

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 seam-

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.

less 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.

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 ($\pm 15^\circ$, $\pm 45^\circ$, 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

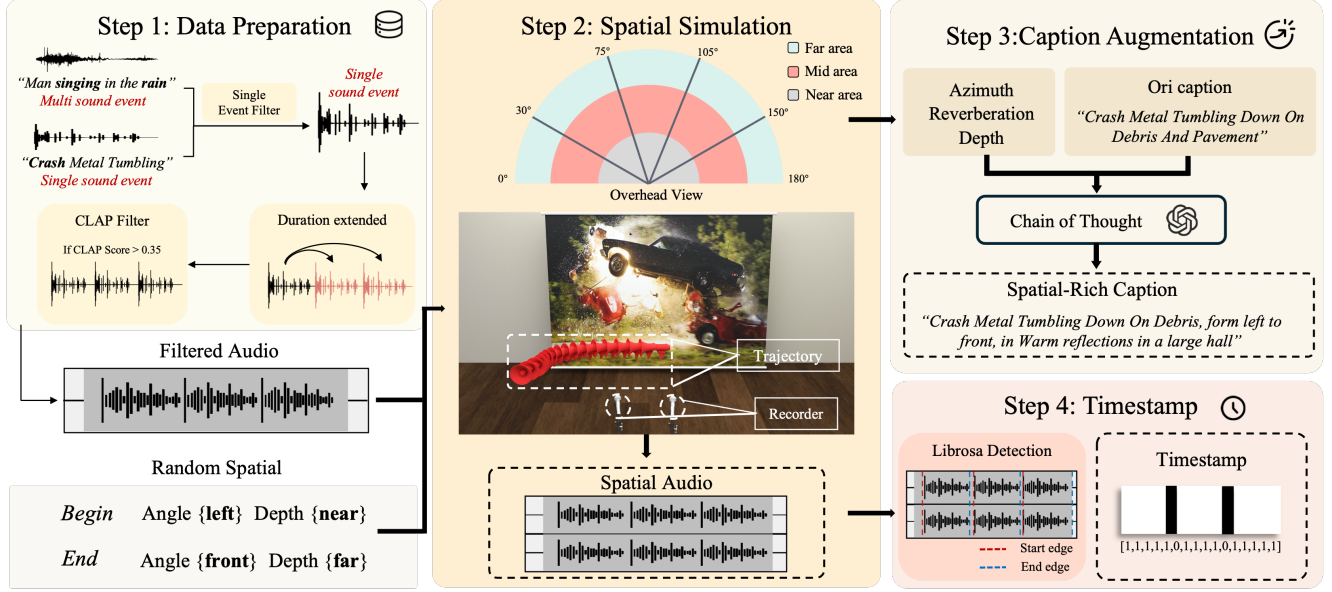


Figure 1. **FilmStereo Dataset Pipeline.** The process begins with sourcing data using randomly sampled parameters to define sound event attributes, followed by a simulated sound design scenario in Step 2 to generate film foley annotations. The resulting data undergoes manual verification to ensure quality 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 presets¹, 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.

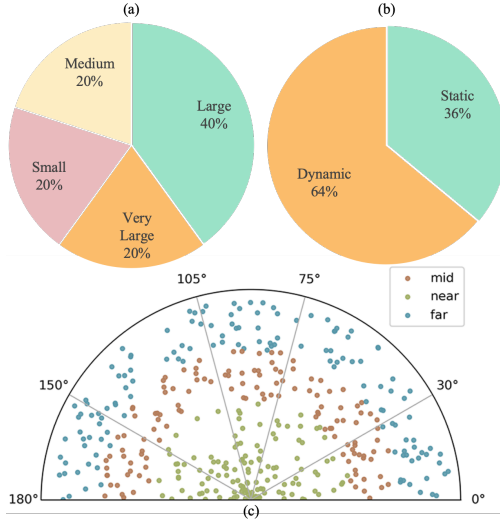


Figure 2. **FilmStereo Distribution Analysis.** (a) Room size distribution. (b) Motion type distribution. (c) Spatial positioning across azimuth angles and depth zones.

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}

- 1: **Stage 1: Foley Analysis**
- 2: **for** $i = 1$ to N **do**
- 3: Extract semantic embeddings: $\mathbf{f}_{\text{sem}} \leftarrow \text{AudioLLM}(a_i)$
- 4: Extract spectral features: $\mathbf{f}_{\text{spec}} \leftarrow \text{VLM}(\text{MelSpec}(a_i))$
- 5: Compute reverberation: $f_{\text{rev}} \leftarrow \text{RT60}(a_i)$
- 6: Measure loudness: $f_{\text{loud}} \leftarrow \text{LUFS}(a_i)$
- 7: Construct feature vector: $\mathbf{f}_i \leftarrow [\mathbf{f}_{\text{sem}}, \mathbf{f}_{\text{spec}}, f_{\text{rev}}, f_{\text{loud}}]$
- 8: **end for**
- 9: **Stage 2: Mixing Planning**
- 10: Initialize mixing plan: $\Pi \leftarrow \emptyset$
- 11: **for** $i = 1$ to N **do**
- 12: Cross-modal validation: $s_{\text{visual}} \leftarrow \text{Align}(\mathbf{f}_i, \mathcal{V})$
- 13: Inter-track balance: $s_{\text{balance}} \leftarrow \text{Compare}(\mathbf{f}_i, \{\mathbf{f}_j\}_{j \neq i})$
- 14: Acoustic quality: $s_{\text{quality}} \leftarrow \text{Evaluate}(f_{\text{rev}}, f_{\text{loud}}, \mathcal{V})$
- 15: Determine operations: $\mathcal{O}_i \leftarrow \text{Diagnose}(s_{\text{visual}}, s_{\text{balance}}, s_{\text{quality}})$
- 16: Update plan: $\Pi \leftarrow \Pi \cup \{(i, \mathcal{O}_i)\}$
- 17: **end for**
- 18: **Stage 3: Specialist Execution**
- 19: Group tracks by operation type:
- 20: $\mathcal{T}_{\text{reverb}} \leftarrow \{i \mid \text{reverb} \in \mathcal{O}_i\}$
- 21: $\mathcal{T}_{\text{eq}} \leftarrow \{i \mid \text{eq} \in \mathcal{O}_i\}$
- 22: $\mathcal{T}_{\text{dyn}} \leftarrow \{i \mid \text{dyn} \in \mathcal{O}_i\}$
- 23: **if** $\mathcal{T}_{\text{reverb}} \neq \emptyset$ **then**
- 24: $\{\theta_{\text{rev}}^i\}_{i \in \mathcal{T}_{\text{reverb}}} \leftarrow \text{ReverbSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{reverb}}}, \mathcal{V})$
- 25: **for** $i \in \mathcal{T}_{\text{reverb}}$ **do**
- 26: $a_i \leftarrow \text{ApplyReverb}(a_i, \theta_{\text{rev}}^i)$
- 27: **end for**
- 28: **end if**
- 29: **if** $\mathcal{T}_{\text{eq}} \neq \emptyset$ **then**
- 30: $\{\theta_{\text{eq}}^i\}_{i \in \mathcal{T}_{\text{eq}}} \leftarrow \text{EQSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{eq}}})$
- 31: **for** $i \in \mathcal{T}_{\text{eq}}$ **do**
- 32: $a_i \leftarrow \text{ApplyEQ}(a_i, \theta_{\text{eq}}^i)$
- 33: **end for**
- 34: **end if**
- 35: **if** $\mathcal{T}_{\text{dyn}} \neq \emptyset$ **then**
- 36: $\{\theta_{\text{dyn}}^i\}_{i \in \mathcal{T}_{\text{dyn}}} \leftarrow \text{DynamicsSpecialist}(\{(a_i, \mathbf{f}_i)\}_{i \in \mathcal{T}_{\text{dyn}}})$
- 37: **for** $i \in \mathcal{T}_{\text{dyn}}$ **do**
- 38: $a_i \leftarrow \text{ApplyDynamics}(a_i, \theta_{\text{dyn}}^i)$
- 39: **end for**
- 40: **end if**
- 41: **Stage 4: Final Mixing**
- 42: $A_{\text{final}} \leftarrow \text{Mix}(\{a_1, a_2, \dots, a_N\})$
- 43: **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 Ta-

Table 4. Generation Quality and Spatial Alignment. \downarrow indicates lower is better, \uparrow indicates higher is better.

Method	FAD \downarrow	ImgBind \uparrow	IoU \uparrow	GCC \downarrow	CRW \downarrow
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

ble 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 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 \uparrow	LSD \downarrow	LE \downarrow	RT60E \downarrow
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

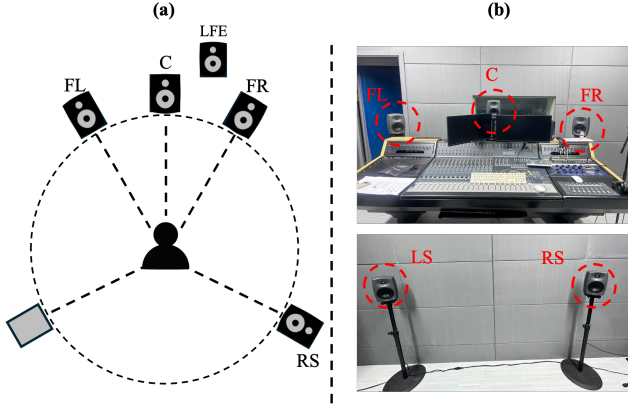


Figure 3. **User study setup.** (a) Standard 5.1 surround sound speaker configuration showing Front Left (FL), Center (C), Front Right (FR), Low Frequency Effects (LFE), Left Surround (LS), and Right Surround (RS) positions. (b) Professional mixing studio environment.



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. Partici-

pants 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

Human Evaluation

Based on classic film clips, we provide six sets of representative cinematic excerpts and generate foley audio for these scenes. You are asked to evaluate different methods of foley generation.

This questionnaire is divided into sections for experts and general users. Questions marked (*) indicate that they require additional responses from experts.

NO.1 - Thelma & Louise

Scene Description:
A vehicle drives from left to right across the screen.

Video Duration: 3s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

NO.2 - Inception

Scene Description:
An explosion occurs from right to left across the screen.

Video Duration: 3s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

NO.3 - The Godfather

Scene Description:
A gunshot travels from the upper right corner to the lower center of the screen.

Video Duration: 3s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

NO.4 - The Godfather

Scene Description:
A footstep sound moves from near to far.

Video Duration: 4s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

NO.5 - Roman Holiday

Scene Description:
A wave sound sweeps from right to left across the screen.

Video Duration: 3s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

NO.6 - The Shining

Scene Description:
An axe chopping sound erupts from the left side of the screen.

Video Duration: 6s

→

A GT Video with Audio

B Ours Audio

C SpatialSonic

D Diff-foley

E FoleyCrafter

Emotion Align
Temporal Align
Spatial Align
Timbre
Immersion

B
C
D
E
F

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.

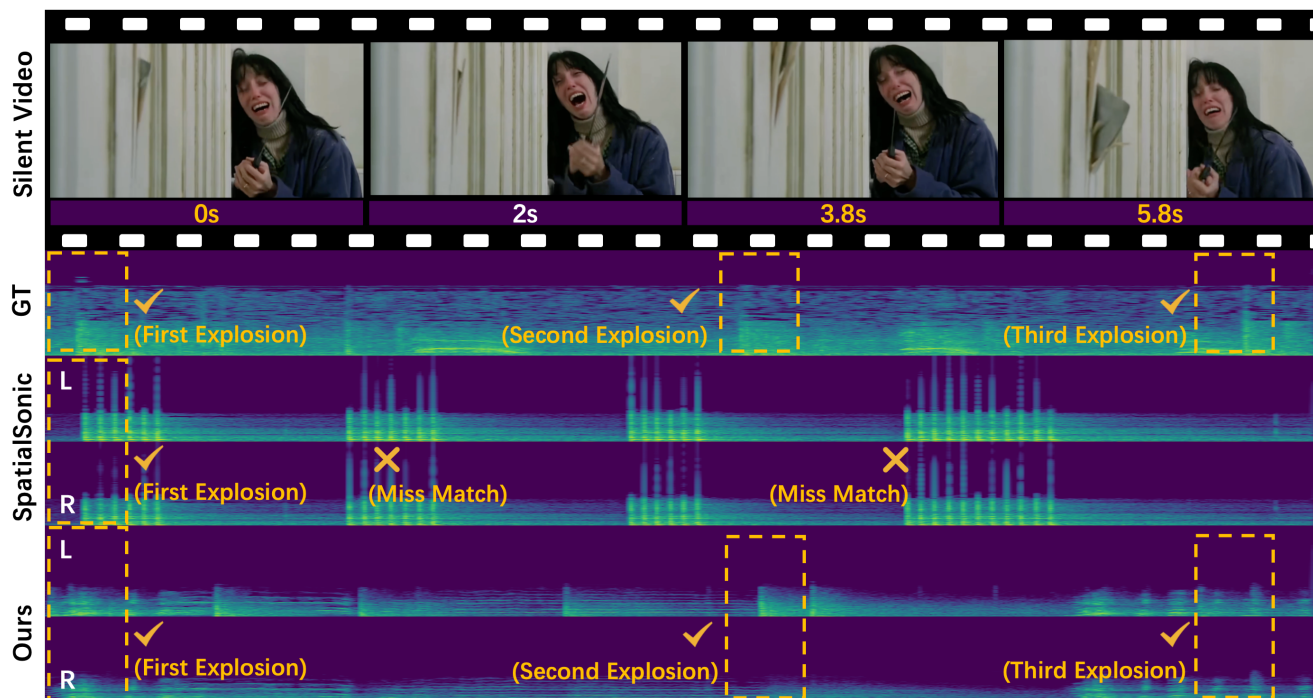


Figure 6. **Temporal Analysis.** Each video frame corresponds to a temporal segment in the spectrogram below. Yellow checkmarks indicate successful audio-visual synchronization. Our method achieves consistent temporal alignment across key events, while baseline methods show varying degrees of synchronization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

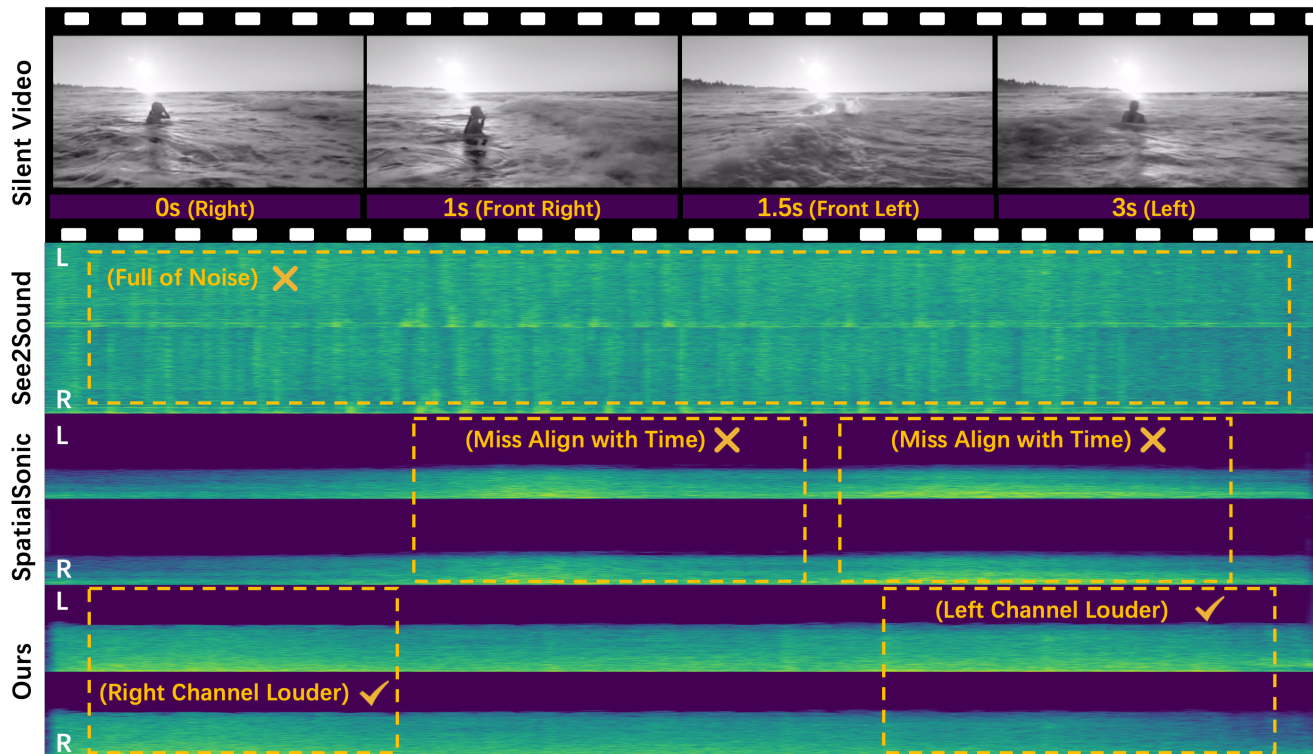


Figure 7. **Spatial Analysis.** Our method demonstrates proper stereo separation with left channel (L) strengthening and right channel (R) weakening, while SpatialSonic (stereo output) shows limited spatial variation despite having two-channel capability.

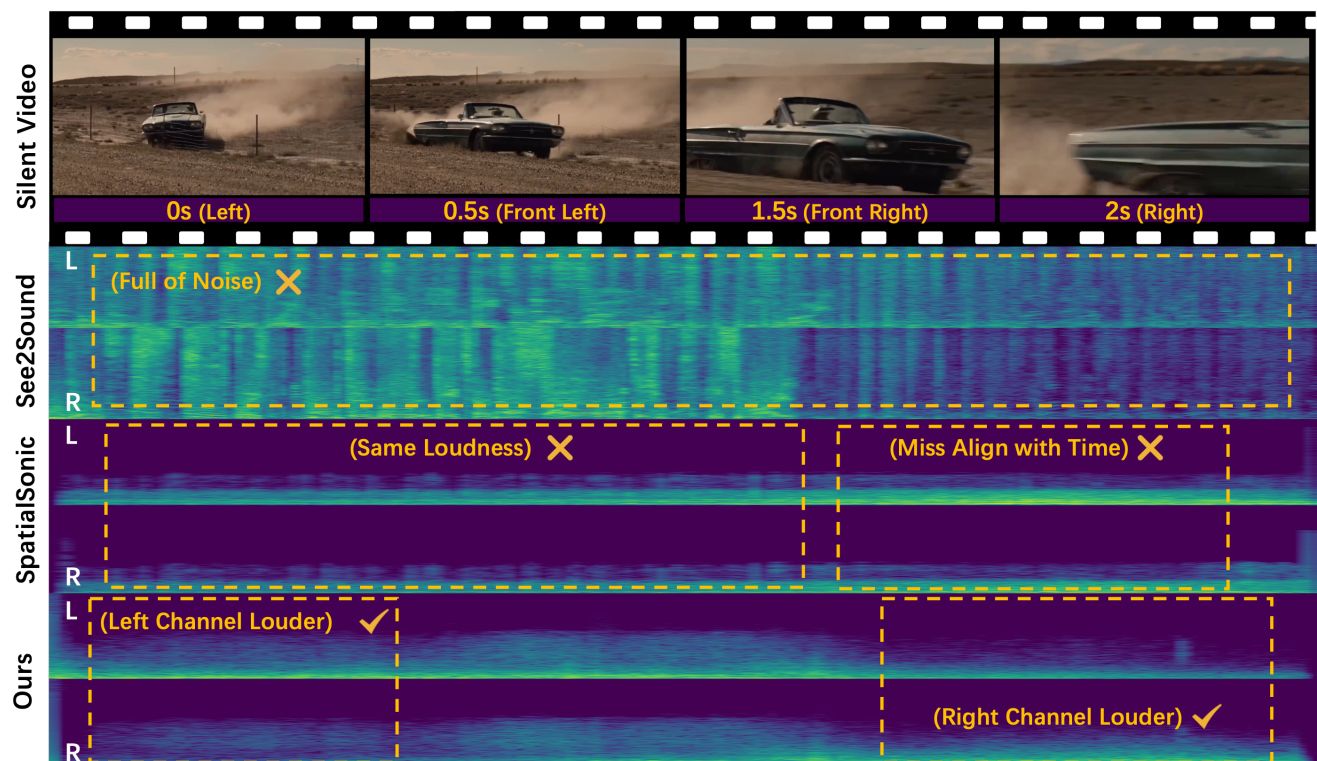


Figure 8. **Spatial Analysis.** Our method demonstrates proper stereo separation with left channel (L) weakening and right channel (R) strengthening, while SpatialSonic (stereo output) shows limited spatial variation despite having two-channel capability.