DEPTHFOLEY: DEPTH-GUIDED VIDEO-TO-AUDIO GENERATION FOR SPATIAL CONSISTENCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Current Video-to-Audio (V2A) approaches have achieved impressive semantic quality but often treat the auditory environment as a spatially static and distance-agnostic plane. Neglecting fundamental distance-dependent acoustic decay, such as sound attenuation and spectral filtering, results in audio that fails to reflect spatial dynamics and shatters the immersive illusion. To address this, we propose **DepthFoley**, a framework that shifts the paradigm from purely data-driven fitting to a depth-guided alignment strategy. By leveraging metric depth estimation, we extract precise metric depth trajectories to guide a pre-trained backbone model via a Depth ControlNet. We further introduce a training-free object-wise inference strategy for multi-object scenes. To ensure spatial consistency without sacrificing generative priors, we introduce Sound Pressure Level (SPL) and Spectrally Perceptual Energy (SPE) losses. Augmented by depth-informed regularization, these objectives guide the model to internalize realistic decay trends while preserving disentangled acoustic dynamics, balancing veridical spatial realism with artistic flexibility. To evaluate this, we introduce DITC and DIGA, two novel metrics experimentally validated to align with human spatial perception. Extensive experiments demonstrate that our architecture achieves superior spatial-acoustic alignment and strong zero-shot generalizability to complex, out-of-distribution environments, establishing a new standard for spatial perceptual realism.

1 INTRODUCTION

Advances in generative AI have enabled high-fidelity video synthesis, driving a commensurate demand for highly realistic auditory experiences. This has propelled Video-to-Audio (V2A) generation, or Foley synthesis Iashin & Rahtu (2021); Sheffer & Adi (2023); Luo et al. (2023); Wang et al. (2024), to the forefront of research. Beyond mere sound effects, a high-quality Foley system is expected to function as a virtual sound engineer: it must not only identify the correct acoustic semantics and synchronize its onset with visual motion, but also faithfully simulate how sound propagates and evolves within the space. Bridging this gap between visual dynamics and auditory perception is essential to transforming "silent films" into truly immersive, spatially-grounded experiences.

Despite progress in visual-semantic alignment, current data-driven paradigms predominantly treat the acoustic environment as a distance-agnostic plane. While state-of-the-art models (Xing et al., 2024; Benita et al., 2025; Zhang et al., 2026; Cheng et al., 2025) excel at semantical and temporal synchronization, they consistently neglect distance-induced acoustic modulations. As illustrated in Fig.1, the auditory signature of a speedboat should diverge sharply as it moves from five to fifty meters away, reflecting both sound pressure attenuation and spectral shifts. By failing to capture this depth-dependent evolution, existing models generate uniformly proximate audio. This lack of spatial grounding creates a cognitive dissonance between dynamic visual depth and static acoustic intensity, severely degrading the immersive experience.

The precise spatial evolution of sound-emitting objects can now be reliably captured through advancements in visual perception. Capitalizing on this, we present **DepthFoley**, a framework that shifts the paradigm from purely data-driven fitting to a distance-guided alignment strategy. Recognizing that simple concatenation of distance cues is insufficient for models to autonomously internalize sound attenuation, we integrate a ControlNet-like (Zhang et al., 2023) architecture into a frozen generative backbone. This design enables the disentangled modulation of spatial attributes while

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

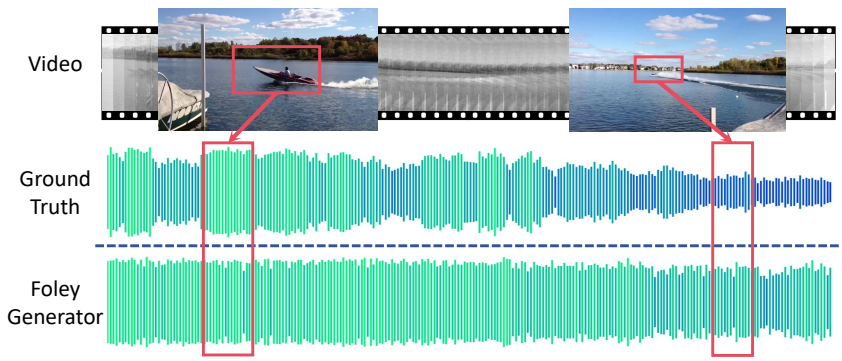


Figure 1: **Illustration of the distance-consistency gap in Foley generation.** As the speedboat moves away (top), the ground-truth audio should exhibit spatial decay (middle). Conventional Foley generators, however, often produce a static amplitude (bottom), leading to a loss of spatial realism and audio-visual misalignment.

preserving original generative priors, allowing for flexible and interactive trajectory editing. Although source-level audio is unavailable during training, at inference we can decompose a multi-object scene into object-centric inputs. We therefore repeatedly apply the shared DepthFoley model using individual object trajectories and compose the resulting audio stems, avoiding a single ambiguous trajectory when objects move differently. Rather than relying on massive datasets to implicitly learn spatial-audio correlations, our approach leverages fundamental acoustic principles, majorly the inverse-square law of distance attenuation, as a directional guide. We incorporate specialized Sound Pressure Level (SPL) and Spectrally Perceptual Energy (SPE) losses, powered by a pre-trained acoustic predictor. These objectives, further refined by Spatially-Anchored Motion Sparsification and Acoustic Slack Constraints, anchor the model's vast generative capacity to veridical spatial trends. By utilizing a relatively small amount of high-quality data as a pivot, the framework establishes a rigorous spatial-acoustic alignment that transcends the limitations of traditional statistical correlation.

To validate our framework, we conduct extensive experiments on curated subsets of VGGSound (Chen et al., 2020) and AVE (Tian et al., 2018), including both single- and multi-object scenes. Using established audio metrics together with the proposed DITC and DIGA, our method consistently improves depth-intensity consistency while maintaining competitive audio quality. Ablation studies further verify the complementary contributions of the Depth ControlNet and dual-loss design. Additional evaluations demonstrate generalization to unseen categories and zero-shot OOD environments, as well as robustness to errors in monocular depth estimation.

In summary, our major contributions are as follows:

- We identify a critical bottleneck in current Foley synthesis: the prevalence of spatially-static and distance-agnostic generation. We overcome this limitation by formalizing a new paradigm of depth-aligned Foley generation, where models are explicitly constrained to synchronize acoustic dynamics with the underlying distance-dependent intensity variations.
- We propose **DepthFoley**, a novel approach that integrates a depth-guided ControlNet into a frozen generative backbone. By introducing spatially grounded objectives: Sound Pressure Level (SPL) and Spectrally Perceptual Energy (SPE) losses, we enable the disentangled modulation of acoustic dynamics, mastering the complex relationship between spatial proximity and acoustic dynamics without compromising generative fidelity. The same trained model further supports object-wise multi-pass inference for multi-object scenes without additional parameters or training.
- Addressing the absence of standardized benchmarks, we introduce evaluation protocols (DITC and DIGA) to quantify spatial-acoustic alignment. Extensive experiments demonstrate that DepthFoley establishes a new state-of-the-art in spatial perceptual realism and zero-shot generalizability, while maintaining competitive overall audio fidelity.

2 Related Work

Foley Generation. Foley generation has been studied under two closely related settings: *Video-to-Audio (V2A)* and *Video-and-Text-to-Audio (VT2A)*. V2A methods generate audio solely from silent video by learning audio-visual correspondence directly from large-scale datasets or via training-free alignment with pretrained generators, Iashin & Rahtu (2021); Sheffer & Adi (2023); Luo et al. (2023); Wang et al. (2024); Xing et al. (2024). In contrast, VT2A approaches explicitly leverages powerful pretrained text-to-audio (T2A) generative models and inject visual information into powerful pretrained text-to-audio models through adapters or ControlNet-style architectures Jeong et al. (2025); Lee et al. (2025); Benita et al. (2025); Zhang et al. (2026). Alternatively, MMAudio Cheng et al. (2025) jointly trains video, text, and audio modalities from scratch using a dedicated multimodal fusion architecture, achieving strong synchronization and audio quality. Despite these advances, existing methods primarily focus on semantic alignment and temporal consistency, while acoustic variations caused by object distance and spatial sound propagation are rarely modeled explicitly, which is the primary focus of our work.

Spatial Audio Generation. Spatial audio research seeks to reconstruct immersive auditory experiences consistent with scene geometry. Prior works predominantly address two sub-fields: audio spatialization, which transforms monaural audio into binaural formats via visual source localization Gao & Grauman (2019); Liu et al. (2024a), and indoor acoustics, which models Room Impulse Responses (RIRs) and reverberation Garg et al. (2021); Saad & Al-Halah (2025). Recently, diffusion models have further advanced First-Order Ambisonics (FOA) Kushwaha et al. (2025); Heydari et al. (2025); Kim et al. (2025), video-guided binaural generation Zhao et al. (2025), and high-quality stereo synthesis Evans et al. (2025). Despite this progress, existing methods focus almost exclusively on directional localization or room-level reverberation. The fundamental impact of source-to-listener distance on acoustic energy and spectral attenuation remains largely unexplored. In contrast, our work explicitly models these distance-aware acoustic dynamics, complementing existing spatialization techniques with geometrically grounded energy and spectral constraints.

Object-aware Distance Estimation from Video. Recent advances in computer vision have enabled the robust extraction of 3D spatial information directly from monocular video. State-of-the-art depth estimation models demonstrate strong capability in reconstructing scene geometry and have progressed from relative depth perception to metric depth estimation, allowing the recovery of absolute physical distance values Yang et al. (2024a;b). In parallel, large-scale vision foundation models for image and video segmentation, such as Segment Anything Ravi et al. (2025), provide accurate and temporally consistent object masks across frames. By combining object-level segmentation with metric depth estimation, it becomes possible to extract per-frame distance trajectories for sounding objects, thereby providing a practical conditioning signal for depth-guided audio generation.

3 DepthFoley

Our proposed method is illustrated in Fig. 2. To synthesize audio that accurately follows the physical laws of sound propagation relative to visual depth in the video, we first estimate object-centric distance trajectories from the video in Sec. 3.1. We then introduce depth-driven audio supervision in Sec. 3.2. Next, we present our framework, which adopts a ControlNet-style adapter, in Sec. 3.3. In Sec. 3.4, we introduce the scaled feature injection during inference, followed by the object-wise multi-pass inference strategy for multi-object scenes in Sec. 3.5.

3.1 Spatial Information Extraction

To extract the sounding object's frame-wise distance trajectory, we employ a two-stream pipeline. Guided by the text prompt T, we first extract a sequence of binary object masks $M \in \{0,1\}^{L\times H\times W}$ from the video V. Concurrently, we estimate pixel-wise metric depth maps $N \in \mathbb{R}^{L\times H\times W}$. By spatially aggregating the metric depth values strictly within these masked regions, we compute the representative distance d_t for each frame t:

$$d_t = \frac{\sum_{i,j} M_{t,i,j} \cdot N_{t,i,j}}{\sum_{i,j} M_{t,i,j} + \zeta}, \tag{1}$$

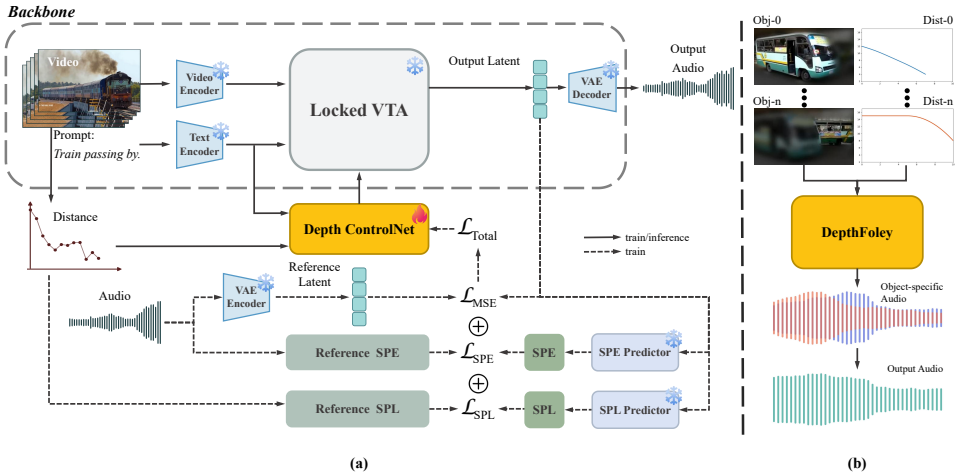


Figure 2: **(a) The proposed DepthFoley architecture.** Our method extends the CAFA baseline by introducing a Depth ControlNet for depth-aware feature modulation. The training process is supervised by SPL and SPE losses, to ensure the model accurately captures the correlation between acoustic properties and distance variations. **(b) Object-wise multi-pass inference.** For each detected object, its object-centric video and distance trajectory are independently processed by the same trained DepthFoley model to generate an object-specific audio stem. The resulting stems are summed to produce the final output audio without additional training.

where ζ is a small constant added for numerical stability. The resulting metric depth sequence $D = \{d_1, \ldots, d_L\}$ serves as the explicit spatial distance condition for the generation modules.

During training, multiple object trajectories are consolidated into a single composite trajectory because only mixed reference audio is available; its construction is detailed in Sec. 4.1. During multi-object inference, Eq. (1) is instead applied independently to each detected object without trajectory merging. In our work, we employ Grounded-SAM2 Liu et al. (2024b); Ren et al. (2024); Ravi et al. (2025) and Depth Anything V2 (Metric Model) Yang et al. (2024b) to obtain the segmentation masks M and depth map N.

3.2 Depth-Driven Audio Supervision

To guide the adapter training in Sec. 3.3 to respect spatial constraints while preserving semantic fidelity, we introduce depth-driven audio supervision. Specifically, we propose two novel objectives: Sound Pressure Level (SPL) loss and Spectrally Perceptual Energy (SPE) loss. Each objective is equipped with a lightweight predictor, pretrained offline and kept frozen during adapter training. We use $t \in \{1, \ldots, L\}$ to index discrete temporal L frames extracted from the video.

Sound Pressure Level (SPL) loss. Sound propagation in an ideal free field follows the *Inverse-Square Law* Kinsler et al. (2000), where acoustic intensity decreases logarithmically with distance. The Sound Pressure Level (SPL) at a distance d_i relative to a reference distance d_j is given by:

$$\text{SPL}_i = \text{SPL}_j - 20 \log_{10} \left(\frac{d_i}{d_j} \right). \tag{2}$$

Inspired by this physical principle, we introduce a lightweight SPL predictor $\mathcal{P}_{\text{SPL}}$, which maps the generated audio latent $\hat{z}_0$ to a temporal SPL trajectory:

$$\hat{S} = \mathcal{P}_{\text{SPL}}(\hat{z}_0), \quad \hat{S} \in \mathbb{R}^L. \tag{3}$$

We define the predicted pairwise SPL difference as $\Delta \hat{S}_{i,j} = \hat{S}_i - \hat{S}_j$. The corresponding distance-derived target is computed from the estimated source distances using the free-field inverse-square law:$\Delta S^{\text{phys}}_{i,j} = -20 \log_{10} \left(\frac{d_i}{d_j} \right)$.

Assisted by $\mathcal{P}_{\text{SPL}}$, our SPL loss learns the relative Distance-SPL correspondence by operating on sampled pairs of temporal indices. These sampled pairs are integrated with two mechanisms to balance rigid spatial constraints with natural audio dynamics:

1) Spatially-Anchored Motion Sparsification: We employ a Spatial Saliency Mask Ω defined by a threshold τ_{dist}:$\Omega = \{(i,j) \mid |\Delta d_{i,j}| > \tau_{dist}\}$, which filters out pairs with negligible distance variation, encouraging the model to focus on spatially meaningful acoustic changes.

2) Acoustic Slack Constraint:

To accommodate source-intrinsic and expressive variations beyond ideal inverse-square attenuation, we introduce an *acoustic slack constraint* with a tolerance margin δ, which relaxes the rigid spatial supervision by allowing small deviations: $\left|\Delta\hat{S}_{i,j} - \Delta S_{i,j}^{phys}\right| < \delta$.

The SPL loss is formulated as:

$$\mathcal{L}_{\text{dir}} = \sum_{(i,j)\in\Omega} \max\left(0, -\Delta\hat{S}_{i,j} \cdot \text{sign}\left(\Delta S_{i,j}^{phys}\right)\right), \tag{4}$$

$$\mathcal{L}_{\text{marg}} = \sum_{(i,j)\in\Omega} \max\left(0, \left|\Delta\hat{S}_{i,j} - \Delta S_{i,j}^{phys}\right| - \delta\right), \tag{5}$$

$$\mathcal{L}_{\text{SPL}} = \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{marg}}\mathcal{L}_{\text{marg}}. \tag{6}$$

Spectrally Perceptual Energy (SPE) Loss. While SPL loss regulates the total energy envelope, it lacks spectral specificity, which may lead the model to amplify high-frequency noise rather than the object's acoustic components. To enforce frequency-aware modulation, we introduce the Spectrally Perceptual Energy (SPE) loss.

We define a lightweight SPE predictor $\mathcal{P}_{\text{SPE}}$, which maps the generated audio latent $\hat{z}_0$ to band-wise spectral energy trajectories:

$$\hat{E} = \mathcal{P}_{\text{SPE}}(\hat{z}_0), \quad \hat{E} \in \mathbb{R}^{B\times L}, \tag{7}$$

where $\hat{E}_{b,t}$ denotes the energy of the b-th frequency band at temporal index t. And to suppress noise-related artifacts, we restrict supervision to high-frequency bands: $\mathcal{B}_{\text{high}} = \{b \mid b \geq b_0\}$.

The SPE loss is defined as:

$$\mathcal{L}_{\text{SPE}} = \frac{1}{|\mathcal{B}_{\text{high}}|\, L} \sum_{b\in\mathcal{B}_{\text{high}}} \sum_{t=1}^{L} \text{SmoothL1}(\hat{E}_{b,t}, E_{b,t}^{gt}). \tag{8}$$

The operator $\text{SmoothL1}(\cdot)$ denotes the Huber loss, applied element-wise, providing robustness to outliers while maintaining stable gradients.

3.3 ControlNet-style Adapter Training

With the distance-driven audio supervision provided by the SPL and SPE losses, we train a ControlNet-style adapter to enable the model to generate depth-informed audio.

ControlNet-style Adapter. Our model takes a text prompt T, silent video V, and metric depth sequence $D = \{d_1, \ldots, d_L\} \in \mathbb{R}^L$ as conditions. Rather than fine-tuning the pretrained generator, we preserve the complete CAFA backbone, including its video-conditioned branch, and introduce a trainable Depth ControlNet A_D. For the l-th DiT block, the conditional hidden state is formulated as:

$$h_{\text{cond}}^{(l)} = \underbrace{\text{Block}_l\left(h^{(l-1)}, c_T\right) + A_V^{(l)}(c_V, c_T)}_{\text{Frozen CAFA}} + \underbrace{A_D^{(l)}(c_D, c_T; \theta_D)}_{\text{Trainable Depth ControlNet}}, \tag{9}$$

where A_V denotes the frozen visual adapter inherited from CAFA, while A_D produces trainable depth-conditioned residuals. During training, only θ_D, including the Depth ControlNet and its output projection, is optimized; the CAFA backbone and A_V remain frozen.

Total Optimization Objective. The total optimization objective for the ControlNet-style adapter combines the spatial regularization terms from Sec.3.2 and the fundamental generation objective, the

standard diffusion noise ϵ_{DC} prediction loss $\mathcal{L}_{\text{diff}}$:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t,c_{all}} \left[\|\epsilon - \epsilon_{\text{DC}}(z_t, t, c_{all})\|_2^2 \right] \tag{10}$$

The final objective is defined as a weighted sum of the diffusion loss, SPL loss, and SPE loss.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{SPL}}\mathcal{L}_{\text{SPL}} + \lambda_{\text{SPE}}\mathcal{L}_{\text{SPE}}. \tag{11}$$

3.4 Scaled Feature Injection

A key advantage of our decoupled architecture is the ability to explicitly disentangle semantic generation from spatial modulation without retraining. During inference, we employ a **Scaled Feature Injection** strategy, formulating the conditional hidden state $h_{\text{cond}}^{(l)}$ of each DiT block as a superposition of the semantic foundation and a weighted spatial residual:

$$h_{\text{cond}}^{(l)} = \text{Block}_l(h^{(l-1)}, c_T) + \mathcal{A}_V^{(l)}(c_V, c_T) + w_d \cdot \mathcal{A}_D^{(l)}(c_D, c_T) \tag{12}$$

Here, the first two terms denote the frozen CAFA backbone, capturing rich acoustic textures but lacking explicit distance dynamics. The third term represents the spatial perturbation induced by our Depth ControlNet, modulated by the **spatial control scale** w_d. The scale w_d controls the trade-off between the frozen backbone's audio prior and depth-aware modulation, enabling adjustable spatial control without retraining.

3.5 Object-wise Multi-pass Inference

For a scene containing K sounding objects, we independently extract an object-centric video $V^{(k)}$, prompt $T^{(k)}$, and distance trajectory $D^{(k)}$. The same trained model is then applied to each object:

$$\hat{x}^{(k)} = G_\theta \left(V^{(k)}, T^{(k)}, D^{(k)}; w_d \right). \tag{13}$$

The resulting stems are summed and globally rescaled only when necessary:

$$x_{\text{sum}} = \sum_{k=1}^{K} \hat{x}^{(k)}, \qquad \hat{x} = \frac{x_{\text{sum}}}{\max\left(1, \|x_{\text{sum}}\|_\infty\right)}. \tag{14}$$

This training-free strategy avoids merging conflicting object trajectories while preserving their generated relative amplitudes. It introduces no additional parameters, with inference cost scaling approximately linearly with K.

4 Experiment

4.1 Experimental Setup

Dataset. We curate 9,500 10-second clips (9,000 for training, 500 for testing) from 20 motion-driven VGGSound categories Chen et al. (2020) (e.g., race cars, aircraft), rigorously filtering out severe occlusions, brief visibility, or conflicting multi-source motions only for training set. Since source-level audio stems are unavailable during training, trajectories from multiple objects are consolidated into a single composite training trajectory by averaging spatially close, co-moving objects or tracking the nearest object when they are separated. This trajectory is used only for mixed-audio training; multi-object inference instead processes each object trajectory independently, as described in Sec. 3.5. Although audio intensity is jointly affected by source-intrinsic actions and spatial proximity, our dataset exhibits a consistent macroscopic distance-dependent attenuation trend. Correlating the mean SPL across distance bins with the inverse-square reference yields a Pearson correlation of 0.37 and a Spearman correlation of 0.40. This aggregate alignment remains observable despite action- and scene-dependent acoustic variations, providing an empirical basis for learning depth-aware acoustic modulation without assuming fixed action-specific intensity patterns. Finally, to ensure robust evaluation across diverse real-world scenarios, we augment the test set with 500 dynamic sequences from AVE Tian et al. (2018), yielding a total of 1,000 test samples.

Baselines. We compare DepthFoley against several state-of-the-art models, namely Seeing&Hearing (S&H) Xing et al. (2024), CAFA Benita et al. (2025), MMAudio Cheng et al. (2025), and FoleyCrafter

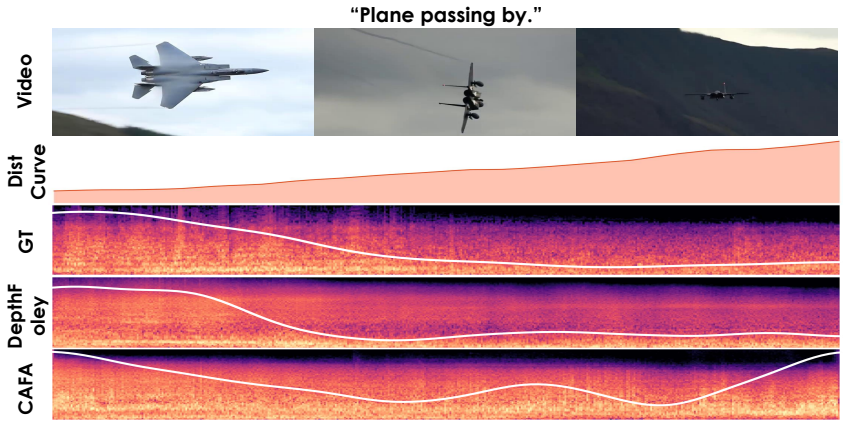


Figure 3: **Qualitative comparison of samples.** As the distance increases according to the scenes shown in the frames, the Root Mean Square (RMS) curves (white lines) on the Mel-spectrograms below illustrate the volume trends. Our method successfully replicates the natural attenuation observed in the Ground Truth (GT), whereas the baseline CAFA exhibits an erroneous volume increase, failing to maintain audio-visual consistency.

(FC) Zhang et al. (2026). For these models, we employ a unified prompt configuration: "The sound of <label>". All methods generate 10-second audio clips for fair comparison. To validate the necessity of our generative spatial-acoustic alignment, we additionally introduce a Post-hoc Scaling baseline, which applies an inverse-square law volume envelope directly to the output of the base CAFA model.

Evaluation Metrics. We follow established metrics to assess general performance. For audio fidelity, we measure perceptual quality and realism using Fréchet Audio Distance (FAD) Roblek et al. (2019), Kullback–Leibler divergence (KL), all computed on features extracted by a pretrained PANNs Kong et al. (2020) network. Furthermore, we evaluate temporal coherence using DeSync (to quantify temporal synchronization between visual actions and audio events) Cheng et al. (2025). We leverage ImageBind (IB)Girdhar et al. (2023)to quantify semantic similarity between the ground truth video and generated audio.

Implementation Details. We use 50-frame depth sequences sampled at 5 FPS, with $\lambda_{\mathrm{SPL}} = 0.025$ and $\lambda_{\mathrm{SPE}} = 0.5$. Complete optimization, predictor-training, and computational details are provided in Appendix E.

4.2 Spatial-Audio Consistency Evaluation

We propose two complementary metrics to quantitatively evaluate our method's unique contribution toward maintaining **spatial-aligned audio-visual trajectories**.

Depth–Intensity Trend Consistency (DITC). DITC measures local consistency between visual distance changes and audio intensity trends. It computes the negated Kendall's τ_b rank correlation between metric distance and audio intensity within temporally segmented intervals, with higher scores indicating stronger agreement with the expected distance-dependent attenuation trend. Segment-wise evaluation reduces reliance on a single global trend across acoustically distinct events.

Depth–Intensity Global Alignment (DIGA). DIGA evaluates macro-level acoustic fidelity by mapping the distance trajectory to an ideal SPL profile using the inverse-square law, then computing its Spearman rank correlation with the generated audio's intensity profile. It measures how consistently the audio follows real-world distance-dependent attenuation.

Further technical details regarding the windowing parameters and the mathematical derivation of these metrics are provided in the Appendix A.

Table 1: **Quantitative evaluation results.** Our method significantly outperforms existing baselines in spatial-audio consistency metrics. In terms of general audio performance, our model remains competitive across multiple metrics while maintaining superior spatial alignment.

Model	Spatial-Audio Consistency		General Performance			
	DITC↑	DIGA↑	FAD↓	KL↓	Desync ↓	IB ↑
MMAudio	0.0491	0.0757	8.41	**1.392**	**0.95**	0.38
FC	0.0333	0.0787	9.89	1.726	1.35	0.29
S&H	0.0407	0.0552	17.01	2.410	1.24	**0.39**
CAFA	0.0571	0.1383	**7.20**	1.665	1.14	0.21
Post-hoc	**0.5684**	0.7544	9.48	1.467	1.01	0.20
Ours	0.5529	**0.7685**	7.35	1.472	1.13	0.22

4.3 Results and Analysis

Objective Evaluation Results. As shown in Table 1, our method outperforms all generative baselines in spatial-acoustic alignment, achieving the highest **DITC** and **DIGA**. Although the non-generative **Post-hoc Scaling** baseline obtains a higher DITC by directly applying an attenuation curve to the waveform, it uses a single global gain envelope for the entire audio mixture. It therefore cannot independently control sources with different trajectories in multi-object scenes and also results in degraded audio fidelity, as reflected by its higher FAD. In contrast, DepthFoley learns distance-aware acoustic modulation within the generative process, providing improved spatial consistency with only a modest fidelity trade-off.

The SPL and SPE objectives provide explicit spatial supervision beyond naive ControlNet conditioning, while the comparable IB score indicates that the capabilities inherited from the CAFA backbone are largely preserved. This modular design can also be integrated with stronger generative backbones. Qualitative results in Fig. 3 further show that DepthFoley follows the observed distance-dependent attenuation trend, whereas CAFA produces inconsistent intensity variations.

Generalizability Analysis. To evaluate generalization across sound categories, we conducted a Leave-5-out cross-validation, training on 15 categories and testing on 5 unseen classes. The model maintained strong spatial alignment (DITC $= 0.5465$, DIGA $= 0.7537$), confirming category-agnostic learning. Furthermore, zero-shot Out-of-Distribution (OOD) tests on 200 diverse indoor internet clips (e.g., infants, bowling) demonstrated that our decoupled architecture successfully extends fundamental spatial priors to complex, non-free-field environments, outperforming the base CAFA model (DITC : 0.463 vs 0.041; DIGA : 0.647 vs 0.062).

Robustness Analysis. Additionally, we assessed the framework's robustness against upstream monocular depth estimation errors. By injecting various forms of controlled noise into the conditioning depth sequences, we confirmed that our model maintains strong spatial-acoustic resilience. It also retains clear depth–intensity consistency under simulated reverberation; at RT60=1.2 s, it achieves DITC/DIGA of $0.415/0.639$, compared with $0.104/0.104$ for CAFA. Detailed experimental configurations and comprehensive quantitative results are provided in the Appendix D.

4.4 Subjective Evaluation

To complement the objective metrics and assess human perception of the generated audio, we conducted a subjective user study focusing on two key dimensions: **Audio Quality (AQ)** and **Spatial-Perceptual Alignment (SPA)**. We randomly selected 30 video clips from the test set. For each clip, we generated audio using our method and the baseline models. We recruited 30 volunteers (aged 20-30, with normal hearing) to participate in the listening test. The details of subjective evaluation is presented in the Appendix. For each video, participants were asked to rate the performance on a 5-point Likert scale (1: Poor, 5: Excellent) based on the following criteria:

Audio Quality (AQ): The overall fidelity and realism of the audio, as well as its semantic and temporal alignment with the video content.

Spatial-Perceptual Alignment (SPA): How accurately the audio intensity and timbre adheres to the spatial cues and physical laws relative to the visual scene.

Table 2 reports the Mean Opinion Scores (MOS) with 95% confidence intervals. Our method achieves the highest AQ and SPA scores, with a particularly clear improvement in spatial alignment. These results indicate that depth-conditioned modulation improves perceived spatial consistency while maintaining high subjective audio quality.

Validation of Proposed Objective Metrics. To establish DITC and DIGA as reliable proxies for human perception, we evaluated their correlation with human Spatial-Perceptual Alignment (SPA) ratings across the test set. The analysis yielded strong positive correlations for both metrics: DITC achieved a Spearman's ρ of 0.57 and Pearson's r of 0.71, while DIGA reached $\rho = 0.60$ and $r = 0.61$ (all $p < 0.05$). These statistically significant results validate that our objective metrics reliably capture the nuances of human spatial hearing, rather than serving as mere numerical artifacts.

Table 2: **Subjective evaluation results.** Our method achieves the highest scores in both dimensions. Notably, the pronounced advantage in SPA significantly enhances the perceived authenticity of the generated Foley.

Model	SPA↑	AQ↑
MMAudio	3.25 ± 0.42	3.62 ± 0.41
FC	2.36 ± 0.41	2.91 ± 0.37
S&H	1.56 ± 0.30	1.77 ± 0.34
CAFA	2.67 ± 0.45	2.98 ± 0.40
Post-hoc	2.82 ± 0.39	3.01 ± 0.33
Ours	$\mathbf{4.18 \pm 0.42}$	$\mathbf{4.01 \pm 0.34}$

Table 3: **Ablation study of the proposed components.** We evaluate the contribution of each module toward the final performance. The results demonstrate that the full integration yields the best improvements.

Method Components			Metrics		
DC	SPL	SPE	DITC ↑	DIGA ↑	FAD ↓
×	×	×	0.1177	-0.1050	9.48
✓	×	×	0.0528	0.0635	34.86
×	✓	✓	0.1377	-0.1550	11.30
✓	✓	×	0.4637	0.7482	11.86
✓	×	✓	0.0739	0.1073	10.60
✓	✓	✓	**0.5529**	**0.7685**	**7.35**

4.5 Ablation Study

We evaluate five architectural and objective variants to validate our framework (Tab.3). Pure data-driven fine-tuning and naive feature concatenation fail to establish robust spatial alignment. SPL and SPE supervision without DC cannot yield strong spatial alignment, verifying DC as an essential structural anchor. Isolated DC (no spatial losses) hurts FAD and fails to improve DITC/DIGA, as raw depth cues disrupt generative priors. Removing SPL greatly reduces spatial-acoustic consistency, while omitting SPE degrades general audio quality. The ablations expose a fidelity-consistency tradeoff: any missing module weakens one objective. The complete DC+SPL+SPE pipeline best balances both requirements for distance-aware Foley synthesis. We additionally ablate multi-object inference on the same test set, varying only the handling of multi-object videos: the previous composite trajectory, the first obj trajectory, or the proposed object-wise multi-pass inference. Details are provided in Appendix I.

5 Conclusion and Discussion

We presented **DepthFoley**, which introduces depth-conditioned residual control and spatially grounded objectives into a frozen Foley generator. The resulting model improves depth-aware acoustic consistency while retaining competitive audio quality, and supports training-free object-wise inference for multi-object scenes. DITC and DIGA further provide complementary measures of local and global spatial alignment.

Limitation. DepthFoley uses free-field attenuation as a first-order prior. Although the RIR stress test shows that its learned depth-intensity trend remains observable under simulated reverberation, the model does not explicitly estimate or control room-acoustic properties such as DRR or RT60. Moreover, the lack of source-separated training audio requires a composite trajectory during training, which may introduce ambiguous supervision for conflicting source motions; object-wise inference avoids trajectory merging but depends on reliable segmentation and scales approximately linearly with the number of objects.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing. We did not use generative AI tools to ggenerate synthetic data sets, help develop theoretical models or conceptual frameworks, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to create or edit software code. All AI-assisted text was reviewed, revised, and verified by the authors. AI-assisted code was manually inspected and tested against the expected outputs, while all reported results were obtained from experiments conducted and verified by the authors. The authors independently checked the technical claims, citations, and interpretations presented in the paper. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our subjective evaluation involved adult volunteers performing minimal-risk perceptual judgments of generated audio. Before participation, all volunteers were informed of the study purpose and procedure and provided consent. Their responses were collected anonymously, and no personally identifiable or sensitive information was recorded. Our experiments use publicly available audiovisual research datasets and do not identify, profile, or infer sensitive attributes of individuals. Nevertheless, improved video-to-audio generation may facilitate deceptive audiovisual content or unauthorized media manipulation. We therefore encourage the use of provenance and synthetic-media detection mechanisms and plan to release the code and pretrained weights under a non-commercial license. Because the framework relies on pretrained vision and audio models, it may also inherit dataset biases or perform unreliably in underrepresented scenes; it should therefore not be used in safety-critical applications without further validation.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce DepthFoley throughout the main paper and appendix. Sections 3.1- 3.5 describe the depth-trajectory extraction pipeline, model architecture, training objectives, scaled feature injection, and object-wise multi-pass inference. Section 4.1 specifies the dataset construction, train–test splits, baselines, evaluation protocol, training hyperparameters, and computational resources. Appendix A gives the complete definitions of DITC and DIGA, while Appendix B details the architectures and training procedures of the frozen SPL and SPE predictors. Additional experiments on trajectory controllability, input-depth robustness, guidance strength, ground-truth analysis, subjective evaluation, and multi-object inference are documented in Appendices C- I. To support further verification and extension, we will release the source code, model configurations, pretrained checkpoints, data-processing scripts, and evaluation implementation upon publication.

REFERENCES

Roi Benita, Michael Finkelson, Tavi Halperin, Gleb Sterkin, and Yossi Adi. Cafa: a controllable automatic foley artist. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 5178–5193, 2023.

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

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

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

Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021.

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

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

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference (BMVC)*, 2021.

Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17590–17598, 2025.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. ViSAGe: Video-to-spatial audio generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lawrence E Kinsler, Austin R Frey, Alan B Coppens, and James V Sanders. *Fundamentals of acoustics*. John wiley & sons, 2000.

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

Saksham Singh Kushwaha, Jianbo Ma, Mark R. P. Thomas, Yapeng Tian, and Avery Bruni. 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)*, pp. 1–5, 2025.

Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam. Video-foley: Two-stage video-to-sound generation via temporal event condition for foley sound. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3813–3825, 2025.

Miao Liu, Jing Wang, Xinyuan Qian, and Xiang Xie. Visually guided binaural audio generation with cross-modal consistency. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7980–7984, 2024a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE Transactions on Audio, Speech and Language Processing*, 33:458–471, 2024c.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Dominik Roblek, Kevin Kilgour, Matt Sharifi, and Mauricio Zuluaga. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, pp. 2350–2354, 2019.

Mahnoor Fatima Saad and Ziad Al-Halah. How would it sound? material-controlled multimodal acoustic profile generation for indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12232–12241, 2025.

Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 247–263, 2018.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2024.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7151–7161, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, Bin Liu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134(1):46, 2026.

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

A Details of metric.

Depth-Intensity Trend Consistency (DITC). DITC evaluates whether the audio intensity follows the expected monotonic trend with respect to the depth trajectory. Since audio signals may contain heterogeneous acoustic events that break global monotonicity, we compute the metric in a piecewise manner. Given an audio waveform, we first extract BEATs Chen et al. (2023) embeddings $E = \{\mathbf{e}_t \in \mathbb{R}^D\}_{t=1}^{T}$ and apply linearly penalized segmentation (implemented by the `ruptures` package) to obtain M temporal segments $\{[s_m, e_m]\}_{m=1}^{M}$. Let a_t denote the sound pressure level (SPL) in dB averaged over a ± 100 ms window centered at time t, and let d_t denote the metric depth at time step t. For each segment, we compute Kendall's rank correlation coefficient

$$\tau_m = \tau\big(\{a_t\}_{t=s_m}^{e_m}, \{d_t\}_{t=s_m}^{e_m}\big),$$

and define the segment score as $\text{DITC}_m = -\tau_m$, since intensity is expected to decrease as depth increases. The final DITC score is obtained by a length-weighted aggregation:

$$\text{DITC} = \sum_{m=1}^{M} w_m \text{DITC}_m, \quad w_m = \frac{e_m - s_m + 1}{T}.$$

Higher DITC indicates stronger consistency between depth variation and audio intensity attenuation.

Depth-Intensity Global Alignment (DIGA). While DITC evaluates local monotonic consistency in a piecewise manner, DIGA measures the global alignment between the audio loudness curve and the spatially expected distance-based attenuation trend. We compute the Spearman rank correlation coefficient between the SPL curve and the metric depth sequence:

$$r_t = -\log(d_t + \epsilon), \qquad \text{DIGA} = \rho\left(\{a_t\}_{t=1}^{T}, \{r_t\}_{t=1}^{T}\right), \tag{15}$$

where $r_t = -\log(d_t + \epsilon)$ denotes the distance-derived attenuation reference and $\rho(\cdot, \cdot)$ is the Spearman rank correlation. Higher DIGA indicates stronger global alignment between audio intensity and the expected distance-dependent attenuation trend.

B Details of predictors

B.1 SPL predictor

Architecture To provide spatially grounded auxiliary supervision, we train a lightweight neural network to predict sound pressure level (SPL) curves from audio latent representations. The SPL predictor is trained on a large-scale paired dataset consisting of approximately 200,000 audio latent sequences and their corresponding SPL trajectories, enabling the model to robustly capture statistically consistent relationships between latent audio representations and sound intensity variations.

The predictor takes as input an audio latent sequence of shape (C, T), where C=64 denotes the latent channel dimension and T=216 is the temporal length. The network architecture is deliberately kept simple to emphasize temporal intensity modeling rather than detailed spectral reconstruction. Concretely, it comprises three one-dimensional convolutional layers operating along the temporal axis, each followed by a ReLU activation. The first layer preserves the channel dimensionality, while subsequent layers progressively reduce it to C/2, allowing for effective temporal feature aggregation with limited model capacity.

After temporal feature extraction, a 1×1 convolution is applied to project the intermediate representation to a single-channel temporal signal, yielding a frame-wise SPL estimate at the latent temporal resolution. To match the temporal resolution of the SPL annotations, which are defined over 50 frames, the predicted sequence is finally resized using linear interpolation along the temporal dimension.

By training on a large corpus of paired audio latents and SPL curves, the predictor learns to map high-level audio representations to smooth and spatially meaningful intensity trajectories. This design encourages the model to focus on global energy variations consistent with sound propagation, while avoiding sensitivity to high-frequency artifacts or timbral details. In practice, the resulting SPL predictor provides stable and effective auxiliary supervision for enforcing spatial–visual consistency during audio generation.

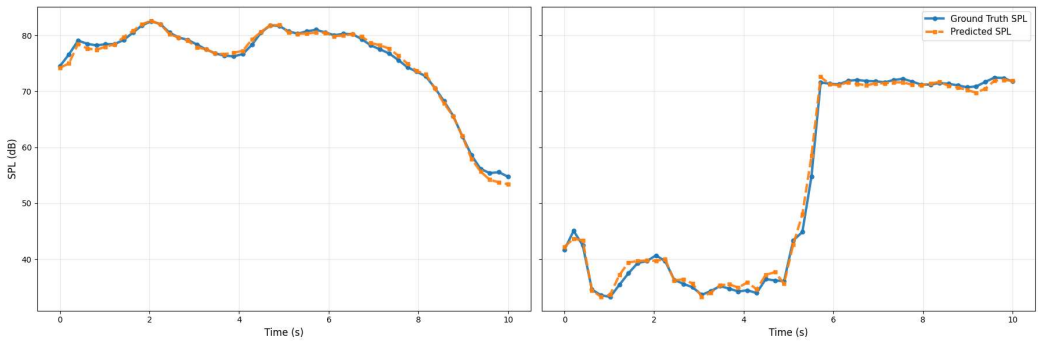


Figure 4: **SPL Prediction vs. Ground Truth.**Comparison between the ground-truth sound pressure level (SPL) and the predictions produced by the SPL predictor. The curves illustrate the temporal evolution of SPL over time, where the predicted SPL closely follows the ground-truth trajectory, demonstrating that the proposed SPL predictor can accurately predict SPL variations from the input audio features, making it suitable for subsequent training and supervision.

Effect As illustrated in Fig.4, we visualize the comparison between the predicted SPL curves and the ground-truth SPL annotations. The results show that the SPL predictor is able to accurately recover the temporal evolution of sound pressure level from the input audio features, providing reliable supervision for subsequent model training.

B.2 SPE PREDICTOR

Architecture We design a lightweight temporal prediction network to map audio latent representations to linear-band spectral energy trajectories, which are later used to construct the SPE loss. Given a latent audio feature sequence $\mathbf{Z} \in \mathbb{R}^{C\times L}$, where C denotes the latent channel dimension and L the temporal length, the predictor outputs a band-wise spectral energy representation $\hat{\mathbf{E}} \in \mathbb{R}^{B\times L}$, with B=32 linear frequency bands.

The network adopts a fully convolutional temporal architecture. Specifically, the input latent features are first projected into a hidden feature space using a 1×1 convolution. This is followed by a stack of temporal convolutional blocks with exponentially increasing dilation factors, enabling the model to capture temporal dependencies at multiple time scales without increasing computational cost. Each temporal block consists of a dilated 1D convolution, group normalization, and a SiLU activation, followed by a residual connection to facilitate stable optimization.

Finally, another 1×1 convolution projects the hidden features to the target band-energy dimension, producing the predicted linear-band energy sequence aligned with the latent temporal resolution.

To construct training targets, we compute short-time Fourier transforms from raw audio signals and aggregate spectral power into 32 linear frequency bands. The resulting band-energy sequences are temporally interpolated to match the latent sequence length and transformed into the log domain for numerical stability. A temporal mean normalization is further applied to remove global energy bias and emphasize relative spectral variations.

The predictor is trained using a Smooth L1 loss between the predicted and ground-truth band-energy representations, masked to account for variable-length sequences. This design ensures that the SPE predictor focuses on modeling the temporal evolution of spectral energy rather than absolute magnitude, making it well suited as an auxiliary supervision signal in our framework.

Effect. As illustrated in Fig.5, the proposed SPE predictor is able to accurately recover linear-band spectral energy distributions from audio latent features. The predicted spectral energy closely matches the ground-truth signals extracted from raw audio, indicating that the predictor effectively captures temporally aligned spectral characteristics required for subsequent training.

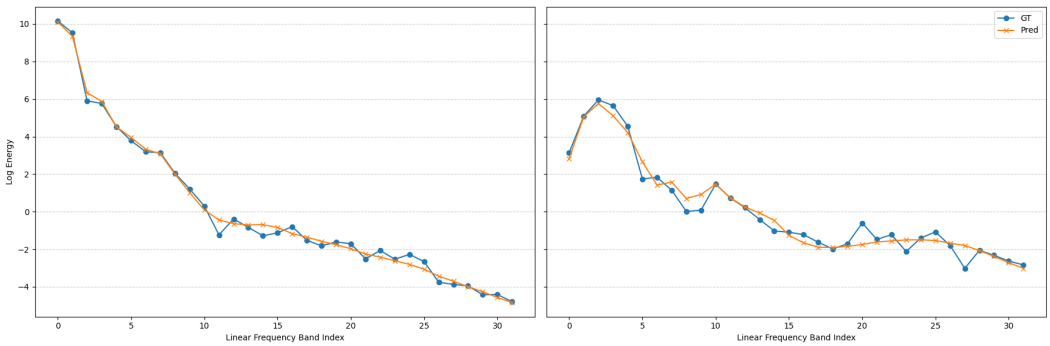


Figure 5: **SPE Prediction vs. Ground Truth.**Comparison between ground-truth (GT) and predicted linear-band spectral energy at different time steps. The SPE predictor infers spectral energy distributions from audio latent features that closely match the ground truth computed from raw audio, demonstrating its ability to accurately model temporally aligned spectral energy variations for subsequent training.

C Details of controllability

To further explore the controllability of our framework, we conduct extensive experiments on distance trajectory editing. We apply three types of perturbations to the original distance curves: **Local Modification**, **Swap**, and **Reverse**.

Table 4: **Evaluation of Controllability across various distance trajectories.** Our method demonstrates robust spatial-perceptual alignment even under extreme distance manipulations. Notably, as the counterfactual distance diverges from the visual source, a marginal reduction in audio fidelity is observed, indicating that extreme spatial forcing can perturb the intrinsic acoustic textures of the generated Foley.

Control Type	DITC↑	DIGA↑	FAD↓
Reverse	0.2624	0.3751	9.85
Swap	0.2801	0.4103	9.88
Local Modification	0.2798	0.4083	9.71
Original	**0.4879**	**0.7698**	**7.63**

Local Modification. For local modification, we alter a contiguous temporal segment of the original distance curve while preserving the global trajectory. Specifically, for each curve of length 50, a random subsequence covering 20%–40% of the timeline is selected, and its values are locally adjusted in a smooth manner. This operation modifies short-term distance dynamics without changing the overall temporal ordering, enabling controlled local deviations from the original metirc depth trajectory.

Swap. For the swap operation, distance curves are randomly exchanged across different videos while keeping the video identities unchanged. Concretely, we shuffle the complete 50-frame distance curves among videos, so that each video is paired with a distance trajectory originating from another sample. This setting intentionally introduces a global mismatch between visual semantics and spatial distance cues.

Reverse. For reverse editing, the temporal order of each distance curve is inverted. That is, the 50 distance values of each video are reversed along the time axis, producing a spatially implausible trajectory that contradicts the original motion direction while preserving the marginal distance distribution.

Table 5: **Robustness to different types of depth noise.** Resilient to scale bias and sparse errors, our model degrades mildly only under severe distance-dependent noise, thanks to the SPL loss's slack margin avoiding overfitting to precise depth values.

Noise type	Params	FAD↓	DITC↑	DIGA↑
Scale Bias	$\sigma = 0.05$	7.62	0.4884	0.7140
	$\sigma = 0.10$	7.57	0.4883	0.7146
	$\sigma = 0.20$	7.65	0.4888	0.7160
Distance-Dependent	SNR=30dB	7.43	0.4644	0.6926
	SNR=20dB	7.76	0.3988	0.6179
	SNR=10dB	8.22	0.3033	0.4928
Sparse Error	5%	7.56	0.4815	0.7080
	10%	7.44	0.4782	0.7002
	20%	7.61	0.4598	0.6832

As shown in Tab.4, compared to the baseline, our model consistently preserves high spatial consistency metrics across all edited scenarios, demonstrating strong robustness in precise acoustic modulation. Although a slight degradation in audio quality is observed under all three perturbation settings, this effect primarily stems from the inherent conflict between the original visual semantic context and the altered metric depth trajectories. Such conflicts are consistent with prior expectations, as enforcing substantial deviations from the ground-truth visual priors poses a fundamental challenge for cross-modal alignment. Nevertheless, the results collectively validate that our model offers reliable and fine-grained spatial editability, enabling high-fidelity Foley generation that strictly adheres to user-defined spatial constraints.

D Details of Robustness Analysis.

We evaluate robustness from two complementary perspectives: (1) errors in the estimated depth condition and (2) reverberation introduced by the acoustic environment. The former examines sensitivity to upstream visual perception, whereas the latter evaluates whether the generated depth–intensity relationship remains observable after room coloration.

D.1 Robustness to Depth Estimation Noise

As shown in Tab.5, we primarily tested the model's performance under three types of disturbances:

1. **Scale Bias Noise:** Simulates global scaling errors by applying a uniform multiplicative perturbation to the entire depth sequence. Three levels are used: weak (σ = 0.05), medium (σ = 0.10), and strong (σ = 0.20).
2. **Distance-Dependent Noise:** This noise models the physical characteristic that depth estimation becomes less reliable at greater distances. Three levels are defined using SNR: 30 dB (weak), 20 dB (medium), and 10 dB (strong).
3. **Sparse Frame Error:** This noise simulates transient depth estimation failures caused by factors such as reflections or occlusions. The corruption rates are 5% (weak), 10% (medium), and 20% (strong), without altering overall sequence structure.

Experimental results indicate that Scale Bias Noise and Sparse Frame Error have negligible impact on the evaluation metrics, while Distance-Dependent Noise introduces mild degradation in audio quality and spatial coherence under strong perturbations.

Overall, our pipeline remains robust to corrupted depth inputs. This robustness is largely attributed to the slack margin design in the SPL loss, which encourages the model to capture general distance trends rather than overfit to precise depth values. As a result, minor fluctuations can be effectively absorbed without causing abrupt or unnatural volume variations.

Table 6: Robustness to simulated reverberation. Levels 1–3 correspond to RT60 values of 0.2, 0.6, and 1.2 s, respectively.

Method	DITC↑			DIGA↑		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
CAFA	0.116	0.110	0.104	0.106	0.105	0.104
DepthFoley	**0.459**	**0.438**	**0.415**	**0.672**	**0.656**	**0.639**

Table 7: **Training resource comparison.** Comparison of model scale, training data, and computational resources, demonstrating that strong spatially consistent audio generation can be achieved with relatively limited training resources.

Model	Params	Train Data Scale	Batch Size	GPU hours	Hardware
CAFA	2.16B	200k	16	48	A100
MMAudio	1.03B	180k+951k	—	304	H100
see&hear	415M	—	—	—	RTX3090
FC	1.22B	200k+1.5k+103k	128	—	A100
DepthFoley(Ours)	1.11B	9k	4	48	RTX4090

D.2 Robustness to Simulated Reverberation

We further examine whether the learned depth-intensity relationship remains observable under room reverberation. We apply the same simulated room impulse responses to the outputs of CAFA and DepthFoley, considering RT60 values of 0.2, 0.6, and 1.2 s. This experiment evaluates robustness to post-generation room coloration rather than explicit control of room-acoustic properties.

As shown in Tab 6, DepthFoley exhibits a gradual decline as reverberation increases but remains substantially above CAFA at every RT60 level. These results indicate that its learned depth–intensity trend remains robust under simulated room effects. However, this stress test does not imply explicit estimation or control of room-acoustic properties such as DRR or RT60, which remains a direction for future work.

E Details of Training Resources

As shown in Tab.7, compared to existing baselines, our model DepthFoley is notably lightweight in terms of training data, requiring only 9k training samples. Our model is trained for 150 epochs on a single NVIDIA GeForce RTX 4090 GPU with a batch size of 4, requiring approximately 48 GPU hours, utilizing the CAFA optimizer configuration and gradient clipping (max norm 1.0). Our training objective incorporates the proposed spatial constraints with weights $\lambda_{\text{SPL}} = 0.025$ and $\lambda_{\text{SPE}} = 0.5$. To ensure robust supervision, the SPL and SPE predictors are pre-trained on 200K latent-audio pairs from VGGSound, establishing a stable mapping from latent manifolds to distance-grounded energy representations.

These results highlight the data efficiency of our approach, demonstrating that strong spatially consistent audio generation can be achieved with relatively limited training resources.

F Effect of Distance Guidance Weight w_d

We provide a quantitative trade-off analysis illustrating the impact of the distance guidance weight w_d on both Audio Fidelity (FAD) and Spatial Consistency (DITC/DIGA). The results, summarized in Tab. 8.

As w_d increases from 0.1, both audio quality (FAD) and spatial consistency (DITC/DIGA) improve significantly, reaching an optimal balance at $w_d = 0.4$. Beyond this threshold, audio quality begins to degrade, while further increasing w_d fails to provide additional gains in spatial perception, with metrics either stagnating or fluctuating slightly. This suggests that while moderate spatial constraints enhance alignment, excessively high weights interfere with the model's semantic generation and

Table 8: **Effect of distance guidance weight w_d.** Performance improves with increasing w_d and peaks at 0.4, after which audio quality degrades and spatial gains saturate

w_d	FAD↓	DITC↑	DIGA↑
0.1	20.97	0.4311	0.6191
0.2	16.53	0.4763	0.6780
0.3	12.98	0.4659	0.6776
0.4	**7.63**	**0.4879**	**0.7698**
0.5	9.06	0.4749	0.7054
0.6	13.06	0.4851	0.7061

audio naturalness. This empirical evidence justifies our selection of $w_d = 0.4$ as a robust balance that preserves high-fidelity audio generation while ensuring physically grounded spatial realism. The guidance-weight analysis is conducted under the single-pass composite-trajectory protocol so that only w_d varies across settings.

G Details of Ground Truth

We further evaluate the ground truth on the test set, obtaining a DITC of 0.3750 and a DIGA of 0.5252, which are notably higher than those of all compared baselines.

The primary reason our model performs better on these metrics is that VGGSound is sourced from publicly available YouTube videos. In many outdoor scenes, the audio often contains a mixture of multiple sources, including voice-overs, background noise, and other environmental sounds, which can adversely affect these metrics. In contrast, the audio generated by our model tends to be more focused on the primary sound source with fewer interfering components, leading to improved performance under these evaluation criteria.

Applying AudioSepLiu et al. (2024c) to remove environmental noise from the GT audio increases its DITC to 0.4941 and DIGA to 0.6684. This verifies that distance-based physical patterns are indeed present in real-world audio but are often entangled with noise. Due to this entanglement, DITC and DIGA are designed to evaluate spatial-acoustic alignment against theoretical physical laws, rather than using raw GT audio as a reference.

H Details of Subjective Evaluation.

We conducted a structured user study with 30 volunteers (aged 20–30, with normal hearing) to evaluate our framework against five state-of-the-art baselines: MMAudio, FoleyCrafter, Seeing&Hearing, CAFA, and Post-hoc. The evaluation suite comprised 30 distinct video-audio groups, which were distributed using a cyclic allocation strategy where each participant assessed 15 groups; by organizing these into six unique combinations assigned to five participants each, we ensured that every sample received an equal number of evaluations. Prior to the formal session, participants underwent a calibration phase where they were presented with a ground-truth (GT) reference as Fig.6, serving as a dual 5-point anchor for both Audio Quality (AQ) and Spatial-Perceptual Alignment (SPA), to standardize the scoring criteria. For each assigned task like Fig.7, participants compared the six anonymized results presented in a randomized order to eliminate selection bias, with the flexibility to replay the videos as many times as necessary. Scores were collected on a 5-point Likert scale and subsequently aggregated for statistical analysis, including the calculation of Mean Opinion Scores (MOS) and 95% confidence intervals to verify the significance of our model's performance.

The subjective evaluation involved minimal risk. Prior to participation, all volunteers were informed about the study's purpose and provided explicit informed consent. All collected rating data was strictly anonymous, participation was voluntary, informed consent was obtained, and no personally identifiable information was collected.



Figure 6: A representative sample page of the Human Study (Ground Truth).

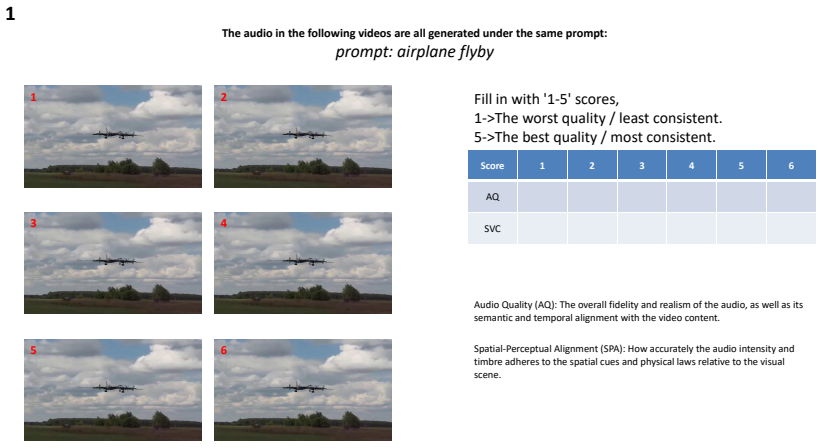


Figure 7: A sample evaluation page from the human study interface.

I Details of Multi-Object Inference Ablation

We evaluate different multi-object inference strategies using the original test set and the same trained DepthFoley checkpoint. The prompts, random seeds, and evaluation protocol are kept unchanged. Single-object videos are processed identically in all settings; only the inference strategy for videos containing multiple objects is modified. We compare the following three settings:

- **Composite trajectory.** Multiple object trajectories are consolidated into a single trajectory using the previous strategy: spatially close, co-moving objects are averaged, while the nearest object is tracked when the objects are separated. The resulting composite trajectory is used for a single inference pass.
- **First Obj trajectory.** Only the trajectory of the first detected object, is retained as the depth condition. A single inference pass is performed without explicitly modeling the remaining objects.
- **Object-wise multi-pass (Ours).** Each detected object is independently processed using its object-centric video and individual distance trajectory. The resulting audio stems are

Table 9: Ablation of multi-object inference strategies on the original test set. Single-object inference remains unchanged across all settings.

Multi-object Strategy	DITC↑	DIGA↑	FAD↓	KL↓
Composite trajectory	0.4879	**0.7698**	7.63	1.569
First Obj trajectory	0.4448	0.6876	9.46	1.698
Object-wise multi-pass (Ours)	**0.5529**	0.7685	**7.35**	**1.472**

summed in the waveform domain and globally rescaled only when necessary to prevent clipping.

Because the three settings differ only in their treatment of multi-object videos, their performance on single-object samples remains unchanged. We report all metrics over the complete original test set. DITC and DIGA follow the same evaluation protocol as Sec. 4.2, while FAD and KL are computed on the final output audio. All generated outputs are evaluated against the same test-set reference trajectories used in Sec. 4.2, ensuring direct comparability across inference strategies.

Object-wise multi-pass inference improves local depth-intensity consistency and audio-distribution metrics while maintaining essentially unchanged global alignment.