FoleyDesigner: Immersive Stereo Foley Generation with Precise Spatio-Temporal Alignment for Film Clips

Mengtian Li[1,2] Kunyan Dai[1] Yi Ding[1] Ruobing Ni[1] Ying Zhang[1]
Wenwu Wang[3†] Zhifeng Xie[1,2†]

[1]Shanghai Film Academy, Shanghai University
[2]Shanghai Engineering Research Center of Motion Picture Special Effects
[3]University of Surrey, UK

Abstract

Foley art plays a pivotal role in enhancing immersive auditory experiences in film, yet manual creation of spatio-temporally aligned audio remains labor-intensive. We propose ***FoleyDesigner****, a novel framework inspired by professional Foley workflows, integrating film clip analysis, spatio-temporally controllable Foley generation, and professional audio mixing capabilities. FoleyDesigner employs a multi-agent architecture for precise spatio-temporal analysis. It achieves spatio-temporal alignment through latent diffusion models trained on spatio-temporal cues extracted from video frames, combined with large language model (LLM)-driven hybrid mechanisms that emulate post-production practices in film industry. To address the lack of high-quality stereo audio datasets in film, we introduce* ***FilmStereo****, the first professional stereo audio dataset containing spatial metadata, precise timestamps, and semantic annotations for eight common Foley categories. For applications, the framework supports interactive user control while maintaining seamless integration with professional pipelines, including 5.1-channel Dolby Atmos systems compliant with ITU-R BS.775 standards, thereby offering extensive creative flexibility. Extensive experiments demonstrate that our method achieves superior spatio-temporal alignment compared to existing baselines, with seamless compatibility with professional film production standards. The project page is available at* https://gekiii996.github.io/FoleyDesigner/.

1. Introduction

Stereo Foley refers to the art of creating and recording sound effects with spatial information, where sounds are positioned across the left and right audio channels to create a sense of directionality and space. In film production, the stereo Foley serves critical narrative and immersive functions. Film Foley artists must synchronize sounds with on-screen actions with frame-level precision while tracking the spatial movement of visual elements. The spatial positioning of sounds is designed according to the script and directorial vision to guide the audience's attention, convey emotions, and enhance dramatic moments. This requires careful control over timing, spatial placement, and sonic qualities to maintain audience immersion and support storytelling. However, general stereo Foley generation methods cannot meet these film production requirements.

Despite the importance of stereo Foley in film production, current audio generation methods fall short of meeting professional requirements. Existing approaches can

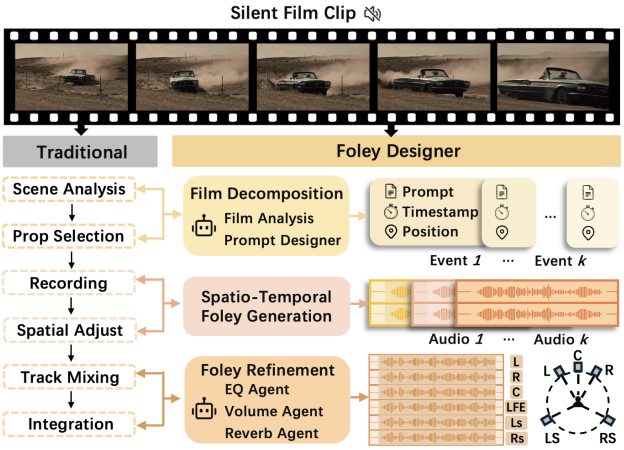

Figure 1. **FoleyDesigner Overview.** The left column detailing the actual steps of a human Foley designer. The right column presents the corresponding simulated functional modules of FoleyDesigner, showcasing outputs at each phase, resulting in a soundtrack suitable for film use.

†Corresponding authors.

be categorized into three groups, each with critical limitations. First, monaural generation methods such as AudioLDM2 [23], Tango 2 [27], and Make-an-Audio 2 [16] produce high-quality sound effects from text prompts using latent diffusion models, but they lack spatial dimensions entirely, rendering them unsuitable for stereo Foley. Second, stereo generation methods like Stable Audio [11], which performs diffusion on waveform representations, can produce stereo audio but lacks precise spatial localization control. Similarly, SpatialSonic [32] and See2Sound [7] generate stereo audio from images or text but cannot achieve frame-level temporal alignment with visual content. Third, monaural-to-stereo conversion methods such as Sep-Stereo [42] and Mono to Binaural [29] depend on pre-existing monaural sources, limiting their flexibility and requiring additional production steps.

Generating film-quality Foley audio requires addressing three technical challenges: (1) *Densely overlapping sound events.* Film scenes contain multiple simultaneous spectrally-temporally overlapping sound sources. Single-pass generative models cannot disentangle these complex acoustic scenes, resulting in incomplete or muddled audio outputs that fail to capture the layered nature of professional Foley. (2) *Precise spatio-temporal grounding.* Current conditioning mechanisms lack explicit grounding in visual spatial cues and temporal dynamics. Text-based approaches provide only coarse directional descriptions that cannot specify continuous spatial trajectories or frame-accurate timing. Image-based methods lack temporal information and cannot capture the dynamic movement of sound sources. (3) *Professional acoustic quality.* Generated audio exhibits acoustic inconsistencies that fail to meet film production standards. Mismatched reverberation between sound events, spectral masking in overlapping frequencies, and loudness imbalances cause important information to be buried in the mixed sounds, degrading cinematic immersion. These limitations prevent current audio generation systems from being used in professional film Foley workflows.

To address these challenges, we present **FoleyDesigner**, a novel framework that integrates professional Foley production pipelines to generate film-quality stereo audio from silent film clips. FoleyDesigner begins with fine-grained Foley decomposition, where Tree-of-Thought reasoning with multi-agent verification analyzes visual content and script context to produce Foley scripts that decompose complex scenes into layered sound events. These decomposed events are then processed through spatio-temporal Foley generation, where we introduce a novel spatio-temporal injection mechanism that conditions a Diffusion Transformer on sound event trajectories extracted from visual tracking, achieving frame-accurate spatio-temporal alignment with visual motion. Finally, Foley refinement and professional mixing employs a multi-agent framework where specialized diagnostic agents identify acoustic inconsistencies through complementary analysis tools and determine mixing parameters, applying professional audio engineering knowledge to ensure reverberation coherence, spectral clarity, and balanced dynamics before upmixing to 5.1 channel surround formats for film applications, as shown in Figure 1.

To support FoleyDesigner, we construct **FilmStereo**, the first spatial audio dataset specifically designed for film Foley generation. Unlike existing audio-visual datasets that lack spatial metadata or contain only coarse temporal alignment, FilmStereo provides stereo recordings with precise temporal annotations and spatial positioning information across eight common Foley categories. FilmStereo enables data-driven training of spatially grounded audio generation models and establishes a benchmark for evaluating film-quality Foley synthesis.

Our contributions are summarized as follows:

- We present **FoleyDesigner**, the first framework that integrates professional Foley production workflows through decomposition, generation, and refinement stages, and introduce **FilmStereo**, a large-scale stereo audio dataset with precise temporal and spatial annotations across eight common Foley categories.
- We propose a spatio-temporal injection mechanism that conditions diffusion transformers on visual tracking trajectories for precise alignment, and introduce a multi-agent framework with Tree-of-Thought reasoning for Foley script validation and automated audio production, promoting automated film Foley production.
- We demonstrate practical application into current film production workflows, where FoleyDesigner generates high-quality multi-channel (e.g. 5.1) surround soundtracks with professional film standards, improving production efficiency while maintaining professional quality.

2. Related work

2.1. Monaural Audio Generation

Monaural audio generation has seen significant progress across **text-to-audio (T2A)**, **image-to-audio (I2A)**, and **video-to-audio (V2A)** tasks. In T2A, autoregressive models like AudioGen [20] treat audio synthesis as conditional language modeling, while diffusion-based models such as AudioLDM [22], Tango2 [27], and Make-an-Audio2 [16] leverage latent diffusion to produce audio from text prompts. For I2A, methods like ImageHear [31], CLIP-Sonic [9], and V2A-Mapper [34] utilize CLIP features to generate audio from static images. V2A approaches, including Frieren [35] and MMAudio [4], enhance semantic and temporal coherence with video inputs. Recent foley generation work includes Spotlighting [15], MultiFoley [3], VideoFoley [21], CondFoleyGen [10], FoleyCrafter [41],

and DiffFoley [26]. However, these methods face limitations for film production. T2A and I2A approaches lack precise audio-visual synchronization required for frame-accurate foley. V2A methods remain limited to monaural output and struggle with spatial positioning. Foley generation systems produce single-channel audio often misaligned with spatial information in the visual scene.

Our method addresses these limitations by integrating spatial cues directly into the generation process, enabling stereo foley synthesis with precise audio-visual synchronization tailored for film production workflows.

2.2. Spatial Audio Generation

Spatial audio generation has evolved from early stereo techniques [28] to modern deep learning approaches. **Stereo audio localization** research, such as Yang and Zheng [39] and Cao et al. [2], focuses on analyzing interaural time and level differences for accurate sound positioning in single- and multi-source scenarios, but primarily addresses analysis and separation rather than controllable generation. **Mono-to-stereo conversion** methods conditioned on visual inputs [12, 25] and depth cues [29] convert existing mono audio to stereo representations, but rely on pre-existing audio sources, fundamentally limiting creative flexibility. Generative models like MusicGen [6] and Stable Audio [11] generate stereo audio natively but without explicit spatial control mechanisms. Recent advances including SpatialSonic [32], See2Sound [7], and OmniAudio [24] improve spatial precision through text, image, or video conditioning. However, these methods neglect frame-level temporal alignment essential for synchronizing audio with visual events in film applications, and their architectures lack integration pathways required for professional post-production workflows. In contrast, FoleyDesigner provides an end-to-end pipeline that unifies stereo generation, explicit spatial control, and 5.1-channel simulation, designed specifically to meet the demands of film post-production workflows.

3. Method

Unlike prior work that generates audio holistically without explicit scene decomposition or spatial control, we address three critical challenges: (1) densely overlapping sound events, (2) lack of audio-visual grounding control, and (3) acoustic inconsistencies in raw audio. FoleyDesigner integrates professional Foley workflows through three sequential stages, as illustrated in Figure 2. We first employ Tree-of-Thought reasoning to decompose silent film clips into verified Foley scripts specifying multiple temporally-layered sound events, enabling generation of densely overlapping soundscapes. Each event is then synthesized as stereo audio through a Diffusion Transformer conditioned on textual descriptions and spatial cues extracted from visual tracking, providing explicit grounding in visual dynamics. Finally, professional audio processing including reverberation, equalization, and dynamics control refines acoustic quality and upmixes stereo tracks to multi-channel surround formats suitable for cinematic production.

3.1. Fine-Grained Film Decomposition for Foleys

Challenges. (1) Single-pass generative models cannot generate sound events simultaneously in densely overlapping scenarios, resulting in incomplete audio outputs. (2) Effective Foley design requires capturing physically observable events and inferring narrative elements that depend on script context.

To address these challenges, we decompose Foley generation into two agent-orchestrated modules that construct a verified Foley script through progressive refinement.

FilmScribe converts silent video $\mathcal{V}$ into structured Foley script $\mathcal{T}$ containing visual descriptions and sound event specifications. A generator agent produces initial script $\mathcal{T}^{(0)}$ from $\mathcal{V}$, while a validator agent verifies accuracy and completeness. The system iteratively refines through:

$$\mathcal{T}^{(k+1)} = \text{Generator}(\mathcal{V}, \text{Feedback}(\mathcal{T}^{(k)}, \mathcal{V})), \quad (1)$$

where k denotes the iteration step, until $\text{Validator}(\mathcal{T}, \mathcal{V}) \rightarrow \text{True}$. This closed-loop verification ensures visual-audio alignment and sound event completeness.

FoleyScriptWriter integrates film script $\mathcal{F}$ and structured text $\mathcal{T}$ to produce hierarchical Foley script $\mathcal{S} = \{(e_i, l_i)\}$, where e_i denotes the i-th sound event and $l_i \in \{\text{fg}, \text{bg}\}$ specifies its layer assignment. By decomposing densely overlapping events individually, each sound event can be generated sequentially, avoiding incomplete or muddled outputs. While visual analysis captures physically observable events, film script $\mathcal{F}$ enables inference of narrative elements not directly observable from visual alone. We employ Tree-of-Thought (ToT) [40] reasoning over directed search graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ denotes candidate script nodes and $\mathcal{E}$ represents refinement edges (Algorithm 1).

(1) *Expand.* Starting from root node encoding $(\mathcal{V}, \mathcal{F})$, an expansion agent generates child nodes representing candidate scripts grounded in Foley design principles: diegetic source separation, semantic alignment with visuals, and emotional modulation matching film tone.

(2) *Score.* Each candidate receives the score from $\text{Score}(\mathcal{S}, \mathcal{V}, \mathcal{F})$ which evaluates visual-audio correspondence, foreground-background separation, and film tone consistency. Misalignments, ambiguous layering, or emotional conflicts reduce respective scores.

(3) *Optimization.* Search terminates immediately when $\text{Score}(\mathcal{S}, \mathcal{V}, \mathcal{F}) > \tau$. Otherwise, the refinement is performed as follows:

- *Refinement*: For correctable issues, such as overlapping

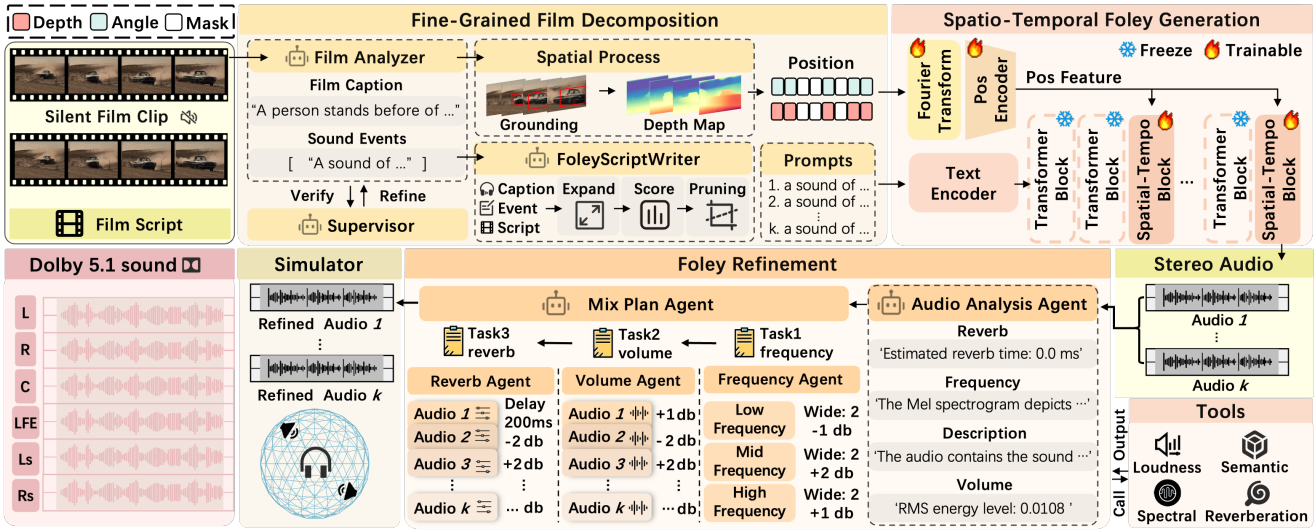

Figure 2. **FoleyDesigner Architecture.** Our pipeline for automated Foley generation consists of three stages, (1) *Fine-Grained Film Decomposition*: analyzes silent video and generates hierarchical Foley scripts; (2) *Spatio-Temporal Foley Generation*: produces spatially-controlled stereo audio using DiT-based diffusion conditioned on visual cues; (3) *Foley Refinement*: applies multi-agent processing to refine audio quality and generate 5.1 surround output.

Algorithm 1 Tree-of-Thought Foley Script Generation

Require: Video $\mathcal{V}$, structured text $\mathcal{T}$, film script $\mathcal{F}$, branching factor k, beam size b, depth limit $d_{\max}$, threshold τ
Ensure: Optimal Foley script $\mathcal{S}^*$

```
Initialize search graph G = (N, E) with root n_0 = (V, T, F)
Initialize candidate set C ← {n_0}, depth d ← 0
while d < d_max and max_{n∈C} Score(S_n, V, F) < τ do
    Expand: For each n ∈ C, generate k children via refinement or regeneration
    Score: Evaluate Score(S_n, V, F) = w1 s_align + w2 s_layer + w3 s_emotion
    Prune: Retain top-b nodes by score → C
    d ← d + 1
end while
return S* = arg max_{n∈C} Score(S_n, V, F)
```

layers or ambiguous event descriptions, we spawn sub-nodes by applying targeted adjustments.

- *Regeneration*: For fundamental failures, such as emotional misalignment or structural incoherence, we create sibling nodes with revised constraints.

Pruning retains top-k candidates per level. Termination occurs when $\text{Score}(\mathcal{S}, \mathcal{V}, \mathcal{F}) > \tau$, depth exceeds $d_{\max}$, or the branch budget is exhausted. This strategy delivers a Foley script ensuring physical fidelity, perceptual clarity, and narrative coherence for downstream diffusion synthesis.

3.2. Spatio-Temporal Foley Generation

Motivation. Text-conditioned models face limitations for Foley generation: (1) they cannot specify precise spatial cue for sound trajectories, (2) lack grounding in video frame geometry, and (3) cannot ensure frame-accurate temporal alignment. We address this by extracting spatio-temporal cues from video frames and conditioning diffusion generation through spatial and temporal controls.

3.2.1. Spatio-Temporal Cue Extraction

To enable spatial positioning of sound events, we extract depth and azimuth information from N keyframes $\mathcal{K} = \{I_1, I_2, \ldots, I_N\}$ sampled from the input video $\mathcal{V}$. We employ a vision language model (VLM) [1] to localize sound sources by annotating bounding boxes $\mathcal{B} = \{b_1, b_2, \ldots, b_N\}$, where each b_i corresponds to the spatial extent of a sound event in keyframe I_i.

For depth estimation, we leverage [38] to generate depth maps $\mathbf{D}_i \in \mathbb{R}^{H \times W}$ for each keyframe, where H and W denote the frame height and width. For each bounding box b_i, we compute the average depth within the box to obtain a scalar depth value $d_i \in \mathbb{R}$. The azimuth angle θ_i is derived from the horizontal center $x_i \in [0, W]$ of the bounding box:

$$\theta_i = \arctan\left(\frac{x_i - W/2}{d_i}\right) \cdot \frac{180^\circ}{\pi} + 90^\circ, \qquad (2)$$

This formulation maps relative horizontal positions into azimuth angles, with $\theta_i \in [0^\circ, 180^\circ]$ representing the sound source's angular position.

To ensure temporal synchronization, we detect sound event timestamps [37] to generate a binary activation vector $\mathbf{c} = \{c_1, c_2, \ldots, c_T\} \in \{0, 1\}^T$, where T is the total number of video frames and $c_t \in \{0, 1\}$ indicates whether a sound event occurs at frame t. The spatial position sequence $\mathcal{X} = \{\mathbf{x}_i = (d_i, \theta_i)\}_{i=1}^N$ extracted from keyframes is temporally interpolated to match the video frame rate, yielding

$\{\mathbf{x}_t\}_{t=1}^{T}$, and masked by the activation vector:

$$\mathbf{p}_t = c_t \cdot \mathbf{x}_t, \quad \mathcal{P} = \{\mathbf{p}_t\}_{t=1}^{T}, \tag{3}$$

where $\mathbf{p}_t \in \mathbb{R}^2$ encodes the depth and azimuth of the active sound event at frame t.

3.2.2. DiT-Based Conditional Generation

We build upon Stable Audio Open [11], a DiT-based latent diffusion model, conditioning it on text prompt $\mathbf{c}_{\text{text}}$ and spatio-temporal cues $\mathcal{P}$. We introduce a **position-aware injection mechanism** via cross-attention to achieve precise spatio-temporal alignment.

Positional Feature Encoding. To enhance the expressiveness of raw spatio-temporal cues, we apply Fourier feature transformation [33] to each position vector:

$$\gamma(\mathbf{p}_t) = [\cos(2\pi \mathbf{B}\mathbf{p}_t); \sin(2\pi \mathbf{B}\mathbf{p}_t)] \in \mathbb{R}^{2m}, \tag{4}$$

where $\mathbf{B} \in \mathbb{R}^{m\times 2}$ is a random projection matrix sampled from $\mathcal{N}(0, \sigma^2)$, m is the number of frequency bands, and $[\cdot;\cdot]$ denotes concatenation.

To incorporate temporal activation information, we modulate the Fourier features with the binary mask c_t:

$$\tilde{\gamma}(\mathbf{p}_t) = c_t \cdot \gamma(\mathbf{p}_t) + \epsilon \cdot \gamma(\mathbf{p}_t), \tag{5}$$

where $\epsilon = 0.1$ is a small constant that preserves weak positional information during inactive frames.

We design a convolutional encoder to process the modulated features, matching the temporal compression ratio r of the audio latent space to ensure positional embeddings maintain temporal alignment with audio latents. Each downsampling block consists of 1D convolution, group normalization, and SiLU activation. The final positional embeddings are obtained as:

$$\mathbf{E}_{\text{pos}} = \text{PosEncoder}(\{\tilde{\gamma}(\mathbf{p}_t)\}_{t=1}^{T}) \in \mathbb{R}^{T'\times d_{\text{emb}}}, \tag{6}$$

where $T' = T/r$ is the compressed temporal dimension and d_{emb} is the positional embedding dimension.

Injection Blocks. Our diffusion backbone processes noisy latent representations $\mathbf{z}_\ell \in \mathbb{R}^{T'\times d_{\text{latent}}}$ through transformer blocks, where d_{latent} denotes the latent feature dimension of the diffusion model, distinct from d_{emb}. To inject spatio-temporal control, we insert an **injection block** after every four standard DiT blocks at layers $\ell \in \{3, 7, 11, 15, 19, 23\}$. Each injection block performs cross-attention between latent features and layer-normalized positional embeddings:

$$\mathbf{z}'_\ell = \text{InjBlock}(\mathbf{z}_\ell, \text{LN}(\mathbf{E}_{\text{pos}})), \tag{7}$$

where $\text{InjBlock}(\cdot)$ is a transformer block with cross-attention. This design allows the model to integrate spatial awareness across different depths of the network.

3.3. Foley Refinement and Professional Mixing

Challenges. (1) *Acoustic inconsistency*: mismatched reverberation or frequency characteristics break film immersion; (2) *Spectral masking*: overlapping frequencies muddy sonic clarity; (3) *Loudness imbalance*: improper volume levels cause important Foley to be buried.

To address these challenges, we introduce a multi-agent post-processing framework that emulates the collaborative framework of professional Foley teams.

Foley Analysis. To diagnose acoustic issues beyond semantics, the agent combines perceptual and quantitative tools. For each generated Foley track a_i, the agent extracts composite feature representation $\mathbf{f}_i = [\mathbf{f}_{sem}, \mathbf{f}_{spec}, f_{rev}, f_{loud}]$, where $\mathbf{f}_{sem}$ denotes semantic embeddings from an Audio-LLM [5], $\mathbf{f}_{spec}$ represents spectral patterns from Mel-spectrograms analyzed through a VLM [1], f_{rev} is the computed reverberation time, and f_{loud} is the measured integrated loudness. This combines semantics with objective measurements for comprehensive acoustic diagnosis.

Mixing Planner. The planner agent performs track-wise diagnosis through cross-modal validation, inter-track balance analysis, and quality assessment. It produces a mixing plan $\mathbf{\Pi} = \{(i, \mathcal{O}_i)\}_{i=1}^{N_{\text{track}}}$, where $\mathcal{O}_i \subseteq \{\texttt{reverb}, \texttt{eq}, \texttt{dyn}\}$ specifies required operations for track i.

Specialist Execution. The plan is dispatched to three specialist agents that operate collaboratively to determine concrete processing parameters. The *Reverberation Specialist* analyzes spatial relationships and determines reverberation parameters matched to scene properties. The *Equalization Specialist* examines spectral overlap to determine frequency band adjustments, minimizing spectral masking and maintaining sonic clarity. The *Dynamics Specialist* evaluates relative loudness levels and determines gain adjustments to prevent important Foley elements from being buried. Together, these specialists ensure that individual adjustments contribute to overall acoustic coherence.

Multi-Channel Surround Upmixing. To meet film Foley standards, we extend stereo output to 5.1 surround format $\{\text{FL}, \text{FR}, \text{C}, \text{SL}, \text{SR}, \text{LFE}\}$ following ITU-R BS.775 configuration [18]. We adopt a channel-wise upmixing strategy that preserves spatial dynamics without simulating room acoustics. Stereo channels $\mathbf{s}_L$ and $\mathbf{s}_R$ are directly mapped to front left (FL) and front right (FR) channels. Center (C), surround left (SL), and surround right (SR) channels are derived from weighted mixes of the stereo signal, with distinct coefficients applied to simulate spatial positions while maintaining energy balance. The low-frequency effects (LFE) channel is produced by low-pass filtering the full mix below 120 Hz:

$$\mathbf{s}_{\text{LFE}}(t) = \text{LPF}(\mathbf{s}_{\text{mix}}(t),\ 120\,\text{Hz}), \tag{8}$$

where $\mathbf{s}_{\text{mix}}(t) = \mathbf{s}_L(t) + \mathbf{s}_R(t)$. This produces 5.1-channel

output that maintains spatial and temporal accuracy of the stereo source while meeting professional film standards.

4. Dataset: FilmStereo

To enable controllable spatial audio generation for film production, we construct **FilmStereo**, a stereo audio dataset integrating spatial captions, temporal annotations, and stereo audio. Existing datasets lack this combination, providing temporal annotations without spatial information or spatial audio without temporal alignment. Figure 3 illustrates our construction pipeline.

Collection. We collect audio from public repositories across 8 categories (23 subcategories). The preprocessing includes: (1) filtering multi-event samples, (2) spectral denoising (-40 dB), (3) loop-padding to 8–10s, and (4) CLAP-based verification ($\tau = 0.35$). The resulting FilmStereo dataset contains 166 hours across 14,784 samples.

Spatial Simulation. We model azimuth using five frontal regions ($\pm 15°$, $\pm 45°$, $0°$) aligned with human localization acuity and depth using three zones (near-field: 0–2m, mid-field: 2–5m, far-field: >5m) based on psychoacoustic principles. Using gpuRIR [8] with 16–18 cm interaural distance, we generate room impulse responses for static and dynamic sources, with balanced distributions across object sizes, motion types, and spatial positions. Environmental reverberation is applied using presets.

Annotation. Spatial captions are generated by GPT-4 with chain-of-thought prompting, integrating sound descriptions with spatial parameters (azimuth, depth, reverberation). For temporal alignment, we detect amplitude peaks in denoised audio to identify event onsets, applying adaptive thresholds based on signal-to-noise ratio. Interval thresholds determine whether segments qualify as distinct sound events, generating start and end timestamps to ensure frame-accurate synchronization with visual content during film post-production workflows.

5. Experiments

Configuration. The complete training process of the framework is divided into two stages: (1) stereo mel-spectrogram VAE training, and (2) DiT-based diffusion model training with spatio-temporal control injection. All stages use FilmStereo datasets with learning rate 3×10^{-5}, batch size 8, on NVIDIA A6000 GPUs.

Metrics. We evaluate our framework in the following three aspects: (1) *Audio Quality*: Inception Score (IS) [30], KL Divergence [14], Fréchet Audio Distance (FAD) [19], and CLAP score [36] assess the perceptual quality and semantic alignment of generated audio; (2) *Spatio-Temporal Alignment*: GCC-MAE and CRW-MAE [32] measure spatial localization accuracy, Fréchet Stereo Audio Distance (FSAD) [32] evaluates stereo quality, and Intersection over Union (IoU) quantifies temporal precision; (3) *Cinematic Foley Quality*: ImageBind Score [13] measures audio-visual semantic coherence. AV-Sync [17] evaluates synchronization accuracy. For professional film foley evaluation, we introduce **Sonic Richness Score (SRS)** and **Cinematic Clarity Score (CCS)**. We use audio-capable MLLMs [5] to evaluate sonic layering diversity and perceptual separation quality in professional film contexts.

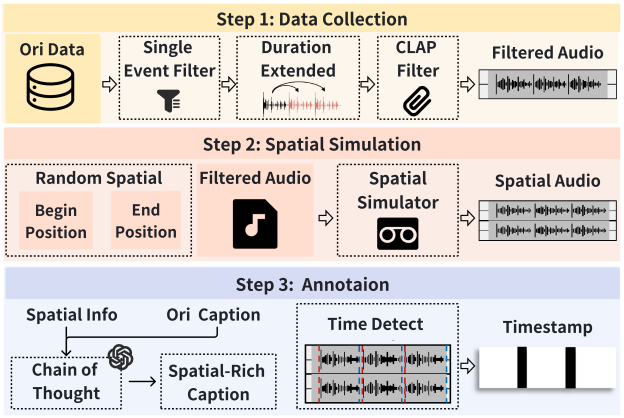

Figure 3. **FilmStereo Construction.** Our four-step pipeline for creating spatially and temporally annotated audio data. (1) Audio filtering and extension, (2) Spatial simulation with random positioning, (3) Spatial-rich caption generation via chain-of-thought, and (4) Temporal annotation with event detection.

5.1. Quantitative Results

Audio Quality. To evaluate the generated audio quality, we convert stereo audio to monaural signals by averaging the two channels, then assess the quality using standard metrics. Our method achieves the highest CLAP score of **0.679** and lowest FAD of **1.88**, outperforming SpatialSonic by **1.0%** and **2.6%**, and Stable Audio [11] by **14.3%** and **20.7%**. This demonstrates semantic alignment and perceptual quality for Foley sound design. While our IS score is lower than SpatialSonic, IS measures sample diversity rather than quality or semantic accuracy. Our focus on spatial coherence and text-audio alignment produces contextually appropriate outputs. Results are summarized in Table 1.

Table 1. Audio Quality. Metrics include Inception Score, KL Divergence, Fréchet Audio Distance, and CLAP score for audio quality assessment. **Best** and second-best results are highlighted. ↓ indicates lower is better, ↑ indicates higher is better.

Method	IS ↑	KL ↓	FAD ↓	CLAP ↑
Stable Audio [11]	10.50	1.86	2.37	0.594
SpatialSonic [32]	**13.79**	1.37	1.93	0.672
Ours	12.36	**1.40**	**1.88**	**0.679**

Spatio-Temporal Alignment. To evaluate spatial accuracy and temporal alignment, we assess stereo audio out-

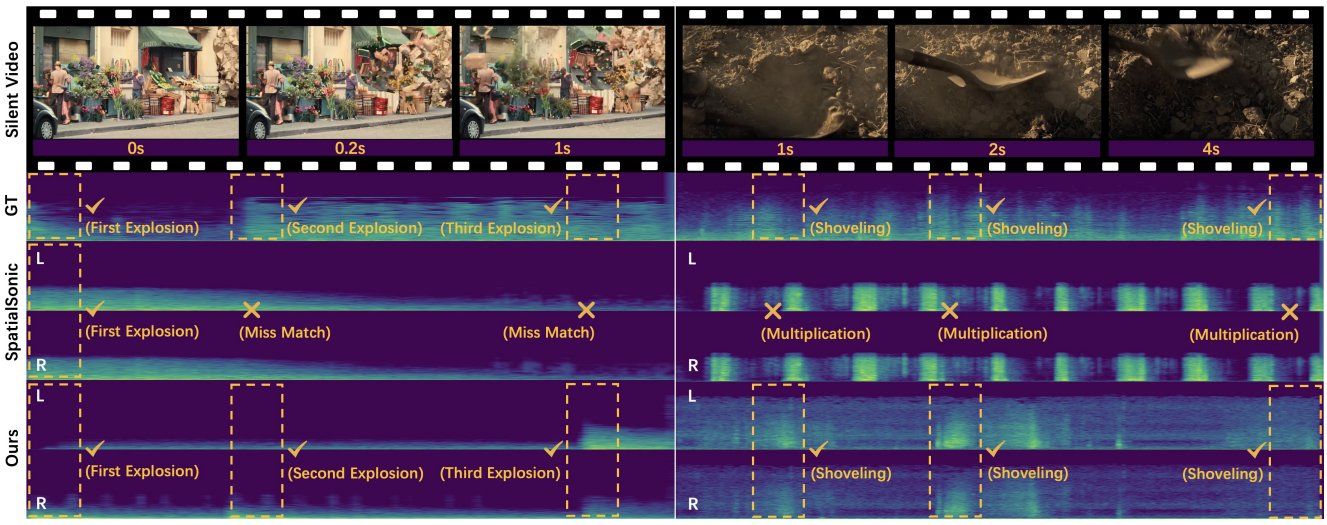

Figure 4. **Qualitative Results.** Qualitative comparison showing temporal alignment between video events and generated audio across two scenarios. Left: explosion sequence with three distinct events. Right: excavation scene with repetitive shoveling actions. Checkmarks indicate successful synchronization, while crosses mark temporal misalignment.

Table 2. Spatio-Temporal Alignment Results. Metrics include GCC and CRW for spatial accuracy, FSAD for stereo quality, and IoU for temporal alignment. **Best** and second-best results are highlighted. ↓ indicates lower is better, ↑ indicates higher is better.

Method	GCC ↓	CRW ↓	FSAD ↓	IoU ↑
Stable Audio [11]	61.17	51.44	0.343	24.5
See2Sound [7]	60.03	51.17	0.291	21.3
SpatialSonic [32]	49.20	36.87	0.163	27.8
Ours	**48.79**	**34.23**	**0.138**	**32.2**

put using metrics in Table 2. FoleyDesigner achieves the best performance across metrics, demonstrating the effectiveness of our position-aware injection mechanism. For spatial accuracy, our method achieves the lowest GCC MAE (**48.79**) and CRW MAE (**34.23**), outperforming SpatialSonic by **0.8%** and **7.2%**. The FSAD score (**0.138**) confirms high-quality stereo imaging with channel separation. For temporal alignment, FoleyDesigner achieves the highest IoU score (**32.2**), improving by **15.8%** over SpatialSonic. These results validate that our position-aware injection mechanism is effective in incorporating spatial positioning and temporal dynamics for Foley sound synthesis.

Table 3. Film Foley Performance. Evaluation on film clips covering audio-visual synchronization (ImageBind Score, AV-Sync) and cinematic quality (Sonic Richness Score, Cinematic Clarity Score). **Best** and second-best results are highlighted.

Method	IB ↑	SRS ↑	CCS ↑	AV-Sync ↑
Stable Audio [11]	0.216	5.31	5.8	0.512
See2Sound [7]	0.105	3.03	3.0	0.601
SpatialSonic [32]	0.251	5.91	4.5	0.545
Ours	**0.402**	**8.27**	**6.2**	**0.726**

Film Foley Performance. To evaluate the effectiveness of our method in real film scenarios, we assess the generated audio across multiple cinematic dimensions using film clips, as shown in Table 3. FoleyDesigner achieves the best performance across all metrics. Our method obtains the highest ImageBind Score (**0.402**) and AV-Sync score (**0.726**), outperforming SpatialSonic by 60.2% and 33.2% respectively, attributed to the position-aware injection mechanism that enables explicit grounding in visual spatial cues and temporal dynamics. The superior SRS (**8.27**) and CCS (**6.2**) scores, improving by 39.9% and 37.8%, validate our film-oriented design. The sonic richness stems from fine-grained film decomposition, while the cinematic clarity results from multi-agent refinement that optimizes acoustic balance across tracks through reverberation, equalization, and dynamics processing.

5.2. Qualitative Results

We present qualitative comparisons on two representative film scenarios, demonstrating **superior temporal synchronization** and **semantic alignment** compared to existing approaches. As shown in Figure 4, the visualization displays silent input clips, ground truth spectrograms, SpatialSonic baseline outputs, and our results.

In the explosion sequence, our method accurately captures the timing and intensity dynamics of three explosions with precise temporal correspondence to visual events. In contrast, SpatialSonic correctly aligns with the first explosion but fails to maintain temporal accuracy for subsequent events, showing temporal drift with mismatches for the second and third explosions. Similarly, for the excavation scene, SpatialSonic exhibits "Multiplication" artifacts, generating repetitive sounds that do not correspond to the actual

visual events, whereas our method successfully generates appropriate shoveling sounds that align with the digging actions shown in the video frames. These qualitative results validate that our approach achieves superior performance in both synchronization accuracy and acoustic quality across different film scenarios. Comprehensive quantitative results and details are provided in Supplementary Materials.

5.3. Ablation Study

To validate the effectiveness of spatio-temporal cues (STC) in FoleyDesigner, we conduct an ablation study comparing the full model against a baseline without STC. The results are shown in Table 4. The baseline configuration without STC shows higher errors across all evaluation metrics. Incorporating STC yields substantial improvements: GCC reduces by **21.3%** and CRW decreases by **38.8%**, demonstrating critical gains in spatio-temporal alignment. The FAD improves by **12.1%**, reflecting better perceptual audio quality.

These results validate that spatio-temporal cues are essential for generating high-quality Foley audio. The consistent improvements across all metrics demonstrate the effectiveness of our position-aware injection mechanism design in achieving superior performance.

Table 4. Ablation study results on the FilmStereo dataset. STC refers to spatio-temporal cues including trajectory information and spatial positioning prompts. **Best** results are highlighted.

Configuration	GCC ↓	CRW ↓	FSAD ↓	FAD ↓
w/o STC	62.02	55.89	0.297	2.14
Full Model	**48.79**	**34.23**	**0.138**	**1.88**

5.4. Human Evaluation

To assess the perceived quality of generated foleys, we conducted offline (5.1 surround) and online (stereo) human evaluations. Offline tests involved 12 participants evaluating the 5.1 output of FoleyDesigner, while the online study involved 53 participants using stereo playback. For baseline comparisons, we evaluated against See-2-Sound [7], Stable Audio [11], and SpatialSonic [32]. Participants rated each method across five dimensions: (1) *immersion*; (2) *emotional alignment* with scene atmosphere; (3) *temporal alignment* with visual events; (4) *spatial alignment* relative to visual sources; (5) *timbre consistency* with film content.

Our method demonstrates performance across all dimensions in both evaluation settings. FoleyDesigner achieves the highest preference rates in emotional alignment (61% online preference) and immersion (58% online preference), indicating that our spatio-temporal cues capture spatial accuracy and semantic and affective qualities of film audio. The preference distributions are presented in Figure 5, showing advantages over existing approaches across all evaluation criteria. These results validate that our method generates spatially-aware foley audio with enhanced realism and user experience compared to baselines.

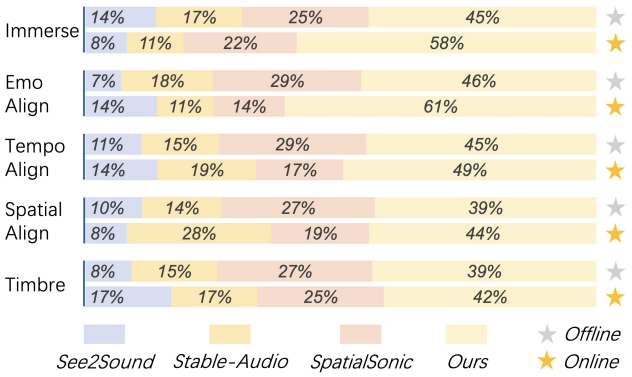

Figure 5. **Human Evaluation Results.** We compared the selection ratio of four methods from five perspectives: (1) Immerse, (2) Emo Align, (3) Tempo Align, (4) Spatial Align, and (5) Timbre.

6. Conclusion

We have presented **FoleyDesigner**, a novel framework for generating spatio-temporally aligned stereo audio for silent film clips. Drawing inspiration from professional Foley workflows, our framework decomposes complex acoustic scenes into hierarchical scripts through Tree-of-Thought reasoning. It integrates multi-stage modules including film clip perception, stereo audio generation, and multi-agent mixing. Crucially, our spatio-temporal conditioning mechanism extracts depth and azimuth cues from visual tracking and injects them via position-aware cross-attention, ensuring frame-accurate synchronization with on-screen movements.

To support stereo Foley generation, we introduced the annotated **FilmStereo** dataset. Experimental results demonstrate that FoleyDesigner achieves superior performance in audio quality, spatial alignment, and temporal accuracy compared to existing methods. Furthermore, it demonstrates strong potential for diverse real-world applications, including film post-production and immersive sound design for virtual reality, supporting efficient and controllable Foley generation.

Despite these advances, we observe that generation performance can degrade in scenes with densely overlapping concurrent sound events (e.g., simultaneous footsteps, object interactions, and ambience), which occasionally leads to spatial localization errors. To address this limitation, our future work will focus on enhancing visual understanding through more robust multi-object tracking and hierarchical spatial reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62402306), the Natural Science Foundation of Shanghai (Grant No. 24ZR1422400), the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2024B72), and the Shanghai Natural Science Foundation (Grant No. 25ZR1401100).

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 4, 5

[2] Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D Plumbley. An improved event-independent network for polyphonic sound event localization and detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 885–889. IEEE, 2021. 3

[3] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18781, 2025. 2

[4] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Taming multimodal joint training for high-quality video-to-audio synthesis. *arXiv preprint arXiv:2412.15322*, 2024. 2

[5] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 5, 6

[6] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720, 2023. 3

[7] Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. SEE-2-SOUND: Zero-shot spatial environment-to-spatial sound. *arXiv preprint arXiv:2406.06612*, 2024. 2, 3, 7, 8

[8] David Diaz-Guerra, Antonio Miguel, and Jose R. Beltran. gpurir: A python library for room impulse response simulation with gpu acceleration. *Multimedia Tools and Applications*, 80(4):5653–5671, 2020. 6

[9] Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley. CLIPSonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2023. 2

[10] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023. 2

[11] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2, 3, 5, 6, 7, 8

[12] Rishabh Garg, Ruohan Gao, and Kristen Grauman. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*, 2021. 3

[13] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 6

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. 6

[15] Feizhen Huang, Yu Wu, Yutian Lin, and Bo Du. Spotlighting partially visible cinematic language for video-to-audio generation via self-distillation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 1170–1178. International Joint Conferences on Artificial Intelligence Organization, 2025. Main Track. 2

[16] Jia-Bin Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *ArXiv*, abs/2305.18474, 2023. 2

[17] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024. 6

[18] ITU-R. Multichannel stereophonic sound system with and without accompanying picture. Technical report, International Telecommunication Union, 2012. Recommendation ITU-R BS.775-3. 5

[19] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *ArXiv*, abs/1812.08466, 2018. 6

[20] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022. 2

[21] Junwon Lee, Jaekwon Im, Dabin Kim, and Juhan Nam. Video-Foley: Two-Stage Video-To-Sound generation via temporal event condition for foley sound. *arXiv preprint arXiv:2408.11915*, 2024. 2

[22] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 2

[23] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. 2

[24] Huadai Liu, Tianyi Luo, Qikai Jiang, Kaicheng Luo, Peiwen Sun, Jialei Wan, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, Shiliang Zhang, Zhijie Yan, Zhou Zhao, and Wei Xue. Omniaudio: Generating spatial audio from 360-degree video. *ArXiv*, abs/2504.14906, 2025. 3

[25] Miao Liu, Jing Wang, Xinyuan Qian, and Xiang Xie. Visually guided binaural audio generation with cross-modal consistency. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7980–7984. IEEE, 2024. 3

[26] 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. 3

[27] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572, 2024. 2

[28] Tobias May, Steven Van De Par, and Armin Kohlrausch. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):1–13, 2010. 3

[29] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2151–2160, 2022. 2, 3

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 2234–2242, Red Hook, NY, USA, 2016. Curran Associates Inc. 6

[31] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[32] Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*, 2024. 2, 3, 6, 7, 8

[33] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 5

[34] 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*, pages 15492–15501, 2024. 2

[35] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2024. 2

[36] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 6

[37] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li. SonicVisionLM: Playing sound with vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26866–26875, 2024. 4

[38] 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, 2024. 4

[39] Qiang Yang and Yuanqing Zheng. Deepear: Sound localization with binaural microphones. *IEEE Transactions on Mobile Computing*, 23(1):359–375, 2022. 3

[40] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 3

[41] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *ArXiv*, abs/2407.01494, 2024. 2

[42] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, page 52–69, Berlin, Heidelberg, 2020. Springer-Verlag. 2

FoleyDesigner: Immersive Stereo Foley Generation with Precise Spatio-Temporal Alignment for Film Clips

Supplementary Material

1. FilmStereo Dataset

Current audio datasets predominantly focus on monaural sound, overlooking the pivotal role of stereophonic audio in enhancing film immersion. As summarized in Table 1, existing datasets typically lack stereophonic recordings or precise temporal annotations. This limitation compels sound designers to manually craft spatial effects from monaural sources, incurring substantial creative and computational overhead. To address this gap, we introduce the **FilmStereo** stereo dataset, which integrates stereo audio, spatial captions, timestamps.

Dataset	Duration (hours)	Temporal	Spatial
LAION-Audio	4.3K	✗	✗
AudioCaps	110	✗	✗
VGG-Sound	550	✗	✗
AudioTime	15.3	✓	✗
CompA-order	1.5	✓	✗
Simstereo	116	✗	✓
FAIR-Play	5.2	✗	✓
MUSIC	23	✗	✓
FilmStereo (ours)	166	✓	✓

Table 1. **Comparison of existing datasets.** FilmStereo provide temporal and spatial annotations simultaneously.

1.1. Data Collection and Preparation

To support film foley generation tasks, we organized our dataset into eight common sound categories, subdivided into 23 subcategories based on typical sound sources used in film post-production. As shown in Table 2, the dataset covers diverse sound types ranging from human actions and environmental sounds to mechanical and impact sounds. Data were collected from publicly available repositories with samples distributed across these categories to ensure comprehensive coverage of film audio requirements.

The data preprocessing pipeline involved multiple quality control steps to ensure dataset integrity. Initially, we filtered out samples with captions indicating multiple concurrent sound events to ensure the validity of subsequent spatial simulations. The acquired audio clips were typically noisy, so we set a threshold of -40 dB to filter out silent segments and applied spectral subtraction denoising algorithms to each recording. Short-duration sounds (less than 2 seconds) were extended to 8–10 seconds through seamless loop-padding using overlap-add techniques, promoting temporal consistency. To verify semantic alignment between audio content and textual descriptions, we employed the CLAP model to compute text-audio similarity scores, discarding samples with scores below $\tau = 0.35$. This rigorous preprocessing resulted in a curated collection of high-quality audio samples suitable for spatial audio simulation.

Category	Percentage (%)
Ambience & Environments	4.70
Bio & Organic	12.87
Cultural & Abstract	21.87
Foley & Physical Interactions	19.01
Designed & Abstract	17.95
Dynamic Systems	6.88
Impact & Destruction	12.59
Mechanical & Technology	4.13

Table 2. **Percentage distribution.** The table shows the proportion of audio clips belonging to each of the eight major sound design categories.

1.2. Spatial Audio Simulation

Guided by insights from psychoacoustic research, we modeled two fundamental spatial perception attributes: azimuth and depth. The azimuth was discretized into five frontal regions (±15°, ±45°, 0°) to correspond with the acuity of human lateral localization capabilities. Depth perception was nonlinearly categorized into three distinct zones—near-field (0–2 meters), mid-field (2–5 meters), and far-field (greater than 5 meters)—based on frequency-dependent attenuation patterns commonly observed in professional Foley practice. These derived acoustic parameters were employed to simulate stereo audio in a perceptually realistic manner that aligns with human auditory perception.

Building upon established room acoustics simulation techniques, we utilized gpuRIR for the room impulse response generation process. During this simulation, we assumed a standard interaural distance of 16–18 cm, without explicitly accounting for the head shadow effect or head-related transfer functions (HRTFs), as these factors were deemed secondary for the scope of this study. The simulation pipeline comprised two distinct stages to handle different types of sound sources. For static sound sources, we performed random sampling of azimuth and depth parameters within the defined ranges. For dynamic sources, trajectories

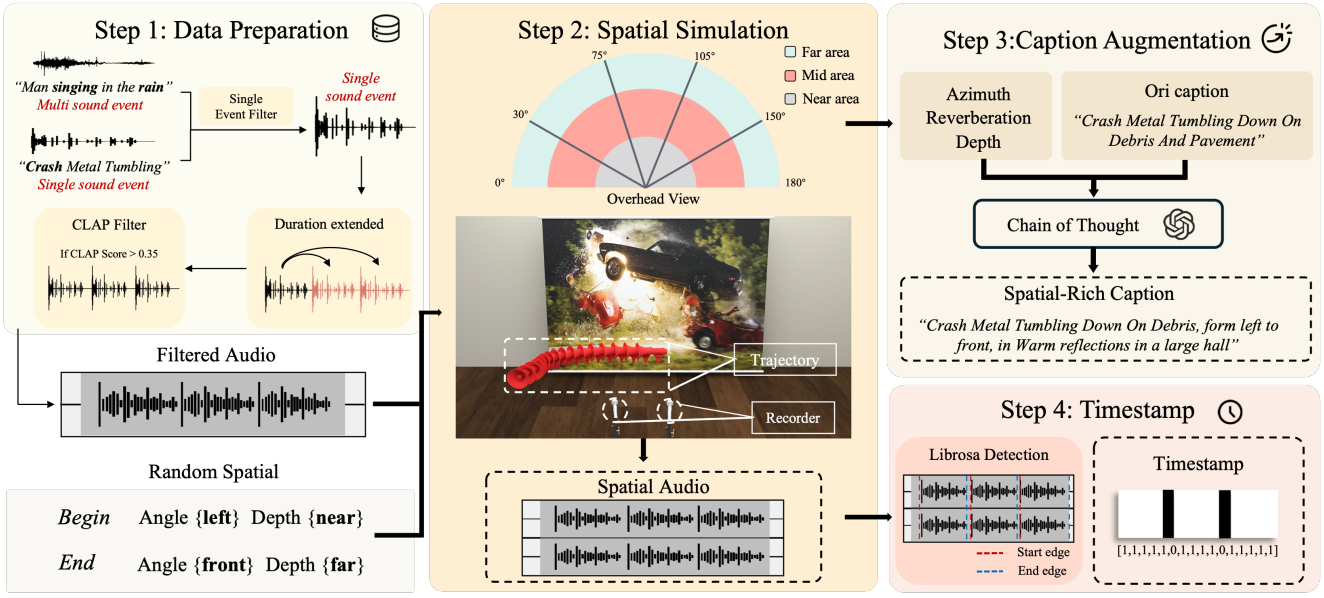

Figure 1. **FilmStereo Dataset Pipeline.** The process begins with sourcing data using randomly sampled parameters to define sound event attributes, followed by a simulated sound design scenario in Step 2 to generate film foley annotations. The resulting data undergoes manual verification to ensure quality and accuracy.

were computed based on spatial position parameters, with intermediate positions refined using linear interpolation to ensure smooth spatial transitions. After the initial spatial simulation, we applied environmental reverberation effects using VST3 plugin presets[1], including hall, room, chamber, and plate reverbs, to match film environmental contexts and enhance the realism of the generated audio.

1.3. Annotation

To enhance the utility of audio datasets for film design applications, we transformed raw captions into spatially-aware descriptions enriched with comprehensive spatial and acoustic information. Sound designers typically infer sound event types, spatial positions, and reverberation effects based on the specific requirements of a film sequence, necessitating detailed spatial descriptions. We extracted parameters such as azimuth, depth, and reverberation effects directly from the audio processing operations performed during the spatial audio simulation pipeline. Using GPT-4, we employed a chain-of-thought prompting strategy to generate captions that seamlessly integrate sound event descriptions with their corresponding spatial and acoustic properties, ensuring alignment with professional sound design workflows.

To align the rhythm of sound with visual elements in film audio production, we introduced temporal annotations comprising precise start and end timestamps that define the exact timing of sound events within each audio clip. We detected amplitude peaks in the denoised audio signals to identify the onset of sound events, applying adaptive thresholds to filter out background noise and silent segments based on these detected peaks. Additionally, we established interval thresholds to determine whether a segment qualifies as a distinct sound event, thereby generating corresponding start and end timestamps with millisecond precision. This temporal annotation process ensures that the generated audio can be precisely synchronized with visual content during film post-production workflows.

1.4. Dataset Statistics

The complete FilmStereo dataset contains 42.3 hours of stereo audio data distributed across 14,784 samples spanning the eight primary categories. As illustrated in Figure 2, the dataset exhibits balanced distributions across multiple dimensions. In terms of object size representation, large objects constitute 40% of the dataset, while medium, small, and very large objects each account for 20%. The motion characteristics show a predominance of dynamic sounds (64%) over static sounds (36%), reflecting the dynamic nature of film audio. The spatial positioning analysis reveals a comprehensive coverage across the frontal hemisphere, with sounds distributed across near-field (green), mid-field (brown), and far-field (blue) distances, ensuring realistic spatial diversity that mirrors typical film audio sce-

[1] https://www.voxengo.com/group/free-vst-plugins-download/

narios.

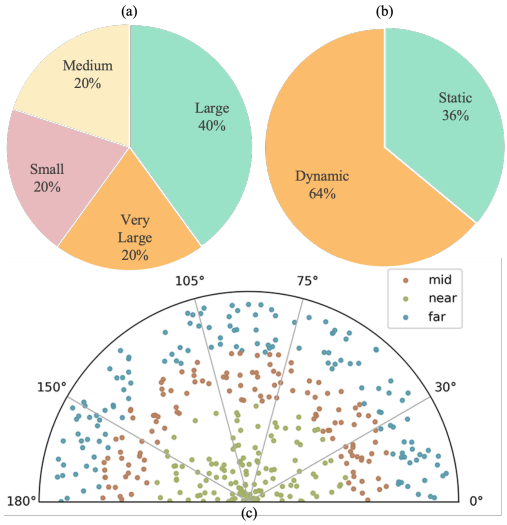

Figure 2. **FilmStereo Distribution Analysis.** (a) Room size distribution. (b) Motion type distribution. (c) Spatial positioning across azimuth angles and depth zones.

2. Implementation Details

2.1. Multi-Agent Refinement

The complete pseudocode for our multi-agent Foley refinement pipeline is presented in Algorithm 1. This algorithm implements the professional mixing framework described in Section 3.3 of the main paper, emulating the collaborative workflow of professional Foley teams through a four-stage process.

Mixing Planning. The planner agent conducts track-wise diagnosis by evaluating three critical aspects for each track. First, cross-modal validation aligns audio features with visual context to detect semantic inconsistencies between the generated sound and the visual content. Second, inter-track balance analysis examines relative loudness and spectral overlap across all tracks to identify potential masking issues where one sound obscures another. Third, acoustic quality assessment matches the reverberation characteristics and frequency distribution to the scene's spatial properties, such as indoor versus outdoor environments. Based on these diagnostic scores, the planner determines the required operations for each track, constructing a structured mixing plan that guides subsequent specialist processing.

Specialist Execution. The execution stage dispatches operations to three specialist agents, each focusing on a specific aspect of audio processing while maintaining awareness of inter-track relationships. The Reverberation Specialist analyzes the spatial context from visual features and determines appropriate reverberation parameters for tracks requiring spatial treatment, adjusting reverb ratio, room size, and damping values to match the acoustic environment depicted in the scene. The Equalization Specialist examines spectral overlap between tracks and determines frequency band adjustments across low, mid, and high frequency ranges to minimize masking effects and ensure each sound element occupies its appropriate spectral space. The Dynamics Specialist evaluates relative loudness levels across tracks and determines gain adjustments in decibels for balanced mixing, ensuring foreground elements maintain prominence while background sounds provide appropriate ambience without overwhelming the mix. Each specialist receives all tracks requiring its operation type simultaneously, enabling coordinated processing that ensures consistency across the entire mix rather than treating tracks in isolation.

2.2. Computational Cost and Inference Time

We provide the runtime breakdown for generating a 3-second stereo Foley clip on a single NVIDIA RTX A6000 GPU in Table 3. While slower than end-to-end models, our method targets professional post-production, where precision, multi-track decomposition, and human-in-the-loop controllability are prioritized over real-time generation speed.

Table 3. Inference Time Analysis. Time consumption in seconds (s) for generating a 3s stereo clip on a single A6000 GPU.

Stage	Time (s)	Component
1. Visual Analysis	2s	VLM
2. Script Decomposition	34s	LLM Agents
3. Audio Generation	8s	DiT Diffusion
4. Foley Refinement	64s	LLM Agents
Total	**108s**	**vs. End-to-End (∼5s)**

3. Additional Quantitative Evaluations

To provide a more comprehensive understanding of Foley-Designer's capabilities, we present additional quantitative experiments including extended baseline comparisons and an ablation study on our multi-agent framework.

3.1. Extended Baseline Comparisons

We evaluated additional state-of-the-art models, specifically Diff-Foley and FoleyCrafter. As shown in Table 4, our method achieves competitive generation quality compared to mono baselines. While FoleyCrafter shows strong mono performance, it lacks spatial control (indicated by GCC and CRW metrics). Furthermore, to investigate the impact of training data, we fine-tuned the mono baseline Tango on our FilmStereo dataset. Adapting its mono-native architecture proved suboptimal compared to our specialized design, yielding a higher FAD and inferior spatial alignment met-

Algorithm 1 Multi-Agent Foley Refinement and Professional Mixing

Require: Generated Foley tracks $\{a_i\}_{i=1}^{N}$, visual context $\mathcal{V}$, Foley script $\mathcal{S}$
Ensure: Refined and mixed Foley audio A_{final}

Stage 1: Foley Analysis
for $i = 1$ to N **do**
  Extract semantic embeddings: $\mathbf{f}_{\text{sem}} \leftarrow \text{AudioLLM}(a_i)$
  Extract spectral features: $\mathbf{f}_{\text{spec}} \leftarrow \text{VLM}(\text{MelSpec}(a_i))$
  Compute reverberation: $f_{\text{rev}} \leftarrow \text{RT60}(a_i)$
  Measure loudness: $f_{\text{loud}} \leftarrow \text{LUFS}(a_i)$
  Construct feature vector: $\mathbf{f}_i \leftarrow [\mathbf{f}_{\text{sem}}, \mathbf{f}_{\text{spec}}, f_{\text{rev}}, f_{\text{loud}}]$
end for
Stage 2: Mixing Planning
Initialize mixing plan: $\Pi \leftarrow \emptyset$
for $i = 1$ to N **do**
  Cross-modal validation: $s_{\text{visual}} \leftarrow \text{Align}(\mathbf{f}_i, \mathcal{V})$
  Inter-track balance: $s_{\text{balance}} \leftarrow \text{Compare}(\mathbf{f}_i, \{\mathbf{f}_j\}_{j \neq i})$
  Acoustic quality: $s_{\text{quality}} \leftarrow \text{Evaluate}(f_{\text{rev}}, f_{\text{loud}}, \mathcal{V})$
  Determine operations: $\mathcal{O}_i \leftarrow \text{Diagnose}(s_{\text{visual}}, s_{\text{balance}}, s_{\text{quality}})$
  Update plan: $\Pi \leftarrow \Pi \cup \{(i, \mathcal{O}_i)\}$
end for
Stage 3: Specialist Execution
Group tracks by operation type:
$\mathcal{T}_{\text{reverb}} \leftarrow \{i \mid \text{reverb} \in \mathcal{O}_i\}$
$\mathcal{T}_{\text{eq}} \leftarrow \{i \mid \text{eq} \in \mathcal{O}_i\}$
$\mathcal{T}_{\text{dyn}} \leftarrow \{i \mid \text{dyn} \in \mathcal{O}_i\}$
if $\mathcal{T}_{\text{reverb}} \neq \emptyset$ **then**
  $\{\theta^i_{\text{rev}}\}_{i \in \mathcal{T}_{\text{reverb}}} \leftarrow \text{ReverbSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{reverb}}}, \mathcal{V})$
  for $i \in \mathcal{T}_{\text{reverb}}$ **do**
    $a_i \leftarrow \text{ApplyReverb}(a_i, \theta^i_{\text{rev}})$
  end for
end if
if $\mathcal{T}_{\text{eq}} \neq \emptyset$ **then**
  $\{\theta^i_{\text{eq}}\}_{i \in \mathcal{T}_{\text{eq}}} \leftarrow \text{EQSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{eq}}})$
  for $i \in \mathcal{T}_{\text{eq}}$ **do**
    $a_i \leftarrow \text{ApplyEQ}(a_i, \theta^i_{\text{eq}})$
  end for
end if
if $\mathcal{T}_{\text{dyn}} \neq \emptyset$ **then**
  $\{\theta^i_{\text{dyn}}\}_{i \in \mathcal{T}_{\text{dyn}}} \leftarrow \text{DynamicsSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{dyn}}})$
  for $i \in \mathcal{T}_{\text{dyn}}$ **do**
    $a_i \leftarrow \text{ApplyDynamics}(a_i, \theta^i_{\text{dyn}})$
  end for
end if
Stage 4: Final Mixing
$A_{\text{final}} \leftarrow \text{Mix}(\{a_1, a_2, \ldots, a_N\})$
return A_{final}

rics. This confirms that simply fine-tuning mono models is insufficient for spatial Foley generation.

3.2. Multi-Agent Framework Ablation

We performed an ablation study to validate the necessity of our multi-agent refinement framework, as shown in Table 5. In complex scenes, the single-stage baseline often missed background events, whereas our multi-agent framework significantly improved Event Recall. Crucially, objective acoustic metrics demonstrate the impact of our Mixing Specialists. While the RT60 Error shows a slight increase due to the inherent complexity of estimating precise reverberation from 2D visual cues in open environments, this trade-off is significantly outweighed by the substantial improvements in spectral clarity (LSD) and dynamic balance (Loudness Error) achieved by our Equalization and Dynamics Agents.

Table 4. Generation Quality and Spatial Alignment. ↓ indicates lower is better, ↑ indicates higher is better.

Method	FAD ↓	ImgBind ↑	IoU ↑	GCC ↓	CRW ↓
Diff-Foley	1.85	0.383	31.80	-	-
FoleyCrafter	**1.69**	**0.430**	32.00	-	-
Tango (Fine-tuned)	2.26	0.328	28.60	54.82	45.36
Ours	1.88	0.402	**32.20**	**48.79**	**34.23**

Table 5. Ablation Study on Agents Framework. Best results are highlighted. ER: Event Recall (%); LSD: Log-spectral distance (dB); LE: Loudness Error (LU); RT60E: RT60 Error (s).

Method	**ER↑**	**LSD↓**	**LE↓**	**RT60E↓**
Single-Stage	68.5%	34.20	7.63	**0.92**
Ours	**84.2%**	**18.42**	**2.86**	1.08

4. User Study Details

4.1. Experimental Setup

Our user study was conducted through both offline and online evaluations to comprehensively assess the perceived quality of generated foley audio.

Offline Evaluation. We recruited 12 participants with normal hearing to conduct perceptual evaluation in a professional audio mixing studio with controlled acoustic conditions. The evaluation environment featured a standard 5.1 surround sound system configured according to ITU-R BS.775-3 specifications, as illustrated in Figure 3.

The listening room was acoustically optimized with controlled reverberation time and minimal background noise. Participants were positioned at the sweet spot, maintaining equal distance from all speakers. The audio playback system utilized monitors with flat frequency response to ensure accurate sound reproduction, as shown in Figure 4.

To ensure the reliability of our subjective metrics, we strictly followed a training session prior to the formal evaluation. Specifically, we provided explicit anchor samples and reference videos with high and low consistency to calibrate participants' perception of Timbre Consistency and Spatial Alignment.

Online Evaluation. We conducted an online

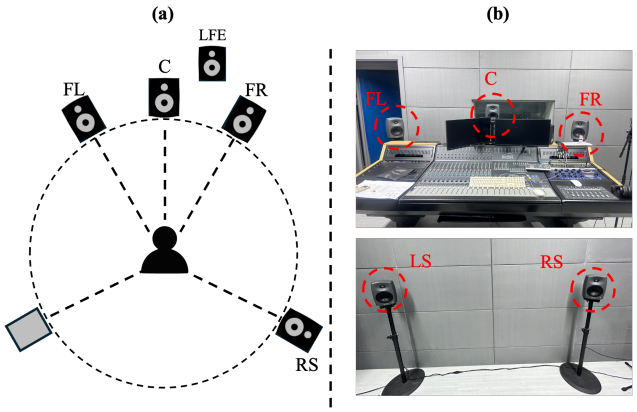

Figure 3. **User study setup**. (a) Standard 5.1 surround sound speaker configuration showing Front Left (FL), Center (C), Front Right (FR), Low Frequency Effects (LFE), Left Surround (LS), and Right Surround (RS) positions. (b) Professional mixing studio environment.

Figure 4. **Participants** conducting the perceptual evaluation in the professional mixing studio environment.

questionnaire-based evaluation with 53 participants, categorized into two groups: 23 film audio professionals (43.4%) and 30 non-professionals (56.6%). Participants evaluated stereo audio samples through a web-based interface. For baseline comparisons, we evaluated our FoleyDesigner against three state-of-the-art methods: See2Sound, Stable-Audio-Open, and SpatialSonic.

4.2. Questionnaire Details

Our online human evaluation was conducted through an questionnaire with 53 participants, categorized into two groups: 23 film audio professionals (43.4%) and 30 non-professionals (56.6%). The questionnaire was designed to evaluate four different audio generation methods across multiple criteria. Figure 5 presents the questionnaire we designed for FoleyDesigner.

For Non-Professional Participants: The questionnaire consisted of five evaluation tasks using film clips. Participants were asked to select the best performing audio sample among four options based on the following criteria:

- **Timbral Matching**: Which audio has the highest compatibility between sound timbre and video content?
- **Spatial Consistency**: Which audio demonstrates the highest consistency between sound spatial positioning and video?
- **Temporal Alignment**: Which audio shows the highest consistency between sound timing and video?
- **Emotional Coherence**: Which audio has the highest compatibility between sound emotion and video content?
- **Immersiveness**: Which audio provides the strongest sense of immersion?

For Professional Participants: The questionnaire included two additional evaluation tasks using film clips, with extended criteria including:

- All criteria from the non-professional evaluation
- **Audio Layering**: Which audio demonstrates the clearest hierarchy between primary and secondary sound layers?
- **Detail Processing**: Which audio shows more refined detail processing?

Each participant watched the original film clips and then evaluated the four generated audio samples across all specified dimensions using a matrix single-choice format.

4.3. Statistical Significance Analysis

Since our user study is preference-based, we performed a Chi-Square Goodness-of-Fit Test to evaluate the statistical significance of the results. The analysis demonstrates a highly significant preference distribution ($p < 0.001$). Furthermore, a post-hoc Binomial Test confirms that our method significantly outperforms the second-best baseline with $p < 0.001$, indicating a robust consensus among raters.

5. Case Study

We conduct comprehensive qualitative analysis through two distinct case studies to evaluate temporal synchronization and spatial audio positioning capabilities across methods with different output channel configurations.

5.1. Temporal Synchronization Analysis

Figure 6 demonstrates the temporal alignment performance across different methods producing various audio formats. Each video frame in the input sequence corresponds to a specific temporal segment in the spectrogram visualization below. The yellow checkmarks indicate successful audio-visual synchronization points where the generated audio content accurately aligns with the visual events and their expected acoustic counterparts.

Our method, which outputs 5.1 Dolby surround audio, achieves great temporal consistency with checkmarks appearing at synchronization points throughout the sequence. This demonstrates that our approach successfully captures

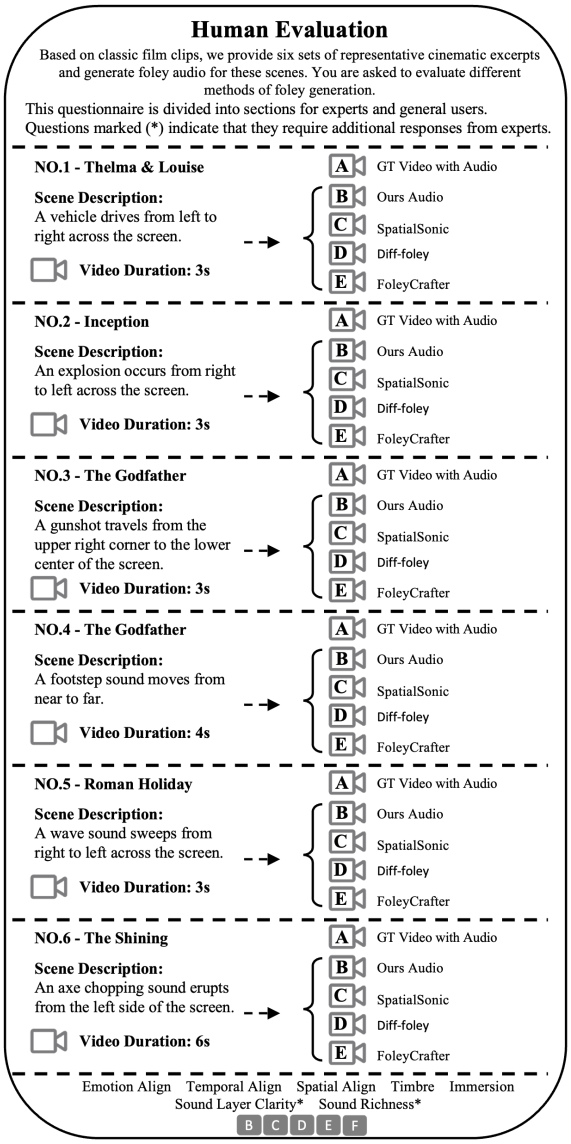

Figure 5. **Questionnaire details**. This is the survey questionnaire we designed and used in the user study.

the timing of film sound events and generates corresponding audio that maintains precise temporal alignment.

As further evidenced by the quantitative results in Table 4, the baseline methods show varying temporal performance. SpatialSonic shows partial synchronization success, while the mono output methods (Diff-Foley and FoleyCrafter) demonstrate different temporal behaviors: Diff-Foley shows limited temporal coherence, missing several key synchronization opportunities, while FoleyCrafter achieves better alignment but still produces inconsistent temporal patterns compared to our multi-channel approach.

5.2. Spatial Audio Positioning Analysis

Figures 7 and 8 present spatial audio analysis demonstrating bidirectional spatial positioning capabilities. Figure 7 shows a scene from Roman Holiday with ocean waves moving left-to-right, while Figure 8 demonstrates right-to-left movement.

Our method generates 5.1 Dolby surround audio with clear spatial positioning in both directions. In the left-to-right case, the left channel (L) gradually strengthens while the right channel (R) weakens. Conversely, in the right-to-left case, the left channel weakens while the right channel strengthens. This bidirectional channel variation accurately reflects the spatial movement of sound sources across scenes.

In contrast, SpatialSonic produces stereo output but exhibits limited spatial variation in both scenarios, with less pronounced channel differences that fail to capture the spatial movement. See2Sound generates audio lacking spatial information.

These complementary examples demonstrate our method's robust spatial audio positioning across different movement directions.

5.3. Discussion

These case studies demonstrate our method's effectiveness in both temporal synchronization and spatial audio positioning. The temporal analysis shows consistent alignment with visual events, while the spatial analysis reveals appropriate channel separation that corresponds to visual movement. The comparison suggests that effective spatial audio generation requires not only multi-channel output capability but also proper modeling of spatial relationships in the audio generation process.

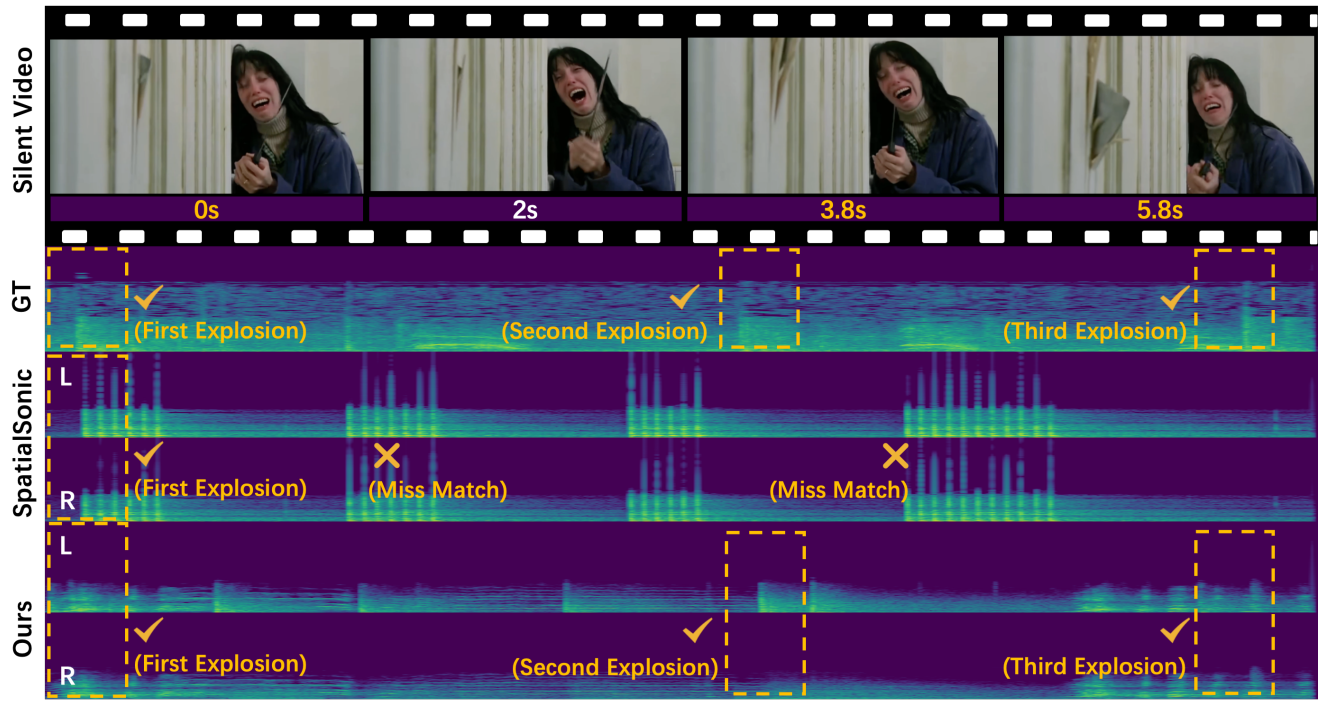

Figure 6. **Temporal Analysis.** Each video frame corresponds to a temporal segment in the spectrogram below. Yellow checkmarks indicate successful audio-visual synchronization. Our method achieves consistent temporal alignment across key events, while baseline methods show varying degrees of synchronization failure regardless of their output channel configuration.

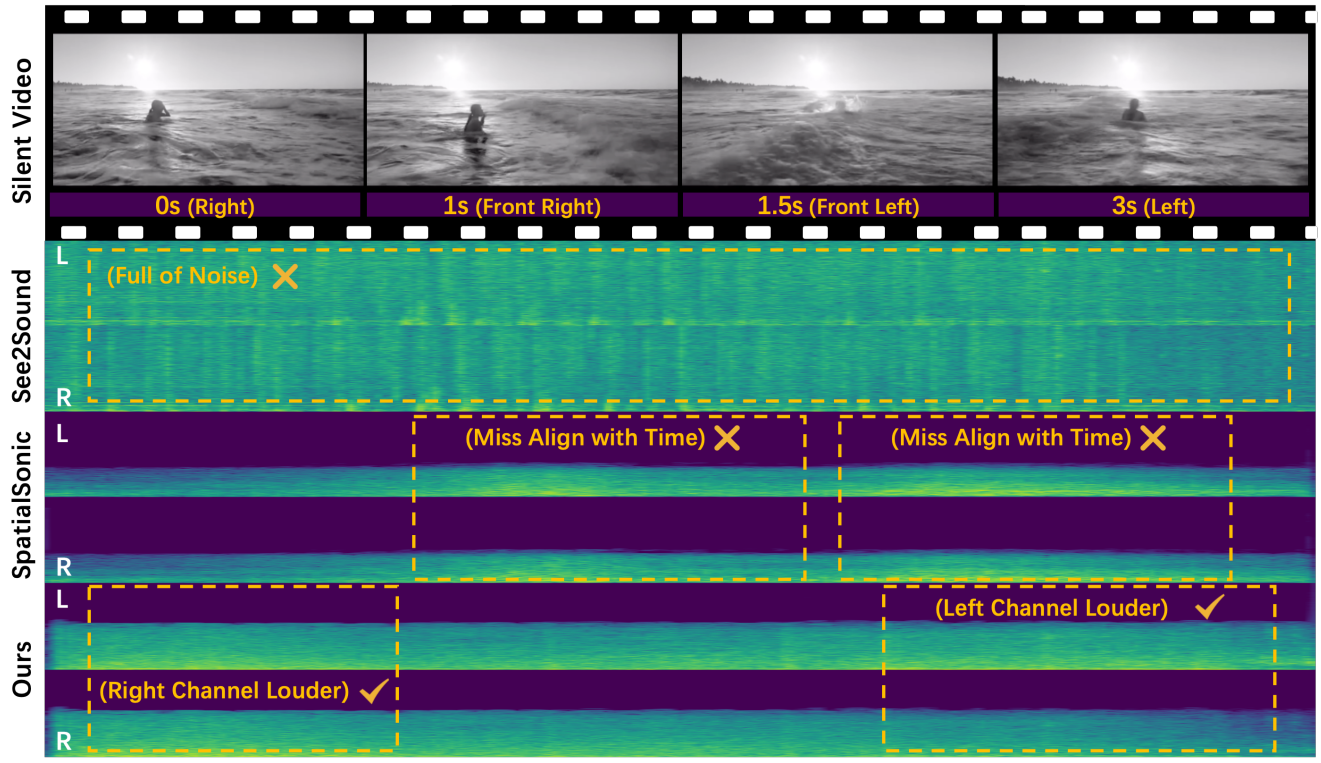

Figure 7. **Spatial Analysis.** Our method demonstrates proper stereo separation with left channel (L) strengthening and right channel (R) weakening, while SpatialSonic (stereo output) shows limited spatial variation despite having two-channel capability.

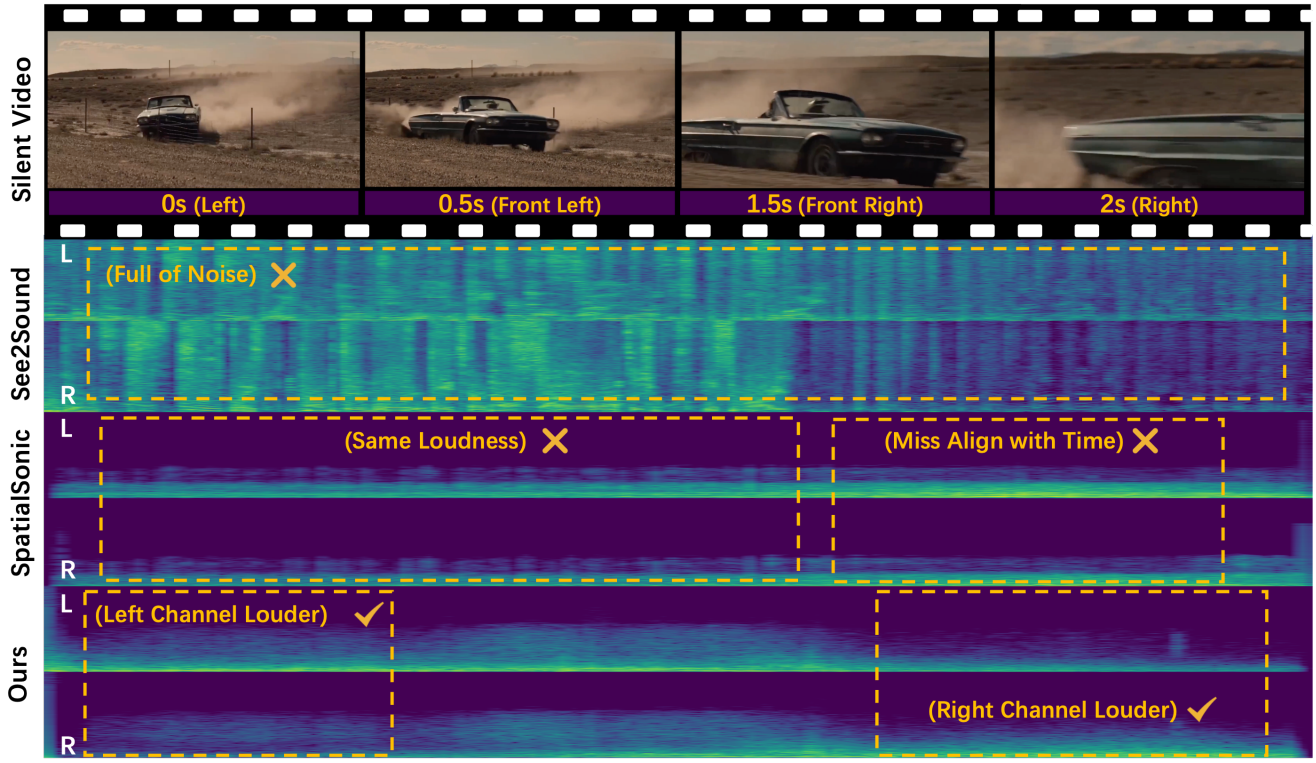

Figure 8. **Spatial Analysis.** Our method demonstrates proper stereo separation with left channel (L) weakening and right channel (R) strengthening, while SpatialSonic (stereo output) shows limited spatial variation despite having two-channel capability.