



FoleyCrafter: Bring Silent Videos to Life with Lifelike and Synchronized Sounds

Yiming Zhang¹ **Yicheng Gu²** **Yanhong Zeng^{1,†}** **Zhening Xing¹**
Yuancheng Wang² **Zhizheng Wu^{2,1}** **Kai Chen^{1,†}**

¹Shanghai Artificial Intelligence Laboratory ²Chinese University of Hong Kong, Shenzhen

Homepage: <https://foleycrafter.github.io/>

Abstract

We study Neural Foley, the automatic generation of high-quality sound effects synchronizing with videos, enabling an immersive audio-visual experience. Despite its wide range of applications, existing approaches encounter limitations when it comes to simultaneously synthesizing high-quality and video-aligned (*i.e.*, semantic relevant and temporal synchronized) sounds. To overcome these limitations, we propose FoleyCrafter, a novel framework that leverages a pre-trained text-to-audio model to ensure high-quality audio generation. FoleyCrafter comprises two key components: the semantic adapter for semantic alignment and the temporal controller for precise audio-video synchronization. The semantic adapter utilizes parallel cross-attention layers to condition audio generation on video features, producing realistic sound effects that are semantically relevant to the visual content. Meanwhile, the temporal controller incorporates an onset detector and a timestamp-based adapter to achieve precise audio-video alignment. One notable advantage of FoleyCrafter is its compatibility with text prompts, enabling the use of text descriptions to achieve controllable and diverse video-to-audio generation according to user intents. We conduct extensive quantitative and qualitative experiments on standard benchmarks to verify the effectiveness of FoleyCrafter. Models and codes are available at <https://github.com/open-mmlab/FoleyCrafter>.

1 Introduction

Foley, a key element in film and video post-production, adds realistic and synchronized sound effects to silent media [8]. These sound effects are the unsung heroes of cinema and gaming, enhancing realism, impact, and emotional depth for an immersive audiovisual experience. Traditionally, skilled Foley artists painstakingly create, record, and process sound effects in specialized studios, making it a labor-intensive and time-consuming process [2]. Despite advancements in recent video-to-audio generation, achieving Neural Foley, which requires synthesizing high-quality, video-aligned sounds that are semantically related and temporally synchronized with the videos, remains challenging [20].

State-of-the-art approaches for Neural Foley in video-to-audio generation can be categorized into two main groups, as illustrated in fig. 1. The first group involves training a video-to-audio generative model on a large-scale paired audio-video dataset [4, 17, 20, 25]. However, the audio quality of such datasets crawled from the Internet can be subpar, with issues like noise and complex environmental sounds recorded in the wild, which hinder the production of high-quality sounds [28, 30]. To address this, the second group of approaches (fig. 1-(b)) adopts a two-stage process. They first translate video into text using video captioning or embedding mapping techniques and then employ a pre-trained text-to-audio model [28, 30, 31]. Leveraging the well-trained text-to-audio generator, these

[†] denotes corresponding author.

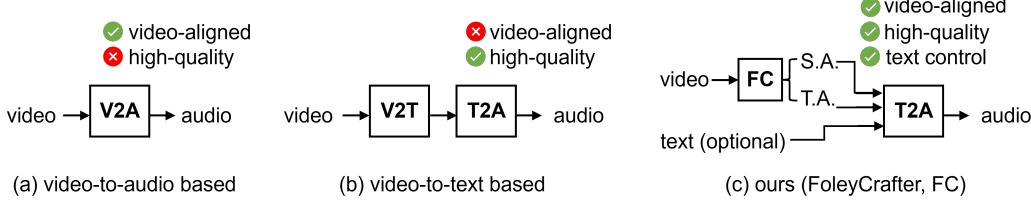


Figure 1: (a) (Video-to-audio) V2A methods struggle with audio quality due to noisy training data, while (b) video-to-text (V2T) methods encounter difficulties in producing synchronized sounds. Our model, FoleyCrafter (FC), integrates a learnable module into a pre-trained Text-to-Audio (T2A) model to ensure audio quality while enhancing video-audio alignment with the supervision of audios.

methods achieve impressive sound quality. Nonetheless, effectively bridging the gap between video and text while preserving nuanced details is challenging. As a result, these methods may produce unsynchronized sounds due to the suboptimal translated text conditions.

To achieve both high-quality and video-aligned sound generation, we present FoleyCrafter, which breathes life into silent videos with realistic and synchronized sound effects. As depicted in fig. 1-(c), the core of FoleyCrafter is an innovative pluggable module that can be integrated with a pre-trained text-to-audio (T2A) model, optimized with the supervision of audios. Specifically, FoleyCrafter comprises two main components: a semantic adapter for semantic alignment and a temporal controller for temporal synchronization. The semantic adapter introduces parallel cross-attention layers into the backbone of the T2A model. It takes as input the extracted video features, allowing FoleyCrafter to generate audio conditioned on the video without relying on explicit text. The temporal controller, on the other hand, is engineered to refine temporal synchronization. The temporal controller features an onset detector and a timestamp-based adapter. The onset detector is trained to predict sound occurrences and silences from videos. The timestamp-based adapter then refines this by aligning audio generation with the predicted timestamps, ensuring sound and silence are synchronized with the video’s temporal flow. During training, we train the semantic adapter and temporal controller with video-audio correspondent data, while fixing the text-to-audio base model to preserve its established audio generation quality. After training, FoleyCrafter can generate high-quality sounds for videos with semantic and temporal alignment in a flexible and controllable way.

We conduct extensive experiments to evaluate the performance of FoleyCrafter in terms of audio quality and video alignment, both semantically and temporally. Our experiments include quantitative analysis, qualitative comparison, and user studies, all of which demonstrate that FoleyCrafter has achieved state-of-the-art results. Additionally, we have showcased the controllability of FoleyCrafter through text prompts, allowing for a more fine-grained and versatile application of the model. Our main contributions can be summarized as follows:

- We present a novel Neural Foley framework that generates high-quality, video-aligned sound effects for silent videos, while also offering fine-grained control through text prompts.
- To ensure both semantic and temporal alignment, we design a semantic adapter and a temporal controller, significantly improving video alignment.
- We validate the effectiveness of FoleyCrafter through extensive experiments, including quantitative and qualitative analyses. Our results show that FoleyCrafter achieves state-of-the-art performance on commonly used benchmarks.

2 Related Work

Diffusion-based Audio Generation. Latent diffusion models have significantly advanced audio generation [18, 19, 24]. AudioLDM pioneers open-domain text-to-audio generation using a latent diffusion model [18, 19]. Tango improves text-to-audio generation with an instruction-tuned LLM FLAN-T5 [7] as the text encoder [12]. Make-an-Audio tackles complex audio modeling using spectrogram autoencoders instead of waveforms [16]. Li et al. conduct comprehensive ablation studies to explore effective designs and set a new state-of-the-art with the proposed Auffusion [32]. In this paper, we introduce FoleyCrafter, a module that extends state-of-the-art text-to-audio generators to support video-to-audio generation while preserving the original text-to-audio controllability.

Video-to-Audio Generation. Foley artistry is a crucial audio technique that enhances the viewer’s auditory experience by creating and recording realistic sound effects that synchronize with visual content [8]. Early Neural Foley models mainly focus on generating sounds tailored to a specific genre or a narrow spectrum of visual cues, underscoring the potential of deep learning to innovate sound effect creation for videos [5, 6, 22, 37]. Despite recent advancements in large-scale generative models [16, 18], generating high-quality and visually synchronized sounds for open-domain videos remains a challenge [9, 10, 20, 21, 26, 28].

State-of-the-art video-to-audio approaches can be categorized into two groups. The first group focuses on training video-to-audio generators from scratch [17, 20, 25]. Specifically, SpecVQGAN [17] employs a cross-modal Transformer [27] to auto-regressively generate sounds from video tokens. Im2Wav [25] conditions an autoregressive audio token generation model using CLIP features, while Diff-Foley [20] improves semantic and temporal alignment through contrastive pre-training on aligned video-audio data. However, these methods are limited by the availability of high-quality paired video-audio datasets. An alternative approach is to utilize text-to-audio generators for video Foley. Xing et al. [31] introduce an optimization-based method with ImageBind [13] for video-audio alignment, while SonicVisionLM [30] generates video captions for text-to-audio synthesis. Wang et al. note the limitations of caption-based methods and propose V2A-Mapper to translate visual embeddings to text embedding space [28]. Nevertheless, effectively bridging the gap between video and text while preserving fine-grained temporal cues remains a significant challenge. In contrast, we introduce FoleyCrafter, integrating a learnable module into text-to-audio models with end-to-end training, enabling a high-quality, video synchronized and high-controllable Foley.

3 Approach

In this section, we introduce the framework of FoleyCrafter. We introduce related preliminaries about Audio Latent Diffusion Models (ALDMs) [18, 19] and conditioning mechanisms in section 3.1. We then delve into the key components of FoleyCrafter in section 3.2. The semantic adapter generates audio based on visual cues and text prompts, while the temporal controller improves temporal synchronization with the input video. We also outline the training process for each component and explain how FoleyCrafter can be used to generate foley for silent videos in section 3.3.

3.1 Preliminaries

Audio Latent Diffusion Model. The latent diffusion model (LDM) has achieved remarkable advancements in text-to-audio generation, as demonstrated by recent studies [12, 18, 19, 32]. In this model, the audio waveform is initially transformed into a mel-spectrogram representation. Subsequently, a variational autoencoder (VAE) encodes the mel-spectrogram into a latent representation denoted as z . The LDM’s UNet is trained on this latent space to generate z by denoising normally distributed noise ϵ . The predicted latent z is then reconstructed by the VAE into a mel-spectrogram, which is finally transformed into a waveform using a vocoder.

A latent diffusion model consists of two main processes: the diffusion process and the denoising process. In the diffusion process, a clean latent representation z undergoes step-by-step noise addition until it reaches an independently and identically distributed noise. Such a process is denoted as,

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (1)$$

where $\bar{\alpha}_t$ is the noise strength at t timestep. The UNet is trained to estimate the added noise at a given timestep t using the following optimization objective:

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim N(0, 1), t, c} [\|\epsilon - \epsilon_\theta(z_t, t, c)\|] \quad (2)$$

where x represents the mel-spectrogram in the ALDM, z_t corresponds to the latent representation of the mel-spectrogram at timestep t , and c denotes the condition information.

Conditioning Mechanisms. There are two kinds of condition mechanisms mainly used in ALDM, *i.e.*, MLP-based mechanism [12, 18] and cross-attention-based mechanism [32, 19]. In the MLP-based mechanism, the time step is mapped to a one-dimensional embedding and concatenated with the text embedding as the conditioning information. This one-dimensional condition vector is then merged with the UNet’s feature map through MLP layers. In contrast, the cross-attention-based

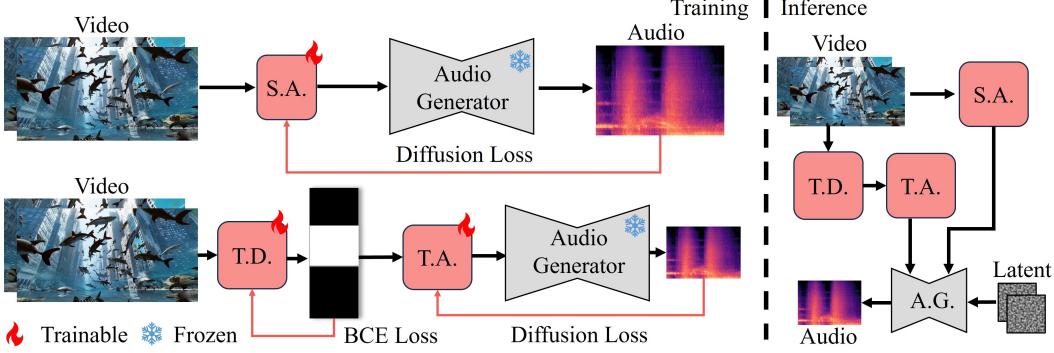


Figure 2: **The overview of FoleyCrafter.** FoleyCrafter is built upon a pre-trained text-to-audio (T2A) generator, ensuring high-quality audio synthesis. It comprises two main components: the semantic adapter (S.A.) and the temporal controller, which includes a timestamp detector (T.D.) and a temporal adapter (T.A.). Both the semantic adapter and the temporal controller are trainable modules that take videos as input to synthesize audio, with audio supervision for optimization. The T2A model remains fixed to maintain its established capability for high-quality audio synthesis.

mechanism utilizes a cross-attention operation in each block of the UNet backbone. This mechanism demonstrates improved alignment with conditions and allows for more flexible and fine-grained controllable generation. It has been widely adopted in recent works [19, 32]. The cross-attention mechanism can be represented as follows:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d}} \right) \cdot V, \quad (3)$$

$$\text{where } Q = W_Q \cdot \varphi(z_t), \quad K = W_K \cdot \tau(c), \quad V = W_V \cdot \tau(c), \quad (4)$$

where φ denote the flattening operation and τ is the condition embedding encoder and W_Q , W_K and W_V is the learnable projection matrices. In this study, we adopt a cross-attention mechanism to integrate textual and visual cues, aligning with recent state-of-the-art ALDMs [19, 32].

3.2 FoleyCrafter

FoleyCrafter is a modular framework that can be integrated with a pre-trained text-to-audio (T2A) model, typically trained on large-scale high-quality audio datasets such as FreeSounds [11, 32]. This design ensures both high-fidelity and diverse audio synthesis while allowing direct audio supervision for optimizing FoleyCrafter. As shown in fig. 2, the framework consists of two main components: the semantic adapter and the temporal controller, specifically designed for semantic alignment and temporal alignment, respectively. During training, only the FoleyCrafter components are trainable, optimizing with the supervision of ground truth audio, while the weights of the T2A model remain fixed. In the following sections, we provide more detailed explanations of each component.

3.2.1 Semantic Adapter

To efficiently extract semantic features from the input video and incorporate them into the pre-trained text-to-audio generator, we employ a visual encoder along with decoupled parallel cross-attention layers. We demonstrate the overview of the semantic adapter in fig. 3.

Visual Encoder. The CLIP encoder has demonstrated its effectiveness as a semantic extractor for visual information [23]. In our approach, we follow previous works [24, 33] and extract visual embeddings from each frame of the input video using the CLIP image encoder. To align these embeddings with the text-to-audio generator, we employ several learnable projection layers. This process can be expressed as:

$$V_{emb} = \text{MLP}(\text{AvgPooling}(\tau_{vis}(v))). \quad (5)$$

Here, v represents the input video, τ_{vis} denotes the CLIP image encoder, and AvgPooling refers to the average pooling of the extracted CLIP features across frames.

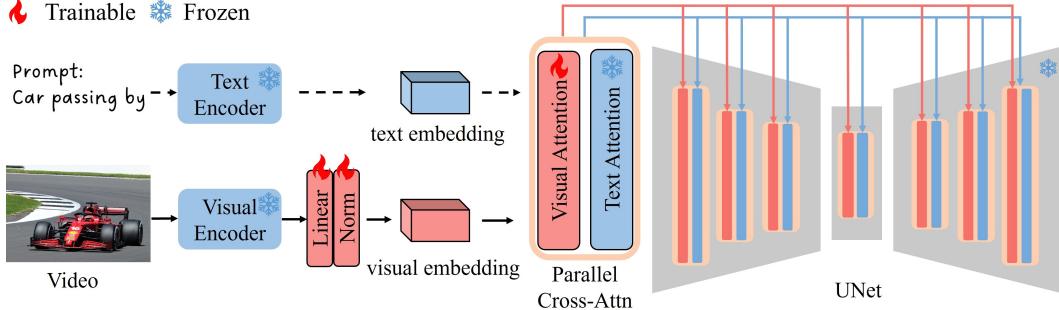


Figure 3: **The overview of semantic adapter.** Semantic adapter employs a pre-trained visual encoder with several learnable layers to extract video embeddings that align better with the text-to-audio generator. Then, it integrates trainable visual-cross attention mechanisms alongside text-based ones, ensuring semantic alignment with the video without compromising text-to-audio generation.

Semantic Adapter. To incorporate the extracted video embedding into the pre-trained text-to-audio generator without compromising its original functionality, we introduce visual-conditioned cross-attention layers alongside the existing text-conditioned cross-attention layers. In this approach, visual and text embeddings are separately fed into their corresponding cross-attention layers. The outputs of the new and original cross-attention layers are then combined using a weight parameter, λ . The parallel cross-attention can be denoted as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK_{txt}^T}{\sqrt{d}}\right) \cdot V_{txt} + \lambda \cdot \text{softmax}\left(\frac{QK_{vis}^T}{\sqrt{d}}\right) \cdot V_{vis}, \quad (6)$$

$$\text{where } K_{txt} = W_K^{txt} \cdot T_{emb}, V_{txt} = W_V^{txt} \cdot T_{emb}, \quad (7)$$

$$K_{vis} = W_K^{vis} \cdot V_{emb}, V_{vis} = W_V^{vis} \cdot V_{emb}, \quad (8)$$

where T_{emb} and V_{emb} represent the extracted text embeddings and video embeddings, respectively. W_K^{txt} and W_V^{txt} correspond to the pre-trained projection layers in the text-conditioned cross-attention layers, which remain fixed during training. On the other hand, W_K^{vis} and W_V^{vis} are newly introduced learnable projection layers used to map the visual embedding to a space that aligns better with the condition space of the pre-trained text-to-audio generator.

During the training of the semantic adapter, we initialize the vision-conditioned cross-attention layers from the text-conditioned ones. As shown in fig. 3, we train the newly added projection layers after the visual encoder and the vision-conditioned cross-attention layers using ground truth audio as supervision. Meanwhile, we keep the text encoder and the text-to-audio generator fixed. The optimization objective can be expressed as:

$$\mathcal{L} = \mathbb{E}_{x, \epsilon \sim N(0, 1), t, c} [\|\epsilon - \epsilon_\theta(z_t, t, T_{emb}, V_{emb})\|]. \quad (9)$$

We observed a related work, IP-Adapter [33], which also incorporates parallel cross-attention layers to inject image information into a pre-trained text-to-image generator. However, our semantic adapter differs in that it introduces a third modality, namely video, into a text-to-audio generator. To effectively capture the visual cues and align them with the condition space of the text-to-audio generator, we randomly drop the text condition during training in the majority of cases (approximately 90%). This design enables the semantic adapter to effectively capture visual cues for audio generation while retaining the capability of combining text prompts for more controllable video-to-audio generation.

3.2.2 Temporal Controller

We have observed that the semantic adapter captures video-level alignment without precise temporal synchronization for each frame. To address this limitation, we have developed the temporal controller to enhance the temporal synchronization. As shown in fig. 4, the temporal controller utilizes a timestamp detector to predict sound and silence intervals within a given video. These predicted timestamp conditions are then fed into a temporal adapter, which controls the generation of audios.

Timestamp Detector. The timestamp detector takes video frames as input and predicts a binary time condition mask, which indicates the presence or absence of sound effects in the target audio.

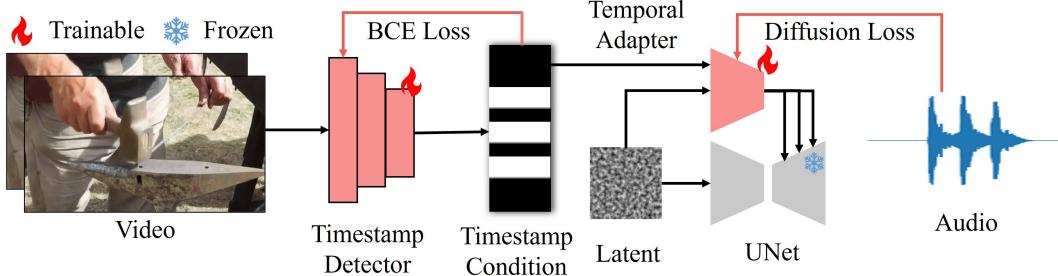


Figure 4: **The overview of the temporal controller.** The temporal module consists of a timestamp detector and a temporal adapter for improved video-audio synchronization. The timestamp detector predicts sound and silence labels based on the video, optimized using ground truth audio event timestamps. The temporal adapter, initialized from UNet encoder blocks, encodes the timestamp condition and injects synchronization information into the UNet decoder.

We train the timestamp detector using binary cross-entropy loss, leveraging sound event labels. The training process can be expressed as:

$$\mathcal{L}_{BCE}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (10)$$

where N represents the number of samples, y denotes the ground truth timestamp label, and \hat{y} represents the prediction. After training, the timestamp detector is capable of predicting the sound and silent timestamp mask, which is then passed to the temporal adapter for further processing.

Timestamp-based Adapter. The timestamp-based temporal adapter shares the same architecture as the UNet encoder of the text-to-audio generator, following the design of ControlNet [35]. Specifically, the temporal adapter utilizes the predicted timestamp condition mask to guide sound generation at the target timestamp. It takes the timestamp mask and the same latent input as the original UNet, and the output is then added as a residual to the output of the original UNet. During training, we only train the replicated UNet blocks using the same optimization objective as the diffusion model. We train the temporal adapter by providing the timestamp labels of sound events from Audioset Strong [14] as input, with the corresponding audio as the target.

3.3 Implementation details

For the semantic adapter, we follow [33] to use a linear projection for clip visual embedding to better align with text representation and expand the embedding length to four. Then we modify all the cross-attention to parallel cross-attention for visual conditions. We train semantic adapter on the VGGSound for 164 epochs with a batch size of 128 on 8 NVIDIA A100 GPUs. For the temporal controller, we train the timestamp detector and temporal adapter separately. The timestamp detector is trained on the dataset with higher relevance audio-visual i.e. AVSync15 [34] and Countix-AV [36] for 23 epochs. The temporal adapter is trained on the AudioSet Strong [14] for 80 epochs. Following, the two components in temporal controller can be combined together for inference and evaluation.

4 Experiments

4.1 Experimental settings

Baselines. We conducted comprehensive evaluations of FoleyCrafter by comparing it with state-of-the-art approaches, namely SpecVQGAN [17], Diff-Foley [20], V2A-Mapper [28], and Seeing-and-hearing [31]. Both quantitative and qualitative comparisons were employed. SpecVQGAN generates audio tokens autoregressively based on extracted video tokens. Diff-Foley utilizes contrastive learning for synchronized video-to-audio synthesis with its CAVP audio and visual encoder. V2A-Mapper translates visual CLIP embeddings to CLAP space, enabling video-aligned audio generation using a pre-trained text-to-audio generator. Seeing-and-hearing [31] propose using ImageBind [13] as a bridge between visual and audio, leveraging off-the-shelf audio and video generators for multimodal generation. Please note that the results of V2A-Mapper and Seeing-and-hearing were obtained directly from the authors, as their source codes and models are not publicly available at this time.

Table 1: Quantitative evaluation in terms of semantic alignment and audio quality. Specifically, FoleyCrafter achieves state-of-the-art performance in terms of Mean KL Divergence (MKL) [17], CLIP [29] and FID [15] on standard benchmarks, *i.e.*, VGGSound [4] and AVSync15 [34]. We report the results with error bars calculated from ten times of evaluation with random seeds.

VGGSound [4]	MKL↓	CLIP↑	FID↓
SpecVQGAN(BN Inception) [17]	4.337±0.001	5.079±0.023	65.37±0.01
SpecVQGAN(ResNet) [17]	3.400±0.001	5.876±0.016	32.01±0.005
Diff-Foley [20]	3.318±0.011	9.172±0.110	29.03±0.61
V2A-Mapper [28]	2.654	9.720	24.16
FoleyCrafter (ours)	2.561 ±0.011	10.70 ±0.121	19.67 ±0.05
AVSync15 [34]	MKL ↓	CLIP ↑	FID ↓
SpecVQGAN(Inception) [17]	5.339±0.077	6.610±0.014	114.44±1.31
SpecVQGAN(ResNet) [17]	3.603±0.006	6.474±0.077	75.56±1.43
Diff-Foley [20]	1.963±0.007	10.38±0.008	65.77±0.01
Seeing and Hearing [31]	2.547	2.033	65.82
FoleyCrafter (ours)	1.497 ±0.006	11.94 ±0.217	36.80 ±2.67

Table 2: Quantitative evaluation in terms of temporal synchronization. We report onset detection accuracy (Onset ACC) and average precision (Onset AP) for the generated audios on AVSync [34], which provides onset timestamp labels for assessment, following previous studies [20, 30]. We report the results with error bars calculated from ten times of evaluation with random seeds.

Method	Onset ACC↑	Onset AP↑
SpecVQGAN(Inception) [17]	16.81±1.14	64.64±4.28
SpecVQGAN(ResNet) [17]	26.74±2.35	63.18±0.72
Diff-Foley [20]	21.18±0.08	66.55±0.10
Seeing and Hearing [31]	20.95	60.33
FoleyCrafter (ours)	28.48 ±1.84	68.14 ±1.03

Evaluation Metrics. We employed several evaluation metrics to assess semantic alignment and audio quality, namely Mean KL Divergence (MKL) [17], CLIP similarity, and Frechet Distance (FID) [15], following the methodology of previous studies [20, 28, 31]. MKL measures paired sample-level similarity by calculating the mean KL-divergence across all classes in the test set. CLIP Score compares the similarity between the input video and the generated audio embeddings in the same representation space. For this, we employed Wav2CLIP [29] as the audio encoder and CLIP [23] as the video encoder, as done in previous works [28, 25]. FID assesses the distribution similarity to evaluate the fidelity of the generated audio. Additionally, we follow Du et al. [10, 30] and adopt onset detection accuracy (Onset Acc) and onset detection average precision (Onset AP) to evaluate the generated audios, using the onset ground truth from the datasets.

4.2 Comparison with State-of-the-art

Quantitative Comparison. We present a quantitative comparison of semantic alignment and audio quality on both the VGGSound [4] and AVSync15 [34] datasets, as shown in table 1. The VGGSound dataset consists of 15,446 videos sourced from YouTube, encompassing a wide range of genres. The results indicate that FoleyCrafter achieves superior semantic alignment with visual conditions and provides higher audio fidelity. Furthermore, we report the results for temporal synchronization on the AVSync15 dataset [34], as displayed in table 2. The AVSync15 dataset is a carefully curated collection of video-audio pairs with strong video-audio alignment and onset detection labels. This characteristic makes it a reliable benchmark for evaluating synchronization. The results in table 2 demonstrate that when combined with the temporal controller, FoleyCrafter achieves state-of-the-art performance in terms of temporal synchronization.

Qualitative Comparison. We provide the visualization of generated audio for qualitative comparison on the AVSync15 [34] in fig. 5. It can be observed that FoleyCrafter generates sound at the most accurate time aligned with visual cues, closely resembling the pattern of the ground truth audio.

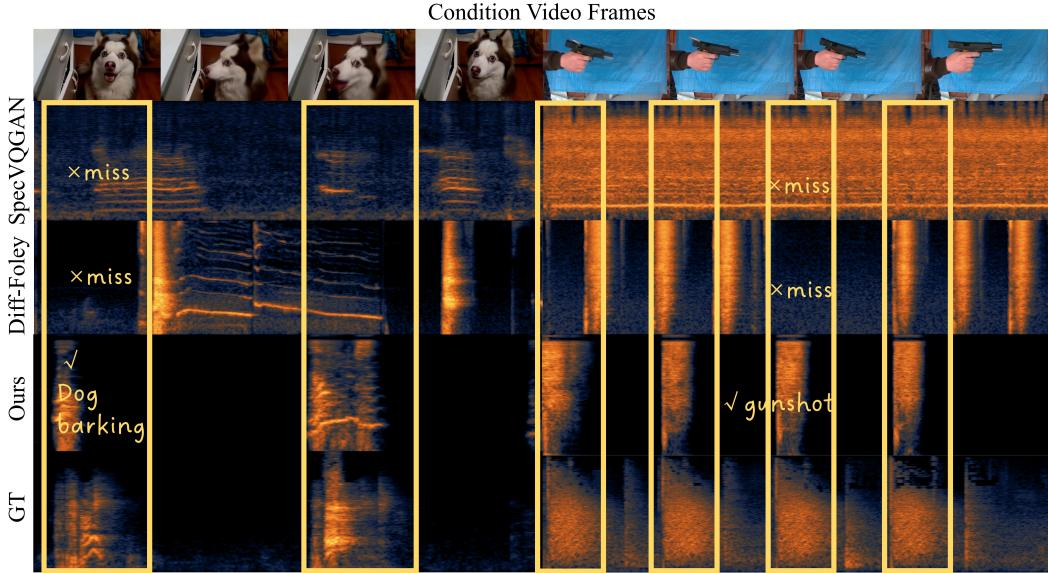


Figure 5: **Qualitative comparison.** As shown in the first case, both SpecVQGAN and Diff-Foley fail to capture the onset of the gunshot sound. In contrast, FoleyCrafter generates the gunshot sound synchronized with the video, showcasing its superior temporal alignment capability.

However, SpecVQGAN tends to introduce more noise, while Diff-Foley often generates more or fewer sound events compared to the ground truth. We provide more results in the Appendix.

One notable advantage of FoleyCrafter is its compatibility with text prompts, allowing for more controllable Foley. We present visualization results of audio generation conditioned on both a video and a text prompt in fig. 6. For instance, when the text prompt describes "high pitch," the corresponding value for high-frequency increases compared to when the prompt describes "low pitch." Moreover, FoleyCrafter can also be utilized with negative prompts to prevent the generation of unwanted sounds. In the third case shown in fig. 6, the visual cues depict a horse running on the beach. By setting the negative prompt as "wind and noise" during inference, the generated audio successfully removes the sound of wind and other environmental noise, resulting in a clear sound of hooves. We provide more results in the Appendix.

4.3 Ablation Study

We conduct ablation studies to validate the effectiveness of semantic adapter and temporal controller. For semantic adapter, we compare the audio-visual relevance of generated samples using different methods of video information injection, as shown in table 3. We consider several baselines for comparison. First, we use a captioner model that utilizes a video-text captioning model [1] to generate text descriptions as inputs to the text-to-audio generator. Second, we directly feed the visual embedding into cross-attention as the text prompt embedding, without any training. Third, we fine-tune the cross-attention module to adapt it to the visual embedding (referred to as ‘visual*’ in table 3). We observed that the caption-based method struggles to capture all the details in the video, resulting in sub-optimal generation results with visual captioning. Using the visual embedding with or without fine-tuning UNet both fail to generate audio that is relevant to the input video. We attribute this to the significant distortion of the original text-to-audio generation framework when incorporating visual information.

For temporal controller, we compare the temporal synchronization performance of FoleyCrafter with and without the module. The results in table 4 demonstrate that the absence of the temporal controller leads to a noticeable decline in onset precision. This decline can be attributed to the fact that the semantic adapter is only capable of capturing video-level semantic information without accurate

Table 3: Ablation on semantic adapter.

Method	MKL \downarrow	CLIP Score \uparrow	FID \downarrow
Captioner	2.331	9.177	67.40
Visual	5.383	2.133	99.77
Visual*	5.821	2.778	95.78
Ours	1.497	11.94	36.80

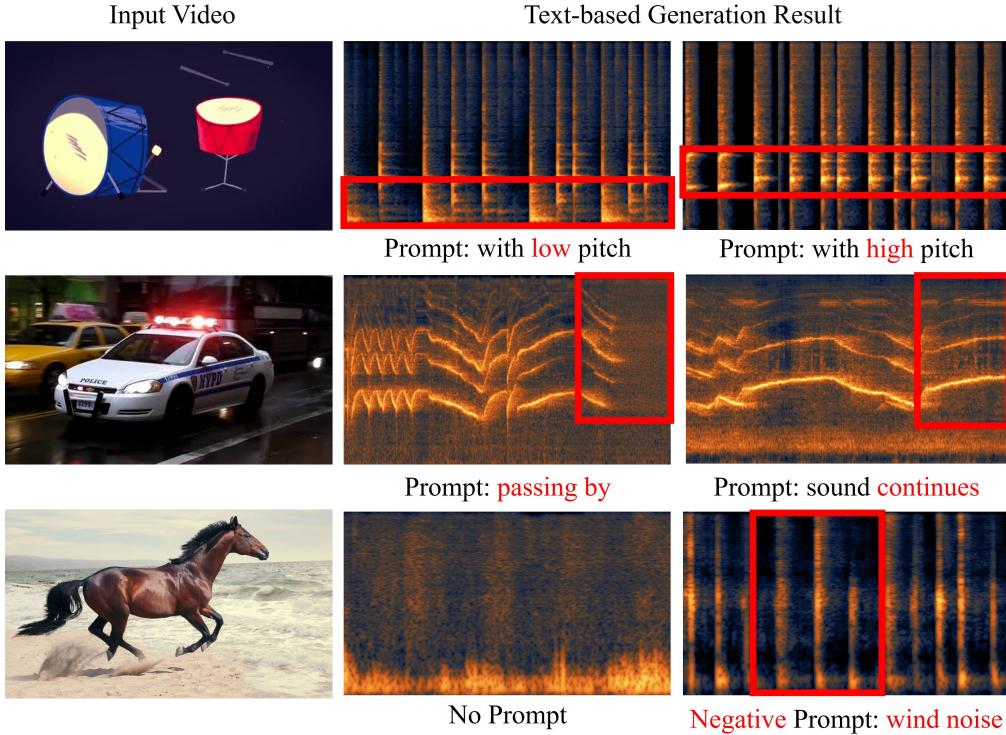


Figure 6: **Video-to-audio generation with text prompts.** FoleyCrafter enhances controllability in video-to-audio generation through text prompts. In the first case, providing a prompt for "high pitch" increases the corresponding value for the drum video. In the third case, a negative prompt like "wind noise" can be used during inference to prevent the generation of wind noise for the horse video.

synchronization features. As a result, it tends to synthesize relevant sounds but with random onset timestamps, leading to a lack of precise temporal alignment.

5 Limitations and Broader Impact

Limitations. Firstly, although the inclusion of the temporal controller enhances the synchronization between the generated audio and the input video, its performance can be ultimately limited by the capabilities of the temporal detector. Second, the effectiveness of the temporal detector is contingent upon the availability of strong and relevant training data. When dealing with more complex visual scenes, predicting the onset event for accurate synchronization becomes challenging due to the scarcity of training data in those specific contexts.

Broader Impact. FoleyCrafter facilitates text-based video-to-audio generation, enabling the generation of sound effects for silent videos and providing control through user prompts. However, it is crucial to acknowledge the potential misuse of such technology for generating fake content on video platforms or social platforms. Users and researchers are strongly advised to exercise caution and carefully screen the use of such technologies to ensure responsible and ethical application.

6 Conclusion

In this paper, we introduce FoleyCrafter for adding sound effects to silent videos. Unlike existing methods that either train a video-to-audio generator from scratch or use video-to-text translation followed by text-to-audio generation, FoleyCrafter is a pluggable module seamlessly integrated into a text-to-audio generator. This integration ensures high-quality audio generation while synchronizing with the video content. FoleyCrafter leverages two key components, namely semantic adapter for semantic alignment and temporal controller for temporal synchronization. Extensive experiments on standard benchmarks demonstrate the effectiveness of FoleyCrafter.

7 Acknowledgement

This project is supported by the National Key R&D Program of China (No. 2022ZD0161600) and partially supported by NSFC (project 62376237) and Shenzhen Science and Technology Program ZDSYS20230626091302006.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014.
- [3] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Audio-visual synchronisation in the wild. *arXiv preprint arXiv:2112.04432*, 2021.
- [4] 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.
- [5] Kan Chen, Chuanxi Zhang, Chen Fang, Zhaowen Wang, Trung Bui, and Ram Nevatia. Visually indicated sound generation by perceptually optimized classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.
- [6] 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, 2020.
- [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.
- [8] Wikipedia contributors. Foley (filmmaking), 2024. Accessed on May 11, 2024.
- [9] Hao-Wen Dong, Xiaoyu Liu, Jordi Pons, Gautam Bhattacharya, Santiago Pascual, Joan Serrà, Taylor Berg-Kirkpatrick, and Julian McAuley. Clipsonic: Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2023.
- [10] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023.
- [11] Freesound Project. Freesound. <https://freesound.org/>, 2024. Accessed: 2024-04-12.
- [12] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned lilm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.
- [13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.
- [14] Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE, 2021.
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [16] 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.

- [17] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *The 32st British Machine Vision Virtual Conference*. BMVA Press, 2021.
- [18] 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. In *Proceedings of the 40th International Conference on Machine Learning*, pages 21450–21474, 2023.
- [19] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.
- [20] 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, 2023.
- [21] Shentong Mo, Jing Shi, and Yapeng Tian. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*, 2024.
- [22] 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*, pages 2405–2413, 2016.
- [23] 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*, pages 8748–8763. PMLR, 2021.
- [24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [25] 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)*, pages 1–5. IEEE, 2023.
- [26] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.
- [27] 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.
- [28] 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, pages 15492–15501, 2024.
- [29] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.
- [30] Zhifeng Xie, Shengye Yu, Mengtian Li, Qile He, Chaofeng Chen, and Yu-Gang Jiang. Sonicvisionlm: Playing sound with vision language models. *CVPR*, 2024.
- [31] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.
- [32] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *arXiv preprint arXiv:2401.01044*, 2024.
- [33] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [34] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. Audio-synchronized visual animation. *arXiv preprint arXiv:2403.05659*, 2024.
- [35] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

Table 5: **User study.** We evaluated the performance of three metrics of different models i.e. semantic and temporal alignment and generation quality.

Method	Semantic	Temporal	Quality
SpecVQGAN	20.29	21.74	20.29
Diff-Foley	20.59	29.41	27.94
V2A-Mapper	44.00	44.00	42.67
FoleyCrafter (ours)	71.23	67.92	69.34

- [36] Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14070–14079, 2021.
- [37] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3550–3558, 2018.

A Appendix

Overview. The appendix includes the following sections:

- **appendix A.1.** Details of training datasets.
- **appendix A.2.** Details of the user study.
- **appendix A.3.** More qualitative results.

Video Result. We also present video results in a separate supplementary file sourced from Sora.

A.1 Details of Training Dataset

FoleyCrafter consist of two key components: semantic adapter and temporal controller which are trained separately. For the training of semantic adapter we use VGGSound [4] as the training set. VGGSound is an audio-visual dataset containing approximately 199,176 videos sourced from YouTube with annotated label classes indicating the video content. We add the prefix 'The sound of' to the label to form the prompt for generation. We train the timestamp detector on the AVSync15 [34]. AVSync15 is a carefully curated dataset from the VGGSound Sync [3] dataset, which contains 1500 strongly correlated audio-visual pairs, making it a high-quality dataset for timestamp detection. We use the AudioSet Strong [14] to train the temporal adapter, which contains 103,463 videos with the sound event and the corresponding timestamp labels.

A.2 Details of user study

To further obtain subjective evaluation results, we conduct a user study. We randomly selected the VGGSound test results generated by different methods for the questionnaire. A total of 20 participants answered our questions. As shown in fig. 7, each question contains audios generated by two methods, one is our method and the other is the baseline e.g. SpecVQGAN [17] Diff-Foley [20] and V2A-Mapper [28]. We ask participants to select the one that has better semantic alignment, temporal alignment, and generation quality. Then the preference score can be calculated as

$$Score = \frac{S}{A} \quad (11)$$

where S is the number of times the method has been selected and A is the appearance times of that method. A higher score means the greater performance of FoleyCrafter. Results can be found at table 5. FoleyCrafter is preferred by users in all three metrics.

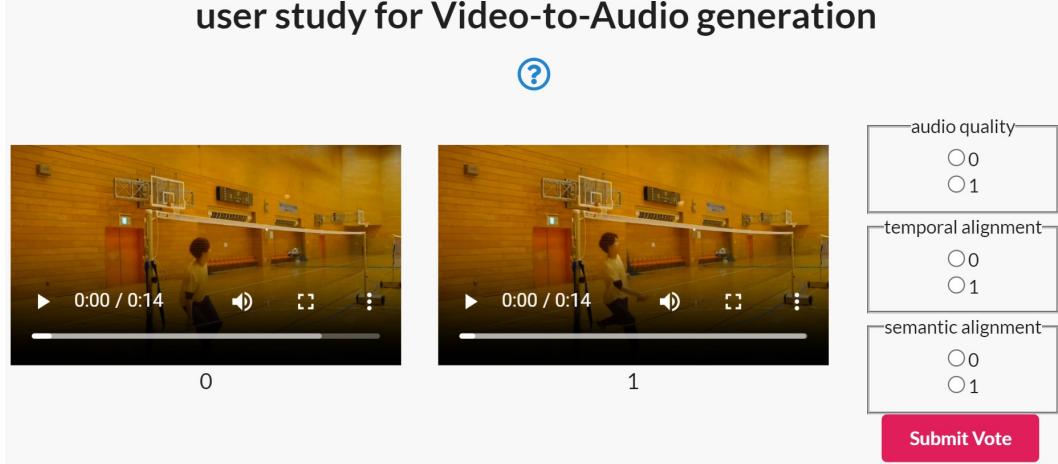


Figure 7: Screenshot of User Study.

A.3 More qualitative results

Foley Generation for Generated Videos. FoleyCrafter is an effective Foley generation tool which can also be used for movie and generated video. Herein, we take the Sora video as example and provide the audio results generated by FoleyCrafter. In the foley process, semantic adapter can directly utilize the rich visual information, which helps FoleyCrafter generate appropriate sound effects for the visual subjects and environment shown in the generated videos.

Text-based video to audio generation. FoleyCrafter achieve text-based video-to-audio generation through parallel cross-attention in semantic adapter. Benefiting from this module, FoleyCrafter can utilize both visual information and text prompts to generate audio. Extra text-based video-to-audio generation results are illustrated in fig. 8 and attached in a separate supplementary file.

Temporal Synchronization Comparison. The temporal controller enhances the temporal alignment in generated audios with visual cues. To show the synchronization ability of FoleyCrafter, we show more intuitive comparison results between FoleyCrafter and other methods as shown in fig. 9. Video results are also provided in a separate supplementary file.

Video to Audio generation on various genres. FoleyCrafter can generate audio for a wide variety of videos. In the supplementary file, we provide generated audio-visual pairs from the VGGSound test cases. The type of video contains realistic video, games, and animation. The main visual objects in the video are people, animals, musical instruments, etc. It fully demonstrates the excellent video-to-audio generation capabilities of FoleyCrafter.

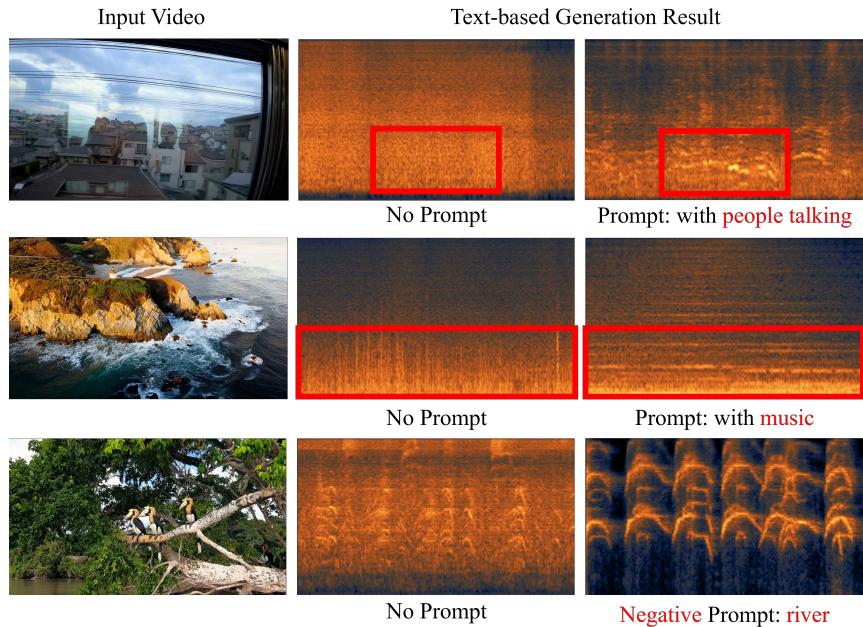


Figure 8: Extra examples on text-based video to audio generation.

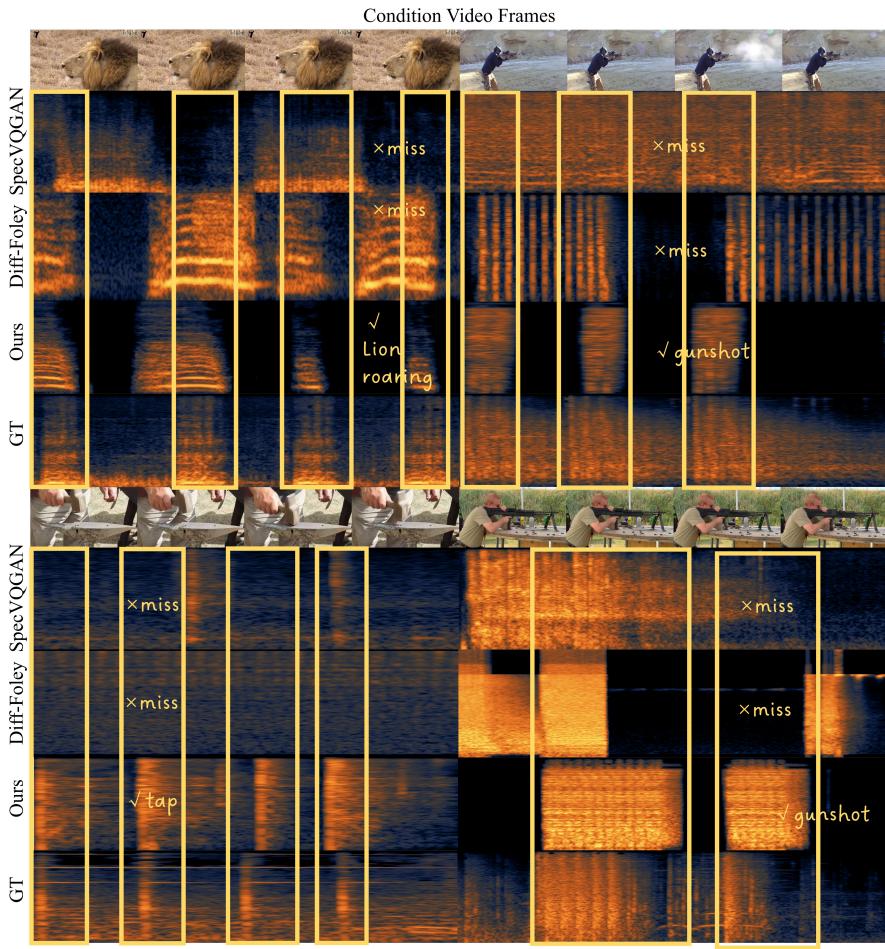


Figure 9: Extra examples on temporal alignment comparison.