablation study of PhysWave. All variants use the same generative backbone and caption condition. Angular errors are reported in degrees, and InvSq Err. is reported in dB.

Model	Control	Static (°) ↓	Moving (°) ↓
ID-D	Static text	1.35	–
ID-P	Static param.	1.12	–
SM-D	Circular text	–	29.22
SM-P	Circular param.	–	14.32
PhysWave	Text + Waypoints	1.73	4.65

Table 2: Reference comparison with published results from prior FOA generation methods. ID and SM denote ImmerseDiffusion and SonicMotion, respectively. D and P denote descriptive and parametric conditioning.

Model	FD ↓	FAD_{CLAP} ↓	CLAP ↑	KL ↓	IS ↑
AudioLDM	31.67	0.33	0.33	2.37	6.96
AudioLDM 2	24.13	0.12	0.33	2.11	8.62
Make-An-Audio	15.44	**0.11**	0.37	1.93	8.79
Make-An-Audio 2	**13.86**	0.14	**0.40**	1.73	10.85
Stable Audio Open	30.84	0.31	0.30	2.30	**11.09**
SonicMotion	24.80	0.26	0.32	1.88	8.55
PhysWave	21.22	0.21	0.33	**1.66**	8.31

Table 3: Audio quality comparison on the omnidirectional W channel.

of SonicMotion are $29.22°$ with descriptive conditioning and $14.32°$ with parametric conditioning. Although SonicMotion is evaluated on circular trajectories while PhysWave supports more general waypoint trajectories, this comparison suggests that PhysWave provides accurate localization under a broader trajectory-control setting. For static sources, PhysWave obtains $1.73°$, which is close to the published static-source results of ImmerseDiffusion. These results suggest that PhysWave improves controllable moving-source localization while maintaining strong static-source accuracy.

Audio quality. We evaluate semantic alignment and perceptual quality on the omnidirectional W channel, following the FOA evaluation protocol of Templin et al. (2025). Since standard text-to-audio metrics are mainly designed for monaural audio, the W channel provides a natural way to evaluate the non-directional acoustic content of FOA audio. We compare PhysWave with state-of-the-art monaural text-to-audio models. Since these baselines do not generate FOA audio, we spatialize each generated monaural clip into FOA using the same trajectory-conditioned rendering pipeline described in Section 3. As shown in Table 3, PhysWave achieves competitive audio quality. It obtains the best KL score (1.66), while its FD (21.22), FAD_{CLAP} (0.21), CLAP score (0.33), and IS (8.31) remain comparable to strong monaural baselines. These results show that PhysWave adds controllable FOA spatialization while preserving the audio quality of the generated content.

Ablation study. We ablate the spatial conditioning interface and the physics consistency losses in Table 1. First, we compare waypoint conditioning with a text-based trajectory condition, where each trajectory is converted into a detailed natural-language description. Replacing waypoints with trajectory text sharply reduces spatial fidelity: static angular error increases from $2.05°$ to $7.33°$, and moving angular error increases from $6.50°$ to $17.62°$. This shows that free-form text is not precise enough to specify fine-grained source motion, including direction, distance, and temporal changes. In contrast, waypoints provide explicit geometric conditions, while the LLM parser bridges user text and structured waypoint inputs. Second, the two physics losses improve different aspects of spatial fidelity. The direction loss reduces the moving angular error from $6.50°$ to $4.81°$ and the static angular error from $2.05°$ to $1.75°$, while having limited effects on the inverse-square metrics. The distance loss mainly improves distance consistency, increasing InvSq Corr. from 0.48 to 0.74 and reducing InvSq Err. from 5.15 dB to 3.78 dB. The full PhysWave model combines both losses, achieving the best static angular error ($1.73°$), moving angular error ($4.65°$) and strong distance consistency (InvSq Corr. $= 0.73$). Audio quality remains stable

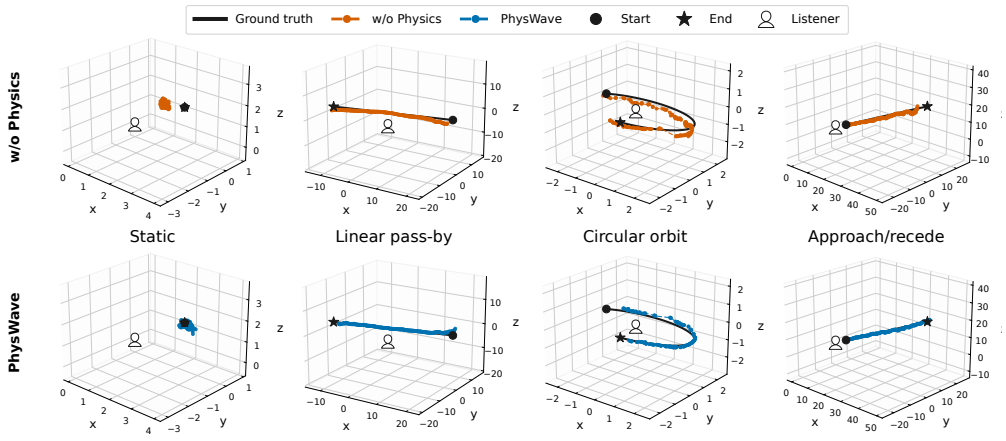

Figure 3: Qualitative trajectory comparison between PhysWave with and without physics-consistency losses.

Parser	Spatial Faith. ↑	Semantic Pres. ↑
Qwen3.5-0.8B	1.89 ± 0.20	3.40 ± 0.38
Llama-3.2-3B-Instruct	2.40 ± 0.24	4.18 ± 0.29
Qwen3.5-4B	3.88 ± 0.28	4.72 ± 0.17

Table 4: LLM parser evaluation on spatial captions. Scores are reported as mean ± standard deviation.

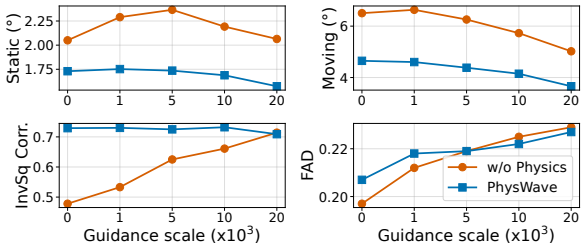

Figure 4: Effect of inference-time physics guidance.

across variants, indicating that the auxiliary physics losses improve spatial fidelity without degrading content quality.

LLM parser evaluation. We evaluate the LLM spatial parser that converts natural-language spatial descriptions into waypoint-caption conditions. Using GPT-4o to score spatial faithfulness and semantic preservation, we find that Qwen3.5-4B performs best among the evaluated open-source parsers, with 3.88 ± 0.28 and 4.72 ± 0.17, respectively. We therefore use it for descriptive control, while the same waypoint-caption representation also supports direct waypoint input. The detailed protocol is provided in Appendix D.

Inference-time physics guidance. We further examine whether the same physics-consistency losses can guide sampling without retraining. At selected denoising steps, we compute the differentiable losses ($\mathcal{L}_{\text{dist}}$ and $\mathcal{L}_{\text{dir}}$), and use their gradients to steer the sample toward spatially consistent FOA audio. As shown in Figure 4, larger guidance scales generally reduce static and moving angular errors and improve InvSq Corr., especially for "w/o physics". FAD_{CLAP} increases mildly at larger scales, indicating a trade-off between spatial correction and audio quality. These results show that the proposed physics-consistency losses can also serve as inference-time guidance for spatial refinement.

Visualization. Figure 3 visualizes source trajectories estimated from generated FOA audio, comparing PhysWave with and without physics-consistency losses during training. Without physics losses, the generated trajectories roughly follow the target motion but show noticeable deviations, especially for static localization and circular motion. In contrast, PhysWave produces trajectories that better align with the ground truth across all motion types. This qualitative trend is consistent with Table 1, where the physics losses reduce angular error and improve distance consistency. Additional trajectory visualizations are provided in Appendix G.

6 Conclusion

We presented *PhysWave*, a physics-guided latent diffusion model for controllable text-to-FOA generation. PhysWave combines waypoint-caption conditioning with differentiable direction and distance priors, improving spatial consistency while maintaining competitive audio quality. The same priors also support training-free inference-time refinement, further demonstrating the benefits of explicit acoustic priors for spatial audio generation.

7 Limitations

PhysWave currently focuses on single-source trajectory control in a free-field setting. It does not model room effects such as reverberation, occlusion, and multi-path propagation, which require additional acoustic modeling beyond the direct-path simulation used in this work. Extending physics-guided FOA generation to multi-source scenes and realistic room acoustics is a useful direction for future work.

References

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. 2018. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466. IEEE.

Jan-Hendrik Bastek, WaiChing Sun, and Dennis Kochmann. 2025. Physics-informed diffusion models. In *The Thirteenth International Conference on Learning Representations*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. 2024. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Linfeng Feng, Lei Zhao, Boyu Zhu, Xiao-Lei Zhang, and Xuelong Li. 2025. Audiospa: Spatializing sound events with text. *arXiv preprint arXiv:2502.11219*.

Michael A Gerzon. 1985. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023a. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pages 3590–3598.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023b. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*.

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

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023a. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. 2025. Visage: Video-to-spatial audio generation. In *13th International Conference on Learning Representations, ICLR 2025*, pages 14239–14259. International Conference on Learning Representations, ICLR.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Shoichi Koyama, Juliano GC Ribeiro, Tomohiko Nakamura, Natsuki Ueno, and Mirco Pezzoli. 2025. Physics-informed machine learning for sound field estimation: Fundamentals, state of the art, and challenges. *IEEE Signal Processing Magazine*, 41(6):60–71.

Saksham Singh Kushwaha, Jianbo Ma, Mark RP Thomas, Yapeng Tian, and Avery Bruni. 2025. Diffsage: End-to-end spatial audio generation using diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning*, pages 21450–21474. PMLR.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883.

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. 2025. Omniaudio: Generating spatial audio from 360-degree video. *Preprint*, arXiv:2504.14906.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *Preprint*, arXiv:2404.09956.

David G Malham and Anthony Myatt. 1995. 3-d sound spatialization using ambisonic techniques. *Computer music journal*, 19(4):58–70.

Fangyuan Mao, Jilin Mei, Shun Lu, Fuyang Liu, Liang Chen, Fangzhou Zhao, and Yu Hu. 2026. Pid: physics-informed diffusion model for infrared image generation. *Pattern Recognition*, 169:111816.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2024. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–15.

Marco Olivieri, Xenofon Karakonstantis, Mirco Pezzoli, Fabio Antonacci, Augusto Sarti, and Efren Fernandez-Grande. 2024. Physics-informed neural network for volumetric sound field reconstruction of speech signals. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):42.

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, and Gangshan Wu. 2025. In-the-wild audio spatialization with flexible text-guided localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1989–2001.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707.

Juliano GC Ribeiro, Shoichi Koyama, Ryosuke Horiuchi, and Hiroshi Saruwatari. 2024. Sound field estimation based on physics-constrained kernel interpolation adapted to environment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4369–4383.

Paul Rosu, Muchang Bahng, Erick Jiang, Rico Zhu, and Vahid Tarokh. 2025. A pde-informed latent diffusion model for 2-m temperature downscaling. *arXiv preprint arXiv:2510.23866*.

Dule Shu, Zijie Li, and Amir Barati Farimani. 2023. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972.

Parthasaarathy Sudarsanam, Sebastian Braun, and Hannes Gamper. 2025. Foa tokenizer: Low-bitrate neural codec for first order ambisonics with spatial consistency loss. *arXiv preprint arXiv:2510.22241*.

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

Christian Templin, Yanda Zhu, and Hao Wang. 2025. Generating moving 3d soundscapes with latent diffusion models. *arXiv preprint arXiv:2507.07318*.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Lei Zhao, Sizhou Chen, Linfeng Feng, Jichao Zhang, Xiao-Lei Zhang, Chi Zhang, and Xuelong Li. 2026. Dualspec: Text-to-spatial-audio generation via dual-spectrogram guided diffusion model. *IEEE Transactions on Multimedia*.

Franz Zotter and Matthias Frank. 2019. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer.

A Comparison with Prior Works

Table 5 compares PhysWave with prior audio generation methods in terms of output format, conditioning input, spatial source support, trajectory control, and physics guidance. Compared with monaural and stereo/binaural methods, PhysWave directly generates FOA spatial audio. Compared with existing FOA generators, PhysWave supports waypoint-based trajectory control for both static and moving sources, going beyond the static-source setting of ImmerseDiffusion and the circular-motion setting of SonicMotion. PhysWave also integrates explicit acoustic priors into the generation process to improve spatial consistency.

Method	Format	Conditioning	Spatial Source		Trajectory Control	Physics-Guided
			Static	Moving		
AudioLDM (Liu et al., 2023)	Mono	Text	–	–	–	✗
AudioLDM 2 (Liu et al., 2024)	Mono	Text	–	–	–	✗
Make-An-Audio (Huang et al., 2023b)	Mono	Text	–	–	–	✗
Make-An-Audio 2 (Huang et al., 2023a)	Mono	Text	–	–	–	✗
TANGO (Ghosal et al., 2023b)	Mono	Text	–	–	–	✗
TANGO 2 (Majumder et al., 2024)	Mono	Text	–	–	–	✗
Stable Audio Open (Evans et al., 2025)	Stereo	Text	–	–	–	✗
SpatialSonic (Sun et al., 2024)	Stereo	Text/Image + Azimuth	✓	✓	Azimuth trajectory	✗
AudioSpa (Feng et al., 2025)	Binaural	Text + Mono Ref.	✓	✗	–	✗
DualSpec (Zhao et al., 2026)	Binaural	Text	✓	✗	–	✗
TAS (Pan et al., 2025)	Binaural	Text + Mono Ref.	✓	✓	Flexible location	✗
ViSAGe (Kim et al., 2025)	FOA	Video	✓	✓	Video-driven	✗
OmniAudio (Liu et al., 2025)	FOA	360 Video	✓	✓	Video-driven	✗
Diff-SAGe (Kushwaha et al., 2025)	FOA	Category + Location	✓	✗	–	✗
ImmerseDiffusion (Heydari et al., 2025)	FOA	Text / Spatial Params.	✓	✗	–	✗
SonicMotion (Templin et al., 2025)	FOA	Text / Spatial Params.	✓	✓	Circular	✗
PhysWave (Ours)	**FOA**	**Text + Waypoints**	✓	✓	**General waypoints**	✓

Table 5: Functional comparison with prior audio generation methods. Mono, Stereo, Binaural, and FOA denote output formats. ✓ means supported, ✗ means not supported, and "–" means not applicable. For SpatialSonic, trajectory control refers to azimuth-level control, while PhysWave supports listener-relative waypoint trajectories with azimuth, elevation, and distance. Physics-guided means that explicit acoustic priors are used during generation.

Family	Ratio	Sampling Range	Description
Static	50.0%	$\theta \in [-180^\circ, 180^\circ]$, $\phi \in [-35^\circ, 35^\circ]$, $r \in [0.5, 5]$ m	Fixed direction and distance
Linear pass-by	$\approx 16.7\%$	$v \in [1, 25]$ m/s, $d_{\min} \in [1, 8]$ m	Coupled direction and distance changes
Circular	$\approx 16.7\%$	$r \in [0.5, 5]$ m, angular span $\geq 30^\circ$	Direction changes at near-constant distance
Approach/recede	$\approx 16.7\%$	$v \in [2, 6]$ m/s, $r \in [1, 60]$ m	Distance changes with fixed direction

Table 6: Trajectory sampling configuration. Each retained monaural clip is rendered once as a static source and once as a moving source.

B Dataset Construction Details

B.1 Source Audio Preparation

We construct the source pool from three public captioned audio datasets: AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), and Clotho (Drossos et al., 2020). Since public captioned audio can contain noisy captions, weak audio-text matches, or overlapping events, we first prioritize clips whose captions describe a dominant acoustic event, following the single-source pre-selection strategy used in prior spatial-audio dataset construction (Sun et al., 2024). All retained clips are converted to mono and resampled to 16 kHz. We then apply energy-based activity detection to extract a 10-second segment with sufficient active content. Clips with less than 1 second of active content are discarded. If a clip is shorter than 10 seconds, we pad it with zeros. If it is longer than 10 seconds, we select the 10-second window containing the largest number of active frames. We apply EBU R128 loudness normalization to a target level of -14 LUFS and clip the normalized waveform to $[-1, 1]$. We then compute audio-caption similarity using a CLAP model and discard clips with scores below 0.3. For Clotho, where each clip has up to five captions, we score all captions and keep the highest-scoring one. As shown in Table 7, the final source pool contains 157,409 training clips and 5,260 test clips. Each source clip is rendered once as a static source and once as a moving source, resulting in 314,818 training FOA examples and 10,520 test FOA examples.

B.2 Trajectory Sampling

For each retained monaural clip, we render two spatial versions: one static source and one moving source. The static version keeps a fixed azimuth, elevation, and distance throughout the clip. The moving version is sampled from three motion families:

```
{
  "Task":
  "You are a spatial audio caption parser. Convert the input scene text into strict JSON for spatial
      audio generation. Do not output waypoint arrays; a deterministic program will convert the
     parsed trajectory into 10 waypoints.",

  "Coordinate system":
  "az in [-180,180], 0=front, +90=left, -90=right, +/-180=back; el is elevation in degrees, 0=
     horizon; r is distance in meters; time is in seconds.",

  "Output schema": {
    "duration": 10.0,
    "events": [{
      "text": "<acoustic content only, no spatial words>",
      "t_start": <float>,
      "t_end": <float>,
      "trajectory": {
        "type": "static  linear  arc  approach  recede",
        "start": {"az": <float>, "el": <float>, "r": <float>},
        "end": {"az": <float>, "el": <float>, "r": <float>},
        "control_points": [
          {"time": <float>, "az": <float>, "el": <float>, "r": <float>}
        ],
        "direction": "clockwise  counterclockwise  null",
        "turns": <float>,
        "speed": "slow  medium  fast"
      },
      "inferred": [<string>]
    }]
  },

  "Rules": [
    "Output JSON only.",
    "Use static for fixed sources; linear for motion from one position to another.",
    "Use approach/recede for distance change; use arc for circular or orbiting motion.",
    "If time is unspecified, use t_start=0.0 and t_end=10.0.",
    "Defaults: az=0, el=0, normal r=8.0, close r=2.0, very close r=1.0, far r=25.0.",
    "If a value is inferred, list its field path in inferred."
  ],

  "Input": "{SCENE_TEXT}"
}
```

Figure 5: Prompt used by the LLM spatial parser. The parser outputs a compact trajectory representation, which is converted into the $M = 10$ waypoint condition used by the diffusion model.

Data Source	Train		Test	
	Clips	Duration (h)	Clips	Duration (h)
AudioCaps	46,696	129.7	4,239	11.8
WavCaps	107,850	299.6	–	–
Clotho	2,863	8.0	1,021	2.8
Total monaural	157,409	437.2	5,260	14.6
Rendered FOA	314,818	874.5	10,520	29.2

Table 7: Statistics of source data after preprocessing. Each retained monaural clip is rendered twice, once as a static source and once as a moving source.

linear pass-by, circular motion, and approach/recede motion. All interpolated trajectories are stored as frame-level trajectory samples (t, θ, ϕ, r) at 0.1 s intervals, where θ is azimuth, ϕ is elevation, and r is the source-listener distance. Table 6 summarizes the sampling configuration. Since each clip produces one static and one moving example, static sources account for 50% of the rendered FOA samples. The moving examples are split nearly uniformly across the three motion families, so each moving family contributes about 16.7% of all rendered FOA samples.

C LLM Spatial Parser

C.1 Parser Prompt

We use an LLM spatial parser to convert a natural-language spatial caption into a structured intermediate representation. The parser separates the non-spatial acoustic content from spatial trajectory attributes and outputs them in JSON format. A deterministic post-processing module then converts the parsed trajectory into the $M = 10$ waypoint condition used by the diffusion model. The full prompt is shown in Figure 5.

C.2 Example Parser Outputs

Figure 6 shows example parser outputs. For compactness, we only show the main parsed fields.

```
{
  "Input":
  "Ocean waves crashing as water trickles and splashes,
   approaching from the left, moving from farther away
   to a closer distance.",

  "Output": {
    "text": "Ocean waves crashing as water trickles and
      splashes.",
    "trajectory": {
      "type": "approach",
      "start": {"az": 90, "el": 0, "r": 25},
      "end": {"az": 90, "el": 0, "r": 2}
    }
  }
}
```

```
{
  "Input":
  "Burping and a man speaking, passing from the
   front-right to the back-left, passing at a normal
   distance.",

  "Output": {
    "text": "Burping and a man speaking.",
    "trajectory": {
      "type": "linear",
      "start": {"az": -45, "el": 0, "r": 8},
      "end": {"az": 135, "el": 0, "r": 8}
    }
  }
}
```


Figure 6: Example outputs of the LLM spatial parser. Each output keeps the non-spatial acoustic content and converts the spatial description into a parametric trajectory.

D LLM-as-a-Judge Evaluation

We use an LLM-as-a-judge protocol to evaluate the spatial parser. For each spatial caption, a candidate parser generates a structured trajectory representation, which is converted into a fixed-length waypoint condition by deterministic post-processing. The judge receives the original spatial caption, the parsed acoustic caption, and the converted waypoints. It then assigns two integer scores from 1 to 5: *spatial faithfulness*, which measures whether the waypoints match the described spatial layout and motion, and *semantic preservation*, which measures whether the parsed caption preserves the acoustic content without spatial leakage. We evaluate 200 randomly sampled spatial captions using GPT-4o with temperature set to 0. All candidate parsers are evaluated on the same caption set using the same judge prompt, and each example is scored once. The judge is instructed to output only the two integer scores and a short justification, where the justification is used only for inspection and is not included in the reported metrics. The full judge prompt is shown in Figure 7.

E Model Details

E.1 FOA VAE Training

We train a VAE-based neural audio codec to compress FOA waveforms into a continuous latent space for diffusion modeling. The autoencoder follows a DAC-style 1D convolutional architecture, with the input and output channels changed to the four FOA components (W, X, Y, Z). All audio is represented at 16 kHz. We train the VAE on 2.05-second crops, with a downsampling ratio of 1024 and a latent channel dimension of 64. After training, the VAE is frozen and used as the latent encoder and decoder for the diffusion model. Related work has also studied FOA-specific neural audio representations for spatially consistent tokenization (Sudarsanam et al., 2025).

Training objective. The VAE objective combines spectral reconstruction, adversarial training, latent regularization, and FOA spatial consistency:

$$\mathcal{L}_{\text{VAE}} = \frac{\lambda_{\text{mrstft}}}{4} \sum_{c\in\{W,X,Y,Z\}} \mathcal{L}^{c}_{\text{mrstft}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}. \tag{15}$$

Here, $\mathcal{L}^{c}_{\text{mrstft}}$ denotes the multi-resolution STFT loss computed on FOA channel $c \in \{W, X, Y, Z\}$. The terms $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{fm}}$ are the adversarial and feature-matching losses from the discriminator, and $\mathcal{L}_{\text{kl}}$ regularizes the VAE bottleneck. To better preserve FOA spatial structure, we add $\mathcal{L}_{\text{dir}}$ to improve spatial reconstruction consistency. We set $\lambda_{\text{mrstft}} = 1.0$, $\lambda_{\text{adv}} = 0.1$, $\lambda_{\text{fm}} = 5.0$, $\lambda_{\text{kl}} = 10^{-6}$, and $\lambda_{\text{dir}} = 0.1$.

Optimization. We train the autoencoder on 2.05-second crops from the training split. The autoencoder and discriminator are optimized with AdamW using betas $(0.8, 0.99)$ and weight decay 10^{-3}. The autoencoder uses learning rate 1×10^{-4}, while the discriminator uses learning rate 2×10^{-4}. Both learning rates are decayed with an exponential scheduler with decay factor 0.999996. We train on two NVIDIA RTX 3090 GPUs, with a batch size of 12 and train for 120K steps.

Reconstruction quality. Table 8 reports FOA VAE reconstruction quality on the test split. The

```
{
  "Task":
  "You are evaluating an LLM parser for a spatial audio system.",

  "Original spatial caption":
  "{SPATIAL_CAPTION}",

  "Parser output":
  "- Text caption: {PARSED_CAPTION}
   - Waypoints (t in seconds, azimuth/elevation in degrees,
     distance in meters): {WAYPOINTS}",

  "Coordinate convention":
  "- azimuth 0 deg = front
   - +45 deg = front-left; +90 deg = left; +135 deg = back-left
   - -45 deg = front-right; -90 deg = right; -135 deg = back-right
   - +/-180 deg = behind/back
   - elevation + is above and elevation - is below
   - distance defaults: very close ~= 1m, close ~= 2m,
     normal ~= 8m, far ~= 25m

   Use this convention exactly; do not swap left and right.
   If a caption does not specify exact timing, do not penalize
   the parser for using the full 0-10s range.",

  "Scoring":
  "Score each criterion on a 1-5 integer scale.",

  "Spatial faithfulness":
  "5 = all spatial and temporal details are accurate;
   4 = minor inaccuracies but overall faithful;
   3 = one major spatial aspect is wrong;
   2 = multiple spatial aspects are wrong;
   1 = waypoints do not match the described layout.",

  "Semantic preservation":
  "5 = acoustic content fully preserved with no spatial leakage;
   4 = acoustic content preserved with minor spatial leakage;
   3 = some acoustic content is lost or spatial leakage is clear;
   2 = acoustic content is strongly degraded or spatial leakage is
       severe;
   1 = acoustic content is lost or mixed with spatial details.",

  "Output format":
  "Return only this JSON object:
   {
     \"spatial_faithfulness\": <int 1-5>,
     \"semantic_preservation\": <int 1-5>,
     \"justification\": \"<one short sentence per criterion>\"
   }"
}
```

Figure 7: Prompt used for LLM-as-a-judge parser evaluation. The judge compares the original spatial caption with the parsed acoustic caption and converted waypoints.

VAE preserves waveform quality while maintaining a low angular reconstruction error of $1.92°$.

E.2 Diffusion Training

We train the latent diffusion model on top of the frozen FOA VAE. Each 10-second FOA clip is encoded into a latent sequence with 64 channels. The denoising network is a DiT with hidden dimension 768, 24 transformer layers, and 12 attention heads, and predicts the velocity target. Text captions are encoded by T5-base with a maximum length of 128 tokens. The interpolated trajectory features are encoded by a Q-Former-style encoder with 160 input frames, 16 learnable query tokens, hidden size 768, 8 attention heads, and MLP hidden size 256. The text tokens, waypoint, and temporal tokens are used as conditioning inputs to the DiT.

Table 8: FOA VAE reconstruction quality.

Model	STFT↓	Mel↓	L1(θ) ↓	L1(ϕ) ↓	Δ_{angle} ↓
FOA VAE	1.44	1.11	$0.87°$	$1.45°$	$1.92°$

Optimization. We train the diffusion model with AdamW using a learning rate 1×10^{-5}, betas $(0.9, 0.99)$, and weight decay 10^{-3}. The learning rate follows a linear warm-up followed by cosine decay to 1×10^{-6}. EMA weights with decay 0.9999 are maintained during training. Classifier-free conditioning dropout is applied with probability 0.1. The model is trained with batch size 64 per process using bfloat16 mixed precision and distributed data parallel training. The reported PhysWave checkpoint is trained for 128K optimization steps.

Physics-consistency loss. In addition to the latent denoising loss, we apply a physics-consistency loss to decoded waveform estimates during training:

$$\mathcal{L}_{\text{phys}} = \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}. \tag{16}$$

We use $\lambda_{\text{dir}} = 1.0$ and $\lambda_{\text{dist}} = 0.05$. The loss is computed on 8 sampled denoising examples per batch, using 40 ms analysis frames. The loss is applied from the start of training and weighted by the denoising SNR.

F Evaluation Metric Definitions

We provide implementation details for the evaluation metrics used in Section 5. All audio quality metrics are computed on the omnidirectional W channel of the generated FOA signal, since standard text-to-audio metrics are defined for single-channel or general audio evaluation.

Audio quality metrics. We follow standard text-to-audio evaluation protocols and report Fréchet Distance (FD), CLAP-based Fréchet Audio Distance (FAD_{CLAP}), Inception Score (IS), Kullback–Leibler divergence (KL), and CLAP score. FD, IS, and KL are computed using a pretrained PANNs classifier (Kong et al., 2020). Specifically, FD measures the Fréchet distance between generated and reference audio embeddings, while IS and KL are computed from the class-posterior distributions predicted by PANNs. FAD_{CLAP} is computed in the CLAP embedding space (Wu et al., 2023). CLAP score is computed as the cosine similarity between the CLAP audio embedding of the generated W

channel and the CLAP text embedding of the input caption.

Spatial fidelity metrics. We evaluate spatial fidelity using direction consistency and distance consistency. For direction consistency, we estimate the frame-level direction of arrival (DoA) from the generated FOA signal. For each frame k, we compute the FOA intensity vector

$$\hat{\mathbf{I}}_k = [\langle W_k X_k \rangle,\ \langle W_k Y_k \rangle,\ \langle W_k Z_k \rangle]^\top, \quad (17)$$

where $\langle \cdot \rangle$ denotes averaging within the frame. The estimated azimuth and elevation are

$$\begin{aligned} \hat{\theta}_k &= \operatorname{atan2}(\hat{I}_{y,k}, \hat{I}_{x,k}), \\ \hat{\phi}_k &= \operatorname{atan2}\left(\hat{I}_{z,k}, \sqrt{\hat{I}_{x,k}^2 + \hat{I}_{y,k}^2}\right). \end{aligned} \quad (18)$$

Given the ground-truth direction (θ_k, ϕ_k), we compute the spherical angular error using the haversine formula. Let

$$a_k = \sin^2\left(\frac{\Delta\phi_k}{2}\right) + \cos(\phi_k)\cos(\hat{\phi}_k)\sin^2\left(\frac{\Delta\theta_k}{2}\right), \quad (19)$$

where $\Delta\theta_k = \hat{\theta}_k - \theta_k$ and $\Delta\phi_k = \hat{\phi}_k - \phi_k$. The DoA error is

$$\Delta_{\text{angle},k} = 2 \cdot \operatorname{atan2}\left(\sqrt{a_k}, \sqrt{1 - a_k}\right). \quad (20)$$

We report the mean Δ_{angle} in degrees over active frames.

For distance consistency, we evaluate whether the generated signal follows the trajectory-conditioned inverse-square attenuation profile defined by the ground-truth distance trajectory. We first compute the generated W-channel energy at each frame:

$$E_k = \langle \hat{W}_k^2 \rangle. \quad (21)$$

We define the active-frame set $\mathcal{K}$ using a relative energy gate on the generated W-channel energy:

$$\mathcal{K} = \{k \mid E_k \geq \tau \max_j E_j\}. \quad (22)$$

We set $\tau = 10^{-3}$ in all experiments. The target inverse-square profile is defined as

$$q_k = \frac{1}{r_k^2}, \quad (23)$$

where r_k is the ground-truth source-listener distance. Since absolute loudness depends on the source content, we compare normalized log-energy profiles. For any positive sequence u_k, we define

$$\widetilde{u}_k = \log(u_k + \epsilon) - \frac{1}{|\mathcal{K}|}\sum_{j \in \mathcal{K}} \log(u_j + \epsilon), \quad (24)$$

where the normalization is computed over active frames.

We report two distance metrics. InvSqErr measures the RMS difference between the normalized generated energy profile and the normalized target profile:

$$\text{InvSqErr} = \frac{10}{\log 10}\sqrt{\frac{1}{|\mathcal{K}|}\sum_{k \in \mathcal{K}}\left(\widetilde{E}_k - \widetilde{q}_k\right)^2}. \quad (25)$$

InvSqCorr measures the Pearson correlation between the two normalized profiles over active frames:

$$\text{InvSqCorr} = \text{corr}_{k \in \mathcal{K}}(\widetilde{E}_k, \widetilde{q}_k). \quad (26)$$

Lower InvSqErr and higher InvSqCorr indicate better agreement with the specified distance attenuation cue.

G Additional Visualizations

We provide additional visualizations of PhysWave's spatial behavior. Figure 8 compares the generated log-energy envelope with the target inverse-square profile over time. Figure 9 shows additional trajectory examples across the four motion families.

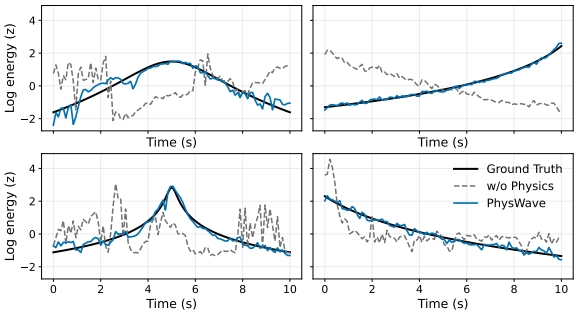

Figure 8: Energy envelope comparison. PhysWave better follows the target inverse-square log-energy profile, while the model without physics losses shows larger temporal deviations.

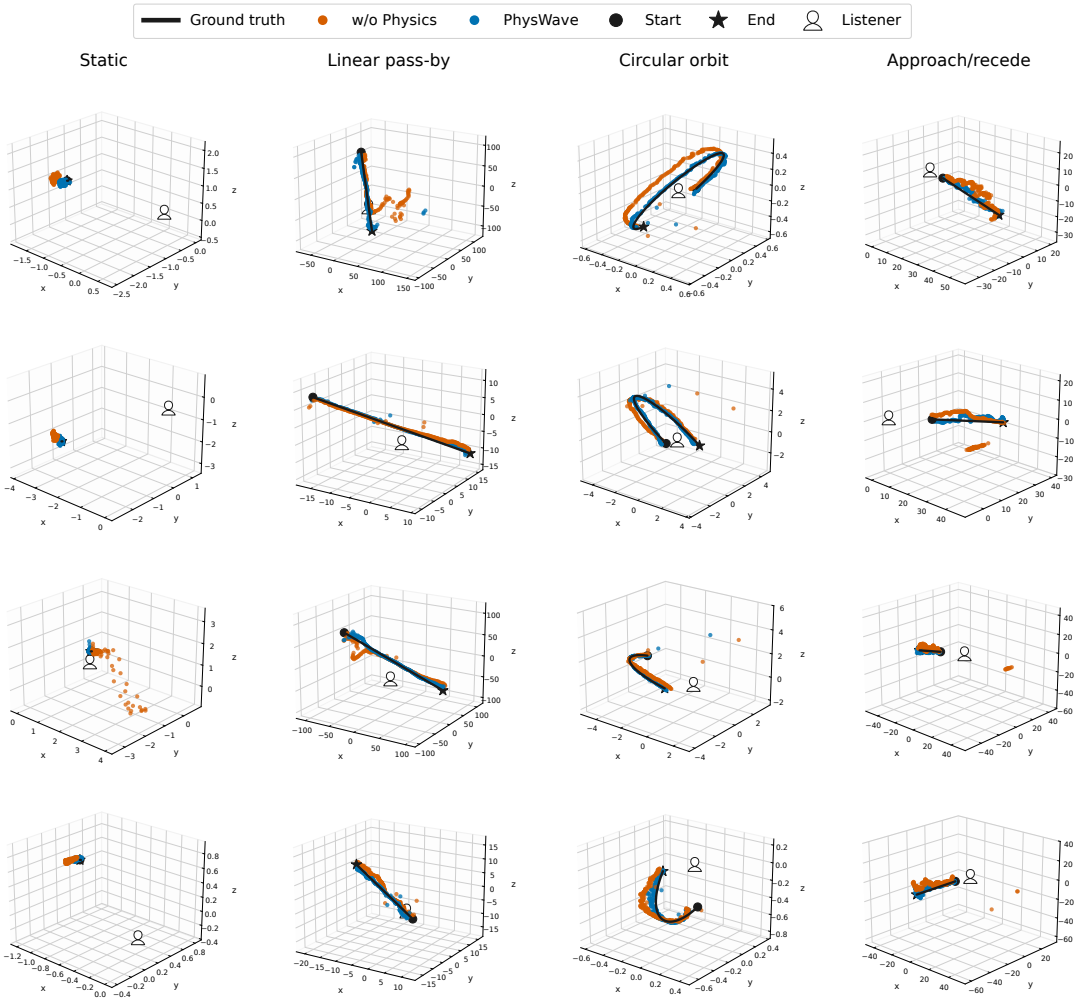

Figure 9: More qualitative trajectory comparison between PhysWave with and without physics-consistency losses.