EVENT-GUIDED MICRO-MOTION AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional Video-to-Audio (V2A) generation typically suffers from temporal aliasing, as standard frame videos are "blind" to the high-frequency micro-motion essential for physically aligned acoustic synthesis. To bridge this gap, we propose MicroFoley, a generative framework re-encoding inter-frame motion into high-frequency event streams. Specifically, by explicitly disentangling static appearance cues from subtle event-driven micro-dynamics, and by grounding temporal activation in physically plausible interaction patterns aligned with human auditory perception, our method successfully eliminates acoustic hallucinations. We introduce a Bidirectional Mamba architecture to model continuous acoustic evolution from causal onsets to acausal decays, which subsequently introduces an Optimal Transport-Conditional Flow Matching (OT-CFM) decoder optimized via a designable Peak-Aware Contrastive Alignment objective. Furthermore, we construct FoleyBench, a novel audio-video-event dataset (160h) integrating simulation and real-world event streams including 450 categories with our processing pipeline. Extensive results demonstrate that our method significantly outperforms state-of-the-art baselines in generation fidelity and synchronization, achieving optimized inference efficiency. This work project is available at `https://anonymous.4open.science/api/repo/MicroD/file/index.html`.

1 INTRODUCTION

Background. Synthesizing high-fidelity and synchronized audio from silent video is important for multimodal creation. V2A has rapidly evolved from early visual-sound paradigms (Owens et al., 2016; Zhou et al., 2018) and GAN-based generation modelling (Chen et al., 2020b). This trend demonstrated the feasibility of cross-modal synthesis but suffered from limited acoustic diversity, to autoregressive models such as SpecVQGAN (Iashin & Rahtu, 2021) and IM2WAV (Sheffer & Adi, 2023) that model audio over discrete codebooks. More recently, Latent Diffusion Models (LDMs) (Rombach et al., 2022) and Conditional Flow Matching (CFM) (Lipman et al., 2022) have become dominant generative paradigms (Luo et al., 2023; Cheng et al., 2025; Ruan et al., 2023; Liu et al., 2023; Evans et al., 2024), leveraging large-scale pretraining to improve generation quality.

Challenges. Despite these advances, existing V2A methods remain fundamentally constrained by standard RGB sensing. Naturally, sound production is a continuous physical process governed by motion dynamics and vibrations (Niwa et al., 2023; Feng et al., 2022), whereas standard videos sample these dynamics as sparse frames, losing sub-frame motion information. Although common visual proxies (Luo et al., 2023; Ruan et al., 2023; Cheng et al., 2025; Wang et al., 2025a) can capture macroscopic motion, they inevitably average out crucial sub-frame fluctuations into coarse 2D displacements, missing continuous kinetic variations. Consequently, representative models (Cheng et al., 2025; Luo et al., 2023; Wang et al., 2024b; Kushwaha & Tian, 2025) remain "blind" to the inter-frame physics governing acoustic transients, often reproducing coarse high-energy moments while overlooking fine-grained motion cues required for continuous and synchronized audio.

Motivation. We draw inspiration from the recent event-based Video Frame Interpolation (VFI) efforts (Tulyakov et al., 2021; 2022; Kim et al., 2023). Compared with frame-based sensors, event sensors operate asynchronously and respond to pixel-wise logarithmic intensity changes, capturing rapid impacts and subtle motion beyond the temporal resolution of standard videos (Gallego et al., 2022). Since such sub-pixel motion has constituted an important physical source of sound (Niwa et al., 2023), we leverage event-based interpolation to transform inter-frame appearance changes into a high-frequency and sparse motion representation for visual conditioning. The resulting event-

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

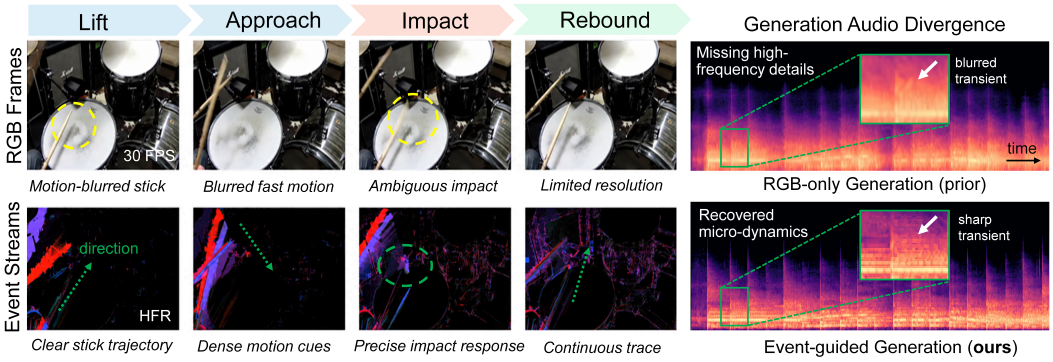


Figure 1: Generated audio conditioned on standard RGB frames versus HFR event priors. (*Left*) Video frames suffer from motion blur and ambiguous impact timing, while event streams provide dense micro-motion cues with clearer trajectories and contact responses. (*Right*) These HFR cues enable sharper transients and more physically aligned and high-fidelity audio generation.

based High-Frame-Rate (HFR) physical priors, offering higher temporal resolution than standard RGB frames (Yin et al., 2025), provide guidable micro-dynamics for high-quality V2A generation.

Contributions. Guided by these insights, we propose MicroFoley, a physically grounded V2A generative framework that leverages event priors to disentangle static material semantics from dynamic micro-motion mechanics, as illustrated in Figure 1. Since subtle acoustic triggers, such as friction and vibration, often manifest as sub-pixel motion (Chen et al., 2024), we employ an event-centric representation to capture these fine-grained dynamics. In particular, MicroFoley introduces a Spatial Residual Pyramid ResNet (SRP-ResNet) with an Adaptive Silence Injector (ASI) to extract micro-motion features while enforcing silence during non-contact periods. The resulting physical priors, together with material semantics, are modeled by a Bidirectional Mamba architecture to extract continuous acoustic evolution, followed by a classic OT-CFM backbone enabling efficient inference for audio synthesis. We further introduce a Peak-Aware Contrastive Alignment (PACA) objective to align generated audio with transient physical impacts. As existing datasets (Owens et al., 2016; Lee et al., 2021; Chen et al., 2020a; Huh et al., 2025) lack the neuromorphic event streams required for micro-motion modeling, we construct FoleyBench, a Sim-to-Real dataset comprising 160 hours and 450 categories of audio-visual data with both synthetic and real-world event streams.

2 Related Work

Video-to-Audio Generation. This line of V2A has progressed from early object interaction-based works (Owens et al., 2016; Zhou et al., 2018) to deep generative frameworks (Chen et al., 2020b; Iashin & Rahtu, 2021; Sheffer & Adi, 2023), with LDMs (Liu et al., 2023; Rombach et al., 2022) largely driving the recent advancements. Established researches align latent representations via contrastive pretraining (Luo et al., 2023; Zhang et al., 2026; Mei et al., 2024; Wang et al., 2024a; Kushwaha & Tian, 2025; Chen et al., 2025; Lee et al., 2025), quantized vectors (Wang et al., 2025a), or Flow Matching (FM) (Cheng et al., 2025; Wang et al., 2024b). While maintaining semantic coherence, these methods fundamentally rely on standard RGB videos, suffering from temporal aliasing and missing high-frequency micro-interactions. To address this, we combine VFI with V2E to encode estimated inter-frame motion as high-frequency event priors for audio synthesis.

Generative Flow Matching. While Diffusion Models (Ho et al., 2020; Song et al., 2020; Peebles & Xie, 2023) rely on discrete sampling, the FM (Wang et al., 2025b; Cheng et al., 2025) and CFM backbone-based model (Le et al., 2023) continuous probability paths via ODEs. However, standard FM/CFM formulations often yield curved, inefficient trajectories, compromising generation quality and inference speed. To circumvent this, we adopt OT-CFM (Kornilov et al., 2024; Tong et al., 2023; Chen et al., 2026), utilizing *optimal transport* to enforce energy-efficient, straight integration paths. This straight-path characteristic naturally aligns with our continuous event interpolation priors, enabling the synthesis of audio that effectively follows physical synthesized trajectories.

Event Streams Representation. Event vision (Gallego et al., 2022) provides microsecond-level temporal resolution, yet generative applications face a severe data bottleneck since standard audio-

visual benchmarks (Chen et al., 2020a; Lee et al., 2021; Huh et al., 2025) lack synchronized event streams. Bio-inspired simulators like the Video-to-Event (V2E) (Hu et al., 2021) bridge this gap by converting frame sequences into asynchronous event streams that mimic neuromorphic sensing. While current event-based learning endeavors (Tulyakov et al., 2022; 2021; Zhu et al., 2022; Kim et al., 2023; Chen et al., 2024; Yin et al., 2025) focus exclusively on visual tasks, we aim to repurpose these high-frequency representations (Yin et al., 2025; Niwa et al., 2023) for visual-audio generation. To mitigate the temporal discreteness and background interference of RGB videos, we employ V2E (Hu et al., 2021) to generate simulated event streams from RGB videos, an approach widely adopted to reduce the gap between existing datasets and real-world event camera recordings.

3 Proposed Framework

3.1 Physically-Enhanced Motion Encoding

The high-fidelity audio synthesis task requires distinct physical priors: the material stiffness determining timbre and the kinetic energy governing temporal activation. Following prior works (Zhang et al., 2026; Chen et al., 2025), we propose a dual-stream encoding scheme that explicitly disentangles these factors before temporal fusion. The framework is illustrated in Figure 2.

Event-Driven Silence Injection. While material defines timbre by characterizing the acoustic identity of sound, event streams characterize sound onset dynamics by providing precise temporal cues of physical interactions. A critical failure mode in V2A is unintended audio during irrelevant motion (Mei et al., 2024). We address this challenge via an *Adaptive Silence Injector* that acts as a soft physical gate. Given the voxelized event tensor $\mathbf{V}_t$ at frame t simulated via V2E (Hu et al., 2021), instead of performing explicit the VFI experience (Tulyakov et al., 2021; 2022), which brings redundancy, we construct the *SRP-ResNet* (He et al., 2016) to extract active spatiotemporal motion features directly from the event streams, as detailed in Figure 5 and Appendix B.2.

To prevent hallucinated sounds from silent motion, we aggregate the spatially gated *SRP-ResNet* features by the global average pooling component to obtain the motion representation $\mathbf{f}_{mot}^{(t)}$ at frame t. We then compute the frame-wise acoustic probability $\alpha_t \in [0, 1]$ to represent the likelihood of sound emission. This probability is obtained by pooling the event features with $\mathrm{Pool}(\cdot)$ to measure concentration, followed by a multi-layer perceptron F_{mlp} and sigmoid activation σ. We further introduce a learnable silence token $\mathbf{v}_\varnothing$ for the null acoustic state and combine it with the motion representation to obtain the frame-wise physical condition $\mathbf{c}_{phys}^{(t)}$, defined as:

$$\mathbf{c}_{phys}^{(t)} = \underbrace{\alpha_t \cdot \mathbf{f}_{mot}^{(t)}}_{\text{Active Motion}} + \underbrace{(1-\alpha_t)\cdot \mathbf{v}_\varnothing}_{\text{Silence Prior}}, \quad \alpha_t = \sigma(F_{mlp}(\mathrm{Pool}(\mathbf{V}_t))). \tag{1}$$

This ensures that when the event energy drops below the physical activation threshold ($\alpha_t \to 0$), the condition gracefully collapses to the silence prior $\mathbf{v}_\varnothing$ for eliminating background artifacts.

Material-Acoustic Alignment. To establish visual semantics and physical motion, we employ the frozen pretrained Contrastive Language-Image Pre-training (CLIP) Encoder (Radford et al., 2021) to extract keyframe vision features. These features are projected into the material embedding space by a learnable adapter $\mathcal{A}_\phi$ into a texture embedding $\mathbf{f}_{mat} \in \mathbb{R}^{T\times D}$ for a given video v.

Furthermore, we design a Cross-Modal Material Distillation (CMMD) mechanism to enhance physical consistency, which aligns the visual embedding $\mathbf{f}_{mat}$ with the spectral manifold of its corresponding ground-truth audio waveform a in embedding space. Combining with a frozen Contrastive Language-Audio Pretraining (CLAP) (Elizalde et al., 2023) as the audio encoder $\mathcal{E}_{aud}$, which provides semantic gradients for optimizing this alignment over valid positive video-audio pairs (v, a) through an InfoNCE loss (Oord et al., 2018). The resulting objective $\mathcal{L}_{CMMD}$ is defined as:

$$\mathcal{L}_{CMMD} = -\mathbb{E}_{(v,a)}\left[\log \frac{\exp(\langle \mathbf{f}_{mat}, \mathcal{E}_{aud}(a)\rangle/\tau)}{\sum_j \exp(\langle \mathbf{f}_{mat}, \mathcal{E}_{aud}(a_j)\rangle/\tau)}\right], \tag{2}$$

where the numerator maximizes the similarity of the matched pair, while the denominator pushes apart negative audio samples a_j indexed by j within the batch. The learnable τ controls the temperature distinctiveness of these material clusters, initialized at 0.07, converging to 0.045 range.

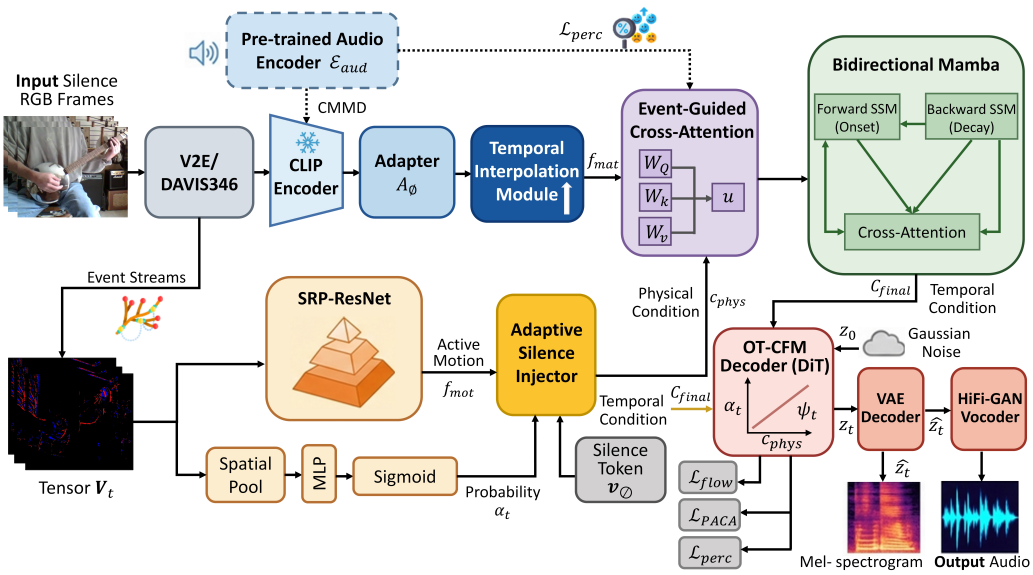


Figure 2: Overview of the proposed MicroFoley. High-frequency events are derived from RGB frames via the V2E (Hu et al., 2021) or the DAVIS346 event camera. Interpolated material features and motion cues are fused via Cross-Attention to drive the Bidirectional Mamba (Bi-Mamba) module, capturing continuous acoustic evolution. The temporal condition guides our optimized OT-CFM decoder (see Section 3.3) to synthesize synchronized audio latents from gaussian noise.

3.2 Bidirectional Temporal Dynamics Modeling

Event-Guided Cross-Attention. Material properties are acoustically relevant only during salient motion. Instead of naive concatenation, we employ an asymmetric Cross-Attention mechanism based on (Vaswani et al., 2017) where motion dictates semantic intake. However, a temporal resolution mismatch exists: the material features $\mathbf{f}_{mat}$ are extracted from sparse keyframes to reduce redundancy, while the physical condition $\mathbf{c}_{phys}$ operates at a high frame rate. To this end, we introduce a Temporal Interpolation module, using nearest-neighbor replication to align $\mathbf{f}_{mat}$ along the time axis with the sequence length T of $\mathbf{c}_{phys}$. We treat the physical condition $\mathbf{c}_{phys} \in \mathbb{R}^{T\times D}$ as the driving *Query*, while the aligned material features $\mathbf{f}_{mat}$ serve as the *Key* and *Value*. These inputs are projected via learnable matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ to compute the fused representation $\mathbf{u}$:

$$\mathbf{u} = \text{LayerNorm}\left(\mathbf{c}_{phys} + \text{Softmax}\left(\frac{(\mathbf{c}_{phys}\mathbf{W}_Q)(\mathbf{f}_{mat}\mathbf{W}_K)^\top}{\sqrt{d}}\right)(\mathbf{f}_{mat}\mathbf{W}_V)\right), \tag{3}$$

where the scaling factor $\sqrt{d}$ normalizes the dot product to stabilize the attention logits and improve training stability. This ensures $\mathbf{u}$ preserves the microsecond-level resolution of the event streams.

Bidirectional Phase Modeling with Bi-Mamba. We leverage the Mamba architecture (Gu & Dao, 2023), rooted in the continuous state-space equation $h'(t)$, which maps a 1-D input $x(t)$ to a latent state $h(t)$ through a linear state-space system parameterized by $\mathbf{A}$ and $\mathbf{B}$ (Gu et al., 2021). To process sampled digital signals, the model discretizes this continuous system using a timescale parameter Δ into a recurrence. Distinguishing itself from static Recurrent Neural Networks, Mamba's State Space Model (SSM) (Zhu et al., 2024) makes Δ, $\mathbf{B}$, and $\mathbf{C}$ input-dependent (selective scan), allowing this module to distinguish between transient impulses and background noise.

To capture both the causal attack (*onset*) and the context-aware release (*decay*), we process the fused sequence $\mathbf{u}$ through bidirectional SSM blocks to obtain the final acoustic condition $\mathbf{C}_{final}$, formulated as follows:

$$\mathbf{C}_{final} = \text{SSM}_{\rightarrow}(\mathbf{u}) \oplus \text{SSM}_{\leftarrow}(\mathbf{u}), \tag{4}$$

where the forward stream $\text{SSM}_{\rightarrow}$ predicts the immediate acoustic response to visual impact, while the backward stream $\text{SSM}_{\leftarrow}$ models the resonance tail grounded in future context. The summation $\oplus$ yields a holistic temporal condition $\mathbf{C}_{final}$ for the generative backbone.

3.3 Optimization with Perceptual Alignment

Flow Matching Objective. We train a neural vector field v_θ, implemented as a Diffusion Transformer (DiT) (Peebles & Xie, 2023), to regress the target velocity, while the holistic temporal condition $\mathbf{C}_{final}$ is injected through Adaptive Layer Normalization (AdaLN) to modulate intermediate features. Following OT-CFM practice (Kornilov et al., 2024), the probability path ψ_t is defined as a straight-line interpolation between the noise $\mathbf{z}_0$ and data $\mathbf{z}_1$, governed by the timestep $t \in [0, 1]$ and a minimum noise standard deviation σ_{min} for numerical stability. Concretely, the training objective $\mathcal{L}_{flow}$ minimizes the expected squared error between the predicted vector field $v_\theta(\psi_t(\mathbf{z}_0), t, \mathbf{C}_{final})$ and the target flow direction $\mathbf{z}_1 - (1 - \sigma_{min})\mathbf{z}_0$ over sampled timesteps and latent pairs.

Peak-Aware Optimization. Standard regression (Cheng et al., 2025; Luo et al., 2023) losses operate uniformly across time, often under-representing high-frequency transients that are perceptually critical. Since impulsive contacts dominate perceived audio-visual synchronization, we treat event-energy peaks as a proxy for physically meaningful acoustic onsets and emphasize errors around these peaks during training. We introduce a Peak-Aware Contrastive Alignment (PACA) objective that uses a normalized event energy profile $\mathcal{E}(t)$, computed from the voxelized event representation, to emphasize transient regions during training. The resulting loss $\mathcal{L}_{PACA}$ is formulated as follows:

$$\mathcal{L}_{PACA} = \mathbb{E}_t \left[\log \left(1 + \lambda_p \mathcal{E}(t) \cdot \|\hat{\mathbf{z}}_t - \mathbf{z}_t\|^2 \right) \right], \tag{5}$$

where λ_p controls the peak-aware temporal re-weighting, and $\hat{\mathbf{z}}_t$ is the generated latent at step t.

Composite Optimization and Waveform Synthesis. To ensure both signal fidelity and semantic realism, we incorporate a CLAP-based perceptual loss $\mathcal{L}_{perc}$ (Elizalde et al., 2023) that maximizes the cosine similarity between generated and ground-truth audio embeddings in the CLAP latent space. The total optimization objective $\mathcal{L}_{total}$ jointly constrains generation across physical, temporal, and semantic manifolds, defined as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{flow} + \mathcal{L}_{PACA} + \lambda_{mat}\mathcal{L}_{CMMD} + \lambda_{sem}\mathcal{L}_{perc}, \tag{6}$$

where λ_{mat} and λ_{sem} balance the auxiliary physical constraints. To regularize the material feature space, we introduce loss $\mathcal{L}_{CMMD}$ that aligns the extracted material embeddings f_{mat} with material priors (see Appendix C.3). During inference, generated audio latent is decoded to a Mel-spectrogram via the VAE decoder (Kingma & Welling, 2013) and vocoded by HiFi-GAN (Kong et al., 2020a).

Table 1: Comparison with existing large-scale audio datasets. We compare dataset scale, category diversity, event-stream availability, audio fidelity, and annotation type. FoleyBench integrates large-scale simulated data with high-fidelity real-world event recordings, providing synchronized audio-video-event data tailored to micro-motion V2A generation.

Dataset	Duration	Clips	Categories	Events	Audio	Annotation
Greatest Hits (Owens et al., 2016)	N/A	977	17	✗	44.1kHz	Impact Sounds
AVE (Tian et al., 2018)	11.5h	4,143	28	✗	Mixed	Event Localization
ACAV100M (Lee et al., 2021)	270kh	100M	N/A	✗	Mixed	Self-Supervision
VGGSound (Chen et al., 2020a)	550h	200k	309	✗	Mixed	In-the-wild Objects
EPIC-SOUNDS (Huh et al., 2025)	100h	78.4k	44	✗	24kHz	Egocentric Actions
FoleyBench-Synth (Ours)	150h	60,000	400	✓	44.1kHz	Micro-Motion
FoleyBench-Real (Ours)	10h	500	50	✓	96kHz	Micro-Motion
FoleyBench (Ours)	160h	60.5k	450	✓	Mixed	Micro-Motion

4 FoleyBench

4.1 Statistics and Comparison

Existing popular audio-visual benchmarks (Chen et al., 2020a; Owens et al., 2016; Lee et al., 2021; Huh et al., 2025) are temporally sparse, lacking the high-frequency micro-dynamics required for physically grounded V2A generation with events. To bridge this, we construct *FoleyBench*, a Sim-to-Real benchmark that balances scale and fidelity. As reported in Table 1, FoleyBench comprises two complementary subsets: *FoleyBench-Synth*, which employs V2E simulator (Hu et al., 2021) to provide adequate event-audio pairs for robust representation learning, and *FoleyBench-Real*, captured with professional event cameras and audio recorders to evaluate real-world generalization.

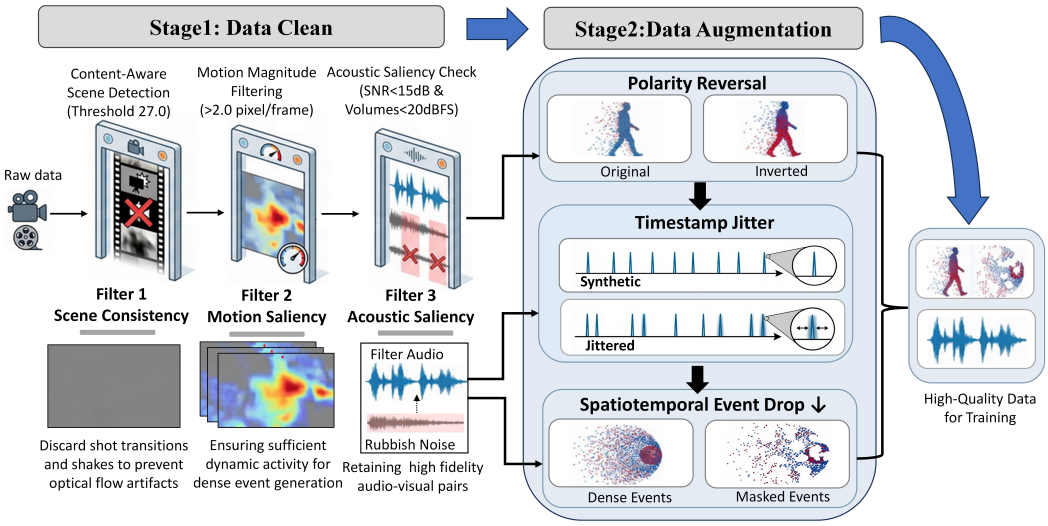


Figure 3: Overview of the designable data processing pipeline. To construct FoleyBench-Synth, we employ a three-stage filtering protocol (*stage 1*) to remove samples with scene transitions (Soucek & Lokoc, 2024), insufficient motion, or low acoustic saliency. We then apply stochastic event augmentations (*stage 2*) to mimic sensor noise and asynchronous latency (Hu et al., 2021; Gehrig et al., 2020), reducing the Sim-to-Real domain gap. Together, these stages yield high-quality and robust audio-video-event pairs for model training and evaluation.

4.2 Data Curation

FoleyBench-Synth. We construct the synthetic data subset from two popular datasets: one consisting of popular high-quality videos selected from VGGSound (Chen et al., 2020a) to cover the diversity of natural scenes, and the other consisting of a collection of physical impact events selected from Greatest Hits (Owens et al., 2016), to support the dataset size and complex scenes application. We employ a gradual data process pipeline in Section 4.3 to simulate the generation of the enhanced event streams with V2E (Hu et al., 2021). This synthetic subset contains approximately 150 hours, covering 400 different physical scenes (e.g., slap, instruments, beat), totaling 60,000 processed clips.

FoleyBench-Real. We further capture synchronized RGB frames and microsecond event streams by leveraging the DAVIS346 neuromorphic event camera, paired with a high-fidelity audio recorder H3-VR. Particularly, we recruit 20 participants to perform high-transient physical interactions across diverse acoustic environments (anechoic to reverberant). The recording protocol explicitly requires variations in interaction speed and intensity to maximize dynamic range. After manual synchronization and quality filtering, this subset contains 500 high-quality audio-video-event triples, totaling approximately 10 hours in length, covering 50 daily real-world scenarios.

4.3 Data Process Pipeline

Data Clean. We establish a filtering protocol on the raw collected videos before feeding them into the V2E simulation illustrated in stage 1 of Figure 3. Above all, creating dense event streams requires strictly continuous motion; therefore, we utilize an adaptive Content-Aware Scene Detection (Threshold 27.0) to aggressively discard clips containing shot transitions, cuts, or sudden camera shakes that would cause catastrophic optical flow artifacts. Since static scenes fail to trigger neuromorphic sensors, we compute the average optical flow magnitude across frames and filtered out samples falling below a Motion Saliency threshold of 2.0 pixels/frame, ensuring that every retained clip possesses sufficient dynamic activity to generate dense event clouds. Finally, to guarantee acoustic saliency , we separate the audio tracks and discard samples with the Signal-to-Noise Ratios (SNR $<$ 15 dB) or effective volumes lower than 20 dBFS, resulting in clean and audio-clear subsets.

Data Augmentation. We design a stochastic augmentation strategy in stage 2 of Figure 3 to mitigate simulation limitations, preventing the network from overfitting to the determinism of the simulator and bridging the domain gap between synthetic events and noisy real-world recordings. Note that we calibrate the basic positive and negative contrast thresholds of V2E to match the empirically

measured nominal characteristics of DAVIS346 under typical indoor lighting conditions. Specifically, we apply Polarity Reversal with a probability of $p = 0.5$. To model temporal uncertainty in event sensing, we introduce Timestamp Jitter, injecting Gaussian noise ($\mu = 0$, $\sigma = 20\,\mu$s) into the event timestamps. Furthermore, we implement Spatiotemporal Event Drop by randomly masking 10%–20% of active events to simulate sensor dead pixels and transmission losses, forcing the model to learn robust acoustic features from incomplete visual cues.

For the overall FoleyBench evaluation, we adopt a standard 80/20% split for both Synth and Real, using 80% of the samples in each subset for model training and the remaining 20% for inference.

5 Experiment and Results

5.1 Implementation

Datasets and Training. We conduct experiments on three complementary data sources: VGGSound (Chen et al., 2020a), Greatest Hits (Owens et al., 2016), and our collected FoleyBench-Real. The former two datasets constitute FoleyBench-Synth, for which event streams are generated through our processing pipeline. We further evaluate on FoleyBench-Real captured by our device in the real world to validate the Sim-to-Real generalization. We train MicroFoley from scratch using PyTorch Distributed Data Parallelism (Li et al., 2020) on 2 NVIDIA H100 (80GB) GPUs. We leverage BF16 mixed-precision training (Micikevicius et al., 2017) for memory-efficient and stable optimization. All trainable components are optimized using the AdamW optimizer (Loshchilov & Hutter, 2017) ($\beta_1 = 0.9, \beta_2 = 0.999$) with an effective batch size of 128 over a total of 100 epochs. We initialize the base learning rate at 2×10^{-5} and subsequently adjust it via a cosine annealing scheduler (Loshchilov & Hutter, 2016) following a 10-epoch linear warm-up. Regarding the baseline training procedures, we adhere to their official dataset partitioning protocols, and all models are retrained under their respective settings for a fair comparison. More details can be found in Appendix A.

Compared Models. We compete MicroFoley with recent state-of-the-art V2A methods covering complementary generation and synchronization paradigms. Note that ground truth (GT) is additionally reported as a reference upper bound for results. TARO (Ton et al., 2025) builds upon flow-based transformers with timestep-adaptive representation alignment and onset-aware conditioning to improve audio fidelity and event-level synchronization. VAFlow (Wang et al., 2025b) directly models the video-to-audio transformation via cross-modality FM backbone, avoiding the traditional noise-to-audio transport paradigm. Omni2Sound (Dai et al., 2026) presents a unified framework for video-, text-, and video-text-conditioned audio generation with strong generalization across heterogeneous audio-generation settings. We further include the latest physics- and temporally-aware approaches: PAVAS (Hyun-Bin et al., 2026) incorporates estimated object-level physical parameters, including mass, motion trajectory, and velocity, into audio synthesis, while FoleyDirector (Li et al., 2026) introduces structured temporal scripts to provide fine-grained temporal guidance for V2A generation.

Evaluation Metrics. We evaluate performance across three dimensions: generated audio fidelity, semantic alignment, and physical synchronization. For *objective* evaluation, we report Fréchet Audio Distance (FAD) (Kilgour et al., 2018) and Kullback–Leibler divergence (KL) (Copet et al., 2023) to quantify the distribution distance and diversity between generated and reference audio, which all used by the common practice of PANNs (Kong et al., 2020b) as the audio classier. To measure cross-modal semantic consistency, we report ImageBind (IB) similarity (Girdhar et al., 2023) between video and generated audio, together with CLAP similarity (Elizalde et al., 2023) to calculate the semantic consistency of the generated audio compared with reference. Moreover, to assess temporal alignment, we compute the F1 (Li et al., 2026) score by matching detected acoustic onsets with reference acoustic onsets to evaluate temporal synchronization. For *subjective* evaluation, we conduct Mean Opinion Scores (MOS) on the audio fidelity (MOS-F) and the temporal synchronization (MOS-S) by recruiting 35 listeners. More subjective details are provided in Appendix D.

5.2 Comparison of Benchmarking Results

Performance on FoleyBench-Synth. On the synthetic subsets (see Section 4), MicroFoley achieves consistently strong performance across both objective and subjective metrics. On audio fidelity and semantic alignment, our method respectively obtains competitive FAD scores of 1.59 and 1.58 on

Table 2: Benchmarking comparison on VGGSound, Greatest Hits, and FoleyBench-Real. FAD/KL measure audio fidelity, IB/CLAP assess semantic alignment, F1 evaluates onset synchronization, and MOS-F/S report perceptual audio fidelity and synchronization as mean ± 95% confidence interval. The best and second-best results are respectively highlighted in **bold** and underline.

Method	FAD↓	KL↓	IB↑	CLAP↑	F1↑	MOS-F↑	MOS-S↑
VGGSound							
GT	–	–	7.58	–	–	4.70±0.05	4.73±0.05
TARO (Ton et al., 2025)	1.72	0.94	7.16	0.347	0.824	4.10±0.09	4.22±0.08
VAFlow (Wang et al., 2025b)	1.66	0.91	7.13	0.349	0.764	4.13±0.09	3.94±0.10
Omni2Sound (Dai et al., 2026)	1.62	0.87	**7.35**	**0.364**	0.790	4.21±0.07	4.04±0.09
PAVAS (Hyun-Bin et al., 2026)	**1.55**	**0.83**	7.30	0.359	0.814	**4.27±0.07**	4.16±0.08
FoleyDirector (Li et al., 2026)	1.69	0.90	7.21	0.352	0.829	4.16±0.08	4.29±0.07
MicroFoley (Ours)	1.59	0.85	7.32	0.361	**0.838**	4.24±0.07	**4.36±0.06**
Greatest Hits							
GT	–	–	7.55	–	–	4.74±0.05	4.77±0.04
TARO (Ton et al., 2025)	1.71	0.92	7.10	0.344	0.831	4.05±0.09	4.27±0.08
VAFlow (Wang et al., 2025b)	1.79	0.97	7.02	0.339	0.752	3.96±0.10	3.86±0.11
Omni2Sound (Dai et al., 2026)	1.66	0.88	**7.30**	**0.359**	0.794	4.18±0.08	4.04±0.09
PAVAS (Hyun-Bin et al., 2026)	1.61	**0.82**	7.28	0.357	0.822	**4.28±0.07**	4.18±0.08
FoleyDirector (Li et al., 2026)	1.68	0.89	7.18	0.350	0.837	4.13±0.08	4.32±0.07
MicroFoley (Ours)	**1.58**	0.83	7.29	0.358	**0.844**	4.25±0.07	**4.40±0.06**
FoleyBench-Real							
GT	–	–	7.82	–	–	4.77±0.04	4.80±0.04
TARO (Ton et al., 2025)	1.98	1.01	7.05	0.336	0.762	3.91±0.10	4.01±0.10
VAFlow (Wang et al., 2025b)	2.13	1.08	6.94	0.331	0.708	3.82±0.11	3.73±0.11
Omni2Sound (Dai et al., 2026)	1.87	0.93	7.34	0.350	0.741	4.08±0.09	3.92±0.10
PAVAS (Hyun-Bin et al., 2026)	1.76	0.88	7.39	0.355	0.790	**4.33±0.07**	4.11±0.09
FoleyDirector (Li et al., 2026)	1.90	0.96	7.20	0.346	0.801	4.05±0.09	4.18±0.08
MicroFoley (Ours)	**1.47**	**0.79**	**7.52**	**0.378**	**0.885**	4.31±0.06	**4.43±0.06**

VGGSound and Greatest Hits, while maintaining strong IB and CLAP similarities. More importantly, MicroFoley achieves the highest F1 scores of 0.838 and 0.844, together with the best MOS-S scores of 4.36 and 4.40. These results have demonstrated that continuous event-guided conditioning can effectively preserve fine-grained motion cues for temporally coherent audio synthesis. We further visualize the generated Mel-spectrograms of results in Figure 4 with selective samples.

Sim-to-Real Generalization. To evaluate the V2A domain gap, we train and evaluate all models on FoleyBench-Real. MicroFoley demonstrates strong robustness under real-world neuromorphic observations, achieving the best FAD (1.47), KL (0.79), IB (7.52), and CLAP (0.378). Notably, it obtains the highest F1 score of 0.885 and MOS-S of 4.43, while maintaining a competitive MOS-F of 4.31. These consistent improvements indicate that the disentangled semantic and physical streams enable MicroFoley to retain reliable micro-motion cues under appearance variations and domain shifts, yielding more robust physical and temporal alignment in real-world V2A generation.

5.3 Ablations

Event-Guided Cross-Modal Fusion. As demonstrated in Table 3, relying solely on RGB semantics via frozen CLIP module (CLIP Only) suffers from temporal aliasing, inherently limiting F1 performance correlated with audio-visual synchronization. Conversely, exclusively using micro-motions (Event Only) captures impact timings but lacks sufficient acoustic semantics, leading to degraded FAD and CLAP. Furthermore, the naive concatenation (Hybrid + Naive Concat) cannot fully establish fine-grained spatiotemporal correspondences between the two modalities. Our Event-Guided Cross-Attention (+ Cross-Attn.) can effectively address this by using high-frequency event micro-motions as queries to dynamically retrieve material semantics, bridging their representation gap.

Physical Causality & Dynamics. Table 4 indicates that the Transformer backbone without our Adaptive Silence Injector (w/o ASI) frequently generates spurious sounds in the absence of salient physical motion. The Adaptive Silence Injector (+ ASI) acts as a soft Physical Causality Gate, guid-

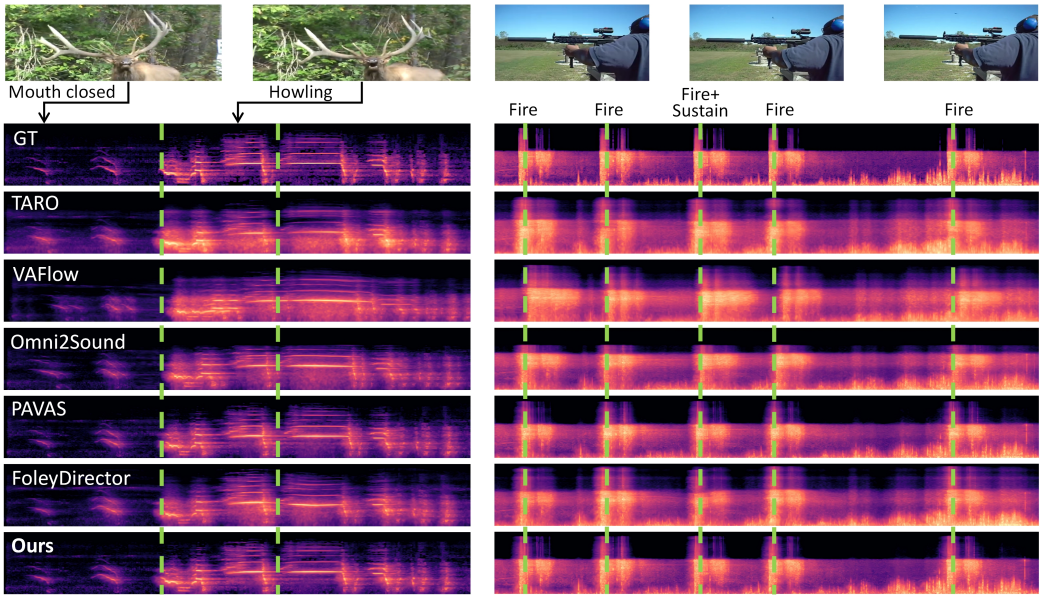


Figure 4: Qualitative comparison of generated audio using Mel-spectrogram. Proposed framework can better preserve transients and sustained structures, yielding temporal patterns aligned with GT.

Table 3: Impact of the modalities and fusion.

Variants	FAD↓	CLAP↑	F1↑
CLIP Only (RGB)	2.24	0.344	0.716
Event Only	2.09	0.318	0.836
Hybrid + Naive Concat	1.73	0.356	0.851
+ Cross-Attn. (Ours)	**1.47**	**0.378**	**0.885**

Table 4: Impact of the Bi-Mamba component.

Variants	FAD↓	CLAP↑	F1↑
Transformer w/o ASI	2.15	0.339	0.774
+ ASI	1.88	0.351	0.817
+ Uni-Mamba	1.64	0.365	0.851
+ Bi-Mamba (Ours)	**1.47**	**0.378**	**0.885**

ing the model toward a silence prior when acoustic activation is weak, thereby suppressing spurious activations and improving both FAD and F1. Moreover, transitioning from a Unidirectional Mamba (+ Uni-Mamba) to our complete Bidirectional Mamba (+ Bi-Mamba) architecture consistently improves audio fidelity and temporal synchronization. Further ablations is reported in Table 14.

Various Training Objectives. Table 5 shows that relying solely on the base CFM objective ($\mathcal{L}_{\text{flow}}$) enables generation but yields limited perceptual quality and cross-modal consistency. We progressively address this limitation by introducing a perceptual loss (+ $\mathcal{L}_{\text{perc}}$) and Cross-Modal Material Distillation (+ $\mathcal{L}_{\text{CMMD}}$), which steadily improve FAD and CLAP by aligning generated acoustics with visual semantics and material properties. We further establish Uniform Temporal Alignment (+ $\mathcal{L}_{\text{UTA}}$) as an ablation baseline that assigns equal importance to all temporal regions, substantially improving F1 score but remaining insensitive to transient energy variations. Therefore, PACA (+ $\mathcal{L}_{\text{PACA}}$) emphasizes micro-dynamic energy peaks as critical physical activation points, further improving both audio fidelity and temporal synchronization within our strategy.

Table 5: Impact of the optimization objectives.

Objectives	FAD↓	CLAP↑	F1↑
$\mathcal{L}_{\text{flow}}$	2.21	0.316	0.738
+ $\mathcal{L}_{\text{perc}}$	1.97	0.348	0.751
+ $\mathcal{L}_{\text{CMMD}}$	1.76	0.364	0.765
+ $\mathcal{L}_{\text{UTA}}$	1.62	0.369	0.836
+ $\mathcal{L}_{\text{PACA}}$ (Ours)	**1.47**	**0.378**	**0.885**

6 Conclusion

We address a fundamental bottleneck in V2A generation: temporal aliasing in the standard RGB video with frames. We present MicroFoley, a physically grounded framework that leverages event streams as a micro-motion bridge to align audio synthesis with high-frequency visual transients. By combining decoupled material-motion representations with Bi-Mamba, OT-CFM, and peak-aware alignment, MicroFoley improves both acoustic fidelity and temporal synchronization beyond traditional semantic-driven approaches. We further construct FoleyBench, a large-scale Sim-to-Real benchmark pairing audio-visual data with synthetic and real-world event streams for V2A. Finally, we advocate physically grounded micro-dynamic modeling for multimodal generation in the future.

AI USE STATEMENT

In this work, generative AI tools were used to assist with writing and language polishing, including grammar checking, wording refinement, and table presentation. AI-assisted tools were also used to support literature retrieval, as well as selected research-related tasks such as discussing experimental settings, refining methodological descriptions, and improving the efficiency of experimental planning and workflow. In addition, generative AI tools assisted with summarizing experimental observations and checking manuscript formatting and template compliance. All AI-assisted outputs were carefully reviewed, verified, and revised by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

This work follows standard research ethics practices in dataset construction, model development, and human evaluation or participation. All datasets were collected and processed in accordance with their applicable usage conditions, with appropriate consideration of privacy and security. For subjective evaluation or participation, all human were informed of the study purpose, procedure, and use of their responses, and participated voluntarily with informed consent; no sensitive personal information was collected. While the proposed method is intended for benign audio-visual generation research, potential misuse may include the creation of misleading or inauthentic audiovisual content. We therefore emphasize responsible use, transparent reporting, and appropriate human oversight, and take full responsibility for the ethical conduct and reporting of this work.

REPRODUCIBILITY STATEMENT

We provide comprehensive materials to facilitate the reproduction and verification of our work. The overall MicroFoley architecture, conditioning mechanisms, and optimization objectives are described in Section 3, while the construction, curation, augmentation, and train/test protocol of FoleyBench are detailed in Section 4. The main training setup, compared methods, and evaluation metrics are reported in Section 5. Additional implementation details are provided in Appendix A, including model configurations, parameter settings, computational profiling, and training objectives. Appendix B further presents the methodological foundations of our framework, including key formulations, derivations, and underlying concepts. Appendix C provides extended technical discussions and background information to facilitate a more comprehensive understanding of the proposed approach. The complete subjective evaluation protocol, including participant setup, rating criteria, and statistical analysis, is documented in Appendix D, while Appendix E reports additional experimental results and controlled ablation studies for the major design choices. Further discussions and supplementary qualitative examples are included in Appendices F and G. In addition, the main code and audiovisual demonstrations are available through the anonymous project website and supplementary materials, supporting reproduction and further investigation of our work.

REFERENCES

Mahmoud Afifi, Konstantinos G Derpanis, Bjorn Ommer, and Michael S Brown. Learning multi-scale photo exposure correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9157–9167, 2021.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3479–3487, 2015.

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

Lin Chen, Jingping Fang, Hairui Liu, Chenyang Xu, Junhao Chen, Xiaorui Li, Weidong Cai, and Xiaoming Chen. A first exploration of neuromorphic ot-cfm for multi-speaker vsr. In *European Conference on Computer Vision*, pp. 133–151. Springer, 2026.

Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating visually aligned sound from videos. *IEEE Transactions on Image Processing*, 29:8292–8302, 2020b.

Yutian Chen, Shi Guo, Fangzheng Yu, Feng Zhang, Jinwei Gu, and Tianfan Xue. Event-based motion magnification. In *European Conference on Computer Vision*, pp. 428–444. Springer, 2024.

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)*, pp. 18770–18781. IEEE, 2025.

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

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.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2026.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

Berthy T Feng, Alexander C Ogren, Chiara Daraio, and Katherine L Bouman. Visual vibration tomography: Estimating interior material properties from monocular video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16210–16219. IEEE, 2022.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2022.

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3583–3592. IEEE, 2020.

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)*, pp. 15180–15190. IEEE, 2023.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1312–1321. IEEE, 2021.

Jaesung Huh, Jacob Chalk, Evangelos Kazakos, Dima Damen, and Andrew Zisserman. Epic-sounds: A large-scale dataset of actions that sound. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):9953–9965, 2025.

Oh Hyun-Bin, Yuhta Takida, Toshimitsu Uesaka, Tae-Hyun Oh, and Yuki Mitsufuji. Pavas: Physics-aware video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14481–14491, 2026.

Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18032–18042. IEEE, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020a.

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

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37:104180–104204, 2024.

Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13529–13539. IEEE, 2025.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

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

Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10254–10264. IEEE, 2021.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.

You Li, Dewei Zhou, Fan Ma, Fu Li, Dongliang He, and Yi Yang. Foleydirector: Fine-grained temporal steering for video-to-audio generation via structured scripts. *arXiv preprint arXiv:2603.19857*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation. In *2024 IEEE 34th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. IEEE, 2024.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Ryogo Niwa, Tatsuki Fushimi, Kenta Yamamoto, and Yoichi Ochiai. Live demonstration: Event-based visual microphone. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4054–4055. IEEE, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2405–2413, 2016.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

AN Rajagopalan et al. Improving robustness of semantic segmentation to motion-blur using class-centric augmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10470–10479. IEEE, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. IEEE, 2022.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10219–10228. IEEE, 2023.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Tomás Soucek and Jakub Lokoc. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11218–11221, 2024.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European conference on computer vision*, pp. 252–268. Springer, 2018.

Tri Ton, Ji Woo Hong, and Chang D Yoo. Taro: Timestep-adaptive representation alignment with onset-aware conditioning for synchronized video-to-audio synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14228–14237. IEEE, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16150–16159. IEEE, 2021.

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric nonlinear flow and multi-scale fusion. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17734–17743. IEEE, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15492–15501, 2024a.

Juncheng Wang, Chao Xu, Cheng Yu, Lei Shang, Zhe Hu, Shujun Wang, and Liefeng Bo. Synchronized video-to-audio generation via mel quantization-continuum decomposition. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3111–3120. IEEE, 2025a.

Xihua Wang, Xin Cheng, Yuyue Wang, Ruihua Song, and Yunfeng Wang. Vaflow: Video-to-audio generation with cross-modality flow matching. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11777–11786. IEEE, 2025b.

Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, and Shuaicheng Liu. Low-light image enhancement with illumination-aware gamma correction and complete image modelling network. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13082–13091. IEEE, 2023.

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, 2024b.

Hao Yin, Shi Guo, Xu Jia, Xudong Xu, Lu Zhang, Si Liu, Dong Wang, Huchuan Lu, and Tianfan Xue. Evmic: Event-based non-contact sound recovery from effective spatial-temporal modeling. *arXiv preprint arXiv:2504.02402*, 2025.

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

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3550–3558. IEEE, 2018.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, and Yonghong Tian. Event-based video reconstruction via potential-assisted spiking neural network. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3584–3594. IEEE, 2022.

A Model Configurations and Architecture Details

To provide a clear and reproducible overview of MicroFoley, we condense the overall architectural hyperparameter, computational profiling, and detailed training configurations in this section. The model approximately comprises 172.2M trainable parameters dedicated to task-specific audio generation utilized events, alongside 85M frozen CLIP ViT-B/32 module (Radford et al., 2021).

A.1 Configurations and Hyperparameter

Table 6 shows the layer-wise configurations and parameter distribution across the key modules. The architecture is unified around a core hidden dimension of $D = 768$, enabling seamless feature alignment across the temporal (Bi-Mamba) and generative (DiT) backbones without necessitating heavy projection bottlenecks.

Table 6: Key model configurations and parameter breakdown of proposed framework. D denotes hidden dimension, H denotes attention heads, and C denotes channels.

Module	Component	Blocks/Layers	Core Configuration	Trainable	Frozen
I. Neuromorphic	VoxelGridEncoder	3	3D Conv, kernel $3 \times 3 \times 3$, $C = \{64, 128, 256\}$	6.2M	0
	EventRefine	1	$1 \times 1 \times 1$ Conv + SiLU, $C = 256$	1.5M	0
	NoiseFilter	1	Spatial-Channel Attention, $D = 128$	0.8M	0
II. Semantic Front-end	CLIP ViT-B/32	12	Transformer, $D = 768$, $H = 12$	0	85.0M
	Adapter & Predictor	2 / 2	Bottleneck MLP ($768 \to 256 \to 768$) / Linear	3.5M	0
III. Dynamic Stream	SRP-ResNet	4	2D Conv + ResBlocks, $C = 256$, Dual-head	9.8M	0
	Adaptive Silence Inj.	2	MLP + Learnable Vector $\mathbf{s}$, $\alpha \in [0, 1]$	2.7M	0
IV. Fusion & Seq.	EG-CA	1	Cross-Attention, $D = 768$, $H = 8$	8.5M	0
	Bi-Mamba	4	Bi-directional SSM, $D = 768$	17.2M	0
V. Generation	OT-CFM DiT	12	Transformer + AdaLN-Zero, $D = 768$, $H = 12$	105.0M	0
	Embeddings	-	Timestep (Sinusoidal) & Condition Projections	3.0M	0
VI. Waveform Recon.	VAE Decoder	4	MLP, Channel Expansion ($80 \to 512 \to 1024 \to 80$)	5.5M	0
	HiFi-GAN Vocoder	4	TransConv + MRF (V1-base config)	8.5M	0
Total	MicroFoley	-	-	172.2M	85.0M

A.2 Computational Profiling

Table 7 reports the inference efficiency on 2 NVIDIA H100 (80GB) GPUs. Benefiting from the linear $O(T)$ complexity of Bi-Mamba and the immense Tensor Core throughput of the H100s, the model achieves a 5.2-second latency for a 10-second video (250 frames at 25 fps) with 20 ODE sampling steps in the DiT backbone (Peebles & Xie, 2023).

A.3 Training Loss and Objectives

Loss Formulation. The entire architecture is jointly optimized end-to-end using a compound objective function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CFM}} + \mathcal{L}_{\text{PACA}} + 0.1\mathcal{L}_{perc} + 0.1\mathcal{L}_{\text{CMMD}} \tag{7}$$

where $\mathcal{L}_{\text{CFM}}$ denotes the OT-CFM loss, $\mathcal{L}_{\text{PACA}}$ represents the Peak-Aware Contrastive Alignment objective that emphasizes generation fidelity around transient physical impacts driven by event-

Table 7: Inference profiling (FLOPs, Latency, and Memory) for a 10-second Video (250 frames).

Module	FLOPs (G)	Latency (ms)	Mem. (MB)	Overall Efficiency
I. Neuromorphic	45.0	45	650	3D spatiotemporal processing
II. Semantic Front-end	437.5	80	1200	CLIP sequential encoding
III. Dynamic Stream	15.5	35	320	Spatial motion processing
IV. Fusion & Seq.	8.5	60	450	Cross-attention & Bi-Mamba
V. Generation	1850.0	4900	1850	DiT + 20-step ODE solver
VI. Waveform Reconstruction	35.0	80	580	Mel-to-waveform upsampling
Total	2391.5	5200	5050	1.9× faster-than-real-time

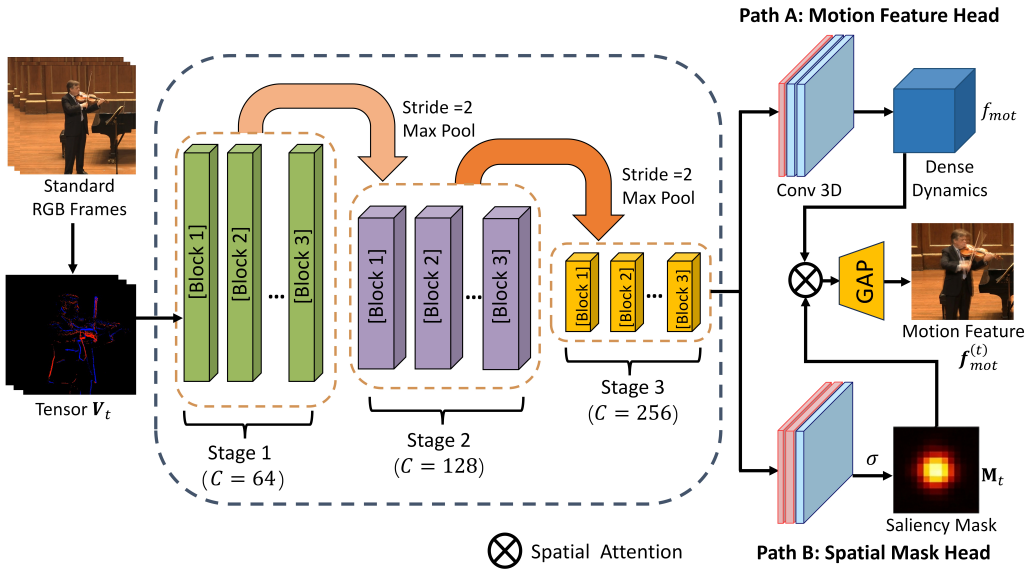


Figure 5: Architecture of SRP-ResNet. Given voxelized event streams $\mathbf{V}_t$, SRP-ResNet employs a three-stage pyramidal hierarchy to extract multi-scale spatiotemporal features with progressively expanded channels ($C = 64 \rightarrow 128 \rightarrow 256$). The backbone branches into a *Motion Feature* Head for dense dynamics $\mathbf{f}_{mot}$ and a *Spatial Mask* Head that predicts a sigmoid saliency mask $\mathbf{M}_t$. Spatially gated motion features are then aggregated via global average pooling (GAP) to obtain $\mathbf{f}_{mot}^{(t)}$, enabling more precise localization of sound-producing motion followed by He et al. (2016).

energy peaks, $\mathcal{L}_{\text{CMMD}}$ enforces visual-acoustic material alignment, and $\mathcal{L}_{perc}$ provides semantic perceptual gradients from a frozen CLAP encoder, supplying for the final loss equation (6).

B Methodological Details

B.1 Event Representation Formulation

Event cameras asynchronously encode pixel-wise brightness variations rather than absolute image intensities (Gallego et al., 2022). Compared with traditional RGB cameras, it can capture full image frames at fixed temporal intervals, event sensors independently respond to local intensity changes at each pixel, producing a sparse stream of asynchronous events with high temporal sensitivity. Following the standard event generation model (Hu et al., 2021), each event e_k is represented as:

$$e_k = (x_k, y_k, t_k, p_k), \tag{8}$$

where (x_k, y_k) denotes the spatial location of the k-th event, t_k is the event timestamp, $p_k \in \{-1, +1\}$ denotes the event polarity, and k indexes individual events in the asynchronous event stream. The polarity indicates the direction of the local brightness change, distinguishing whether the logarithmic intensity increases or decreases at the event location.

Let

$$L(x, y, t) = \log I(x, y, t), \tag{9}$$

where $I(x, y, t)$ denotes the image intensity at pixel location (x, y) and time t, and $L(x, y, t)$ represents the corresponding logarithmic intensity. The logarithmic formulation provides a relative measure of brightness variation and is widely adopted in event-based sensing to model changes across different illumination levels (Gallego et al., 2022). For each pixel, the sensor maintains a memorized reference intensity and continuously monitors the accumulated deviation from this reference. An event is triggered once the magnitude of this change exceeds the corresponding contrast threshold p_k, defined as:

$$p_k = \begin{cases} +1, & L(x_k, y_k, t_k) - L_{\text{mem}}(x_k, y_k) \geq C_+, \\ -1, & L(x_k, y_k, t_k) - L_{\text{mem}}(x_k, y_k) \leq -C_-, \end{cases} \tag{10}$$

where $L_{\text{mem}}(x, y)$ denotes the memorized logarithmic intensity at pixel (x, y) from the most recent event update, and $C_+ > 0$ and $C_- > 0$ are the positive and negative contrast thresholds. Accordingly, $p_k = +1$ indicates an increase in brightness, whereas $p_k = -1$ indicates a decrease. After each threshold crossing, $L_{\text{mem}}(x, y)$ is updated to reflect the new reference level, allowing subsequent events to encode further temporal variations at the same pixel. This mechanism converts continuous visual changes into a sparse sequence of polarity-aware responses concentrated around active motion regions.

Concretely, for FoleyBench-Synth, temporally interpolated RGB observations are converted into asynchronous event sequences by maintaining a memorized logarithmic intensity for each pixel and triggering events through successive contrast-threshold crossings. The resulting polarity-aware events are then aggregated within short temporal intervals into voxel grids $\mathbf{V}_t$, preserving sparse spatial activity and fine-grained temporal variations while providing a regular tensor representation for convolutional processing. These voxelized event features are subsequently fed into SRP-ResNet to extract localized motion-sensitive cues for downstream physical-acoustic modeling.

B.2 Saliency-Guided Motion Representation Aggregation

Given the voxelized tensor $\mathbf{V}_t$, the proposed SRP-ResNet first extracts a shared spatiotemporal feature S_t, which is further processed by two parallel heads to obtain the dense motion feature F_t and the spatial saliency mask $\mathbf{M}_t$ as shown in Figure 5. The feature extraction and saliency-guided modulation process is formulated as:

$$S_t = F_{\text{SRP}}(\mathbf{V}_t), \qquad F_t = H_{\text{mot}}(S_t), \qquad \mathbf{M}_t = \sigma(H_{\text{mask}}(S_t)), \qquad \widetilde{F}_t = \mathbf{M}_t \odot F_t, \tag{11}$$

where $F_{\text{SRP}}(\cdot)$ denotes the SRP-ResNet feature extractor that maps the voxelized event tensor into shared spatiotemporal features. $H_{\text{mot}}(\cdot)$ and $H_{\text{mask}}(\cdot)$ represent the motion feature head and spatial saliency mask head, respectively. Specifically, S_t denotes the shared feature map extracted from the event representation, $F_t \in \mathbb{R}^{C\times H\times W}$ represents the dense motion feature containing localized dynamic responses, and $\mathbf{M}_t$ denotes the spatial saliency mask generated through the sigmoid activation $\sigma(\cdot)$. The operator $\odot$ indicates element-wise multiplication, where $\mathbf{M}_t$ is broadcast along the channel dimension to modulate the motion feature map. This saliency-guided modulation emphasizes acoustically relevant event responses while suppressing irrelevant background activities.

The resulting gated feature map $\widetilde{F}_t$ is further aggregated along the spatial dimensions to obtain the frame-wise motion representation. Specifically, Global Average Pooling as $\text{GAP}(\cdot)$ independently aggregates each feature channel, where c denotes the index of the c-th channel among the total C channels. The corresponding channel-wise response $\mathbf{f}_{\text{mot}}^{(t)}(c)$ is computed as:

$$\mathbf{f}_{\text{mot}}^{(t)}(c) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \widetilde{F}_t(c, i, j) = \text{GAP}\left(\widetilde{F}_t\right), \tag{12}$$

where C denotes the number of feature channels, $c \in \{1, \ldots, C\}$ indexes an individual channel, and H and W represent the spatial dimensions of the feature map. The indices i and j denote the spatial coordinates within each feature channel. The above operation produces the c-th element of the motion representation vector $\mathbf{f}_{\text{mot}}^{(t)} \in \mathbb{R}^C$ by averaging spatial responses within the corresponding channel. By collecting all channel-wise responses over c, the complete motion representation $\mathbf{f}_{\text{mot}}^{(t)}$ is obtained, which aggregates spatially distributed event responses into a compact feature vector while preserving dominant micro-motion patterns.

The resulting $\mathbf{f}_{\text{mot}}^{(t)}$ summarizes acoustically relevant event-driven micro-motion and is subsequently fed into the Adaptive Silence Injector. Although GAP itself is a standard feature aggregation operation, our design performs spatial aggregation after event-driven saliency gating, thereby reducing the influence of background event responses and producing a reliable motion condition for V2A.

C Extended Clarifications

This section provides additional related background and technical clarification to complement the main paper, with the aim of offering a further understanding of our work motivations, design choices, and underlying mechanisms of MicroFoley. In particular, we further elaborate on several aspects that are only briefly discussed in the main text due to space constraints, including event-based representations, temporal modeling, material-aware acoustic reasoning, generative optimization, and silence modeling. These extended discussions provide supplementary context for interpreting the proposed framework and its relation to existing methodologies.

C.1 Event Streams vs. Standard RGB

Temporal Quantization Bottleneck. Existing V2A models (Wang et al., 2025b; Cheng et al., 2025) predominantly rely on dense RGB frames (e.g., 25-30 fps), imposing a rigid temporal quantization limit of $\Delta t \approx 40$ms. This leads to severe motion blur and absolute temporal blind spots during rapid impacts, causing RGB pipelines to produce muffled or desynchronized acoustic onsets. Event streams can operate with microsecond precision ($\approx 100\mu$s), acting as an asynchronous high-pass filter that strictly isolates dynamic kinetic energy from static backgrounds (Gallego et al., 2022).

V2E Simulation for Inter-Frame Dynamics. To leverage these high-frequency advantages on standard video datasets, we employ the V2E simulator (Hu et al., 2021). Unlike standard frame differencing, our VFI-to-V2E pipeline estimates intermediate visual states between consecutive low-frame-rate RGB frames and converts their logarithmic brightness changes into simulated events. The resulting event streams provide temporally dense and sparse representations of inter-frame motion, offering fine-grained physical priors for micro-motion-aware audio generation. Nevertheless, these events remain estimates inferred from the source frames rather than direct observations of physical events unrecorded by the original video.

SRP-ResNet and Sub-Pixel Reconstruction. While V2E provides a temporally densified motion representation, spatial precision is recovered via our SRP-ResNet module illustrated in Figure 5. By processing the accumulated event voxel grids through a specialized residual network equipped with *Pixel Shuffle* (sub-pixel convolution) layers, this module learns to upscale feature maps into high-fidelity *sub-pixel frames*. The mechanism allows the network to resolve minute physical interactions across fabric that manifest as fractional pixel shifts. By explicitly reconstructing these sub-pixel kinetic trajectories, MicroFoley captures the microscopic friction cues necessary for synthesizing crisp and texture-aligned Foley sounds that traditional RGB networks miss, as listed in Table 8.

Table 8: Comparison of visual modalities for foley sound generation. Event streams provide the necessary temporal resolution and sub-pixel sensitivity to micro-motion acoustics.

Modality	Temporal Resolution	Spatial Sensitivity	Acoustic Alignment
Dense RGB (Standard)	Macro ($\sim$ 40ms)	Pixel-level (Quantized)	Muffled / Desynchronized
Neuromorphic Events	Micro ($\sim 100\mu$s)	Sub-pixel Level	Crisp / High-Frequency

We further visualize three representative and challenging motion regimes in Figure 6: rapid large-displacement motion with brief local changes in badminton, dense repetitive hand–surface interactions in drumming, and irregular non-rigid dynamics induced by object–fluid coupling in water splash. Across these cases, event streams highlight localized temporal variations around active regions and provide denser dynamic cues than sparse RGB observations.

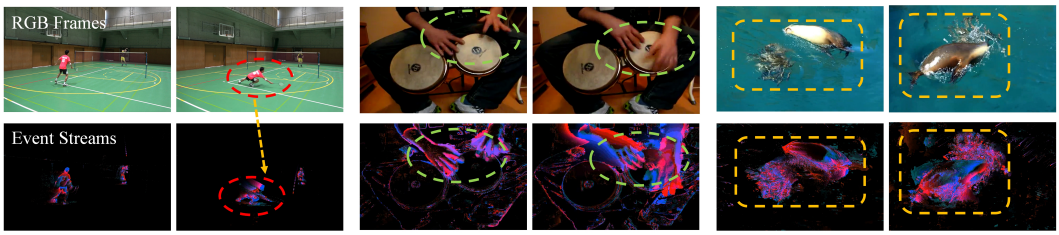


Figure 6: RGB and event-stream comparison across three challenging dynamic scenes. The dashed boxes highlight visually ambiguous regions around the primary sound-producing sources, where rapid motion, repeated contact, or non-rigid (liquid) dynamics are difficult to resolve from sparse RGB observations. In contrast, event streams preserve denser local temporal variations around these active regions. In the event representation, the red events indicate positive brightness changes, while the blue events indicate negative brightness changes.

C.2 Linear Bi-Mamba vs. Quadratic Transformers

In the Fusion & Sequence Modeling stage, temporal interactions are traditionally modeled via standard Transformers. For a sequence of length T, the self-attention mechanism imposes a quadratic computational bottleneck $O(T^2 \cdot D)$, prohibiting long-context modeling under heavy batch sizes.

As demonstrated in Table 9, MicroFoley integrates a *Bi-Mamba* architecture. By replacing the attention matrix with SSM blocks and parallel scan algorithms (Gu & Dao, 2023), Bi-Mamba reduces the complexity to strictly linear with respect to sequence, in contrast to quadratic cost of self-attention.

Table 9: Computational complexity and memory footprint comparison for sequence modeling. Specifically, T: sequence length, D: hidden dimension, N: SSM state dimension.

Architecture	Time Complexity	Memory Complexity	Receptive Field
Standard Transformer	$O(T^2 \cdot D)$	$O(T^2 + T \cdot D)$	Global (Bidirectional)
Causal Mamba	$O(T \cdot D \cdot N)$	$O(T \cdot D \cdot N)$	Causal (Unidirectional)
Bi-Mamba (Ours)	$O(T \cdot D \cdot N)$	$O(T \cdot D)$	Global (Bidirectional)

This architectural leap is precisely what allows our model to process sequences with an effective batch size of 128 on 2 H100 GPUs, offloading the saved FLOPs to the downstream generative backbone without sacrificing bidirectional temporal context.

C.3 Physics-Aware Material Taxonomy

The Semantic-Acoustic Ambiguity. In physical sound synthesis, visual appearance is frequently decoupled from acoustic properties. For instance, a clear plastic cup and a glass cup may share nearly identical visual embeddings in standard vision encoders, yet they produce drastically different acoustic resonances upon impact. Relying solely on implicit audio-visual alignment often leads to physically implausible audio generation.

Physics-Aware Categorization. To explicitly resolve this ambiguity, we meticulously curate a comprehensive 256-class material taxonomy. Rather than classifying generic objects, this taxonomy explicitly categorizes the underlying physical properties that dictate sound emission. As detailed in Table 10, our hierarchy spans from fundamental acoustic priors (e.g., rigid metals, yielding fabrics) to complex composite interactions and microscopic surface textures.

Guiding the CMMD Objective. Rather than introducing an additional material classification objective, we leverage existing material-aware priors from a pretrained material recognition model trained on the *Materials in Context Database* (Bell et al., 2015). The extracted material representation provides complementary visual cues related to acoustic properties and is aligned with the corresponding audio representation through the CMMD objective. This alignment encourages the visual condition to preserve material-related information relevant to sound generation, such as texture and timbral characteristics, without requiring additional material annotations or classification supervision.

Table 10: Overview of the 256-class material taxonomy on CMMD. The comprehensive categorization ensures physically plausible audio generation by grounding visual features.

Category Type	Representative Examples	Role in Semantic Alignment
Basic Materials	Wood, metal, glass, plastic, fabric, ceramic, paper, rubber	Establishes foundational acoustic resonance priors.
Composite Materials	Wooden furniture, metal doors	Captures complex sound mixtures from multi-material interactions.
Surface Properties	Smooth, rough, textured variants	Differentiates subtle friction and contact acoustic textures.
Contextual Objects	Musical instruments, kitchen utensils	Provides high-level semantic context for specialized Foley sounds.

C.4 OT-CFM vs. Standard Diffusion

Mainstream audio generation models (e.g., AudioLDM (Liu et al., 2023)) predominantly rely on Denoising Diffusion Probabilistic Models. While powerful, these models navigate complex, curved Gaussian noise trajectories during sampling, typically necessitating 100 to 1000 inference steps to resolve high-fidelity waveforms. This inherent inefficiency renders them impractical for real-time applications without heavy distillation. MicroFoley fundamentally shifts this paradigm by adopting OT-CFM. Rather than denoising, OT-CFM models a deterministic and straight-line probability path from the prior $\mathbf{x}_0$ to the data $\mathbf{x}_1$, formulated as follows:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1, \quad \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0. \tag{13}$$

By enforcing this mathematically optimal trajectory, we prevent the generative path from meandering. This stability allows our DiT backbone to utilize a standard Euler ODE solver to achieve superior generation quality in merely 20 integration steps, a $50\times$ speedup over traditional diffusion that directly enables our efficient latency.

C.5 Latent Silence Injection vs. Post-processing

A persistent artifact in V2A generation is the hallucination of background noise (e.g., static hiss or ambient hums) during visually silent segments. Prior works attempt to mitigate this via heuristic post-processing, such as hard energy thresholding. However, this often results in abrupt, unnatural acoustic clipping, cutting off natural reverberations or decay tails. MicroFoley resolves this at the latent level through the ASI. Leveraging the spatial attention mask $\mathbf{M}_t$ from the neuromorphic front-end, ASI predicts a frame-wise acoustic activation probability $\alpha_t \in [0, 1]$ and adaptively combines the motion representation $\mathbf{f}_{mot}^{(t)}$ with a learnable silence token $\mathbf{v}_\varnothing$, formulated as follows:

$$\mathbf{c}_{phys}^{(t)} = \alpha_t \mathbf{f}_{mot}^{(t)} + (1-\alpha_t)\mathbf{v}_\varnothing. \tag{14}$$

Compared with binary gating, this soft interpolation provides a smooth transition between active motion and the null acoustic state. The learnable silence token $\mathbf{v}_\varnothing$ serves as a latent silence prior, suppressing spurious acoustic activations while preserving natural temporal continuity.

D Subjective Evaluation Protocol

Given the nuanced nature of micro-motion Foley sounds, objective metrics alone cannot fully capture the perceptual quality and physical plausibility of the generated audio. To rigorously validate MicroFoley against baseline models, we meticulously designed a double-blind subjective evaluation protocol focusing on MOS-F and MOS-S with 1-5 Likert scales and 95% confidence intervals.

D.1 Evaluation Environment and Blind Test Setup

A critical challenge in subjective evaluation is the preservation of its 96kHz studio-quality acoustic characteristics. Standard laptop speakers or low-fidelity output devices inherently mask the high-frequency friction and transient collision cues that our neuromorphic architecture aims to synthesize. Therefore, all participants were strictly mandated to conduct the evaluation using professional closed-back headphones or high-fidelity in-ear monitors in a quiet indoor environment.

Subjective Evaluation Instructions

Please wear headphones for this task. Listen/Watch the samples carefully before rating. Choose the option that best matches your opinion and check the box next to it ✓.

Task 1: MOS-F

Definition: Assess the naturalness of the sound, the absence of artifact hallucinations, and whether the acoustic resonance correctly matches the visual material (e.g., distinguishing wood from metal).

Q: "*How realistic and physically plausible is the generated sound for the given visual material*?"

Scale:
5 Excellent: Indistinguishable from real studio recordings; material physics perfectly matched; zero artifacts.
4 Good: Highly realistic with accurate material representation, but contains imperceptible background noise.
3 Fair: Moderate robotic artifacts or slight material ambiguity (e.g., sounds generic rather than specific).
2 Poor: Severe distortion, overwhelming background noise, or completely incorrect material representation.
1 Bad: Completely unnatural noise, or severe hallucinations during visual silence.

Task 2: MOS-S

Definition: Evaluate the strict temporal alignment between visual kinematic impact or micro-motion and the corresponding acoustic onset.

Q: "*How well does the audio timing align with the visual motion or impacts in the video*?"

Scale:
5 Perfect: Acoustic onset aligns flawlessly with the visual micro-motion or impact.
4 Slight Delay/Advance: Synchronization is generally accurate, with delays only noticeable under extreme scrutiny.
3 Noticeable Offset: Evident lag or advance between the visual impact and the sound generation.
2 Poorly Aligned: Sound often plays during visually static moments; highly distracting.
1 Unrelated: Audio timing bears absolutely no physical relationship to the visual sequence.

Figure 7: The questionnaire interface and rating criteria. Specifically, Task 1 (MOS-F) focuses on the absolute quality, naturalness, and physical plausibility of the audio, while Task 2 (MOS-S) strictly assesses the temporal alignment between visual kinematic events and acoustic onsets.

We implemented a web-based, double-blind testing interface. For each trial, participants were presented with a muted 10-second video clip alongside randomly shuffled audio tracks generated by MicroFoley, competing baselines, and the ground-truth recording. This randomized A/B/n testing methodology ensures the complete elimination of presentation order bias and brand-name prejudice.

D.2 Participant Demographics and Quality Control

The evaluation was conducted with a pool of 35 human evaluators, comprising 15 professional audio engineers/Foley artists and 20 general participants with normal hearing. To guarantee the integrity of the collected data, we integrated stringent *Attention Checks* throughout the evaluation pipeline. Specifically, we embedded trap trials containing either pure white noise or severely out-of-sync audio (temporally shifted by ± 3 seconds). Subjective data from participants who failed to assign minimum scores to these explicit trap trials were automatically discarded.

The final MOS is computed with a 95% Confidence Interval (CI) to quantify statistical significance:

$$\text{MOS} = \frac{1}{N}\sum_{i=1}^{N} s_i, \quad \text{CI}_{95\%} = 1.96 \times \frac{\sigma}{\sqrt{N}}, \tag{15}$$

where N is the number of valid ratings, $s_i \in \{1, 2, 3, 4, 5\}$ is the assigned score, and σ is the standard deviation.

D.3 Grading Criteria and Evaluation Interface

To minimize subjective variance and standardize the evaluation scale, participants were provided with explicit grading anchors and evaluated the samples through a strictly controlled web interface. The evaluation isolates two orthogonal dimensions: MOS-F and MOS-S. Providing these explicit instructions, as illustrated in Figure 7, ensures high inter-rater reliability across different participants.

E Additional Experimental Results

Table 11: Overall benchmarking results on FoleyBench-Synth.

Method	FAD↓	KL↓	IB↑	CLAP↑	F1↑	MOS-F↑	MOS-S↑
GT	–	–	7.56	–	–	4.72±0.05	4.75±0.05
TARO (Ton et al., 2025)	2.14	1.10	7.13	0.317	0.821	4.02±0.09	4.20±0.08
VAFlow (Wang et al., 2025b)	2.09	1.07	7.08	0.320	0.754	4.06±0.09	3.93±0.10
Omni2Sound (Dai et al., 2026)	1.97	1.02	**7.33**	**0.341**	0.791	4.18±0.08	4.04±0.09
PAVAS (Hyun-Bin et al., 2026)	1.93	**0.94**	7.29	0.332	0.819	**4.25±0.07**	4.17±0.08
FoleyDirector (Li et al., 2026)	2.02	1.04	7.20	0.326	0.834	4.13±0.08	4.31±0.07
MicroFoley (Ours)	**1.88**	0.98	7.31	0.335	**0.842**	4.22±0.07	**4.38±0.06**

Overall FoleyBench-Synth Results. We further report the aggregate evaluation on FoleyBench-Synth in Table 11, complementing the main experiment results. Results show the aggregated results on FoleyBench-Synth. MicroFoley achieves the best FAD (1.88), F1 (0.842), and MOS-S (4.38), while remaining competitive in semantic alignment and perceptual audio fidelity. In particular, the strong F1 and MOS-S results further demonstrate the effectiveness of event-guided micro-motion modeling in preserving fine-grained temporal synchronization across diverse synthetic scenarios.

Table 12: Impact of the efficiency and NFE.

NFE	FAD↓	CLAP↑	F1↑	RTF↓
2 Steps	2.81	0.304	0.731	**0.05**×
5 Steps	2.08	0.338	0.806	0.12×
10 Steps	1.65	0.364	0.859	0.25×
20 Steps (Ours)	1.47	0.378	0.885	0.52×
50 Steps	**1.44**	**0.380**	**0.889**	1.25×
Δ vs. 20 Steps	*-0.03*	*+0.002*	*+0.004*	*+0.73*×

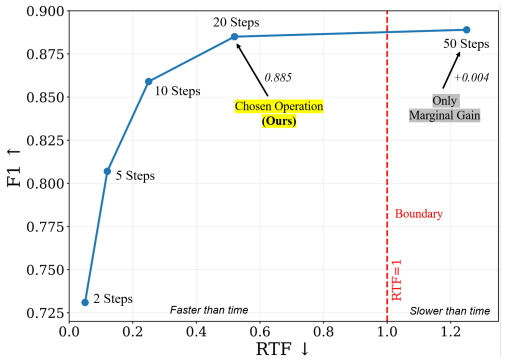
Inference Speed. As reported in Table 12, we evaluate the sampling efficiency of MicroFoley across different Number of Function Evaluations (NFE) settings. Traditional diffusion models typically require many iterative steps due to stochastic sampling along complex generation trajectories. In contrast, our decoder employs OT-CFM to construct a deterministic and near-straight transport path between noise and data distributions. We further evaluate the real-time factor (RTF) to quantify inference efficiency. MicroFoley achieves strong generation quality at only 20 steps with an RTF of 0.52×. Increasing to 50 steps yields only marginal gains in FAD, CLAP, and F1, while increasing the RTF to 1.25×, corresponding to an additional 0.73× inference difference relative to the 20-step setting. This demonstrates that 20 NFE provides a favorable quality-efficiency trade-off, enabling faster-than-real-time generation of ours.

To further characterize the convergence behavior, we analyze the quality–efficiency scaling and normalized marginal gains under different NFE settings in Figure 8. The performance increases rapidly with additional NFE and gradually saturates around 20 steps. The normalized marginal gain further indicates that most improvements are achieved during the early sampling stages, while increasing NFE beyond 20 steps brings only limited benefits. This trend demonstrates that OT-CFM enables efficient convergence without requiring excessive iterative refinement.

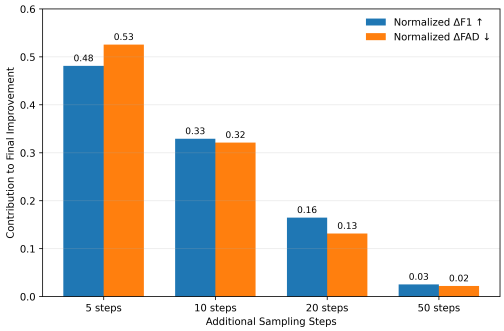
Table 13: Effectiveness of the SRP-ResNet component.

Variants	Gating	FAD↓	CLAP↑	F1↑
3D-CNN	×	2.23	0.326	0.752
3D-CNN	✓	1.96	0.339	0.801
SRP-ResNet	×	1.74	0.356	0.839
SRP-ResNet (Ours)	✓	**1.47**	**0.378**	**0.885**

SRP-ResNet Design. We ablate the impact of the core event backbone SRP-ResNet and the acoustic gating mechanism in Table 13. Voxelized events form sparse polarity grids, where informative responses are concentrated around rapid local intensity variations. These repre-



(a) Quality–efficiency trade-off.



(b) Normalized marginal gain.

Figure 8: Sampling efficiency analysis under different NFE step settings. (a) Quality–efficiency trade-off between generation quality and inference cost. (b) Normalized marginal improvements showing diminishing returns with increasing NFE. Overall, the analysis highlights that MicroFoley achieves an effective balance between generation quality and sampling efficiency.

sentations preserve fine-grained temporal changes and highlight dynamic regions beyond dense RGB observations. Replacing a standard 3D-CNN with SRP-ResNet substantially improves all metrics even without gating, reducing FAD from 2.23 to 1.74 and increasing F1 from 0.752 to 0.839. This indicates that the pyramidal residual representation better preserves sparse micro-dynamic structures instead of smoothing them into coarse motion responses. Independently, introducing the gating mechanism consistently benefits both backbones by selectively emphasizing acoustically active event responses. Combining SRP-ResNet with gating achieves the best FAD, CLAP, and F1 of 1.47, 0.378, and 0.885, demonstrating the complementary roles of micro-motion representation and physical activation modeling.

Further Bi-Mamba Analysis. To isolate the benefit of our temporal backbone from bidirectionality itself, Table 14 compares Uni-Mamba, Bi-Transformer, and Bi-Mamba under the same generation setting. In particular, Bi-Transformer provides only marginal gains over Uni-Mamba, with F1 increasing from 0.851 to 0.852 while CLAP increasing from 0.365 to 0.367. In contrast, Bi-Mamba consistently improves all metrics, respectively reducing FAD to 1.47 and increasing CLAP and F1 to 0.378 and 0.885. These results suggest that the improvement is not explained by bidirectional context alone; the state-space formulation better captures continuous acoustic evolution and fine-grained temporal dependencies in event sequences. This supports our use of Bi-Mamba for modeling continuous acoustic dynamics.

Table 14: Impact of the temporal modeling.

Variants	FAD↓	CLAP↑	F1↑
Uni-Mamba	1.64	0.365	0.851
Bi-Transformer	1.59	0.367	0.852
Bi-Mamba (Ours)	**1.47**	**0.378**	**0.885**

In this setting, Uni-Mamba removes the backward SSM branch, while Bi-Transformer replaces the bidirectional Mamba blocks with Transformer layers (Vaswani et al., 2017) under otherwise identical settings. Bi-Mamba models foley sounds through complementary onset and decay dynamics: the forward SSM captures acoustic attacks, while the backward SSM refines resonance tails. This phase-aware design better matches physical sound evolution than Uni-Mamba or Bi-Transformer.

Dense Optical Flow. To determine whether the improvement of MicroFoley arises merely from introducing dense motion information, we replace the V2E event pipeline with dense RAFT optical flow (Teed & Deng, 2020) at an equivalent temporal rate demonstrated in Table 15, while keeping the remaining framework unchanged. Dense optical flow reaches an F1 of 0.803, whereas the V2E representation substantially improves it to 0.885 while simultaneously reducing FAD from 1.96 to 1.47 and increasing CLAP from 0.333 to 0.378. Although optical flow effectively describes macroscopic displacement, it compresses local intensity variations into smooth motion fields and therefore loses fine-grained transient responses. In contrast, neuromorphic events preserve asynchronous local changes from rapid impacts and subtle interactions, providing fine-grained temporal cues for onset-aligned audio generation beyond standard RGB observations.

Table 15: Impact of the event representation.

Variants	FAD↓	CLAP↑	F1↑
Dense RAFT	1.96	0.333	0.803
V2E Pipeline (Ours)	**1.47**	**0.378**	**0.885**

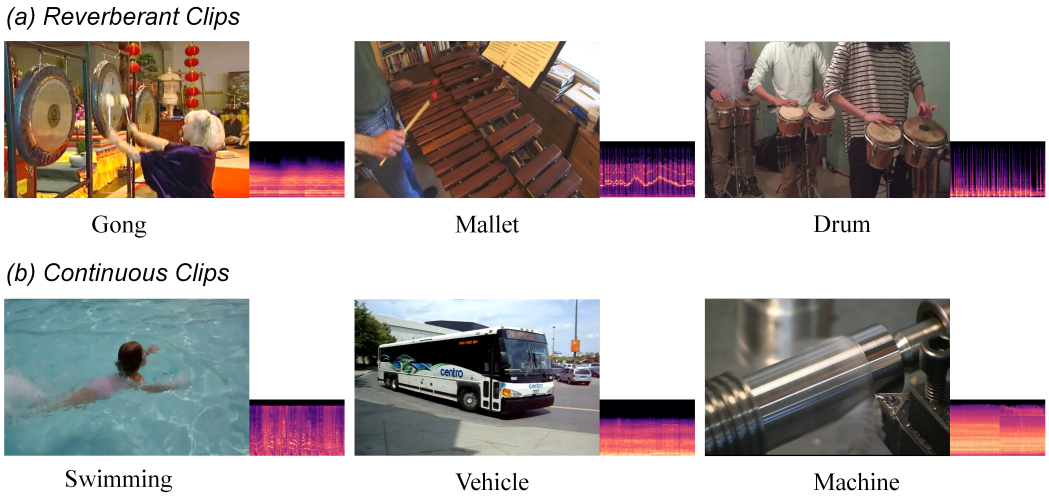


Figure 9: Samples from the class-stratified acoustic evaluation. *Reverberant* clips exhibit transient excitation followed by resonance decay, while *Continuous* clips show sustained acoustic evolution. Related Mel-spectrograms are provided to visualize their shared temporal characteristics, including sustained energy evolution and continuous frequency transitions.

Table 16: Impact of the acoustic classes.

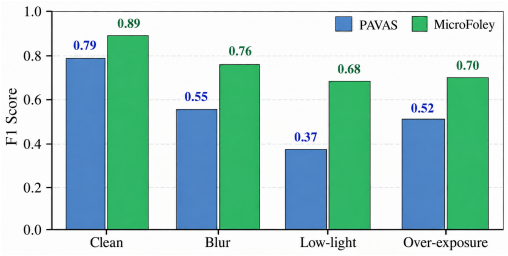
Variants	FAD↓	CLAP↑	F1↑
Reverberant			
w/o ASI	2.15	0.312	0.741
w/o PACA	1.88	0.345	0.792
Full (Ours)	**1.72**	**0.353**	**0.821**
Continuous			
w/o ASI	2.05	0.305	0.712
w/o PACA	1.92	0.337	0.753
Full (Ours)	**1.78**	**0.346**	**0.784**

Class-Stratified Sound Evaluation. We examine whether the proposed physical constraints remain effective beyond short impulsive impacts by stratifying VGGSound (Chen et al., 2020a) into two acoustically challenging subsets: *Reverberant* sounds with pronounced resonance tails and *Continuous* sounds such as liquids and friction, selected clips are shown in Figure 9. Quantitatively, Table 16 shows that removing ASI causes the largest degradation in both categories, with F1 dropping to 0.741 and 0.712, indicating that explicit physical activation remains important even when acoustic energy extends over time. Removing PACA also consistently decreases F1 and CLAP, confirming that transient-aware optimization contributes to temporal alignment without restricting the model to purely impulsive sounds. The complete model achieves the best performance on both reverberant and continuous subsets, respectively reaching F1 scores of 0.821 and 0.784. These results demonstrate that ASI and PACA complement each other across diverse temporal sound structures rather than being specialized only for isolated collision events.

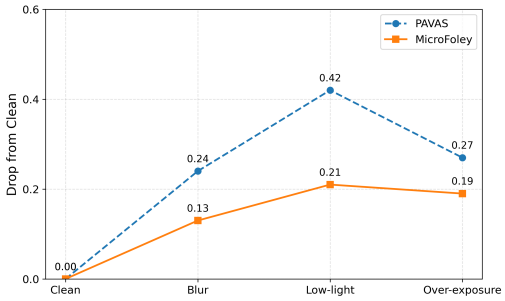
Significantly, ASI regulates the *current physical excitation* rather than directly suppressing the generated audio amplitude. When visible motion becomes weak, the gating probability decreases to prevent spurious new acoustic activations, while previously triggered interactions remain encoded in the temporal representation and can still propagate through the bidirectional sequence model and OT-CFM decoder. Consequently, reverberation and decay tails are preserved as context-dependent acoustic evolution instead of being abruptly truncated by the silence prior.

Visual Degradation Robustness. Event streams asynchronously capture local brightness changes with high temporal resolution, wide dynamic range, and reduced motion blur, making them well suited to fast motion and challenging illumination where RGB frames often degrade (Gallego et al., 2022). Motivated by these characteristics, we further evaluate whether the event-guided motion representation in MicroFoley provides more stable audio generation when the visual input is degraded by common real-world conditions, including *motion blur*, *low illumination*, and *over-exposure*.

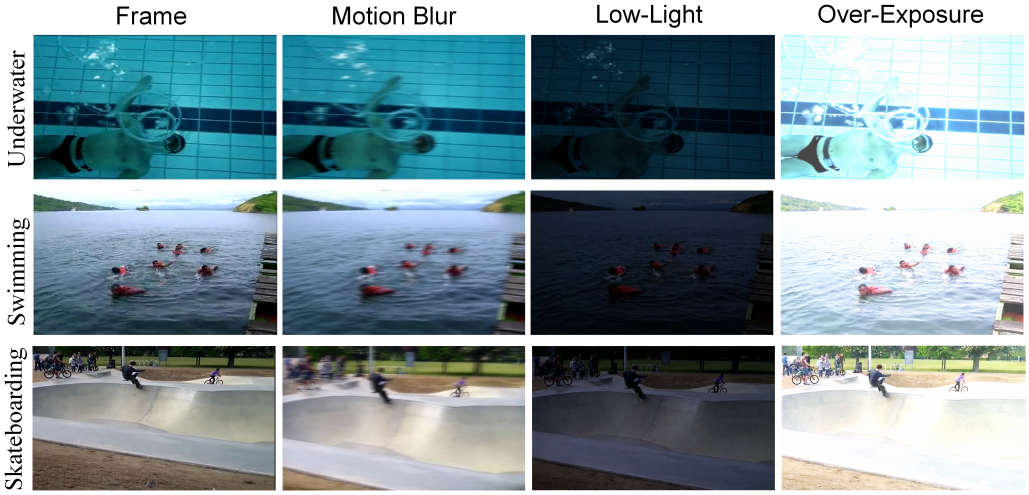
We construct a controlled robustness subset from VGGSound (Chen et al., 2020a) by randomly selecting 20,000 video clips and generating three degraded variants for each clip, resulting in 60,000 additional challenging-condition samples. The original clips are retained as the clean reference. All degradations are applied only to the visual stream, while the original audio remains unchanged. To



(a) Performance under visual degradations.



(b) Robustness against visual corruption.



(c) Representative samples under challenging visual conditions.

Figure 10: Robustness evaluation under challenging visual conditions on VGGSound. (a) Quantitative results across different degradation types. (b) Relative changes compared with clean observations, highlighting the stability under adverse conditions. (c) Selected video samples demonstrating the appearance changes caused by different perturbations.

avoid introducing artificial temporal flickering, the degradation parameters sampled for each clip are fixed across all of its frames. For all clips, the degraded RGB sequence is subsequently processed through the same processing pipeline as illustrated in Figure 3, ensuring that methods are evaluated from the same degraded visual observations. Specifically, we consider:

- **Motion Blur.** We simulate motion-induced image smearing using a directional motion-blur kernel following Rajagopalan et al. (2023). A 9×9 line kernel is applied with its orientation uniformly sampled from $[0°, 360°)$ for each clip and kept fixed throughout the sequence. The kernel is centered and applied with probability 1.0, producing a moderate blur level that suppresses fine object trajectories without severely destroying scene semantics.
- **Low-light**. We simulate insufficient illumination using gamma-based intensity attenuation established by Wang et al. (2023). For normalized RGB intensities $I \in [0, 1]$, we apply $I' = I^{\gamma}$ with $\gamma = 1.8$, followed by a global brightness factor of 0.6. This setting reduces visibility in dark regions while retaining the underlying spatial structure and motion content.
- **Over-exposure**. We simulate excessive exposure through global intensity amplification followed by saturation clipping (Afifi et al., 2021). Specifically, the input intensity is transformed as $I' = \mathrm{clip}(1.8I, 0, 1)$, causing bright regions to become saturated and reducing local appearance contrast while preserving the original temporal sequence.

To assess the robustness of V2A generation under challenging visual conditions based on the previous conditions, we introduce controlled visual degradations, including motion blur, low-light, and over-exposure effects. We select PAVAS (Hyun-Bin et al., 2026), a state-of-the-art physical-aware

method, as the comparison method and examine the performance variations under these adverse scenarios. As shown in Figure 10, both methods suffer from performance degradation as visual quality deteriorates, while MicroFoley consistently exhibits smaller performance drops across different degradation types, demonstrating improved stability under complex visual conditions.

F Further Discussion

F.1 Limitations

Although MicroFoley effectively alleviates temporal aliasing and achieves high-fidelity V2A generation through neuromorphic event representations, several limitations remain.

Event Data Availability and Simulation Gap. The limited availability of real-world neuromorphic event data requires our model to leverage FoleyBench-Synth, where event streams are generated through V2E simulations for scalable representation learning. While stochastic augmentations, including polarity reversal, timestamp jitter, and spatiotemporal event dropping, are introduced to mitigate the Sim-to-Real gap, simulated events cannot fully capture the asynchronous noise patterns and sensing characteristics of real event cameras under extreme illumination or complex motion. Moreover, since V2E interpolates observations between existing RGB frames, it cannot recover physical events absent from the original video sequence. Instead, it provides a structured and temporally densified motion representation with useful inductive biases for acoustic modeling, while genuine sub-frame physical observations are primarily provided by FoleyBench-Real.

Generalization to Unconstrained Videos. The current data filtering strategy prioritizes reliable audiovisual correspondence by removing clips with shot transitions, severe camera motion, or insufficient acoustic quality. While this improves benchmark reliability, it may limit the generalization capability of MicroFoley to unconstrained in-the-wild long videos with complex environmental variations. Furthermore, although class-stratified evaluations demonstrate that ASI and PACA consistently improve reverberant and continuous sound generation, modeling extremely long decay processes or weak sustained interactions remains challenging when the corresponding visual motion cues become subtle or ambiguous.

Material-Aware Semantic Ambiguity. The dual-stream architecture relies on a frozen CLIP encoder to provide material-related semantic cues. Consequently, heavy occlusion or visually ambiguous sounding objects may introduce uncertainty into CMMD alignment and affect the resulting acoustic texture modeling. Improving material representations with stronger physical priors and more comprehensive visual-audio grounding remains an important direction for future work.

F.2 Future Work

While we have made substantial efforts to mitigate these limitations through the construction of FoleyBench-Real, stochastic event augmentation, strict data filtering, and comprehensive robustness evaluations, several promising directions remain for future exploration. Expanding FoleyBench-Real with high-fidelity event-audio recordings covering more complex physical interactions, extreme illumination, and diverse reverberant environments could further reduce reliance on simulated events and enable more rigorous validation of genuine sub-frame acoustic dynamics. Meanwhile, more expressive event simulation and domain adaptation strategies may further narrow the gap between synthetic and real neuromorphic observations. Improving robustness in unconstrained videos through scene-aware dynamic routing and temporal segmentation could enable adaptive processing of shot transitions and dynamic camera motion. For reverberant and sustained sounds, future temporal models may explicitly decouple physical excitation from acoustic propagation and decay, allowing long resonance tails to persist even when visible motion becomes weak.

Beyond these improvements, reducing the computational cost of audio generation remains an important direction. Although OT-CFM already achieves efficient inference with an RTF of $0.52\times$ using 20 sampling steps, one-step flow distillation provides a promising route toward ultra-low-latency Foley generation on edge devices. We hope this work lays the foundation for future event-driven audio generation, where neuromorphic sensing and acoustic modeling are jointly explored to capture fine-grained physical dynamics beyond conventional frame-based perception in the future.

G Additional Qualitative Results

We provide additional qualitative analyses to visually compare the generated acoustic patterns across selective samples and scenarios, as shown in Figure 11 and Figure 12.

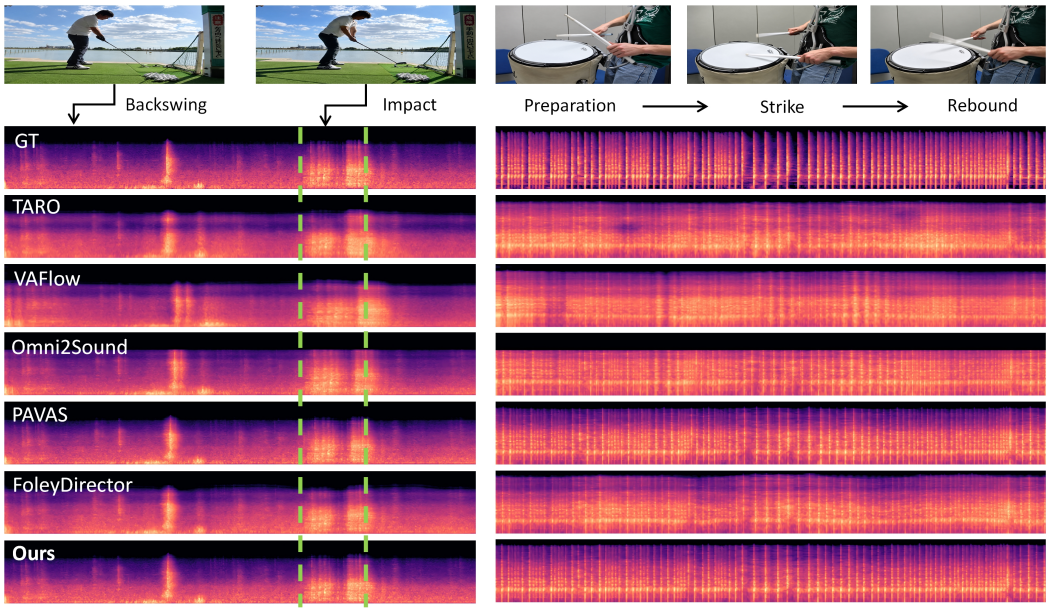


Figure 11: Qualitative Mel-spectrogram comparison on two distinct acoustic scenarios. The golf-swing sample is dominated by a sparse impact event, requiring precise localization of the contact moment, whereas the drum sequence contains dense and repetitive strikes with rapidly varying rhythmic intervals. MicroFoley adapts to both regimes, preserving isolated impact responses in the former while maintaining stable repetitive structures in the latter.

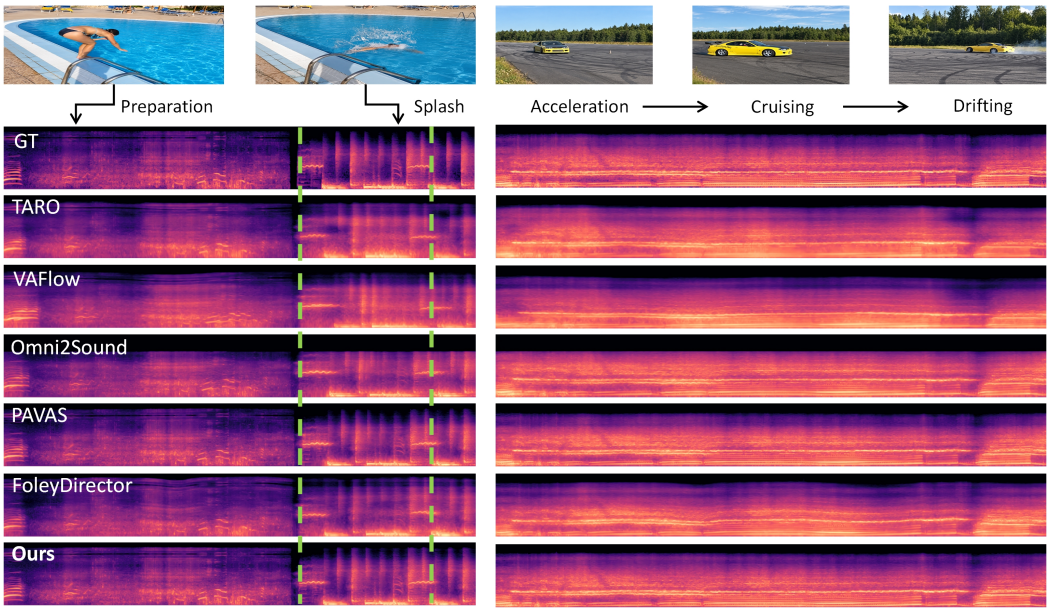


Figure 12: The left and right examples respectively represent impulsive and continuous sound events, including *preparation–splash* and *acceleration–cruising–drifting* phases. These cases further demonstrate that MicroFoley preserves both abrupt acoustic transitions in impact sounds and gradual spectral evolution in continuous motions, producing physically consistent audio patterns across different temporal dynamics.