# Foley Control: Aligning a Frozen Latent Text-to-Audio Model to Video

Ciara Rowles, Varun Jampani, Simon Donné, Shimon Vainer,
Julian Parker, Zach Evans

*Stability AI*

October 24, 2025

**Abstract**

Foley Control is a lightweight approach to video-guided Foley that keeps pretrained single-modality models frozen and learns only a small cross-attention bridge between them. We connect V-JEPA2 video embeddings to a frozen Stable Audio Open DiT text-to-audio (T2A) model by inserting compact video cross-attention after the model's existing text cross-attention, so prompts set global semantics while video refines timing and local dynamics. The frozen backbones retain strong marginals (video; audio given text) and the bridge learns the audio–video dependency needed for synchronization — without retraining the audio prior. To cut memory and stabilize training, we pool video tokens before conditioning. On curated video–audio benchmarks, Foley Control delivers competitive temporal and semantic alignment with far fewer trainable parameters than recent multi-modal systems, while preserving prompt-driven controllability and production-friendly modularity (swap/upgrade encoders or the T2A backbone without end-to-end retraining). Although we focus on Video-to-Foley, the same bridge design can potentially extend to other audio modalities (e.g., speech).

## 1 Introduction

Sound design is central to immersion in film, games, and VR: subtle contact sounds, material cues, and timing-sensitive transients anchor visual events in a coherent perceptual scene. Although recent video-to-audio (V2A) systems have advanced fidelity and semantic coverage, they frequently entail heavy training pipelines or control stacks that limit practicality in production settings [5, 26, 32].

Broadly, prior work splits into two paths. Adapter-based methods (e.g., FoleyCrafter [37]) plug semantics and timing controllers into strong T2A generators, improving alignment without retraining large backbones. By contrast, end-to-end foundation V2A models demand far more data with tightly aligned video–audio pairs to learn the audio prior, cross-modal mapping, and temporal synchrony simultaneously — driving curation of massive paired corpora with heavy filtering (onset heuristics, CLAP/ImageBind screening, alignment checks) and representation-alignment losses [5, 26]. Scale is further hampered by real-world noise: dubbing, off-screen sources, background music, and imprecise timestamps degrade supervision and underrepresent long-tail events. We instead freeze a strong T2A backbone and learn a thin video–audio bridge, attaining competitive alignment with far less data: our corpus uses ˜700k Kinetics–700 clips, whereas HunyuanVideo– Foley trains on ˜100k hours ($\approx 3.0\times10^7$ twelve–second segments), i.e., $\sim 43\times$ more paired data.

At the other extreme, multimodal diffusion transformers (e.g., MMAudio [5], HunyuanVideo-Foley [26] and others [4] [28] [13]) jointly train audio, video, and sometimes text streams end-to-end. Such approaches achieve impressive synchronization and coverage, but at the cost of massive curated datasets, high compute budgets, and reduced modularity.

This paper proposes **Foley Control**, a lightweight framework that targets the same alignment benefits while preserving the practicality of frozen generative backbones. Our key idea is to connect V-JEPA2 [1] video embeddings to a frozen Stable Audio Open DiT [7] by inserting *collaboration layers*—compact, video-conditioned cross-attention modules placed *inside* existing transformer blocks. This placement is deliberate: video cross-attention is applied *after* the model's original text cross-attention, so text prompts first establish high-level semantics and structure, and video then refines temporal

grounding and localized dynamics. By freezing the remaining parameters of the DiT blocks, we retain the strong generative prior learned from large-scale audio–text corpora and focus the trainable capacity on cross-modal synchronization rather than relearning audio generation.

From an architectural perspective, we employ a streamlined integration strategy in which V-JEPA2 embeddings are pooled into a compact grid representation and injected through lightweight cross-attention modules placed inside the frozen DiT blocks. This design ensures that video information refines the audio latent trajectory without altering the established text-conditioning pathway. Rotary position embeddings (RoPE) [27] further enhance temporal grounding by providing ordering signals across modalities, eliminating the need for heavier synchronization mechanisms. The resulting architecture remains compact yet expressive, scaling effectively to longer contexts and diverse scenes while preserving prompt-driven controllability.

Taken together, these elements provide a practical route to high-quality Foley generation: reuse a strong, frozen T2A backbone for audio fidelity and prompt control, and add just enough trainable capacity to align timing and dynamics to the video.

## 2 Related Work

**FoleyCrafter.** Early neural Foley systems learn to synthesize sounds that are semantically and temporally aligned with visual inputs, but often depend on limited audio–visual data and struggle to preserve high audio fidelity. FoleyCrafter [37] addresses this by *plugging* lightweight controllers into a strong text-to-audio backbone, thereby retaining audio quality while improving video–audio alignment. Concretely, it builds on a U-Net–based V2A generator (in the spirit of AuFusion [34]) and employs *multiple* control streams: a *semantic adapter* that injects video/text features throughout the U-Net (early, middle, and late blocks), and a *timestamp/onset controller* that is applied primarily in late layers to sharpen synchronization around transient events. Event timing cues are provided by a *separate* timestamp detection model , whose outputs modulate the diffusion steps to align onsets without altering the pretrained backbone. This division of labor—frozen backbone for fidelity, semantic control across the network, and late-layer timing refinement—yields stronger alignment under modest compute.

**MMAudio.** MMAudio [5] introduces a unified, *from-scratch* multimodal training paradigm that jointly leverages audio–text and audio–video pairs under a conditional flow-matching objective. A hybrid architecture—multimodal DiT blocks followed by audio-only blocks—supports scalable data mixing and strong semantic alignment, while a synchronization module operating via high frame-rate visual features further improves temporal precision. A related approach, HunyuanVideo-Foley [26], scales this paradigm with a massive curated text–video–audio dataset and a dual-stream multimodal diffusion transformer that fuses audio–video attention with text cross-attention. Additionally, it introduces a representation-alignment loss (REPA[36]) that steers the audio DiT's hidden states toward self-supervised audio embeddings, enhancing fidelity and stability, and employs a DAC-style autoencoder for higher-quality waveform reconstruction.

**Stable Audio Open.** Stable Audio Open [7] is a foundation text-to-audio model based on latent diffusion, combining a fully convolutional VAE, T5-based text conditioning, and timing embeddings to enable efficient generation of variable-length 44.1kHz stereo signals up to 95 seconds. Despite operating in a compressed latent space, it achieves state-of-the-art fidelity on both music and sound effects, offering a strong frozen backbone for adaptation in multimodal alignment tasks. Early text-to-audio (T2A) diffusion models such as DiffSound [35], AudioGen [17], AudioLDM [19], and Make-An-Audio [9] established the latent diffusion paradigm for sound synthesis. Stable Audio Open [7] extends this approach with high-fidelity 44.1 kHz generation and strong semantic conditioning, while TangoFlux [10] explores fast flow-matching variants for text-conditioned audio generation.

**V-JEPA2.** V-JEPA2 [1] is a large-scale self-supervised video model designed to learn predictive representations of the physical world from internet-scale video. It extends the joint-embedding predictive architecture (JEPA) by scaling pretraining to over one million hours of video and up to one billion parameters, using a masked feature prediction objective in representation space. Unlike generative approaches that reconstruct pixels, V-JEPA2 focuses on predictable dynamics such as motion trajectories,

yielding stronger representations for action understanding, anticipation, and temporal reasoning. The model achieves state-of-the-art results on motion understanding benchmarks (e.g., 77.3 top-1 accuracy on Something-Something v2) and human action anticipation (39.7 recall-at-5 on Epic-Kitchens-100), while also supporting downstream video question-answering when aligned with large language models. These properties make V-JEPA2 a compelling choice for video-conditioned generative tasks such as Foley synthesis, where fine-grained motion cues and temporal structure are critical

**Other related V2A / audiovisual models.** Several recent works also explore video-to-audio or joint audiovisual generation along different tradeoffs [22, 37, 21, 33, 5, 20, 26].

For instance, **FRIEREN** proposes rectified flow matching in spectrogram latent space to regress a conditional transport vector field, enabling few-step or even one-step audio sampling with strong video-audio alignment [33]. **UniVerse-1** fuses pretrained video and music experts via a stitching-of-experts approach to jointly generate synchronized audio and video [31]. **ThinkSound** frames audio generation as a reasoning process via chain-of-thought, decomposing generation into stages of Foley synthesis, object-centric refinement, and editing, guided by a multimodal LLM [20]. More recently, **DeepSound-V1** also introduces stepwise CoT reasoning in video→audio synthesis [18], and **YingSound** uses a multimodal CoT controller plus conditional flow matching for sound effect generation in few-shot settings [3]. These works complement ours: while they may retrain large joint models or adopt reasoning-based pipelines, our approach uniquely freezes a strong text–audio backbone and learns only a light cross-modal bridge for alignment.

### 2.1 Architectural Positioning of our approach

Our approach differs from prior adapter-based frameworks for adding multi-modality to frozen single modality models such as FoleyCrafter [37] or Stylecodes[25], which attach specialized controllers to a U-Net backbone for alignment. While such modular control can improve synchronization under limited data, it partitions the conditioning pathways – forcing each module to learn its own alignment rather than leveraging the pretrained model's holistic structure.

In contrast, **Foley Control** adopts a more unified transformer-based design that integrates video conditioning directly within the frozen diffusion transformer's existing attention layers. This avoids separate control heads and allows cross-modal signals to propagate through the same representational channels as text, better aligning with modern large-scale pretrained architectures such as Stable Audio Open [7]. Results from large-scale modeling [12] indicate that architectures which enable pretrained components to co-adapt through shared attention tend to harness scale and generalize more effectively than systems with manually partitioned control modules.

At the other end of the spectrum, fully multi-modal diffusion transformers such as MMAudio [5] and HunyuanVideo-Foley [26] extend this idea further by training end-to-end across text, video, and audio streams. These models demonstrate even higher efficiency and expressivity when massive, curated datasets are available, but they require orders of magnitude more paired data and compute to converge. **Foley Control** therefore strikes a middle ground: it retains the scalability and representational advantages of transformer conditioning while remaining data-efficient by freezing the text–audio prior and learning only lightweight video–audio bridges.

## 3 Method

### 3.1 Preliminaries

Our approach builds upon two key components: a frozen audio generation backbone and pretrained video encoders for semantic grounding. We briefly review the necessary background.

**Audio Latent Diffusion.** We adopt the **Stable Audio DiT** [7] as the generative backbone. Stable Audio is a diffusion-based model operating in the latent space of an audio autoencoder, enabling high-fidelity waveform synthesis at a sampling rate of 44.1 kHz. Given conditioning embeddings (e.g., text or duration), the model learns to denoise latent audio representations over a fixed number of timesteps. In our framework, the backbone remains *fully frozen*, ensuring training efficiency and stability.

**Video Representation Learning.** For visual grounding, we leverage **V-JEPA2** [1], a transformer-based video encoder pretrained with predictive objectives. Given a sequence of frames, V-JEPA2 produces spatiotemporal patch-level embeddings, which can be pooled into *tubelets* or spatial grids (e.g., 4×4, 8×8) to capture both global dynamics and localized motion cues. These embeddings serve as *key* tokens for cross-modal alignment with the audio latent sequence.

**Problem Setup.** Formally, given a video segment $\mathcal{V} = \{f_1, \ldots, f_T\}$, our goal is to synthesize an aligned audio waveform $x \in \mathbb{R}^L$, where $L$ corresponds to the clip duration at 44.1 kHz. The video encoder maps $\mathcal{V}$ to a sequence of tokens $\mathbf{v} \in \mathbb{R}^{S_v \times D_v}$, while the audio diffusion model operates on latent sequences $\mathbf{a} \in \mathbb{R}^{S_a \times D_a}$.

## 3.2 Dataset Curation

Training high-quality video-to-audio models requires large-scale, temporally aligned multimodal data. To this end, we constructed a dataset derived from the **Kinetics-700** dataset [6] , which provides a diverse set of human action videos in a wide range of everyday activities. Since not all videos contain meaningful or relevant sound events, we applied a data curation pipeline similar to that used by HunyuanVideo-Foley [26] . Since the dataset was already partitioned into clips, we first filtered out any silent samples from the dataset, we then used ImageBind [8] and Meta Audiobox Aesthetics [29] scores to filter out both low quality and conceptually distinct samples, ensuring high-fidelity and semantically consistent audio–video pairs similar to the filtering strategy of HunyuanVideo-Foley [26].

## 3.3 Framework Overview

Our framework is based on **Stable Audio Open**[7], a diffusion transformer (DiT) model for high-fidelity text-to-audio generation. Raw waveforms $x \in \mathbb{R}^L$ are encoded by an audio VAE into latent sequences $\mathbf{a} \in \mathbb{R}^{S_a \times D_a}$, where $S_a$ is the sequence length and $D_a$ the latent dimensionality. The DiT backbone performs *latent diffusion* using the *v-prediction* parameterization: at a random step $t$, we corrupt the clean latent $\mathbf{a}_0$ to $\mathbf{a}_t$ and train the model $v_\theta(\mathbf{a}_t, t, \text{cond})$ to predict the velocity that guides $\mathbf{a}_t$ back toward $\mathbf{a}_0$. At inference, we integrate the sampler using the predicted velocities to recover $\mathbf{a}_0$, then decode to waveform via the VAE.

**Diffusion Transformer (DiT).** At its core, Stable Audio Open employs a stack of transformer blocks designed for sequence modeling in the latent space. Each block incorporates multi-head self-attention, feed-forward networks, and cross-attention. Text embeddings, obtained from a pretrained T5 encoder [24], are injected through cross-attention, enabling semantic control over the generated audio. This design allows fast parallel sampling and supports long-context audio generation at 44.1 kHz.

**Freezing Strategy.** In our framework, the entire Stable Audio Open backbone—including the VAE encoder/decoder, DiT blocks, and T5 conditioning layers—remains *frozen*. This design choice ensures stable optimization, reduces computational cost, and preserves the strong generative prior acquired from large-scale audio–text pretraining.

**Additional Cross-Attention Layers** To integrate video semantics without disrupting the pre-trained frozen stable audio model, we insert video cross-attention *in every DiT block*, immediately after the backbone's text cross-attention and before the feed-forward network (SA → Tx-CA → **Vid-CA** → FFN). Audio latents act as queries and video tokens as keys/values; the rest of the block (including the text pathway) remains frozen.

**Tiny MLP adapter on the video path.** Before forming $K, V$, we pass the (detached) video features through a lightweight two-layer MLP with GELU and residual addition.

**RoPE scheme** We use standard rotary position embeddings (RoPE) [27]. RoPE is applied independently to audio queries and video keys: each modality computes its own phase from its sequence positions, and the rotations are not shared across modalities. Concretely, after linear projection we rotate $Q$ and $K$ in-place along their last dimension (per head) prior to attention. This preserves relative temporal phase information, helping the model align video motion and audio onsets more precisely.

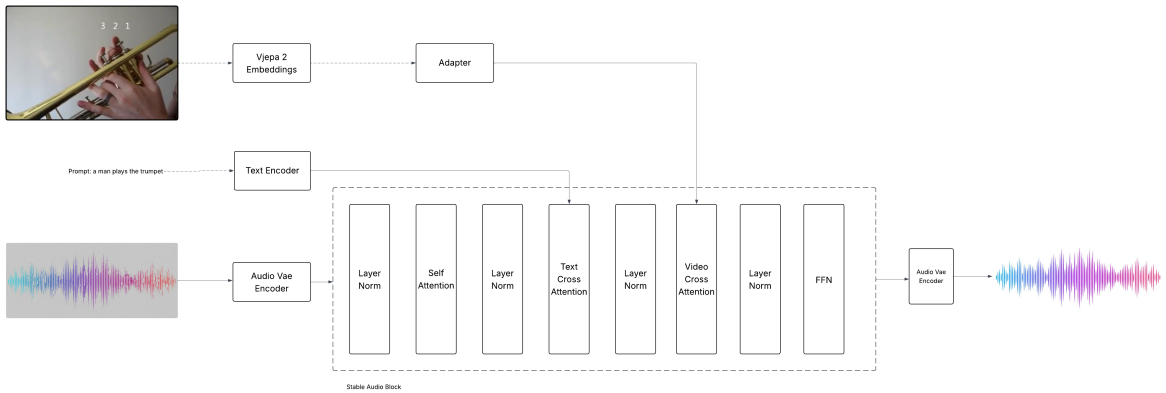


Figure 1: Forward flow: frames → V–JEPA2 → adapter (video vCA), prompt → text encoder (Tx–CA), and noise (latent init) entering the DiT at the same level

While prior work [5, 11] employs specialized synchronization modules (e.g., SyncFormer) for cross-modal alignment, we found RoPE sufficient for stable temporal correspondence without additional alignment networks in our adapter setting.

**Cross-attention placement.** Our placement choice is inspired by Kong et al. [15], which augments each DiT block with an audio cross-attention layer inserted after the text cross-attention. Following this design, we adopt the same ordering for our video cross-attention so that prompts establish global semantics before modality-specific timing and dynamics are injected. Unlike their setup, which employs label-aware RoPE (L-RoPE) to distinguish multiple audio streams, we found standard RoPE sufficient for stable and precise cross-modal alignment in our single-stream video-to-audio configuration.

**Scope of training.** Only the parameters introduced by this sublayer are trainable (video MLP adapter, $W_q$, $W_{kv}$, $W_o$, attention weights, and local norms); all backbone weights, the audio VAE, and the text pathway remain frozen. Unless otherwise stated, each DiT block has its own (non-shared) set of sublayer parameters.

**V–JEPA2 Embedding Pooling.** We condition collaboration cross-attention on V–JEPA2 tokens derived from 16 FPS video streams. For each 4 s segment, we sample 64 frames and encode them with V–JEPA2. To obtain a compact sequence, we *pool each effective frame* into a *single* token (the encoder operates with stride 2, so one effective frame corresponds to two input frames). This results in 32 effective frames per 4 s segment and thus 32 tokens per segment. To bound computational cost, we restrict inputs to a maximum of 12 s and concatenate the segment-level embeddings in temporal order. Originally, we experimented with spatial grids such as 8×8 (64 tokens per frame) and 16×16 (256 tokens per frame), but found that reducing to a single pooled token per frame preserved salient spatial context while substantially improving efficiency and stabilizing optimization.

# 4 Experiments

## 4.1 Experimental Setup

We evaluate our proposed joint audio–video fine-tuning framework on the curated Kinetics-700 dataset (Section 3.2), using the large filtered corpus for pretraining and the high-quality SFT subset for supervised alignment. We train all the models for the experiment with a batch size of 12, using a frozen `StableAudioDiT` backbone and V-JEPA2 embeddings; only the collaboration layers are updated. We adopt the original Stable Audio Open velocity-prediction training setup and apply token-drop regularization with 10% probability. For evaluation, we use the Meta Movie Audio Bench test set dataset.

## 4.2 Ablation Studies

We analyze the impact of video embedding granularity on model performance through a series of controlled ablations.

**Pooling Strategies.** We compare two ways of aggregating V-JEPA2 patch tokens into video tokens: (i) *frame pooling* (1 token per two frames), (ii) *grid8* pooling (8×8 tokens per frame, 64 per frame). Frame pooling offers maximum computational efficiency, while grid-based schemes capture richer spatial and motion cues at higher cost.

We report the KL-PANNs metric computed between generated and ground-truth audio event posteriors on the MovieGenBench test set, without text prompts, to isolate the effect of visual conditioning.

To reduce compute, all ablation runs are trained on a fixed 30% random subset of our curated Kinetics–700 training split; the same subset is used for both pooling variants, with identical hyperparameters , schedules and seed across conditions.

Table 1: Ablation study comparing pooling strategies over training steps using the KL-PANNs metric (lower is better) on the Kinetics-700 validation subset without text guidance.

| Training Steps | Grid8 | Single Pooled Embedding |
|---|---|---|
| 50,000 | 3.220921 | 3.222953 |
| 100,000 | 3.145059 | 3.188678 |
| 200,000 | 3.153171 | 3.194564 |
| 300,000 | 3.104110 | 3.130133 |
| 400,000 | 3.119460 | 3.111351 |

**Results and Discussion.** As shown in Table 1 and Figure 2, the lower-resolution **Single pooled embedding** configuration achieves performance on par with, or slightly better than, the grid8 embedding variant. Despite a substantial reduction in visual token count—and therefore compute and memory use—no meaningful loss in temporal alignment or perceptual fidelity was observed. Across 400k training steps, the metrics differ by less than 0.03, indicating that the mid-level spatial resolution of the single-pooled embeddings captures sufficient motion and context cues for Foley synchronization.

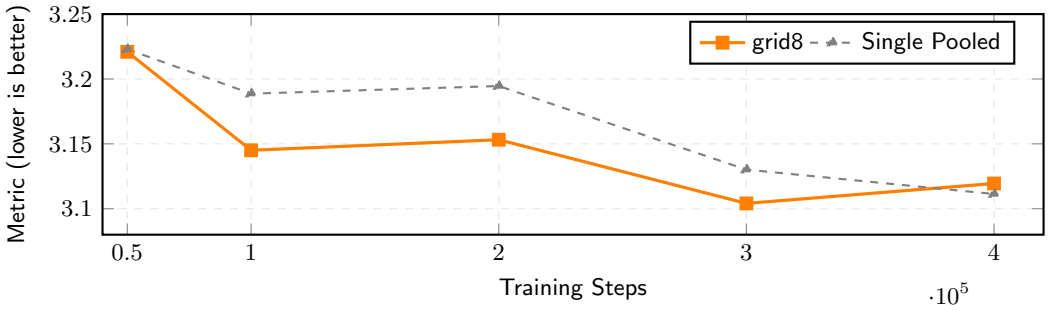


Figure 2: Ablation trends across training steps for different pooling strategies, evaluated with the KL-PANNs metric (lower is better) on the Kinetics-700 validation subset without text guidance. All runs were trained on a fixed 30% subset of the curated training data.

## 4.3 Comparison to Existing Work

We compare our framework against recent state-of-the-art video-to-audio generation systems, including MMAudio [5], HunyuanVideo-Foley [26], ThinkSound, and FRIEREN. All baselines represent distinct strategies for bridging video and audio modalities, ranging from fully joint multimodal diffusion training to modular adapter-based control. To ensure comparability, we evaluated all models under a consistent protocol using the **MovieGenBench**[23] benchmark, which emphasizes long-form cinematic scenes with diverse dynamics and complex soundscapes.

**Evaluation Protocol.** Each method generates audio at 44.1 kHz, conditioned on video frames and corresponding text prompts when supported. For our model, the Stable Audio DiT backbone, VAE, and CLAP text encoder remain entirely *frozen*; only the lightweight collaboration layers and alignment heads introduced in Section 3 are trained. This setup isolates the effect of our proposed video bridge while maintaining a fixed generative prior across all experiments. All systems are evaluated on the same MovieGenBench test split, using synchronized video clips with corresponding ground-truth soundtracks. **For metric computation, preprocessing, and dataset loaders, we used the MMAudio evaluation/testing repository[1] to ensure consistent scoring across methods**.

**Metrics.** Following the official benchmarking suite [26], we report a comprehensive set of perceptual and statistical metrics capturing complementary aspects of generation quality, including Fréchet Distance and KL divergence using PANNs [14] and PaSST [16], cross-modal consistency via ImageBind [8] and synchronization via Synchformer [11]

- **ImageBind Score (IB)** quantifying audio–visual semantic consistency via cosine similarity between audio and image frame embeddings extracted by **ImageBind**[8].
- **Mean KL Divergence (KL)** between classifier-based audio event posteriors, using PaSST (**KL-PaSST**[16]) and PANNs (**KL-PANNs**). Lower values denote better distributional consistency.
- **Fréchet Distance (FD)** between generated and real audio embeddings, computed with three pretrained encoders: PaSST (**FD-PaSST**), PANNs (**FD-PANNs**), and VGGish (**FD-VGG**). Lower values indicate closer alignment between the generated and reference distributions.
- **DeSync Score** evaluating temporal misalignment (in seconds) predicted by **Synchformer**[11]; lower values indicate better synchronization.

**Baselines.**

- **MMAudio** [5]: a fully multimodal diffusion transformer jointly trained on text, audio, and video under a conditional flow-matching objective.
- **HunyuanVideo-Foley** [26]: a large-scale multimodal DiT trained end-to-end on curated text–video–audio data, incorporating representation-alignment losses for temporal precision.
- **ThinkSound** [20]: a modular system combining CoT reasoning with pretrained encoders and a controllable diffusion backbone, designed for robustness to domain variation.
- **FRIEREN** [33]: an autoregressive video-to-sound system emphasizing temporal causality and synchronization through hierarchical attention mechanisms.
- **FoleyCrafter** [37]: an adapter-based approach that enhances synchronization by injecting semantic and temporal controllers into a pretrained text-to-audio backbone, improving alignment without retraining large generative models.

**Fairness and Implementation Details.** All comparisons use identical input video frame rates and duration limits. During evaluation, each model produces a single audio sample per clip without post-processing, ensuring consistency across systems and avoiding any external enhancement or mixing effects. We do not include comparisons against **V-AURA** [30], as the method is constrained to clips of 2.5 seconds in duration, which makes it unsuitable for evaluation on longer-form datasets such as MovieGenBench that emphasize multi-second temporal dependencies and ambient context.

[1] https://github.com/facebookresearch/mmaudio/tree/main/eval

**MovieGenBench.** We test on the **MovieGenBench** dataset [23], which emphasizes cinematic sound design and long-range temporal dependencies. This benchmark evaluates generated audio against a strong text-to-video-with-audio (T2VA) reference model, providing a measure better aligned with large-scale multimodal systems such as MMAudio [5] and HunyuanVideo-Foley [26], which excel at generating ambient, scene-level audio. Because MovieGenBench includes extensive background and environmental textures, it favors models that maintain coherent ambiance and long-horizon consistency rather than isolated transients. In contrast, **VGGSound** [2] consists primarily of short, event-driven Foley-style clips that emphasize localized synchronization and sound event accuracy. We also omit comparisons with **V-AURA** [30], a video-to-audio model limited to generating 2.5-second clips, which makes it unsuitable for long-form benchmarks like MovieGenBench. As shown in Table 2, Foley Control performs competitively under these more demanding, ambient conditions.

**Kling-Foley AudioEval.** To avoid train–evaluation contamination, we do not report results on the Kling-Foley AudioEval benchmark introduced by Wang et al. [32]. Our training corpus included material overlapping or closely related to that evaluation split, which could yield inflated or non-comparable scores.

Table 2: Comparison on the MovieGenBench dataset.

| System | KL-PANNs ↓ | KL-PaSST ↓ | IB ↑ | FD-VGG ↓ | FD-PANNs ↓ | FD-PaSST ↓ | DeSync ↓ |
|---|---|---|---|---|---|---|---|
| **FRIEREN** | 3.58 | 3.89 | 0.14 | 5.65 | 59.04 | 560.91 | 0.30 |
| **MMaudio** | 2.52 | 2.35 | 0.25 | **4.14** | 37.60 | **343.24** | **0.29** |
| **HunyuanVideo-Foley** | 2.58 | 2.11 | **0.30** | 7.00 | 31.28 | 373.62 | 0.31 |
| **Foley Control (ours)** | 2.93 | 2.59 | 0.20 | 5.89 | **31.10** | 383.99 | 0.32 |
| **ThinkSound** | 3.16 | 2.90 | 0.18 | 6.62 | 33.62 | 468.25 | 0.30 |
| **FoleyCrafter** | **1.11** | **1.29** | 0.26 | 6.94 | 40.70 | 493.08 | 0.33 |

**Training Efficiency.** While large multimodal diffusion systems such as HunyuanVideo-Foley train end-to-end for **200k–700k steps** on roughly **100k hours** of curated text–video–audio data using **128×H20 GPUs** and an **effective batch size of 2048**, our Foley Control bridge trains for only **400k steps** with an **effective batch size of 384**. In contrast to MMAudio and ThinkSound, which use a comparable amount of paired audio–video data but additionally rely on extensive *audio-only* pretraining to learn their generative priors, Foley Control requires *no* such auxiliary corpus—leveraging instead a frozen Stable Audio backbone trained independently on text–audio data. Compared to HunyuanVideo-Foley, Foley Control operates with nearly **two orders of magnitude less paired data and compute**, yet achieves competitive synchronization and semantic alignment, underscoring the efficiency of the lightweight cross-modal adapter strategy. *Similarly, FoleyCrafter* [37] demonstrates that adapter-based designs can deliver strong alignment and controllability by keeping a pretrained T2A backbone frozen and learning only compact temporal and semantic controllers, employing a more elaborate adapter architecture built atop a U-Net–based generative model.

# 5 Conclusion

We introduced **Foley Control**, a lightweight bridge that brings video guidance to a frozen text-to-audio generator by inserting compact, trainable collaboration layers after the model's existing text cross-attention. With V-JEPA2 embeddings, token pooling, and RoPE-based ordering cues, our design preserves the strengths of the audio prior and prompt controllability while adding the temporal control needed for Foley.

Across a curated data corpus and evaluation on MovieGenBench, the approach delivers competitive semantic and temporal alignment while training only a small fraction of parameters compared to fully multimodal systems. Ablations show that aggressively pooled video tokens match the performance of denser grid features, substantially reducing compute and memory without degrading synchronization.

Practically, the framework remains modular: encoders or the T2A backbone can be swapped or upgraded without end-to-end retraining, which is attractive for production settings where models evolve, additional control modules are added and latency/VRAM budgets matter.

**Limitations and Future Work.** Our current setup caps video duration and conditions on pooled tokens, which may miss rare fine-grained spatial cues. The method also assumes clean inputs and does not explicitly model spatial (binaural/ambisonic) acoustics or streaming/online alignment. Future work may include adaptive tokenization (learned pooling or budget-aware routing), longer-context conditioning, more varied data, spatial audio generation, robustness to in-the-wild edits and background music, and extending the bridge to other audio modalities such as speech and dialogue.

# References


[1] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[2] 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)*, pages 721–725. IEEE, 2020.

[3] Zihao Chen, Haomin Zhang, Xinhan Di, Haoyu Wang, Sizhe Shan, Junjie Zheng, Yunming Liang, Yihan Fan, Xinfa Zhu, Wenjie Tian, et al. Yingsound: Video-guided sound effects generation with multi-modal chain-of-thought controls. *arXiv preprint arXiv:2412.09168*, 2024.

[4] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, Andrew Owens, and Justin Salamon. Video-guided foley sound generation with multimodal controls. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18770–18781, 2025.

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

[6] DeepMind / CVDFoundation. Kinetics dataset (kinetics-700, kinetics-400, etc.). https://github.com/cvdfoundation/kinetics-dataset.

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

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

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

[10] Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*, 2024.

[11] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329. IEEE, 2024.

[12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[13] Tornike Karchkhadze, Kuan-Lin Chen, Robert Henzel, Alessandro Toso, Mehrez Souden, Joshua Atkins, et al. Stereofoley: Object-aware stereo audio generation from video. *arXiv preprint arXiv:2509.18272*, 2025.

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

[15] Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation, 2025.

[16] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

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

[18] Yunming Liang, Zihao Chen, Chaofan Ding, and Xinhan Di. Deepsound-v1: Start to think step-by-step in the audio generation from videos. *arXiv preprint arXiv:2503.22208*, 2025.

[19] 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.

[20] Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *arXiv preprint arXiv:2506.21448*, 2025.

[21] 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.

[22] Shentong Mo, Jing Shi, and Yapeng Tian. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024.

[23] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

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

[25] Ciara Rowles. Stylecodes: Encoding stylistic information for image generation. *arXiv preprint arXiv:2411.12811*, 2024.

[26] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, Qun Yang, Jin Zhou, and Zhao Zhong. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *arXiv preprint arXiv:2508.16930*, 2025.

[27] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2024.

[28] Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*, 2025.

[29] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[30] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[31] Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

[32] Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, et al. Kling-foley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*, 2025.

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

[34] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[35] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023.

[36] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

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