Spotlighting Partially Visible Cinematic Language for Video-to-Audio Generation via Self-distillation

Feizhen Huang, **Yu Wu**, **Yutian Lin**[*] and **Bo Du**[*]
School of Computer Science, Wuhan University
{feizhenhuang, wuyucs, yutian.lin, dubo}@whu.edu.cn

Abstract

Video-to-Audio (V2A) Generation achieves significant progress and plays a crucial role in film and video post-production. However, current methods overlook the *cinematic language*, a critical component of artistic expression in filmmaking. As a result, their performance deteriorates in scenarios where Foley targets are only partially visible. To address this challenge, we propose a simple self-distillation approach to extend V2A models to cinematic language scenarios. By simulating the cinematic language variations, the student model learns to align the video features of training pairs with the same audio-visual correspondences, enabling it to effectively capture the associations between sounds and partial visual information. Our method not only achieves impressive improvements under partial visibility across all evaluation metrics, but also enhances performance on the large-scale V2A dataset, VGGSound.

1 Introduction

Video-to-Audio (V2A) Generation [Luo *et al.*, 2024; Wang *et al.*, 2024b; Wang *et al.*, 2024a; Du *et al.*, 2023], which generates corresponding audio directly from silent videos, has significant applications in film and video post-production.

During live filming, capturing clean sound is often challenging due to ambient noise interference, the faintness of certain sounds, and other factors. As a result, most sounds must be recreated in post-production. As illustrated in Figure 1, the process of adding relevant and synchronized sound effects to silent videos is known as *Foley* [Ament, 2014]. Traditional Foley requires skilled Foley artists to reproduce different sounds by manipulating various objects. The success of this process heavily depends on the artist's expertise and experience. Moreover, due to the vast array of sound categories and manipulable objects, this process is labor-intensive and time-consuming. This complexity hinders large-scale replication and individual video creation. In contrast, V2A Generation [Luo *et al.*, 2024; Wang *et al.*, 2024b; Wang *et al.*, 2024a; Du *et al.*, 2023] offers an appealing alternative by automatically generating corresponding audio directly from silent videos. This innovative approach alleviates the burden on human labor and offers a more scalable and faster solution for both individuals and companies.

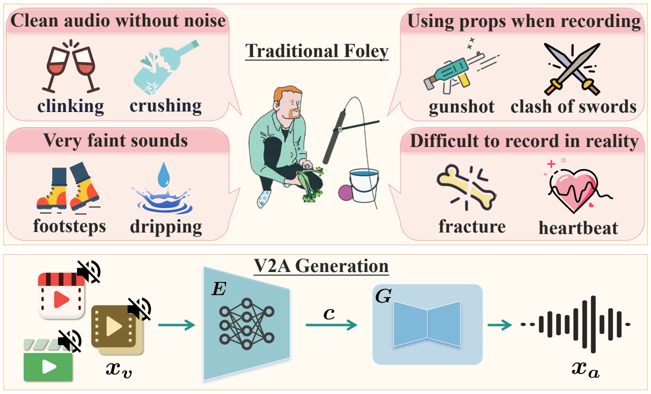


Figure 1: *Foley* is the process of adding sound effects to silent videos, playing an essential role in film/video production due to factors illustrated in the pink boxes. Traditional Foley relies on skilled Foley artists to manually reproduce sounds, whereas V2A Generation can directly generate corresponding audio from silent videos, providing a more efficient and convenient solution.

Significant progress has been made in V2A Generation. SpecVQGAN [Iashin and Rahtu, 2021] leads the way in audio generation for open-domain videos. Diff-Foley [Luo *et al.*, 2024] stands out for its focus on addressing the challenge of audio-visual temporal synchronization, sparking a wave of subsequent research with impressive advancements in model performance [Wang *et al.*, 2024b; Zhang *et al.*, 2024a; Ren *et al.*, 2024], precise temporal alignment [Pascual *et al.*, 2025; Viertola *et al.*, 2024], lightweight designs [Wang *et al.*, 2024a], and text control integration [Du *et al.*, 2023; Xie *et al.*, 2024; Jeong *et al.*, 2024]. However, current methods overlook the role of *cinematic language* [Mercado, 2019], a cornerstone of artistic expression in film and video.

Cinematic language [Mercado, 2019] has the expressive power of storytelling, enabling directors to convey subjective intentions effectively. Camera techniques such as close-ups and camera movements are common to cinematic language. For instance, close-ups highlight specific features of charac-

[*] Corresponding author.

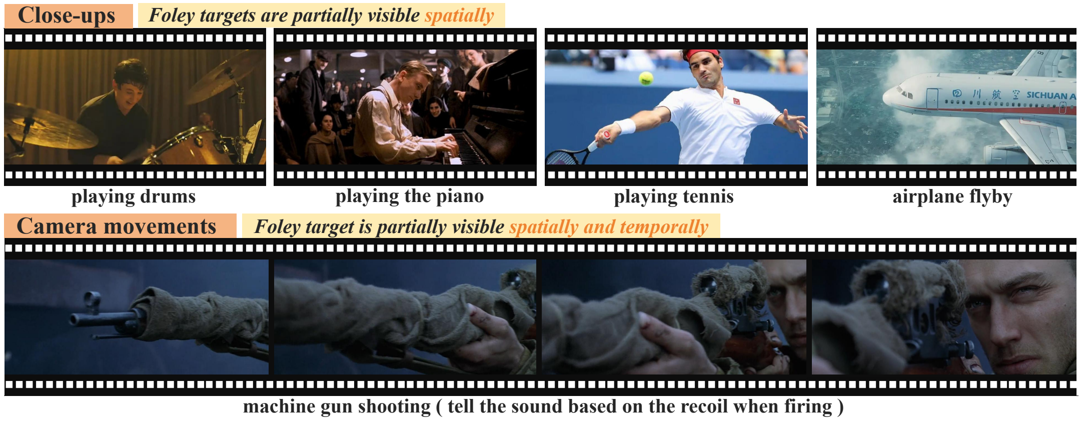

Close-ups
Foley targets are partially visible spatially
playing drums
playing the piano
playing tennis
airplane flyby
Camera movements
Foley target is partially visible spatially and temporally
machine gun shooting (tell the sound based on the recoil when firing)

Figure 2: *Cinematic language* is a fundamental element of artistic expression in film, such as close-ups and camera movements. These camera techniques often create scenarios where Foley targets are only partially visible spatially or temporally. We refer to them as **partial visibility**.

ters or objects by zooming in, while camera movements dynamically introduce or remove them from the frame. These camera techniques aim to portray characters as shown in Figure 2, such as characters playing instruments or engaging in sports, where Foley targets may only be *partially visible in the spatial frame or temporal sequence*. We refer to these situations where Foley targets are partially visible spatially or temporally as **partial visibility**. In these situations, humans can infer sound based on other visual clues and temporal clues, even with incomplete information. However, current V2A methods struggle to handle the challenges posed by partial visibility, leading to poor performance in such scenarios.

State-of-the-art (SOTA) V2A models [Luo *et al.*, 2024; Wang *et al.*, 2024b; Zhang *et al.*, 2024a] typically employ a two-stage training process: large-scale pretraining on audio-visual datasets to learn a robust video encoder, followed by training an audio generator conditioned on the extracted video features. In cinematic language scenarios, partial visibility of Foley targets causes the video encoder to extract inaccurate video features. This misrepresentation results in incorrect audio-visual associations, hindering the generation of corresponding audio. To directly adapt these models for cinematic scenarios, a straightforward approach would be collecting cinematic videos to retrain both the video encoder and the audio generator. However, this approach faces substantial obstacles. High-quality cinematic video clips are not only scarce but also restricted by copyright. Furthermore, training models directly on cinematic videos with incomplete visual information can be problematic. Such outliers may cause the model puzzled, rather than enabling it to effectively learn from partial visual information, ultimately disrupting optimization and resulting in unexpectedly poor performance. Interestingly, current V2A models exhibit strong performance on non-cinematic videos. This suggests that the pre-trained knowledge can be leveraged to bridge the performance gap between non-cinematic and cinematic scenarios.

In this paper, we propose a simple teacher-student framework to capture partial audio-visual clues by constructing paired supervision video, focusing on addressing the challenges of partial visibility in cinematic scenarios. First, we simulate fundamental cinematic language variations to create paired training videos: one with cinematic variations and the other without while preserving consistent audio-visual correspondences. The latter serves as a supervision signal, subtly guiding the model to transfer its prior knowledge to understand the partial visual information. These paired videos are more effective than directly training on collected cinematic videos. Next, we adopt a teacher-student framework to align the video features from these paired training videos, benefiting from the supervision signals. This not only enables the student model to learn the associations between sounds and partial visual clues but also preserves its original performance. Our approach is both efficient and general, requiring neither additional cinematic data nor modifications to the subsequent audio generative model. It achieves impressive improvements under partial visibility across all evaluation metrics and enhances performance on the large-scale V2A dataset, VGGSound [Luo *et al.*, 2024] compared to the baseline. Our contributions are as follows:

- We are the first to focus on cinematic language in the V2A generation area, where the Foley targets are only partially visible.
- We propose an efficient and general teacher-student framework that captures partial audio-visual clues by creating paired training videos while maintaining its original performance.

2 Related Work

2.1 V2A Generation

The progress in V2A Generation research attracts a lot of attention. Early methods [Chen *et al.*, 2020b] typically train separate models for each video category to generate more relevant and higher-fidelity audio, which limits the generalization abilities of models. SpecVQGAN [Iashin and Rahtu, 2021] makes a pioneering effort in audio generation for open-domain videos. Diff-Foley [Luo *et al.*, 2024] stands out by addressing the challenge of audio-visual temporal synchronization, inspiring a wave of subsequent studies with impressive results in enhanced model performance [Wang *et al.*, 2024b; Zhang *et al.*, 2024a; Ren *et al.*, 2024], precise temporal alignment [Pascual *et al.*, 2025; Viertola *et al.*, 2024], lightweight designs [Wang *et al.*, 2024a], and the integration of text control [Du *et al.*, 2023; Xie *et al.*, 2024; Jeong *et al.*, 2024]. Despite these advancements, current methods all neglect cinematic language and perform poorly in such scenarios, which is the focus of our work.

2.2 Partial Visual Clues Perception

Partial visual clues often arise when the predicted targets are partially occluded or out of sight. The ability to perceive partial visual information is a common challenge across various visual tasks. In traditional computer vision tasks, such as image classification and segmentation, models [He *et al.*, 2016; Huang *et al.*, 2017] need to make inferences based on visible portions. For person re-identification, models [Miao *et al.*, 2019; Li *et al.*, 2021] are required to identify the same individual despite variations caused by camera perspectives and occlusions. Similarly, robots need to predict human motion directions and trajectories based on visible human body parts [Zhao *et al.*, 2024]. Scene understanding tasks [Zhan *et al.*, 2020; Lee and Park, 2022] also face this challenge, as models must infer semantic information from incomplete visual clues, especially in applications like autonomous driving. In our research, we encounter similar difficulties related to partial visibility in V2A generation.

2.3 Teacher-Student Methodology

The Teacher-Student methodology, also known as knowledge distillation [Hinton, 2015], involves training a student model to learn from a teacher model. This approach aims to achieve model compression or enhance performance. Self-distillation [Zhu *et al.*, 2018; Xu and Liu, 2019; Zhang *et al.*, 2019; Yun *et al.*, 2019] is a unique variant of the Teacher-Student methodology, where the teacher and student models are different versions or stages of the same model. This characteristic enables the model to learn from itself. In our work, we adopt this practical training methodology, leveraging the prior knowledge of a pre-trained video encoder to guide the learning of audio-visual correlation under partial visibility.

2.4 Data Augmentation

Data augmentation is widely applied to improve model performance and generalization abilities. In image-related fields, data augmentation techniques are widely used in traditional computer vision tasks [He *et al.*, 2016; Huang *et al.*, 2017], such as flipping, rotation, translation, and noise injection. Moreover, video-based data augmentation methods [Gowda *et al.*, 2022] also gain attention in video action recognition. These methods concentrate more on actions and temporal information by altering color [Zhang *et al.*, 2024b] or blending foregrounds and backgrounds [Ding *et al.*, 2022; Li *et al.*, 2023; Wang *et al.*, 2021] from different videos. Our method simulates cinematic language variations to emphasize audio-visual alignment using partial visual clues.

3 Method

An overview of our proposed method is shown in Figure 3. First, we introduce the motivation behind our method in Section 3.1. Next, we propose cinematic language variations f in Section 3.2 and adopt the teacher-student methodology to learn partial visibility in Section 3.3. Finally, in Section 3.4, we describe the latent diffusion models for V2A Generation with cinematic language.

3.1 Motivation

Our method targets the challenge of partial visibility in cinematic language scenarios. To do so, we aim to improve the model's ability to capture audio-visual corrections from partial visual information.

As illustrated in Figure 1, most Video-to-Audio (V2A) models follow a two-stage process: first, a feature extractor E extracts visual features $c = E(x_v)$ from the video x_v. Subsequently, using these visual features c as a condition, a generative model G generates the corresponding audio $x_a = G(c)$. Indeed, the performance of the generative model G relies heavily on the semantic and temporal information embedded in the condition c. Therefore, the quality of the condition c is critical for generating relevant and synchronized audio.

For normal video clips x_v, i.e., *without* cinematic language, the generative model G performs well, indicating that the extracted visual features $c = E(x_v)$ provide accurate audio information, making them ideal generation conditions. However, when x'_v *with* cinematic language, the quality of the generated audio x'_a deteriorates significantly. Given the strong performance of G on non-cinematic videos, we reasonably infer that this degradation stems from the poor generation conditions $c' = E(x'_v)$, rather than the generative model G itself. In other words, when the Foley targets are only partially visible, the video encoder E struggles to capture meaningful audio-visual associations from partial visual clues, thereby failing to provide a suitable generation prior.

Consequently, assisting the video encoder in learning the associations between sounds and partial visual clues and providing better conditions for the generative model G emerges as a natural approach.

3.2 Cinematic Language Variations

We propose Cinematic Language Variations f to simulate partial visibility in cinematic language [Mercado, 2019], helping to establish audio-visual associations in such scenarios. This is more effective than directly training on collected cinematic videos. In cinematic language, close-ups and camera movements are two common camera techniques that cause situations where the Foley targets are partially visible in the spatial or temporal dimension. The cinematic language variations f simulates the process of these two techniques as examples.

Given that close-ups emphasize specific local details of characters or objects, f_{cu} uniformly crops the video frames. The cropping sizes H' and W' are randomly selected within a reasonable range, which are determined as follows:

$$H' = H \times r_h, \quad W' = W \times r_w, \tag{1}$$

where H and W are the original height and width of the video frames, and $r_h, r_w \sim U(a_1, a_2)$. The range $[a_1, a_2]$ should

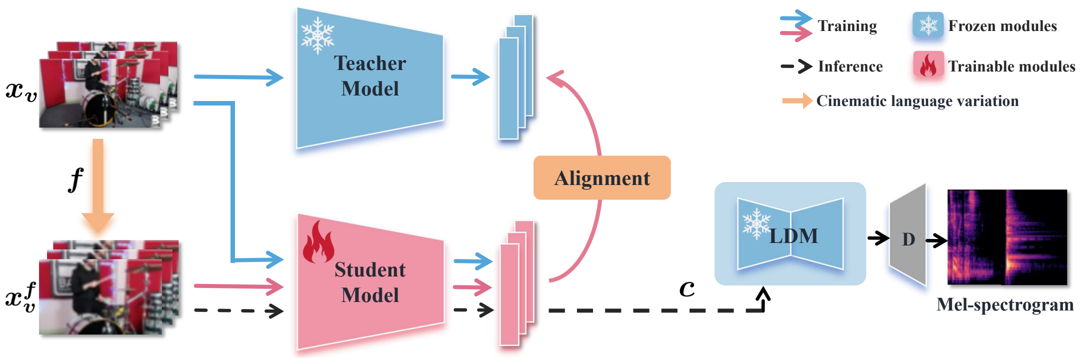


Figure 3: Our approach involves two key components: cinematic language variations f and a teacher-student framework. First, cinematic language variations f create training video pairs (x_v, x_v^f), where x_v^f retains the same semantic and temporal information of Foley target as x_v. By aligning the visual features extracted from x_v^f with those from x_v can facilitate the student model in learning the audio-visual connection with partial visibility. After training, the student model can provide a better generation condition for the subsequent inference stage

be chosen to ensure that the Foley targets are partially visible while retaining sufficient visual information.

To simulate camera movements, f_{cm} utilizes the same cropping size above and shifts the shot along the central axis of the video frames. The shift direction is:

- When $W > H$, shift left or right at random.
- When $H > W$, shift up or down at random.

3.3 Partial Visibility Learning by Self-distillation

We adopt the teacher-student methodology to learn the partial visibility simulated by Cinematic Language Variations f. Given that the pre-trained video encoder contains rich audio-visual priors, we utilize it as the teacher model T to guide the student model S in learning the association between sounds and partial visual clues, as illustrated in Figure 3.

Given video data x_v, by simulating cinematic language variations f, we obtain video x_v^f. Derived from x_v, x_v^f retains the same semantic and temporal information of the Foley target with x_v. They constitute training video pairs (x_v, x_v^f) with the same audio-visual correspondence. The only difference between training video pairs is the partial visibility of the Foley target: it is fully visible in video x_v while partially visible in video x_v^f. Since the Foley target in video x_v is fully visible, the visual features $c_t = T(x_v)$ extracted by the teacher model T accurately encompass sufficient audio information for conditioning. Thus, aligning the features $c_{sf} = S(x_v^f)$ extracted by the student model S with feature c_t can facilitate the student model S in learning the audio-visual connection between sounds and partial visual clues. During training, we also align $c_s = S(x_v)$ with the feature c_t to maintain the performance on the original dataset. The optimization loss L_p is defined as follows:

$$L_p = \cos(c_t, c_{s'}) + \text{MSE}(c_t, c_{s'}), \quad c_{s'} = c_s \text{ or } c_{sf} \tag{2}$$

We introduce k as the proportion of training video clips with cinematic language variations. By aligning the visual features of training video pairs, we can effectively guide the student model S to map videos with partial visibility into the original feature space of the teacher model. This approach not only learns a better generation condition but also requires no modifications to the subsequent generative model.

3.4 Latent Diffusion Models

Latent Diffusion Models (LDMs) [Rombach *et al.*, 2022] are probabilistic generative models that map the data distribution into a low-dimensional latent space, consisting of an autoencoder and a U-Net denoiser. In the V2A task, the latent encoder E encodes the Mel-spectrogram $x \sim p(x)$ into the latent representation $z = E(x)$, and the UNet denoiser ϵ_θ is then trained to reverse the noise addition to generate new latent representations. Under a given condition c, the optimization [Ho *et al.*, 2020; Song *et al.*, 2021] process of LDMs can be defined as follows:

$$L_{LDM} := \mathbb{E}_{z\sim\mathcal{E}(z),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{3}$$

where ϵ represents Gaussian noise, and z_t denotes the latent representation at time step t.

For conditional LDMs, classifier-free guidance (CFG) [Ho and Salimans, 2022] is a widely used alternative to classifier guidance (CG) [Dhariwal and Nichol, 2021]. CFG jointly trains conditional $\epsilon_\theta(x_t, c, t)$ and unconditional $\epsilon_\theta(x_t, t)$ diffusion models by randomly dropping the condition c. During sampling, the noise prediction is calculated as:

$$\hat{\epsilon}_\theta(z_t, c, t) = w\epsilon_\theta(z_t, c, t) + (1 - w)\epsilon_\theta(z_t, t), \tag{4}$$

where w is the guidance scale.

In our study, we build upon the open-source Diff-Foley [Luo *et al.*, 2024], an LDM conditioned on video features extracted from a frozen CAVP video encoder, to achieve Foley in cinematic scenarios.

4 Experiments

Datasets. We conduct our experiments using VGGSound [Chen *et al.*, 2020a], a large-scale audio-visual dataset containing over 200,000 video clips across 309 distinct sound categories. Existing V2A methods utilize VGGSound for both training and evaluation. We follow the original VGGSound train/test split.

Baseline. Given our choice of Diff-Foley [Luo *et al.*, 2024] as the foundation model for our study, we select Diff-Foley as our baseline, which is a leading open-source V2A model. For

Test Set	Method	FAD↓	FD↓	KID(10^{-3}) ↓	KL↓	IS↑	Align Acc(%)↑
VGGSound	Diff-Foley	7.481	25.445	10.71	3.280	11.670	**92.946**
	Ours	**7.173**	**24.532**	**10.05**	**3.235**	**11.835**	91.718
VGG-CU	Diff-Foley	8.926	28.843	12.17	3.824	11.136	74.882
	Ours	**7.825**	**25.661**	**10.45**	**3.438**	**11.601**	**85.071**
VGG-CM	Diff-Foley	9.164	28.394	11.48	3.843	10.665	72.570
	Ours	**8.194**	**25.932**	**10.23**	**3.508**	**11.047**	**81.850**

Table 1: Evaluation results for Video-to-Audio generation across three test sets: the original VGGSound test set, VGG-CU (close-up) test set, and VGG-CM (camera movement) test set. During training, only cinematic language variation f_{cu} is applied to VGGSound [Chen *et al.*, 2020a] training set with $k = 75\%$. Diff-Foley baseline exhibits a notable performance drop on both VGG-CU and VGG-CM test sets, compared to its original performance on VGGSound test set. In contrast, our method significantly outperforms the baseline across all evaluation metrics on these two test sets, suggesting that it learns the audio-visual associations under partial visibility.

simplicity, we adopt the CFG configuration for both generation and comparison, keeping all other settings unchanged.

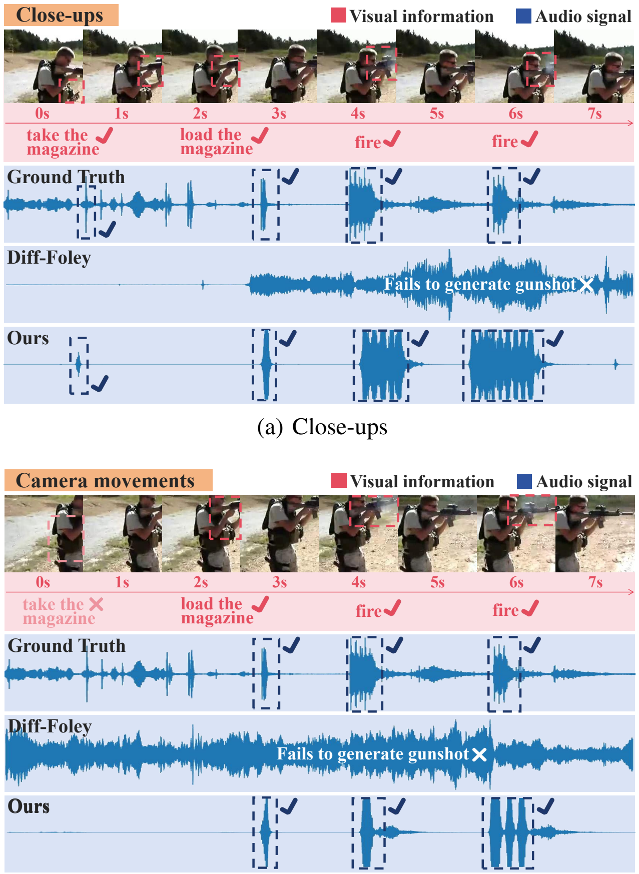
Evaluation Metrics. For evaluation, we adopt evaluation metrics FAD, FD, KID, KL, and ISc from audioLDM [Liu *et al.*, 2023], alongside Align Acc from Diff-Foley [Luo *et al.*, 2024]. FAD, FD, and KID measure the similarity between real and generated audio, while KL assesses the paired similarity in probability distributions. ISc measures the diversity and quality of generated audio, and Align Acc assesses the audio-visual synchronization.

Implementation Details. For model configuration, we employ a pre-trained video encoder from CAVP [Luo *et al.*, 2024] as the teacher model. The student model adopts the same architecture, with weights initialized from the teacher model's pre-trained parameters. The input video clips are sampled at 4 frames per second (FPS), resulting in $T = 4N$ frames for each N-second video clip. Then the input video $x_v \in \mathbb{R}^{T\times 3\times H\times W}$ is extracted by the video encoder into video feature $E_v \in \mathbb{R}^{T\times C}$, with the feature dimension $C = 512$. For training, we only apply cinematic language variation f_{cu} on VGGSound [Chen *et al.*, 2020a] training set with $k = 75\%$, where $a_1 = 0.4$ and $a_2 = 0.6$. The student model is trained for 25 epochs on 4 NVIDIA 4090 GPUs, using the AdamW optimizer with a learning rate of 5×10^{-4} and a total batch size of 32.

To evaluate performance under partial visibility, we create two modified test sets by applying cinematic language variations to the VGGSound [Chen *et al.*, 2020a] test set. Specifically, f_{cu} is used to create VGG-CU (close-ups) test set, and f_{cm} is used to create VGG-CM (camera movements) test set. Following Diff-Foley [Luo *et al.*, 2024], we generate 10 samples per video to ensure reliable evaluation. For simplicity, we use only CFG [Ho and Salimans, 2022] configuration in Diff-Foley, keeping all other experimental settings unchanged, including the DPM-Solver [Lu *et al.*, 2022] Sampler with 25 inference steps and CFG scale $\omega = 4.5$.

4.1 V2A Generation with Cinematic Language

Simulated Cinematic Scene on VGGSound. Table 1 presents the quantitative results across three test sets: the original VGGSound test set, VGG-CU test set (created using f_{cu}), and VGG-CM test set (created using f_{cm}). In cinematic language scenarios, we observe that Diff-Foley [Luo *et al.*, 2024] baseline shows a significant decline in all evaluation metrics on both VGG-CU and VGG-CM test sets compared to its performance on the original VGGSound test set. In contrast, our method outperforms the baseline by a substantial margin across all evaluation metrics on these two test



(a) Close-ups

(b) Camera Movements

Figure 4: The figures show the qualitative results for V2A generation in cinematic language scenarios involving close-ups and camera movements. Taking a machine gun shooting video as an example, the pink dashed boxes and text mark the partial visual information, while the blue parts represent the corresponding auditory signals.

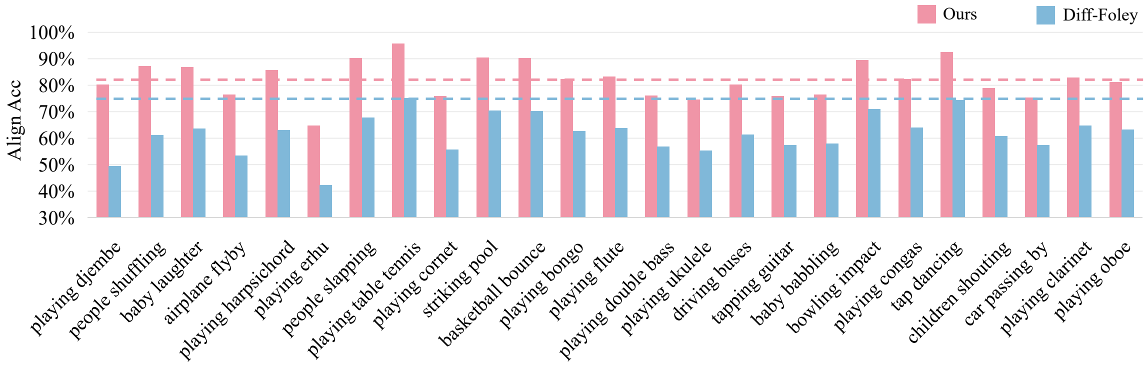


Figure 5: We evaluate the Align Acc metric for each video category on VGG-CU test set and present the top 25 video categories (309 in total) with the most significant improvements. The dashed line represents the overall Align Acc across the entire VGG-CU test set.

prefer Diff-Foley	prefer Ours
158 (25.48%)	462 (74.51%)

Table 2: A human study is conducted using collected 31 YouTube videos with diverse cinematic scenarios from real-world.

sets. Notably, even when trained only on VGG-CU training set, our model exhibits considerable improvement on VGG-CM test set. These results suggest that our method effectively learns the associations between audio and partial visual information. Furthermore, by learning partial visual information, our method achieves improved performance on the original VGGSound test set compared to the baseline.

Figure 4 presents the qualitative results for V2A generation in cinematic language scenarios involving close-ups and camera movements. Taking the machine gun shooting video as an example, partial visual information is highlighted with pink dashed boxes and pink text, while the corresponding auditory signals are represented in blue. As depicted in Figure 4, we can easily identify the machine gun and the moment of the sound from the video frames, even when the gun is partially visible. However, Diff-Foley fails to generate the appropriate gunshot and other sounds at the correct moment. In contrast, our method successfully generates synchronized gunshot audio when the Foley target is partially visible. Notably, in camera movements scenarios shown in Figure 4(b), the visual clues in the 0s and 1s video frames are lost. Our method does not generate sound when the visual clues are completely missing. While in the following video frames, our method correctly generates relevant and synchronized audio. These results suggest that our method correctly learns the audio-visual associations under partial visibility.

Real Cinematic Scene. To evaluate performance in real-world cinematic scenarios, we collect a new set of 31 YouTube videos that feature a wide range of cinematic techniques, ensuring no overlap with the original VGGSound test set. These videos include characteristics such as close-ups, camera movements, zooms, scene transitions, as well as non-partial factors like color and lighting variations. We conducted a human study with 20 participants and the results confirm the superiority of our method as shown in Table 2,

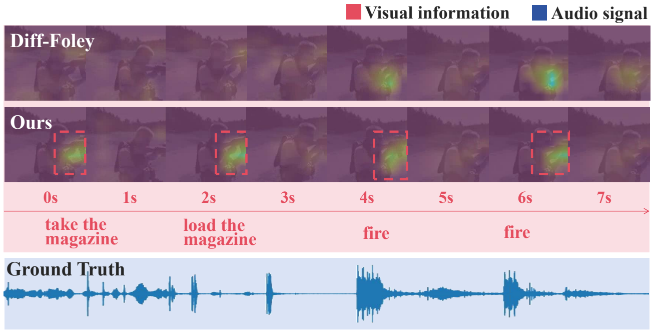


Figure 6: We use the Grad-CAM [Selvaraju *et al.*, 2017] to visualize the model's attention in close-up scenarios. Taking the machine gun shooting video as an example, the pink dashed boxes and text represent the partial visual information, while the corresponding audio ground truth is represented in blue.

4.2 Analysis of Partial Visibility Learning

Visualization Analysis. To better understand what the model learns, we visualize the model's attention areas along the video frames in close-up scenarios. As shown in Figure 6, we observe that the four regions highlighted by the model's attention (denoted by pink dashed lines) are semantically relevant to the partial visual clues. Compared to the ground truth audio, these highlighted regions are also temporally aligned with the corresponding sound. In contrast, the highlighted regions of Diff-Foley's attention are misaligned with the partial visual clues and are not temporally synchronized with the audio. These visualized results demonstrate that our model accurately captures the semantic and temporal association between sound and partial visual clues.

Improvements in Different Video Categories. To further understand where our approach is superior, We evaluate the Align Acc metric for each video category on VGG-CU test set and present the top 25 video categories (309 in total) with the most significant improvements. As shown in Figure 5, these categories primarily include instrumental sounds, ball impact sounds, and vehicle movement sounds that are commonly found in cinematic scenarios. In these categories, Diff-Foley exhibits notably lower audio-visual consistency com-

Test Set	Training f	FAD↓	FD↓	KID(10^{-3}) ↓	KL↓	IS↑	Align Acc(%)↑
VGG-CU	f_{cu}	**7.905**	**26.341**	**10.65**	**3.511**	11.411	82.951
	f_{cm}	8.505	27.403	11.30	3.652	11.645	79.420
	f_{cu}&f_{cm}	8.157	26.460	10.76	3.513	**11.705**	**83.311**
VGG-CM	f_{cu}	**8.311**	**26.861**	**10.59**	3.585	10.918	79.437
	f_{cm}	8.699	27.523	11.19	3.624	**11.418**	79.115
	f_{cu}&f_{cm}	8.653	27.057	11.02	**3.558**	11.311	**80.514**
VGGSound	f_{cu}	**7.174**	**24.686**	**10.17**	**3.242**	11.739	91.398
	f_{cm}	7.306	24.917	10.38	3.257	**11.933**	**91.560**
	f_{cu}&f_{cm}	7.388	24.889	10.33	3.261	11.895	91.119

Table 3: We explore the effect of different cinematic language variations f during training, including f_{cu} (close-ups), f_{cm} (camera movements), and $f_{cu\&cm}$ (a combination of both). The respective proportions are set as $k_{f_{cu}} = 50\%$, $k_{f_{cm}} = 50\%$, and $k_{f_{cu\&cm}} = 66.7\%$ (with $f_{cu} : f_{cm} = 1 : 1$). The evaluation settings remain consistent with the previous experiments.

Test Set	Proportion k	FAD↓	FD↓	KID(10^{-3}) ↓	KL↓	IS↑	Align Acc(%)↑
VGG-CU	50%	7.905	26.341	10.65	3.511	11.411	82.951
	75%	**7.825**	**25.661**	**10.45**	**3.438**	11.601	**85.071**
	100%	7.992	26.135	10.67	3.468	**11.727**	84.121
VGG-CM	50%	8.311	26.861	10.59	3.585	10.918	79.437
	75%	**8.194**	**25.932**	**10.23**	**3.508**	11.047	**81.850**
	100%	8.393	26.366	10.46	3.520	**11.130**	81.279
VGGSound	50%	7.174	24.686	10.17	3.242	11.739	91.398
	75%	**7.173**	**24.532**	**10.05**	**3.235**	**11.835**	**91.718**
	100%	8.036	26.487	10.68	3.425	11.454	86.553

Table 4: We explore the proportion k of training video clips with cinematic language variation f_{cu}.

pared to its average performance. In contrast, our method achieves significantly better results, demonstrating its effectiveness in addressing the challenges of partial visibility in cinematic scenarios.

4.3 Ablation Study

The Impact of Different Variations. As shown in Table 3, we introduce three cinematic language variations during training to assess their impact: f_{cu}, f_{cm}, and $f_{cu\&cm}$. The f_{cu} variation simulates spatial partial visibility, while f_{cm} simulates temporal partial visibility. According to FAD, FD, and KID metrics, only using f_{cu} yields the best improvements across all three test scenarios, suggesting that the generated audio is semantically more aligned with the video. This indicates that simulating spatial partial visibility (f_{cu}) is more effective than simulating temporal partial visibility (f_{cm}) for learning the semantic association between sounds and partial visual clues. On the other hand, in terms of Align ACC, $f_{cu\&cm}$ achieves the best improvements in VGG-CU and VGG-CM test sets, with the generated audio showing better temporal alignment with the videos. This suggests that simulating both spatial and temporal partial visibility ($f_{cu\&cm}$) enhances the model's ability to learn the temporal synchronization between sounds and partial visual clues.

The Proportion of Variation. Table 4 demonstrates the effect of the proportion k of training video clips with cinematic language variation f_{cu}. In VGG-CU and VGG-CM test sets, model performance improves as k increases from 50% to 75%, showing the effectiveness of f_{cu}. When k reaches 100%, where no data from the original dataset is used, the model's performance shows a decline compared to 75%, emphasizing the importance of training with the paired data.

5 Limitations and Broader Impact

Limitations. Our method excels in scenarios with partial visibility such as camera movements and close-ups, but its adaptability and performance across other diverse cinematic languages still require further exploration.

Broader Impact. V2A boosts video production efficiency, offering substantial benefits to creators. However, vigilance and regulations are still needed to prevent potential misuse.

6 Conclusion

We present a simple yet effective method to address the challenge of partial visibility in cinematic language scenarios by providing a better condition for the audio generator. By simulating the cinematic language variations, the student model aligns the video features from training video pairs with the same audio-visual correspondences, which helps learn the associations between sounds and partial visual clues. Experimental results demonstrate the impressive improvements of our approach, not only in partial visibility scenarios but also on the original V2A dataset, VGGSound.

Acknowledgments

This work was partially supported by the Key Research and Development Program of Yunnan Province under Grant 202403AA080002, partially supported by the National Natural Science Foundation of China under Grant 62471344, and partially supported by the National Natural Science Foundation of China under Grant 62372341.

References

[Ament, 2014] Vanessa Theme Ament. *The Foley grail: The art of performing sound for film, games, and animation*. Routledge, 2014.

[Chen *et al.*, 2020a] 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.

[Chen *et al.*, 2020b] 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.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Ding *et al.*, 2022] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9716–9726, 2022.

[Du *et al.*, 2023] 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.

[Gowda *et al.*, 2022] Shreyank N Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. Learn2augment: learning to composite videos for data augmentation in action recognition. In *European conference on computer vision*, pages 242–259. Springer, 2022.

[He *et al.*, 2016] 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*, pages 770–778, 2016.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Iashin and Rahtu, 2021] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 2. BMVA Press, 2021.

[Jeong *et al.*, 2024] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. *arXiv preprint arXiv:2407.05551*, 2024.

[Lee and Park, 2022] Hyunmin Lee and Jaesik Park. Instance-wise occlusion and depth orders in natural scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21210–21221, 2022.

[Li *et al.*, 2021] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2898–2907, 2021.

[Li *et al.*, 2023] Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

[Liu *et al.*, 2023] 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. *Proceedings of the International Conference on Machine Learning*, pages 21450–21474, 2023.

[Lu *et al.*, 2022] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

[Luo *et al.*, 2024] 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, 2024.

[Mercado, 2019] Gustavo Mercado. *The filmmaker's eye: The language of the lens: The power of lenses and the expressive cinematic image*. Routledge, 2019.

[Miao *et al.*, 2019] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 542–551, 2019.

[Pascual *et al.*, 2025] Santiago Pascual, Chunghsin Yeh, Ioannis Tsiamas, and Joan Serrà. Masked generative video-to-audio transformers with enhanced synchronicity. In *European Conference on Computer Vision*, pages 247–264. Springer, 2025.

[Ren *et al.*, 2024] Yong Ren, Chenxing Li, Manjie Xu, Wei Liang, Yu Gu, Rilin Chen, and Dong Yu. Sta-v2a: Video-to-audio generation with semantic and temporal alignment. *arXiv preprint arXiv:2409.08601*, 2024.

[Rombach *et al.*, 2022] 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.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Viertola *et al.*, 2024] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *arXiv preprint arXiv:2409.13689*, 2024.

[Wang *et al.*, 2021] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11804–11813, 2021.

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

[Wang *et al.*, 2024b] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024.

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

[Xu and Liu, 2019] Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5565–5572, 2019.

[Yun *et al.*, 2019] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[Zhan *et al.*, 2020] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3784–3792, 2020.

[Zhang *et al.*, 2019] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019.

[Zhang *et al.*, 2024a] 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.

[Zhang *et al.*, 2024b] Yitian Zhang, Yue Bai, Huan Wang, Yizhou Wang, and Yun Fu. Don't judge by the look: Towards motion coherent video representation. In *The Twelfth International Conference on Learning Representations*, 2024.

[Zhao *et al.*, 2024] Jieting Zhao, Hanjing Ye, Yu Zhan, Hao Luan, and Hong Zhang. Human orientation estimation under partial observation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11544–11551. IEEE, 2024.

[Zhu *et al.*, 2018] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. *Advances in neural information processing systems*, 31, 2018.