

VIDEO-FOLEY: TWO-STAGE VIDEO-TO-SOUND GENERATION VIA TEMPORAL EVENT CONDITION FOR FOLEY SOUND

^b*Junwon Lee*, [#]*Jaekwon Im*, [#]*Dabin Kim*, ^{b#}*Juhan Nam*

^bGraduate School of Artificial Intelligence, KAIST, Republic of Korea

[#]Graduate School of Culture Technology, KAIST, Republic of Korea

ABSTRACT

Foley sound synthesis is crucial for multimedia production, enhancing user experience by synchronizing audio and video both temporally and semantically. Recent studies on automating this labor-intensive process through video-to-sound generation face significant challenges. Systems lacking explicit temporal features suffer from poor controllability and alignment, while timestamp-based models require costly and subjective human annotation. We propose **Video-Foley**, a video-to-sound system using Root Mean Square (RMS) as a temporal event condition with semantic timbre prompts (audio or text). RMS, a frame-level intensity envelope feature closely related to audio semantics, ensures high controllability and synchronization. The annotation-free self-supervised learning framework consists of two stages, Video2RMS and RMS2Sound, incorporating novel ideas including RMS discretization and RMS-ControlNet with a pretrained text-to-audio model. Our extensive evaluation shows that Video-Foley achieves state-of-the-art performance in audio-visual alignment and controllability for sound timing, intensity, timbre, and nuance. Code, model weights, and demonstrations are available on the accompanying website.¹

Index Terms— Video-to-Sound, Controllable Audio Generation, Multimodal Deep Learning

1. INTRODUCTION

Foley sound plays an important role in enhancing immersive experiences by providing synchronized audio with visual content in film, gaming, and VR environments. However, creating sounds that precisely match the timing, intensity, timbre, and nuance of the visual objects is labor-intensive, highlighting the need for automation or assistance [1]. Recent advances in generative AI have encouraged researchers to explore models that learn the cross-modal correspondence and synthesize audio content from the video input.

In the video-to-sound generation, achieving semantic and temporal synchronization between the two modalities is crucial. However, existing studies have not successfully accomplished this dual goal. Early video-to-sound models, such as GAN-based methods [2, 3], aimed to generate audio from video input in an unsupervised manner. They focused on learning the semantic correspondence between audio and visual from datasets of in-the-wild-quality.

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-0022238).

¹<https://jwnlee.github.io/video-foley-demo>

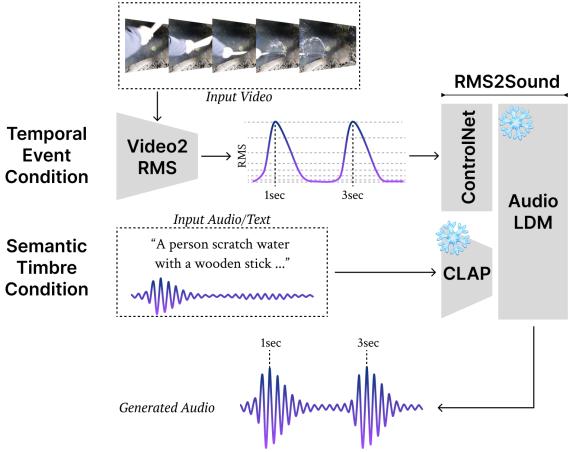


Fig. 1: Overall pipeline of the proposed model, a two-stage Video-to-Sound generation framework. Note that RMS can be extracted from audio waveform numerically. Video2RMS and RMS2Sound part are trained separately.

Subsequent work has proposed controllable video-to-sound generation models that allow for changing timbre with audio prompts [4] or audio-visual correlations [5]. While these approaches showed promising results, they generally suffered from temporal misalignment and low audio quality due to a lack of explicit temporal guidance or low-quality data.

More recent studies have explicitly incorporated temporal information into their models. Diff-Foley [6] utilized a temporal-aware audio-visual joint embedding space to condition the audio-generating diffusion model. However, low visual temporal resolution (4fps) due to high computational costs limited the accuracy of temporal alignment. Other approaches used onset/offset timestamps of sounds in a text to guide audio generation [7, 8]. They trained timestamp detection networks to classify each video frame via supervised learning. However, this method requires human-annotated data, which is costly and often ambiguous in defining a golden standard. Additionally, simply detecting the start and end points of sound events misses many important aspects of audio, such as the volume dynamics of a moving car, which are difficult to represent in text.

We propose **Video-Foley**, a temporal-event-guided Video-To-Sound model for highly synchronized and controllable Foley sound generation (Fig. 1). Our contributions are as follows: 1) We introduce the Root Mean Square (RMS) of audio content as a key temporal feature for audio-visual synchrony. Here, we define RMS as a frame-level energy feature calculated from an audio wave-

form. This captures not only the presence of sound events but also their intensity and temporal change, deeply intertwined with subtle timbre and nuance[9]. RMS-guided systems, along with audio or text prompts, ensure high temporal and semantic synchronization, providing enhanced controllability. 2) We propose a two-stage framework that predicts the RMS of audio waveforms from video (Video2RMS) and generates temporally aligned audio waveforms from the predicted RMS (RMS2Sound). Our model is efficiently trained with video-audio pairs (i.e., general video files) and audio-only data without human annotations, using novel techniques including RMS discretization and RMS-ControlNet leveraging a pretrained text-to-audio model. 3) Through quantitative evaluation, including human surveys, we demonstrate that Video-Foley achieves state-of-the-art performance in both temporal and semantic alignment on the Greatest Hits dataset. The subsequent qualitative analysis and accompanying demo highlight the high controllability in timing, intensity, timbre, and nuance of the audio.

2. PROPOSED METHOD

Video-Foley consists of two components, Video2RMS and RMS2Sound, which are trained separately in a self-supervised manner. Video2RMS predicts the RMS curve from video input, while RMS2Sound generates audio waveforms from the RMS curve along with semantic control with text/audio. We defined RMS as a frame-level amplitude envelope feature of audio waveform defined as follows: for the i -th frame,

$$R_i(x) = \sqrt{\frac{1}{W} \sum_{t=ih}^{ih+W} x^2(t)} \quad (1)$$

where $x(t)$ ($t \in [0, T]$) is the audio waveform, W is a window size and h is a hop size.

2.1. Video2RMS

Video2RMS aims to predict the RMS curve, representing the windowed root mean of squared audio amplitude proportional to intensity, from a sequence of video frames. We introduce two key ideas to tackle this problem. First, we propose to discretize the RMS target and formulate the problem as a classification task. Since non-ambient action-based sounds are transient and sparse, much of the audio remains nearly silent. Our preliminary experiment showed that training with the L2 loss as a regression task led to poor results, as the model tended to predict silence to reach a local minimum. We discretized the continuous RMS curve into equidistant bins after scaling with the μ -law encoding [10], formulated as follows:

$$f(r) = \frac{\ln(1 + \mu|r|)}{\ln(1 + \mu)} \quad (2)$$

where $r \in [0, 1]$ is the RMS value and $\mu + 1$ is the number of discretized bins. Second, we use the label smoothing to mitigate the penalty for near-correct predictions. We adopted the Gaussian label smoothing, frequently used in pitch estimation [11]. The smoothed label y is formulated as follows:

$$y(i) = \begin{cases} \exp\left(-\frac{(c_i - c_{gt})^2}{2\sigma^2}\right) & \text{if } |c_i - c_{gt}| \leq W \ (c_i, c_{gt} \neq 0) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where i is the class index, c_{gt} is the ground-truth class, $\sigma = 1$, and W is the window size determined by the ablation study.

Video2RMS model consists of three 1D-convolutional blocks, two Bi-LSTM layers, and a linear projection head, similar to the visual encoder of RegNet [2]. Each convolutional block includes a

convolution layer, a batch normalization layer, and a ReLU activation layer. As for the input, the BN-Inception network [12], pre-trained on ImageNet, extracts video features frame-wise from RGB images and 2-channel optical flow. The loss function is defined as $L = \sum_i CE(\hat{R}_i, R_i)$ where R denotes the discretized RMS label, $\hat{\cdot}$ is the prediction, and CE is the cross entropy loss.

2.2. RMS2Sound

In order to generate audio that matches both semantic and temporal conditions, we developed RMS-ControlNet, which integrates RMS conditioning controls into a pretrained text-to-audio diffusion model. As shown in Fig. 1, RMS-ControlNet generates audio conditioned on the given RMS and CLAP [13] embedding. Due to CLAP’s ability to extract audio and text embeddings in a joint multimodal space, RMS-ControlNet can generate audio with a target RMS based on both text and audio timbre conditions. We train RMS-ControlNet on audio data only. The training objective is as follows:

$$\mathbb{E}_{x,t,\epsilon} \|\epsilon - f(z_t, t, C(x), R(x))\|_2^2 \quad (4)$$

where x is audio waveform, z is a latent representation of y encoded with a variational autoencoder, z_t is z with t times noise added, C is the CLAP encoder, and R is the RMS calculation. We implement RMS-ControlNet similarly to the original implementation of ControlNet [14]. We freeze the parameters of the text-to-audio model and train only a copy of its encoding layers. To match the feature size of RMS to z_t , we use an additional zero convolution 2D layer.

3. EXPERIMENTS

3.1. Experimental Details

Dataset We used the Greatest Hits dataset[15] with its official train-test split for training and evaluation. The dataset contains 977 videos of a person making sounds with a wooden drumstick on 17 different materials (wood, metal, rock, leaf, plastic, cloth, water, etc.) using two types of actions (hit, scratch). We segmented the videos with denoised audio into 10-second clips without overlap, and resampled to 16kHz for audio and 30fps for video. Each video frame was resized to 344×256 pixels. This resulted in 2,212 training videos (6.14 hours) and 732 test videos (2.03 hours). The training set was used to train Video2RMS, and the test set was used to evaluate both Video2RMS and the entire Video-Foley model. To increase extensibility and applicability, we trained RMS-ControlNet using audio-only data from a variety of sounds, rather than limiting it to hit and scratch sounds. We used the FreeSound dataset[16], which contains about 6,000 hours of audio. All audio was resampled to 16 kHz.

Training The Video2RMS and RMS2Sound models are trained separately but combined during inference. For Video2RMS, RMS was discretized into 64 bins ($\simeq 0.50$ dB granularity), and Gaussian label smoothing was applied ($W = 2$). The model was trained for 500 epochs using a StepLR scheduler (rate 1e-3, step size 100). For RMS2Sound, AudioLDM [17] was used as the backbone text-to-audio model. We initialized the weights of both AudioLDM and ControlNet using an official checkpoint. RMS-ControlNet was trained for 300k steps using the AdamW optimizer. The learning rate started at 1e-4 and was halved every 10k steps. During training, only the parameters of ControlNet were updated.

3.2. Evaluation

To measure the performance of synchronized video-to-sound generation, three main aspects are considered. *Semantic Alignment* eval-

Model	E-L1 ↓	Acc@1 ↑	Acc@5 ↑	Acc@10 ↑
<i>Video2RMS</i>				
disc. RMS (g.t.)	0.00243	1	1	1
Video-Foley (Ours)	0.0450	0.0116	0.241	0.446
w/ label smoothing	0.0431	0.0128	0.238	0.449
random choice (l.b.)	0.2807	0.0154	0.0769	0.154

Table 1: Performance of Video2RMS module. disc. RMS (g.t.): discretized version of ground-truth RMS, l.b.: lower bound.

ates how well the timbre and nuance of sound match the material and action type in the video, *Temporal Alignment* examines the accuracy of the start and end timing of a sound event as well as its intensity changes over time, and *Audio Quality* assesses the overall quality of the audio. Both objective and subjective evaluations are conducted. To match our experiment settings, we resampled generated audios to 16kHz and combined them with the 30fps videos to create 10sec video-audio pairs.

RMS-ControlNet, based on AudioLDM, can use either an audio prompt or a text prompt for timbre conditions. We conducted ablation studies to compare these two prompt methods. For the audio prompt, we simply used ground-truth audio. For the text prompt, we utilized a prompt template: “A person {action} {material} with a wooden stick.” and annotations on material and actions from the Greatest Hits dataset. If there were multiple actions or materials, we made multiple sentences and combined them with “After that.”. If no annotation was available, we used “A person hit something with a wooden stick.” as the default text prompt.

Objective To measure overall audio quality, Frechet Audio Distance (FAD) [18] was used, which is a set-wise distance of audios in embedding space. Given that FAD correlation with human perception is embedding-dependent [19], we used pretrained PANNs wavegram-log-mel [20] and CLAP from Microsoft [13] to extract embeddings through *fadt2*². To measure the semantic alignment between audio and video, FAVID [21] was used, which is the Frechet distance of concatenated video and audio embeddings. Pretrained VGGish [22] and I3D [23] were used for audio and video embeddings, respectively. Additionally, the CLAP [24] score was calculated by measuring the cosine distance between the generated and ground-truth audio in the joint text-audio embedding space.³ Lastly, we used E-L1, the L1 distance between the continuous RMS values of the generated and ground-truth audio as proposed in T-Foley [9], to measure the temporal synchrony of audio and video. For RMS2Sound, all metrics except FAVID were used. E-L1 and class-wise accuracy at k (k=1, 5, 10) measured the RMS prediction performance of Video2RMS, excluding frames where both the ground truth and prediction are silent similar to the previous study [5]. This exclusion is necessary because only a small portion of frames in audio are non-silent for hit/scratch actions, making it easier for the model to predict silence and thus failing to effectively capture the performance in non-silent frames.

Subjective We conducted a human evaluation to assess the perceptual quality of the generated audio in relation to the input video. 20 participants were asked to score the audio on a five-point Likert scale based on four criteria: Material / Action / Temporal Alignment, and Audio Quality. Semantic Alignment was divided into two categories to evaluate how well the sound matches the material type and

²<https://github.com/DCASE2024-Task7-Sound-Scene-Synthesis/fadt2>

³Note that this CLAP model is from LAION, which differs from the Microsoft model used in AudioLDM.

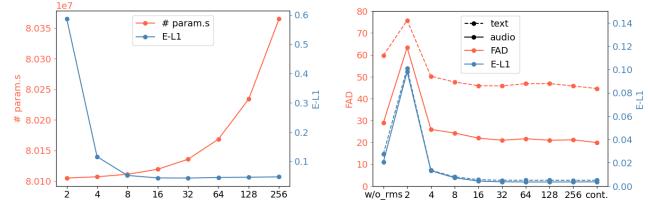


Fig. 2: Performance of Video2RMS (left) and RMS2Sound (right) for different numbers of RMS bins. w/o_rms: without RMS condition (Text-to-Audio[17]), cont.: continuous RMS, no discretization.

action nuance of the sound events in the video. The Mean Opinion Score (MOS) and its 95% confidence interval were calculated. The evaluation survey consisted of 12 questions covering different material-action types. Since CondFoleyGen [5] does not support text prompts, we used audio prompts for the other models to ensure a fair comparison.

4. RESULT

4.1. Analysis on Video2RMS and RMS2Sound

Table 1 demonstrates the performance of the Video2RMS model. The model successfully predicts the RMS curve from the video, given the scores of discretized ground-truth RMS (upper bound due to information loss) and random choice (lower bound). Video2RMS significantly outperforms random choice in accuracy at Acc@5 and Acc@10, indicating that the model accurately predicts bins adjacent to the ground truth, as reflected by the low E-L1 value. Although the model scores lower in accuracy at Acc@1, this is due to its focus on predicting a realistic RMS curve rather than matching the exact magnitude bin. Label smoothing generally improves performance, enhancing both E-L1 and classification accuracy.

The number of bins for the RMS discretization is a critical parameter that significantly affects both Video2RMS and RMS2Sound. To determine the optimal value, we conducted an ablation study, as presented in Figure 2. In Video2RMS, we identified a trade-off between performance and computational cost, as shown on the left; more bins improve temporal synchrony but require more model parameters. As shown on the right, both temporal alignment performance and audio quality in RMS2Sound saturate after bins greater than 64. At 64 bins, we found no performance drop in the quantitative measures when using discretized RMS instead of continuous RMS. Therefore, we set the discretization bins to 64. It is worth noting that the vanilla text-to-audio model without RMS conditioning lagged in both E-L1 and FAD metrics. This implies that RMS conditioning, which aligns with the semantics of the given prompt, helps the model generate higher-quality audio, even when the text-to-audio parameters are frozen.

4.2. Analysis on Video-to-Sound

Quantitative Study Table 2 compares the performance of various video-to-sound systems on the GreatestHits test set. Video-Foley achieved state-of-the-art performance on all objective metrics as well as the human MOS. Notably, it showed a significant performance gap in temporal alignment and semantic material alignment compared to the audio-visual cued model (CondFoleyGen) and the onset-based model (SyncFusion). This suggests the RMS conditioning is superior for video-to-sound generation, because it conveys both

Model	Audio Quality			Temporal Alignment		Semantic Alignment			
	FAD-P ↓	FAD-C ↓	MOS	E-L1 ↓	MOS	CLAP ↑	FAVD ↓	MOS _{material}	MOS _{action}
Ground Truth	0	0	4.57(±0.08)	0	4.83(±0.06)	1	0	4.70(±0.08)	4.90(±0.04)
<i>No Prompt</i>									
SpecVQGAN [†] [3]	101.0	579	-	0.0230	-	0.323	6.42	-	-
Diff-Foley [‡] [6]	87.0	550	2.11(±0.11)	0.0291	1.86(±0.14)	0.403	4.61	1.78(±0.13)	2.38(±0.17)
<i>Audio Prompt</i>									
CondFoleyGen ^{av} [5]	42.2	381	3.10(±0.13)	0.0238	1.93(±0.13)	0.572	1.01	2.36(±0.16)	2.79(±0.17)
SyncFusion[7]	65.9	335	3.10(±0.13)	0.0239	3.10(±0.19)	+0.631	4.50	3.04(±0.18)	3.22(±0.19)
Video-Foley* (Ours)	27.2	187	3.93(±0.12)	0.0090	4.40(±0.11)	0.644	0.80	3.83(±0.15)	4.56(±0.08)
<i>Text Prompt</i>									
SyncFusion[7]	81.6	424	-	0.0292	-	+0.529	5.11	-	-
Video-Foley* (Ours)	66.8	451	-	0.0103	-	0.476	3.28	-	-
Text-to-Audio[17]	59.8	397	2.39(±0.13)	0.0217	2.00(±0.13)	0.443	2.67	2.78(±0.16)	3.21(±0.17)

Table 2: Performance of the proposed Video-Foley and other video-to-sound models on *Greatest Hits* testset. *av*: audio-video paired prompt used, ⁺: same CLAP model for train and evaluation. Regarding train data, [†]: *VGGSound* (~0.4k hr), [‡]: *VGGSound*, *AudioSet* (~1.3k hr), ^{*}: *Greatest Hits* trainset for *Video2RMS* (~6 hr) and *FreeSound* dataset for *RMS2Sound* (~6k hr), otherwise: *Greatest Hits* trainset (~6 hr).

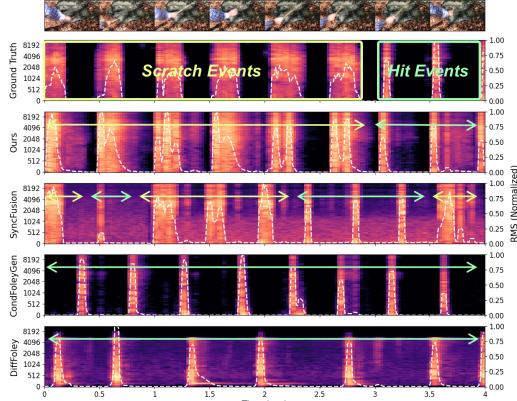


Fig. 3: Controlling timbre and energy transition: Video-Foley generates hit and scratch sounds at desired positions using RMS guidance.

timing and intensity dynamics, providing more detailed information than simple timestamps. Furthermore, this temporal feature can imply the timbre and nuance of the sound through its curve shape, complementing the semantic prompt. Diff-Foley, despite using temporal information for audio-visual joint space training, lagged in performance due to poorer temporal alignment granularity (4fps) and the frequent generation of visually irrelevant sounds, likely caused by noisy in-the-wild training sets. In every aspect, including audio quality, Video-Foley also outperforms AudioLDM[17], the frozen text-to-audio model in RMS2Sound. This suggests that an appropriate RMS condition, well matched with the prompt, can help the model generate higher fidelity audio, consistent with the results in Figure 2. Video-Foley and SyncFusion, trained exclusively with audio prompts, perform better with audio prompts than text. The complexity of describing multiple sound events over 10 seconds with text versus audio may also contribute to this trend.

Qualitative Study Extensive case studies were conducted to demonstrate the performance and controllability of Video-Foley. Our analysis underscores that the intensity level and energy transition in RMS are often intertwined with the timbre and nuance of sound, consistent with the findings of the previous study[9]. We plot the mel-spectrogram and normalized RMS of the generated audio from each model. Figure 3 shows the synergy of complex prompts with RMS. Only Video-Foley generates hit or scratch sounds at the right time.

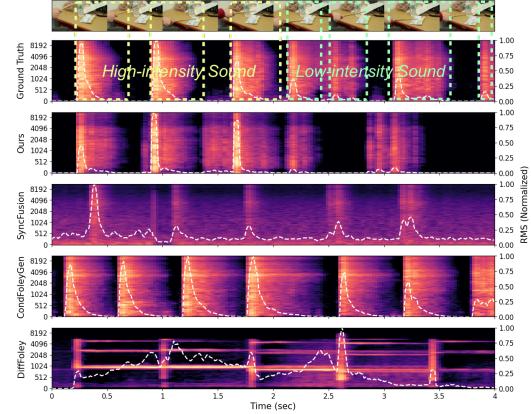


Fig. 4: Controlling intensity and nuance: Video-Foley predicts different levels and shapes of the RMS curve for each sound event.

Our model can distinguish the timing and type of each sound event through RMS, even for complex audio or text prompts with multiple events. Onset-based models only predict when to make a sound but cannot distinguish different timbres for each event. In contrast, ours can control both the timing and the corresponding timbre by modifying the RMS. Figure 4 illustrates the controllability and high audio-visual alignment of Video-Foley. Only ours effectively predicts and recommends the appropriate level and transition curve of RMS, ensuring synchronization with the input video. This includes not only timing but also the intensity and nuance of sound events.

5. CONCLUSION

We propose Video-Foley, a two-stage video-to-sound model using RMS as a temporal feature. Our quantitative and qualitative studies demonstrate that RMS conditioning enhances both temporal and semantic audio-visual synchrony while ensuring high controllability, due to its synergy with audio or text prompts. We believe RMS is an effective control feature, as shown in T-Foley[9]. Video2RMS may provide an excellent initial condition for creators to shape desired sound. More case studies and control examples are provided on our accompanying website. We plan to extend our work to a large-scale, in-the-wild dataset.

6. REFERENCES

- [1] Keunwoo Choi, Sangshin Oh, Minsung Kang, and Brian McFee, “A proposal for foley sound synthesis challenge,” *arXiv preprint arXiv:2207.10760*, 2022.
- [2] Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan, “Generating visually aligned sound from videos,” *IEEE Transactions on Image Processing*, vol. 29, pp. 8292–8302, 2020.
- [3] Vladimir Iashin and Esa Rahtu, “Taming visually guided sound generation,” in *The 32st British Machine Vision Virtual Conference*. BMVA Press, 2021.
- [4] Chenye Cui, Zhou Zhao, Yi Ren, Jinglin Liu, Rongjie Huang, Feiyang Chen, Zhefeng Wang, Baoxing Huai, and Fei Wu, “Varietysound: Timbre-controllable video to sound generation via unsupervised information disentanglement,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.
- [5] Yuxi 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*, 2023, pp. 2426–2436.
- [6] 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*, vol. 36, 2024.
- [7] Marco Comunità, Riccardo F Gramaccioni, Emilian Postolache, Emanuele Rodolà, Danilo Comminello, and Joshua D Reiss, “Syncfusion: Multimodal onset-synchronized video-to-audio foley synthesis,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024, pp. 936–940.
- [8] Zhifeng Xie, Shengye Yu, Qile He, and Mengtian Li, “Son-icvisionlm: Playing sound with vision language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26866–26875.
- [9] Yoonjin Chung*, Junwon Lee*, and Juhan Nam, “T-foley: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis,” in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2024.
- [10] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al., “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, vol. 12, 2016.
- [11] Sangeun Kum and Juhan Nam, “Joint detection and classification of singing voice melody using convolutional recurrent neural networks,” *Applied Sciences*, vol. 9, no. 7, pp. 1324, 2019.
- [12] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International conference on machine learning*. pmlr, 2015, pp. 448–456.
- [13] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, “Clap learning audio concepts from natural language supervision,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.
- [14] 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*, 2023, pp. 3836–3847.
- [15] 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*, 2016, pp. 2405–2413.
- [16] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuxian Zou, and Wenwu Wang, “Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.
- [17] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley, “Audoldm: text-to-audio generation with latent diffusion models,” in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 21450–21474.
- [18] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Frechet audio distance: A metric for evaluating music enhancement algorithms,” *arXiv preprint arXiv:1812.08466*, 2018.
- [19] Modan Tailleur*, Junwon Lee*, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto, “Correlation of frechet audio distance with human perception of environmental audio is embedding dependant,” in *2024 32nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2024.
- [20] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pre-trained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.
- [21] Lucas Goncalves, Prashant Mathur, Chandrashekhar Lavania, Metehan Cekic, Marcello Federico, and Kyu J Han, “Peavs: Perceptual evaluation of audio-visual synchrony grounded in viewers’ opinion scores,” *arXiv preprint arXiv:2404.07336*, 2024.
- [22] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “Cnn architectures for large-scale audio classification,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2017, pp. 131–135.
- [23] Joao Carreira and Andrew Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [24] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2023, pp. 1–5.

A. SUBJECTIVITY IN ONSET ANNOTATION

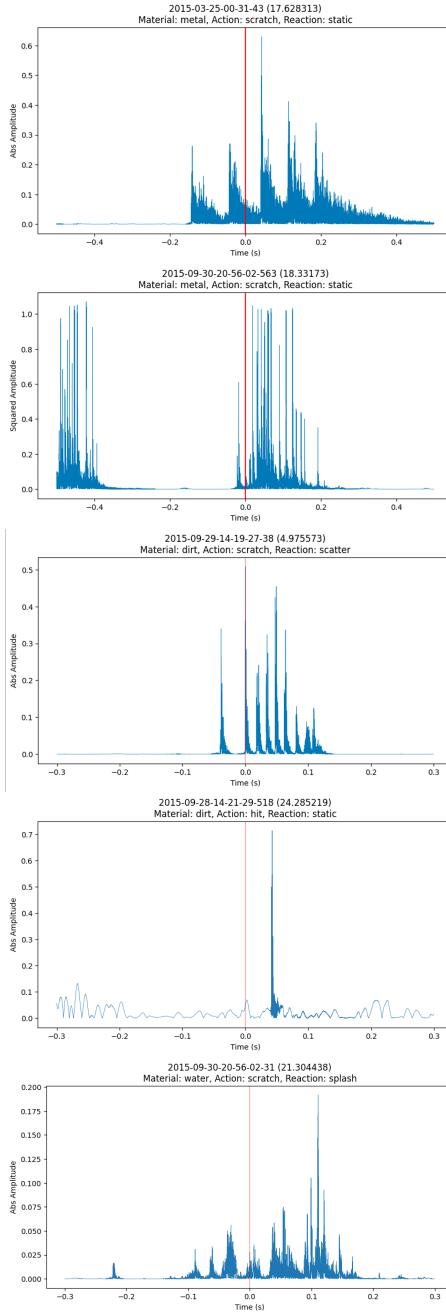


Fig. 5: Onset annotation examples in *Greatest Hits* dataset. The blue curve shows the absolute amplitude of the waveform, and the red vertical line indicates the onset annotation timestamp.

We emphasize that onset annotation is highly subjective and lacks a systematic approach to define it, affecting the quality of the annotations. For some sound events, such as scratching sounds with multiple adjacent attacks or water, wind, and instrument sounds with slow attacks, it can be challenging to define an onset timestamp that aligns with both the waveform envelope and human perception. In other words, humans may not perceive the sound's starting point as

the moment it actually begins to sonify. This could be critical in video-to-sound generation, where precise temporal synchrony is essential. It could potentially degrade model performance and make timestamp-based evaluation unreliable. Figure 5 demonstrates examples from the *Greatest Hits* dataset highlighting the subjectivity of onset annotation. In the dataset, annotations for each sound event are provided, including the onset timestamp, material, action, and reaction. It is important to note that these annotations were manually performed by humans during the dataset's construction. Therefore, utilizing onset annotation involves both labor costs and handling subjectivity issues.

B. DETAILS IN VIDEO-FOLEY

B.1. Video2RMS

B.1.1. Details

For optical flow extraction, pretrained RAFT (Raft_Large_Weights.C_T_SKHT_V2) in pytorch was used.⁴ The checkpoint was pre-trained on FlyingChairs, FlyingThings3D and finetuned on Sintel. RMS was calculated from the audio waveform with a 512 window size and a 128 hop length, following the configuration in T-Foley[9]. By padding $(512 - 128)/2$ values at both ends of the waveform in reflect mode, we obtained 1250 frames. The model was trained with a batch size of 512 using Adam optimizer.

B.1.2. Architecture

The model architecture of Video2RMS utilizes a BN-Inception[12] network pretrained on ImageNet⁵, 1D-convolutional blocks, Bi-LSTM layers, and a linear projection head. A similar architecture has been used in RegNet[2], but it served as a hidden embedding extractor for the GAN generator. VarietySound[4] also employed a similar architecture to extract temporal features, but only used RGB images as input. SpecVQGAN[3] experimented with both pretrained BN-Inception (using RGB images and optical flow input) and ResNet (using RGB images input) to extract the conditioning features for the transformer generator. However, none of these models were explicitly trained by their final output nor used for classification. Additionally, the input video frame rate for these models was 21.5 fps, whereas ours is 30 fps, matching the commercial standard.

Onset/offset detection models[5, 7, 8] perform binary classification for each video frame, typically utilizing a ResNet(2+1)D network followed by an MLP layer. These models use RGB images as input and do not incorporate optical flow. Moreover, the ResNet(2+1)D network is trained, while our pretrained BN-Inception is frozen. Lastly, their video frame rate is significantly lower (15 fps) compared to ours.

B.2. RMS2Sound

We utilize the 'audioldm-s-full' checkpoint from the official repository. To maintain training consistency, we adhere to the original AudioLDM configuration (e.g. audio normalization). The window length and hop length are set to 1024 and 160, respectively. By padding $(1024 - 160)/2$ values at both ends of the waveform in reflect mode, we obtained 1024 frames. When using the predicted

⁴https://pytorch.org/vision/main/models/generated/torchvision.models.optical_flow.raft_large.html

⁵https://yjxiong.blob.core.windows.net/models/bn_inception-9f5701afb96c8044.pth

RMS from Video2RMS, nearest-neighbor interpolation is applied to match the feature size. The generated audio duration is 10.24 seconds. We only use Classifier-Free Guidance (CFG) for prompting and do not apply it to RMS conditions, as we did not observe meaningful performance improvements. The CFG is formulated as follows:

$$\begin{aligned} \hat{f}(z_t, t, C(x), R(x)) \\ = \omega f(z_t, t, C(x), R(x)) + (1 - \omega)f(z_t, t, R(x)) \end{aligned} \quad (5)$$

where ω is a guidance scale, z is a latent representation of x encoded with a variational autoencoder (VAE), z_t is z with t times noise added, C is the CLAP encoder, and R is the RMS calculation. Note that a learned null embedding is used instead when CLAP embedding $C(\cdot)$ is not given to the model f . In our experiment, ω is fixed to 3.5.

C. METRIC FOR OBJECTIVE EVALUATION

Frechet Distance When reference set embeddings r and a generated set embeddings g are given, we calculate the FAD as follows:

$$FAD(r, g) = \|\mu_r - \mu_g\|_2 + \text{tr} \left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g} \right) \quad (6)$$

where μ_x and Σ_x are the mean and covariance matrix of the distribution x . FAVD is calculated similarly, using the concatenated audio and video embedding.

CLAP score First, we extract embeddings from ground-truth e and generated audio \hat{e} in the audio-text joint embedding space of CLAP. Then, the cosine distance between the two embedding vectors $\cos(e, \hat{e})$ is measured.

E-L1 E-L1(Event-L1) which is proposed in T-Foley[9] is defined as the following:

$$E\text{-L1} = \frac{1}{k} \sum_{i=1}^k \|E_i - \hat{E}_i\| \quad (7)$$

where E_i is the ground-truth event feature of i -th frame, and \hat{E}_i is the predicted one. In this paper, RMS is the temporal event feature. For evaluating Video2RMS, E-L1 between the predicted RMS and the ground-truth RMS is measured. In the case of Video-Foley, E-L1 between the RMS extracted from generated audio and ground-truth audio is considered.

PEAVS We found a lack of standardized metrics to measure temporal synchrony in video-audio pairs. While proposing E-L1, we also considered using PEAVS[21]. PEAVS is calculated by a neural model trained to estimate human opinion scores on audio-visual synchrony through regression to maximize correlation with human perception. However, we found that PEAVS poorly aligns with the MOS (Mean Opinion Score) we measured, as shown in Table 3. We suspect two main reasons for this discrepancy. First is the limited scope of training data. Only 200 videos from the AudioSet evaluation split were used as the source of training data. Even though this resulted in 18.2K videos after applying nine synchrony-related distortions at various levels, the training data still represents a very small portion of the entire AudioSet. Therefore, it cannot cover the wide range of in-the-wild video distributions. It may not generalize well to our Greatest Hits data, which features very specific sound events (e.g., hitting or scratching something with a wooden drumstick). Second, the quality of the source audio-visual pairs. AudioSet originates from YouTube videos, which often have poor audio and visual quality and frequently contain off-screen sounds irrelevant to the visual part. This may lead the model to predict better scores when it receives generated audio with poor quality or irrelevant noises. In

Model	PEAVS \uparrow	MOS \uparrow
Ground Truth	1.81	4.83(± 0.06)
<i>No Prompt</i>		
SpecVQGAN [†] [3]	2.64	-
Diff-Foley [‡] [6]	2.25	1.86(± 0.14)
<i>Audio Prompt</i>		
CondFoleyGen ^{av} [5]	1.80	1.93(± 0.13)
SyncFusion[7]	2.21	3.10(± 0.19)
Video-Foley* (Ours)	1.70	4.40(± 0.11)
<i>Text Prompt</i>		
SyncFusion[7]	2.31	-
Video-Foley* (Ours)	2.27	-
Text-to-Audio[17]	1.82	2.00(± 0.13)

Table 3: Comparison between PEAVS and MOS for Temporal Alignment

Table 3, Diff-Foley received a high PEAVS score despite performing poorly in the human MOS score.

Onset Metrics Onset metrics, Onset Acc (accuracy) and Onset AP (average precision), were measured following the protocol in CondFoleyGen[5]. These metrics were calculated with a tolerance of ± 0.1 seconds after applying non-maximum suppression with a window length of 50 ms based on the onset prediction confidence. Moreover, all metrics excluded true-negative cases (both predicted and labeled outputs being non-onset) due to the label imbalance in the dataset: onset labels are much sparser than non-onset labels in a video. This imbalance incentivizes the model to predict non-onset more frequently. Consequently, including all true-negative cases may overrate and exaggerate the model’s performance, making it distinct from actual perceptual performance and ultimately inaccurate.

D. SUBJECTIVE HUMAN EVALUATION

In the qualitative evaluation, we aimed to compare the performance of video-to-sound systems on various video inputs that include diverse contextual information. To achieve this, we created 12 evaluation questions by selecting cases that combine the ‘material’ and ‘action’ categories from the Greatest Hits test dataset. Each question presented the ground truth audio and the audio generated by Video-Foley, SyncFusion, DiffFoley, CondFoleyGen, and AudioLDM in a random order. Specifically, from the dataset’s 18 ‘material’ categories, we excluded ‘None’ and selected six ‘{material}-scratch’ cases where the sound characteristics significantly change by scratching actions. These cases included *plastic-scratch*, *rock-scratch*, *dirt-scratch*, *drywall-scratch*, *gravel-scratch*, and *grass-scratch*. Subsequently, we selected six ‘{material}-hit’ cases from the remaining material categories where the sound characteristics notably change by hitting actions. These cases included *carpet-hit*, *ceramic-hit*, *metal-hit*, *water-hit*, *wood-hit*, and *leaf-hit*. To standardize the length of the sample videos and control evaluator fatigue, we trimmed each video to 4 seconds from the starting point.

Considering that the names of qualitative evaluation criteria related to sound elements can be interpreted differently based on the evaluator’s experience, expectations, and preferences, we provided pre-experiment guidelines as shown in Figure 6. We redefined and introduced the criteria for *Material/Action Alignment* as *Semantic Alignment: Material Type* and *Semantic Alignment: Action Type* to clearly distinguish them from *Temporal Alignment* during the evaluation. We also provided textual guidelines on the evaluation ele-

Evaluation Criteria

Each item in the evaluation sheet presents video sound effects generated by different AI models in a random order for the same silent video input. The evaluation criteria are as follows:

1. **Semantic Alignment: Material Type**
 - Semantic alignment evaluates how well the generated sound matches the event information in the video.
 - For ‘Semantic Alignment: Material Type’, the focus is on assessing whether the timbre of the sound effect naturally aligns with the material (wood, metal, stone, water, etc.) being struck in the video.
2. **Semantic Alignment: Action Type**
 - Semantic alignment evaluates how well the generated sound matches the event information in the video.
 - For ‘Semantic Alignment: Action Type’, the focus is on assessing whether the intensity variation of the sound effect naturally aligns with the type of action (‘hitting’ and ‘scratching’) in the video.
3. **Temporal Alignment**
 - ‘Temporal Alignment’ assesses how well the sound matches the visual events in the video in terms of timing.
4. **Audio Quality**
 - ‘Audio Quality’ evaluates the naturalness and overall quality of the sound.

Fig. 6: Guidelines for evaluation elements

A. Semantic Alignment: Material Type

Even if the generated sound is not perfectly temporally aligned with the video, please give a high score if the sound accurately reflects the material of the struck object.

Evaluation Criteria

- 1: The sound does not reflect the material of the struck object, mostly inaccurate and unconvincing.
- 2: The sound has elements reflecting the material, but is still mostly inaccurate or unnatural.
- 3: The sound reasonably matches the material, though some inaccuracies or unnatural elements are present.
- 4: The sound matches the material well, with only minor inaccuracies or slight unnaturalness.
- 5: The sound perfectly and naturally matches the material of the struck object.



Fig. 7: Explanation of Likert scale scores and video examples for each evaluation criterion

ments and precautions for the four criteria. This included a detailed explanation of the situations corresponding to each point on the Likert scale regarding that each evaluator may interpret the points of the Likert scale differently. Additionally, considering that it might be difficult to fully understand the evaluation elements and criteria for sound and video alignment based solely on text, we provided video sample examples corresponding to the lowest and highest scores for the four evaluation elements. Refer to an example in Figure 7.

E. ABLATION STUDY FOR VIDEO2RMS

E.1. Ablation Study on Using Onset Annotations

To examine the impact of the onset label, onset joint training was conducted. An additional binary cross-entropy term, scaled by a factor of $1e^{-3}$, was added to the loss function. An extra branch head with a linear projection layer was added to the model after the LSTM layer. Onset timestamp annotations provided in the Greatest Hits dataset were only used as labels for supervision. We used onset-related metrics following previous work[5] (refer to Section C).

Table 4 demonstrates the performance of the Video2RMS model along with the previous Video2Onset model[7]. Additional onset supervision does not benefit RMS prediction. There was no significant improvement in RMS prediction (i.e. E-L1 and accuracy), while

Model	E-L1 ↓	Acc ↑	Model	Onset Acc ↑	Onset AP ↑
<i>Video2RMS</i>			<i>Video2Onset</i>		
disc. RMS (g.t.)	0.00243	1	SyncFusion † [7]	0.484	0.501
Video-Foley RMS (Ours)	0.0450	0.0116	Video-Foley onset	0.359	0.653
w/ onset joint train*	0.0448	0.0118	w/ RMS joint train*	0.387	0.837
w/ label smoothing	0.0431	0.0128			

Table 4: Performance of Video2RMS module. †: Reproduced to match our setting, disc.: discretized, g.t.: ground-truth, *: identical model. Onset metrics measured with a tolerance of ± 0.1 s.

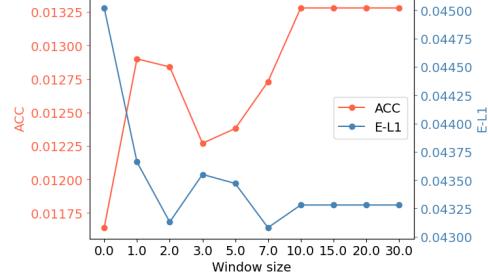


Fig. 8: Ablation study on window size of label smoothing in Video2RMS

the onset prediction performance increased significantly in both accuracy and average precision. This suggests that RMS, despite not requiring any manual annotation, could aid the onset prediction that relies on labeled data. In addition, our BN-Inception-based model and the ResNet(2+1)D-based model[7, 8] have comparable overall performance in the Video2Onset detection task. Our model shows better precision and lower recall, likely due to the optical flow input and global sequential modeling. Refer to Section B.1 for further details.

E.2. Ablation Study on Label Smoothing

Figure 8 illustrates the performance of Video2RMS with different label smoothing window sizes W . We found that $W = 2$ offers the best balance between E-L1 and accuracy. For larger window sizes, the model produces more jitter in the RMS curve. Although performance saturates after $W = 10$ with higher accuracy, this results in a spikier RMS curve and a higher E-L1 value, indicating poorer overall performance. Nevertheless, using Gaussian label smoothing consistently improved performance regardless of the window size.

F. ABLATION STUDY FOR RMS2SOUND

Table 5 summarizes the performance of RMS2Sound on audio and text prompts with ground-truth RMS conditions. The discretized RMS performed comparably to the original continuous RMS in terms of audio quality, semantic similarity, and temporal alignment. In contrast, the vanilla text-to-audio model without RMS conditioning (AudioLDM) underperformed in every metric. This supports our assumption that realistic RMS conditions enhance the overall quality of generated audio.

G. INFERENCE FOR OTHER MODELS

Although some models (i.e., SpecVQGAN and Diff-Foley) do not receive semantic prompts, we included them in the evaluation to compare the performance mainly on temporal alignment. Detailed

Model	FAD-P \downarrow	FAD-C \downarrow	CLAP \uparrow	E-L1 \downarrow
<i>Audio Prompt</i>				
w/o RMS[17]	29.0	194	0.619	0.02076
disc. RMS	21.6	154	0.686	0.00348
cont. RMS	19.9	152	0.657	0.00361
<i>Text Prompt</i>				
w/o RMS[17]	59.8	397	0.443	0.0276
disc. RMS	46.8	333	0.504	0.00496
cont. RMS	44.6	323	0.531	0.00498

Table 5: Performance of RMS2Sound module. w/o RMS: pretrained AudioLDM without RMS condition, disc. RMS: discretized RMS in 64 bins, cont. RMS: continuous RMS.

methods to match the inference settings of different models are explained in the following paragraphs.

SpecVQGAN[3] SpecVQGAN generates 10 seconds of audio from 21.5 fps video. We utilize the official code and checkpoint. The model ‘2021-07-30T21-34-25_vggsound_transformer’ was used.

Diff-Foley[6] Official code and checkpoint were used. As Diff-Foley generates 8-second audio from 4fps video, we made two inferences: one with video frames from 0-8 seconds and another from 2-10 seconds. We then concatenated the entire first segment with the latter 2 seconds of the second segment to produce a 10-second audio. We used the official repository’s default settings for inference (4.5 classifier-free guidance scale, 50 classifier guidance scale, 25 steps with DPM Solver).

CondFoleyGen[5] We use the official code and checkpoint of CondFoleyGen. Although CondFoleyGen generates 2 seconds of audio from 15fps video, the official code was implemented to generate multiples of 2 seconds of audio by adjusting the parameter W_scale . We set W_scale to 5 to generate 10 seconds of audio. The model trained with the Greatest Hits dataset was used.

SyncFusion[7] SyncFusion was trained to generate 5.46 seconds of audio from 15fps video. Using the official code and model checkpoint without augmentation, we generated 5-second audio clips and concatenated them. For text prompts, we used the same text as Video-Foley. Default settings for inference were followed (150 steps with the DDIM sampler).

H. ADDITIONAL CASE STUDY ON VIDEO-FOLEY

As an extension of Section 4.2, we provide additional demonstrations to highlight the performance of Video-Foley. Figures 9a and 9b show examples where Video-Foley successfully predicts the RMS curve for sound events with different timbres, generating synchronized scratching and hitting sounds. Figure 10 illustrates how Video-Foley recommends the appropriate intensity for each sound event, accurately reflecting the video’s nuance to generate sound. These capabilities are due to Video2RMS’s ability to distinguish action types (e.g., hit and scratch), timing, and intensity and predict their corresponding energy transitions, and RMS2Sound’s ability to generate appropriate timbre and nuance at the corresponding timings. Additionally, RMS helps enhance temporal alignment. Figure 11 demonstrates Video-Foley generating hitting sounds synchronized with the video in terms of start and end timing, all without using any human annotation. Refer to Section E.1 for the impact of RMS joint training on the Video2Onset task.

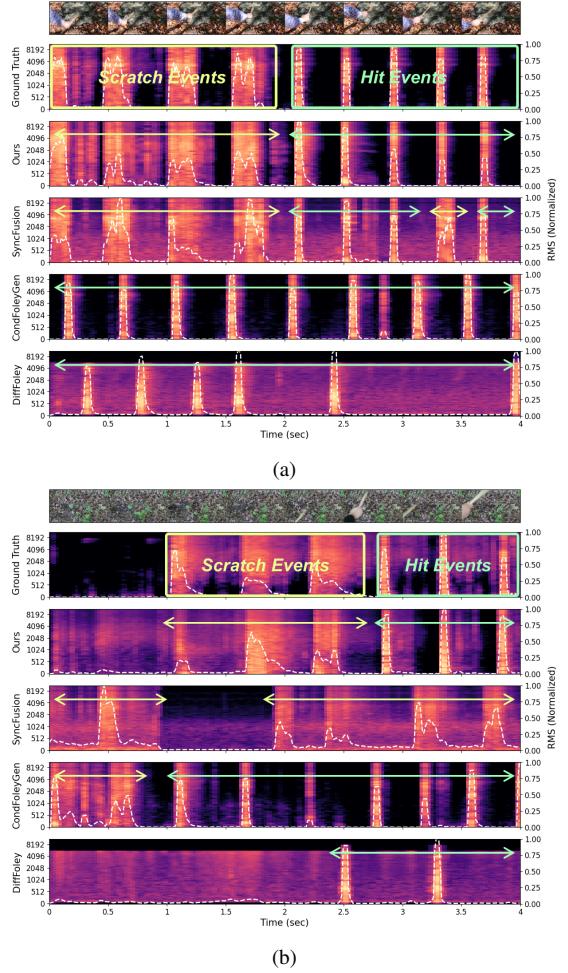


Fig. 9: Controlling timbre and energy transition: Video-Foley generates hit and scratch sound events at desired positions using RMS guidance.

I. UNLOCKING THE POTENTIAL OF RMS-CONTROLNET

RMS-ControlNet, trained for additional temporal event guidance with RMS condition on top of the pretrained Text-to-Audio model (AudioLDM), shows great potential in controllable audio generation tasks. We provide demos to showcase its high controllability, which prior TTA models were not able to achieve.

Figure 12 shows how RMS can be simply and intuitively used for temporal guidance, including timing, nuance, and locational information. With the same text prompt, RMS-ControlNet guides AudioLDM to generate audio that matches different input RMS conditions (A-shaped, monotonic decrease, monotonic increase, and V-shaped) while maintaining audio semantics. Such intensity dynamics are often used in Foley sound generation, which current text-to-audio models struggle to reflect with sufficient temporal accuracy. Figure 13 shows how text prompt can adjust audio semantics along with RMS guidance. Using the same input RMS, users can control audio semantics such as the sound source and timbre (Figure 13a & 13b) and timbre and nuance (Figure 13c). This highlights RMS-ControlNet’s ability to guarantee high controllability in RMS

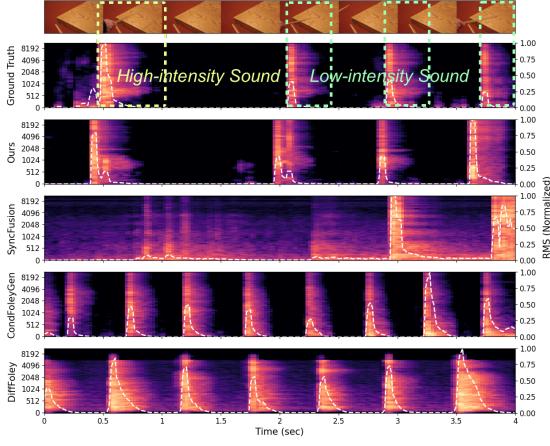


Fig. 10: Controlling intensity and nuance: Video-Foley predicts different levels and shapes of the RMS curve for each sound event.

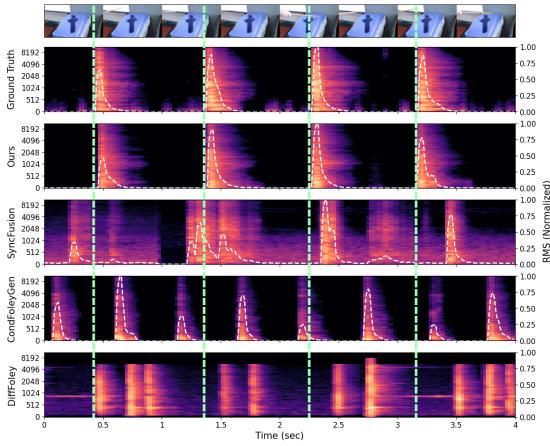


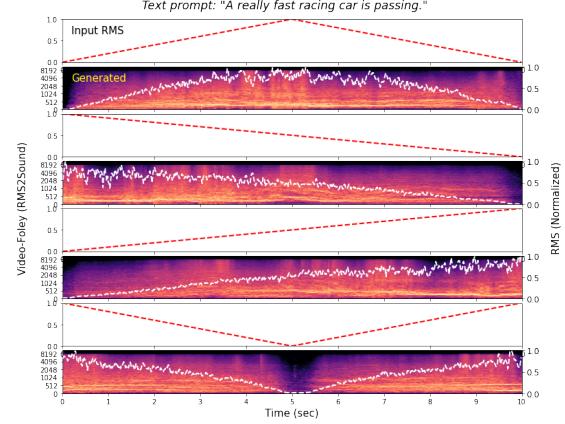
Fig. 11: Controlling temporal alignment: Video-Foley predicts the most accurate start and end timing for each sound event.

guidance for timing and intensity while preserving the power in text-to-audio generation.

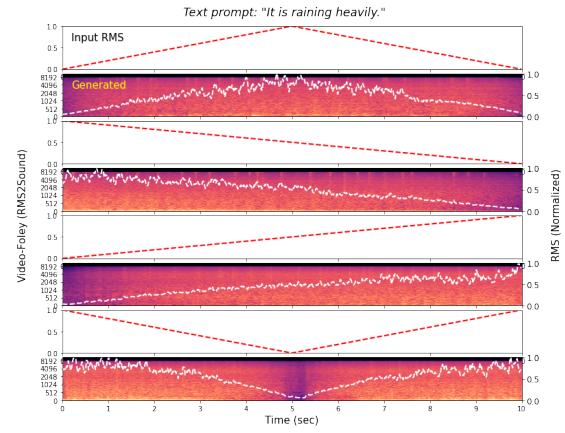
J. BROADER IMPACT

The broader impact of this research project extends across multiple domains, offering significant advancements in both video-to-sound generation and controllable audio generation. In the entertainment industry, particularly in film, gaming, and virtual reality, our work can greatly enhance the efficiency and creativity of Foley sound production, reducing the manual effort required while ensuring precise synchronization of audio with visual elements. This could lead to more immersive and accessible content creation, enabling smaller studios and independent creators to produce high-quality multimedia experiences.

In the field of controllable audio generation, our approach offers a new level of precision and flexibility, empowering users to tailor audio outputs to specific needs, whether for artistic expression, accessibility, or adaptive technology. For instance, this technology could be used to create more realistic soundscapes in virtual environments, improve auditory cues for visually impaired users, or



(a)



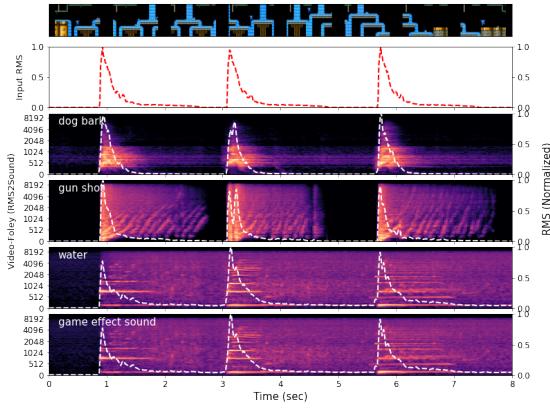
(b)

Fig. 12: Controlling energy transition: RMS-ControlNet can guide the temporal dynamics of sound intensity through RMS while reflecting the given semantic text prompt.

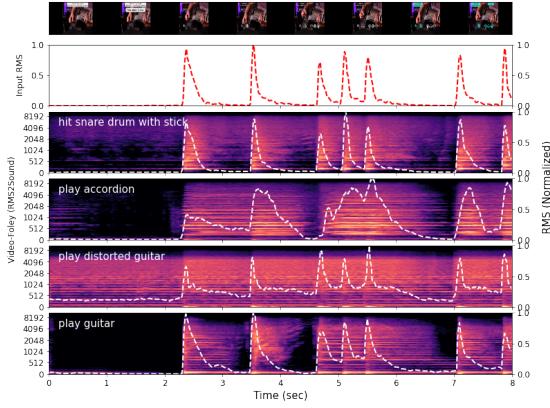
generate educational content with dynamic and contextually relevant audio.

However, the ethical implications of this technology must be carefully considered. The ability to generate highly realistic audio synchronized with video raises concerns related to deepfakes and the potential misuse of this technology for creating deceptive or harmful content. Such misuse could have serious implications for privacy, misinformation, and human rights, as it could be employed to fabricate audio-visual evidence or manipulate public perception.

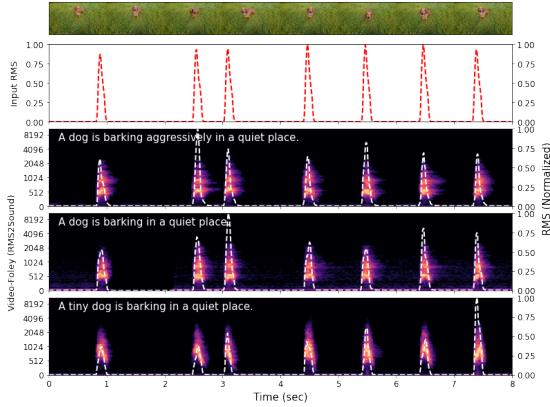
To mitigate these risks, it is crucial to establish ethical guidelines and promote responsible use of this technology. Developers and users should be aware of the potential for abuse and work towards implementing safeguards to prevent unauthorized or malicious use. Additionally, ongoing dialogue with policymakers, ethicists, and the public is essential to ensure that the benefits of this technology are realized while minimizing the potential for harm, ultimately protecting individual rights and maintaining public trust.



(a)



(b)



(c)

Fig. 13: Controlling timbre and nuance: RMS-ControlNet can guide the sound source and timbre through a text prompt while controlling timing and intensity through RMS conditions.