The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Read, Watch and Scream! Sound Generation from Text and Video

Yujin Jeong*, **Yunji Kim**, **Sanghyuk Chun**, **Jiyoung Lee**†

NAVER AI Lab

Abstract

Despite the impressive progress of multimodal generative models, video-to-audio generation still suffers from limited performance and limits the flexibility to prioritize sound synthesis for specific objects within the scene. Conversely, text-to-audio generation methods generate high-quality audio but pose challenges in ensuring comprehensive scene depiction and time-varying control. To tackle these challenges, we propose a novel video-and-text-to-audio generation method, called ReWaS, where video serves as a conditional control for a text-to-audio generation model. Especially, our method estimates the structural information of sound (namely, energy) from the video while receiving key content cues from a user prompt. We employ a well-performing text-to-audio model to consolidate the video control, which is much more efficient for training multimodal diffusion models with massive triplet-paired (audio-video-text) data. In addition, by separating the generative components of audio, it becomes a more flexible system that allows users to freely adjust the energy, surrounding environment, and primary sound source according to their preferences. Experimental results demonstrate that our method shows superiority in terms of quality, controllability, and training efficiency.

Code & demo — https://naver-ai.github.io/rewas
Extended version — https://arxiv.org/abs/2407.05551

1 Introduction

Generative models have developed dramatically, making content creation easier for people. Especially, text-to-video generation models such as Make-a-Video (Singer et al. 2022) and Sora (Brooks et al. 2024) show the impressive emergence of generative models in the video domain, showing remarkable utility in film and advertising. While we are fully immersed in the video content by watching and listening, unfortunately, these generated videos are silent. Generating the sound aligned to a video is a challenging task requiring both a contextual and temporal understanding of the video. Figure 1 shows an example of when text and video controls are required to generate a sound that precisely matches the given scene. Here, the dog is growling while holding a toy in his mouth. A human can imagine the sound of the video; the dog growls lowly, and the growling sounds like the dog is biting something. When the person grips and pulls the toy, the dog will treat the human by growling louder. Finally, when the dog shakes his head, the growling will become louder. If a generative model does not understand the visual information, it will be a random growling sound, not like the dog biting something. If audio is not controlled by text, the generated audio might be only related to the dog, *e.g.*, a barking sound.

Table 1 shows the recent attempts to generate an audio sample from the given video or text. There are two major directions to generate an audio sample from the given video directly. First, there have been studies of a sound effect (SFX) generation with short moments for video editing tasks (Comunità et al. 2024; Du et al. 2023), known as Foley. They are restricted to the pre-defined sound effect classes and can only control discrete information, such as onset. As another attempt, video-to-audio (V2A) generation methods have been proposed (Luo et al. 2024; Xing et al. 2024; Iashin and Rahtu 2021; Sheffer and Adi 2023). However, they still struggle to generate open-domain sounds from multiple objects together. Furthermore, both SFX and V2A methods cannot take text controls, more rich user control. Figure 1 shows the example when there is no text control; a V2A method just generates audio of "barking" rather than "growling" by focusing on the dog in the video.

As another line of research, text-to-audio (T2A) generation has been actively studied (Huang et al. 2023a,b; Liu et al. 2023, 2024; Ghosal et al. 2023). Despite their diverse and high-quality audio generation quality, they lack a temporal understanding of video-only information. Like the example in Figure 1, the text-only condition can make irrelevant audio to the video (*e.g.*, when the dog shakes heads). To tackle the problem, we may need more controllability to the T2A model, such as AudioLDM (Liu et al. 2023). Recently, a few studies (Wu et al. 2023; Guo et al. 2024a; Chung, Lee, and Nam 2024) tried to control the pre-trained AudioLDM more precisely based on ControlNet (Zhang, Rao, and Agrawala 2023). Although they can control the pitch, temporal order, energy, or rhythm of the generated audio, their generation process needs expensive timestamp-wise annotations for each control feature.

More recently, parallel to our study, SonicVisionLM (Xie

*Works done during an internship at NAVER AI Lab.
†Corresponding author: lee.j@navercorp.com

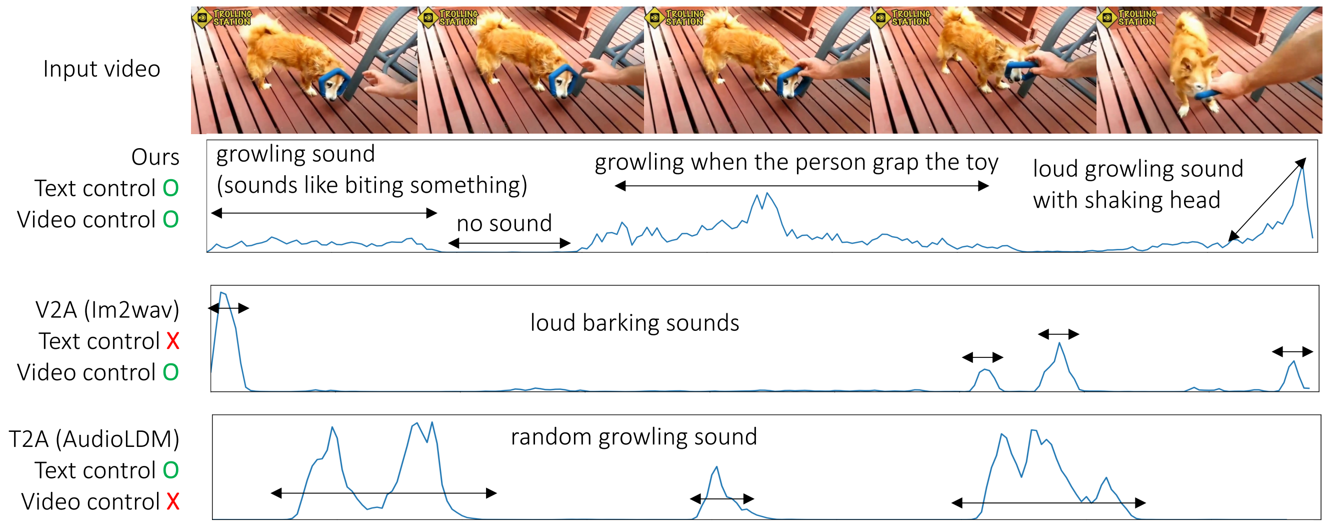

Figure 1: An example of audio generation requiring both text and video control. The text instruction "dog growling" is used for the text control. The video-to-audio (V2A) (Sheffer and Adi 2023) or text-to-audio (T2A) (Liu et al. 2023) generation methods cannot understand the detailed semantics from texts (the dog is growling, not barking) or video (the dog is biting something, and the alignment), respectively.

et al. 2024) and Seeing&Hearing (Xing et al. 2024) incorporate text information, providing users the freedom to generate specific sounds. Although these methods can control audio generation with both vision and language, they still suffer from either limited discrete control (*e.g.*, onset) (Xie et al. 2024), or lacking timestamp-wise control (Xing et al. 2024). Moreover, they require a video-to-text converting process, such as video captioning or feature mapping, for use with the T2A model. This text conversion weakens temporal alignment, leading to the loss of fine-grained temporal details.

In this work, we propose a novel video-and-text-to-audio generation approach, named Read, Watch and Scream (ReWaS), by integrating video as a conditional control for a well-established T2A model. While a text prompt specifies the subject, we additionally employ a control feature extracted from the video. More specifically, our method presents an energy adapter on AudioLDM motivated from ControlNet (Zhang, Rao, and Agrawala 2023), an efficient structure control method for text-to-image generation. Since a video feature does not directly imply the structure of the audio, we estimate the temporal *energy* information, a basic audio structural information, from the video.

The energy operates as a time-varying control to complement the sound according to the dynamics of the given video. As shown in Figure 1, ReWaS successfully understands complex information from both text and video. Here, we define energy as the mean of frequency in each audio frame, which is related to visual dynamics and semantics (Jeong et al. 2023; Guo et al. 2024a). It is relatively simple to estimate from a video rather than complex acoustic features (*e.g.*, mel-spectrograms). Therefore, our energy control facilitates connecting video for T2A model, reflecting strong alignment between audio and video.

We compare our method and other state-of-the-art video-to-audio generation models (Du et al. 2023; Xing et al. 2024; Luo et al. 2024; Iashin and Rahtu 2021; Sheffer and Adi 2023) on two video-audio aligned datasets, VGGSound (Chen et al. 2020) and GreatestHits (Owens et al. 2016). In the experiments, ReWaS outperforms V2A methods in human evaluation for three categories (audio quality, relevance to the video, and temporal alignment between audio and video) with a significant gap (almost +1 point for every category in 5-scale MOS). Also, ReWaS shows a superior audio generation performance quantitatively and qualitatively. Our method shows the best fidelity score (FD), structure prediction (energy MAE), and AV-alignment score on VGGSound. Moreover, we achieve the best AP and energy MAE on Greatest Hits without the use of reference audio samples like CondFoleyGen (Du et al. 2023). As shown in the qualitative study, ReWaS can capture the challenging "short transition" of the video when the skateboarder jumps into the air, and no skateboarding sound appears in the video. It is also possible to generate the sound of a video generated by a general text prompt.

2 Related Work

2.1 Text-to-audio generation

Early work for audio generation was built upon GANs (Kreuk et al. 2023; Dong et al. 2018), normalizing flows (Kim et al. 2020), and VAEs (Van Den Oord, Vinyals et al. 2017). Recently, several studies using diffusion models have shown promising progress on a broad range of acoustic domains. DiffSound (Yang et al. 2023) employs a diffusion-based token decoder for the first time to transfer text features into mel-spectrogram tokens. Make-An-Audio (Huang et al. 2023b), AudioLDM (Liu et al. 2023), AudioLDM2 (Liu et al. 2024), Tango (Ghosal et al. 2023) and Make-An-

Method	General sound?	Text control?	Visual control?	W/o V2T mapping?	Efficient training?
Sound effect generation (Comunità et al. 2024; Du et al. 2023)	✘	✘	✔[†]	✔	✘
Video-to-audio (Luo et al. 2024; Iashin and Rahtu 2021; Sheffer and Adi 2023)	✔[‡]	✘	✔	✔	✘
Text-to-audio (Huang et al. 2023a,b; Liu et al. 2023, 2024; Ghosal et al. 2023)	✔	✔	✘	✔	✘
Text-to-audio & Control (Wu et al. 2023; Guo et al. 2024a; Chung, Lee, and Nam 2024)	✔	✔	✘	✔	✔
Video-to-text & Text-to-audio (Xie et al. 2024; Xing et al. 2024)	✔	✔	✔[†⋆]	✘	✔
Video-and-text-to-audio (**ReWaS**)	✔	✔	✔	✔	✔

[†] Unable to adjust continuous sound variations (*i.e.*, energy). [‡] Hardly generate sounds of multiple subjects together.

[⋆] Taking limited timestamp-wise visual control (*e.g.*, requiring the full timestamp-wise onset annotations, or only able to take a few frames)

Table 1: Comparison of audio generation methods: Can it make a general sound? Can it take text or visual control? Does it need video-to-text (V2T) mapping? and the training efficiency.

Audio2 (Huang et al. 2023a) are well-founded in latent diffusion model (LDM) (Rombach et al. 2022), demonstrating high-quality results with large scale training. A series of LDM predicts mel-spectrograms using a VQ-VAE decoder, and a pretrained vocoder generates raw waveforms from the generated mel-spectrograms. While these methods successfully generate high-quality audio samples for the given text prompt, they are only designed for taking text conditions, unable to understand visual semantics.

Meanwhile, there have been a few attempts based on ControlNet (Zhang, Rao, and Agrawala 2023), an efficient training method for structure control for text-to-image generation. ControlNet utilizes hints (*e.g.*, Canny edge maps, scribbles, depth maps) to provide a structural composition to the generated images. Inspired by this, text-to-audio methods have incorporated ControlNet to accomplish controllable music (Wu et al. 2023) and audio effect (Guo et al. 2024a; Chung, Lee, and Nam 2024) generation. They have provided more explicit and fine-grained control over the generated audio, leading to performance improvement and adherence to the desired characteristics.

However, designing these time-varying controls still requires costly labor for users. To address this challenge, we predict energy control through a given video, which is a practical function for creating SFX, post-production for filmmaking, and utilizing AI-generated silent videos.

2.2 Video-to-audio generation

Existing video-to-audio (V2A) generation methods have focused on achieving two main characteristics: (i) audiovisual relevance and (ii) temporal synchronization. The first stream aims to represent general sound by leveraging datasets such as VGGSound (Chen et al. 2020) and AudioSet (Gemmeke et al. 2017). Given a set of video features, SpecVQGAN (Iashin and Rahtu 2021) learns a transformer to sample quantized representations (*i.e.*, codebook) based on visual features to decode spectrogram. Im2wav (Sheffer and Adi 2023) utilizes rich semantic representations obtained from a pre-trained CLIP (Radford et al. 2021) as sequential visual conditioning for an audio language model, and applies CFG (Ho and Salimans 2022) to steer the generation process. Recently, diffusion-based models have shown the stunning ability to generate high-quality audio (Luo et al. 2024; Xing et al. 2024). DiffFoley (Luo et al. 2024) improves audiovisual relevance by learning temporal and semantic alignment through contrastive learning. However, it necessitates tremendous training data, such as the utilization of both VGGSound and AudioSet for alignment training. Seeing&hearing (Xing et al. 2024) is another diffusion-based model that optimizes the text-to-audio diffusion model, AudioLDM (Liu et al. 2023) by using ImageBind (Girdhar et al. 2023) which learns joint embedding space for six modalities (image, text, audio, depth, thermal, and IMU). However, ImageBind Video Encoder takes only two frames for each video sampled from 2 second, which results in lacking timestamp-wise contol. Therefore, they often struggle to generate temporally aligned sounds at short times in the video (*e.g.*, dog barking, people laughing).

On the other hand, other research works (Comunità et al. 2024; Xie et al. 2024) have focused on creating simplistic SFX (*e.g.*, stick hits) using datasets like CountixAV (Zhang, Shao, and Snoek 2021) and GreatestHits (Owens et al. 2016), which provide fewer classes but more precisely temporal aligned data. CondFoleyGen (Du et al. 2023) trains a Transformer to autoregressively predict a sequence of audio codes for a spectrogram VQGAN, conditioned on the given audiovisual example. Syncfusion (Comunità et al. 2024) predicts a discrete onset label that denotes the beginning of a sound for repetitive actions. Recent SonicVisionLM (Xie et al. 2024) employs a large language model to utilize text as an intermediate product that facilitates user interaction for personalized sound generation. They freeze Tango (Ghosal et al. 2023) and train ControlNet with timestamp estimated by a video for 23 SFX categories exclusively, where the video is converted to sound event timestamp and text. Although they have shown promising results in SFX generation, their timestamp detection module is limited to a single visual object, and they cannot implicit detailed temporal cues in visual content because they use videos to convert them into text. our method generates sounds for various categories from the visual context at the same time.

3 Preliminary

3.1 Text-to-audio latent diffusion model

In this paper, we specifically utilize AudioLDM (Liu et al. 2023) which generates a latent of mel-spectrogram z com-

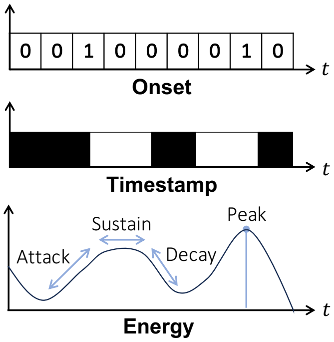

Figure 2: Discrete timestamp annotations vs. Continuous energy.

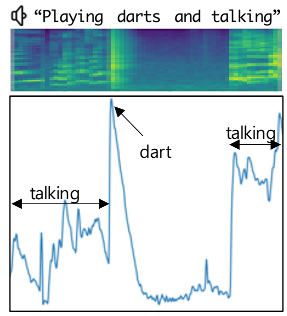

Figure 3: Energy can improve temporal alignment.

puted by VAE (Kingma and Welling 2014). The diffusion model ϵ_θ of AudioLDM is trained to predict the noise added to a given data by minimizing the objective function, $\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0,\epsilon,t} \|\epsilon - \epsilon_\theta(z_t, t, \mathbf{E}_a)\|_2^2$, where ϵ represents the noise added at time t, z_t is noisy latent induced via the forward process and $\mathbf{E}_a$ denotes the embedding of the audio x obtained from the CLAP audio encoder $f_{\text{audio}}(\cdot)$ (Wu* et al. 2023). Here, the model is conditioned by $\mathbf{E}_a$ using classifier free guidance (CFG) (Ho and Salimans 2022).

In the sampling process, the generation starts from a noise z_T sampled from $\mathcal{N}(0, I)$ and the text embedding $\mathbf{E}_y$ from the CLAP text encoder $f_{\text{text}}(\cdot)$. The reverse process conditioned on $\mathbf{E}_y$ generates the audio prior z_0 using the modified noise estimation $\hat{\epsilon}_\theta(z_t, t, \mathbf{E}_y) = (1 + w)\epsilon_\theta(z_t, t, \mathbf{E}_y) - w\epsilon_\theta(z_t, t)$, where w is a guidance weight to balance the audio condition $\mathbf{E}_a$. The VAE decoder decodes the sampled latent z to predict a mel-spectrogram. Finally, the decoded mel-spectrogram is converted to a raw audio sample using the HiFi-GAN vocoder (Kong, Kim, and Bae 2020).

Although AudioLDM enables text-conditional audio generation, it still lacks of understanding of visual contents and their temporal information. This study adds a visual control to the pre-trained AudioLDM. Instead of directly using a visual feature to control, we extract more essential information from the given video, which will be discussed in Section 3.2.

3.2 Video-to-audio with temporal alignment

We assert that a video input can bring principal temporal information that is hard to convey with a text prompt. However, directly injecting temporal information from visual into an audio generation model remains a significant challenge. In contrast, previous works have attempted to generate sound by estimating the onset (Comunità et al. 2024), or audio timestamp (Xie et al. 2024) from videos to improve audiovisual relevance. However, they are limited to producing an unnatural sound for a single object in that discrete conditions cannot serve continuous sound variations.

In this work, we consider *energy*, the averaged mel-spectrogram on the frequency axis, to produce a continuous condition. Figure 2 shows that energy is a continuous time-varying signal, including envelope components of sound such as peak, attack, sustain, and decay. Energy can be obtained cheaply and automatically by computing the frame-level magnitude of mel-spectrograms (Ren et al. 2020). Moreover, we empirically observe that energy can also implicitly improve the temporal alignment of the video. For example, Figure 3 shows energy can contain continuously varying audio information.

4 Method

This paper introduces a novel sound generation method conditioned on text and video, to generate a waveform temporally well aligned with the visual input. Our model consists of two parts: (i) *control prediction*, which intermediately predicts energy control from the video. (Section 4.1) (ii) *conditional sound generation*, which uses the energy control signal as a condition in the diffusion process to generate corresponding audio outputs (Section 4.2), which are both temporally and semantically aligned with text and video.

4.1 Energy control prediction from video

Energy control. ReWaS is based on AudioLDM that uses CLAP embeddings for text and audio alignment. A naïve approach using video as a condition is to align latent space between audio-video-text. Luo et al. (2024) attempted to align tri-modal embeddings in a unified space by large-scale contrastive learning prior to training diffusion models. To more efficiently overcome this challenge, we design an energy control as an intermediate bridge from video to audio. We speculate that energy control brings three advantages: First, the power of audio is intuitively correlated to visual dynamics and semantics (Jeong et al. 2023; Sung-Bin et al. 2023). With the natural fact that people can imagine the power of sound from the size of the instance or distance to the object, we regard audio energy as a visually correlated signal that can be certainly obtained from video. Second, as shown in Ren et al. (2020) and Guo et al. (2024b), energy plays as a structural condition for audio generation. Thus, it is well-suited to parameter-efficient fine-tuning methods such as ControlNet. Finally, using temporal acoustic signals for generating audio needs a skilled user to annotate the pitch, melody, or rhythm for every timestamp. It makes the audio generation phase impractical and difficult for the public to control. Meanwhile, energy is highly related to physical interactions implicated in visual signals; thus, it can be easily estimated from the video. Our approach does not require timestamp-wise fine-grained user control, but automatically estimating energy structure from the given video.

Video embedding. To predict the energy control from video input, we extract features from the pretrained SynchFormer (Iashin et al. 2024) video encoder. We empirically observe that the image encoder (*e.g.*, CLIP (Radford et al. 2021)) is limited to V2A generation, especially from a temporal alignment perspective. We finally take video embedding $\mathbf{E}_v \in \mathbb{R}^{S \times C}$, where S is the number of segments and C is the dimension of latent. The implementation details for this process are described in Appendix.

Training energy control from video. Similar to Ren et al. (Ren et al. 2020), we calculate the energy from the mel-spectrogram by averaging the frequency bins and further

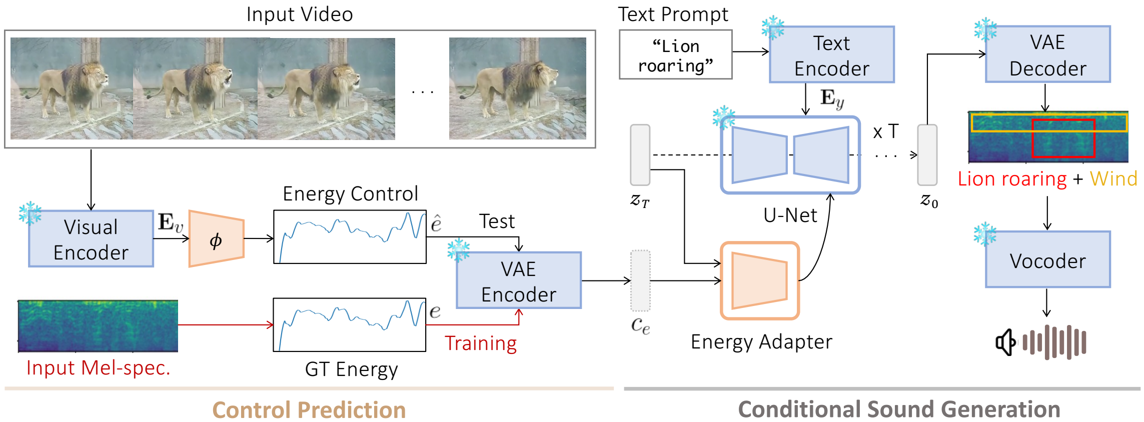

Figure 4: Overall architecture of ReWaS. Our model predicts energy control from a given video, and generates sound with text prompt and control condition. Red lines are used in training only, and replaced to the video-to-energy estimator ϕ in test time.

smoothing the time-sequential energy information. We first transform the raw waveform to the mel-spectrogram, $\texttt{mel} \in \mathbb{R}^{D\times W}$, where D represents the number of mel-frequency bins, and W is the width of the spectrogram following AudioLDM (Liu et al. 2023). However, we empirically observe that the computed energy fluctuates a lot for each temporal window, which hinders stable training. We resolve the issue by taking a smoothing operator. The energy of audio $e \in \mathbb{R}^W$ is defined as $e_a = \texttt{Smoothing}\left(\frac{1}{D}\sum_{d=1}^{D}\texttt{mel}_{w,d}\right)$. We use the second-order Savitzky-Golay filter (Virtanen et al. 2020) with a window length of 9 for smoothing.

We estimate $\hat{e}$ by using a shallow projection module ϕ from the video encoder output (See Figure 4 "Control Prediction"). For efficient training, we resize e_a by taking the nearest-neighbor interpolation to have the same number of segments S as the visual representations. We also can apply the same resize method to video embeddings at inference time. Now, we train our energy control prediction module ϕ by minimizing the following loss function $\mathcal{L}_e = ||\phi(\mathbf{E}_v) - \texttt{Resize}(e)||_2^2$.

The output $\hat{e}$ of the projection module is used for energy control at inference time. We train ϕ separately to diffusion models for training efficiency. In addition, our energy estimation module is not specialized for generation models, thus our energy control can be utilized in other ways.

4.2 Conditional sound generation

Adding control signal. To reflect the energy control signal, we train the energy adapter following ControlNet (Zhang, Rao, and Agrawala 2023). The weights of the energy adapter are initialized from pretrained parameters of diffusion models, and connected to AudioLDM with zero convolution layers. Compared to training audiovisual alignment into the latent space in diffusion model (Luo et al. 2024; Xing et al. 2024), our adapter takes the benefit of robust fine-tuning speed (*e.g.*, Luo et al. (2024) uses 8 A100 GPUs for 140 hours for feature alignment and LDM tuning, whereas we use 4 V100 GPUs for total 33 hours). To add the control feature for z_t, the energy control e_a is duplicated by the number of mel-filterbanks, and transferred to the VAE encoder for the purpose of encoding, followed by a fully-connected layer and SiLU (Elfwing, Uchibe, and Doya 2018). This latent control feature c_e is added to the z_0, where z_0 is an audio prior obtained from the VAE encoder. Thus, given a text embedding $\mathbf{E}_y$ and latent control feature c_e, we train energy adapter by optimizing the following objective: $\mathcal{L}_c = \mathbb{E}_{z_0,t,\mathbf{E}_y,c_e,\epsilon\sim\mathcal{N}(0,1)}||\epsilon - \epsilon_\theta(z_t, t, \mathbf{E}_y, c_e)||_2^2$. During training, we randomly drop $\mathbf{E}_y$ with the probability 0.3 for better controls. We denote that $\mathcal{L}_c$ and $\mathcal{L}_e$ are optimized separately.

Sound generation. We use DDIM (Song, Meng, and Ermon 2020) to generate sound from the noise. The reverse sampling process is conditioned on both text and video. We replace e to $\hat{e} = \phi(\mathbf{E}_v)$ at inference. Once mel-spectrogram is generated by the VAE decoder, it can be transformed into a raw waveform using the pre-trained vocoder (Kong, Kim, and Bae 2020) as explained in Section 3.1. We also note that conditioning on video increases the total inference time by only 3%.

5 Experiments

5.1 Experimental settings

Datasets. For a fair comparison with existing baselines, we train the control prediction module and the adapter in the conditional sound generation module on VGGSound (Chen et al. 2020). VGGSound is a large-scale dataset containing ≈200k video clips, accompanied by corresponding audio tracks. The dataset covers 309 classes of general sounds, roughly categorizing them into acoustic events, music, and people. The videos are sourced from YouTube, providing a diverse and realistic corpus. Since the VGGSound includes plentiful general sound examples, ReWaS trained on the VGGSound enables general-purpose sound generation for real-world scenarios. We randomly sampled 3K videos to construct VGGSound test subset. To evaluate temporal alignment accuracy, we use Greatest Hits (Owens et al. 2016) test set including the videos of hitting a drumstick with materials. Since Greatest Hits samples have a distinct

Model	FD↓	FAD↓	MKL↓	CLAP↑	MAE↓	AV-align↑	# TP↓
SpecVQGAN	26.63	5.57	3.30	0.1336	0.1422	0.2851	379M
Im2wav	16.87	5.94	2.53	0.4001	0.1310	0.2763	365M
Diff-Foley	21.96	6.46	3.15	0.4010	0.1571	0.2059	859M
Seeing&Hearing	20.72	6.58	**2.34**	**0.5805**	0.1668	0.1858	-
ReWaS (Ours)	**15.24**	**2.16**	2.78	0.4353	**0.1149**	**0.3008**	**204M**

Table 2: Performance comparison on VGGSound (Chen et al. 2020) with reproduced five seconds audio samples. "Energy" and "TP" denote energy MAE and number of the trainable parameters.

Model	Acc↑	AP↑	MAE↓
CondFoleyGen	**23.94**	60.24	0.1520
ReWaS (Ours)	19.15	**63.28**	**0.1398**

Table 3: Performance comparison on Greatest Hits (Owens et al. 2016). We use material types as text prompts, while CondFoleyGen uses both reference audio and video as inputs.

Model	Audio Quality ↑	Relevance ↑	Temporal Alignment ↑
SpecVQGAN	2.76	2.50	2.64
Im2wav	2.97	3.18	3.01
Diff-Foley	2.89	2.97	2.98
ReWaS (Ours)	**3.70**	**4.04**	**3.68**

Table 4: Human evaluation of V2A methods on audio quality, audiovisual relevance, and temporal alignment with 5-scale MOS.

audio property compared to the other audio samples, we fine-tune ReWaS on the Greatest Hits training samples.

Baselines. We compare ReWaS against open-source V2A generation approaches in priority, SpecVQGAN (Iashin and Rahtu 2021), Im2wav (Sheffer and Adi 2023) and Diff-Foley (Luo et al. 2024), which are trained on the VGGSound and AudioSet datasets. Furthermore, we compare Seeing&Hearing (Xing et al. 2024), which optimizes a pre-trained AudioLDM during the inference stages by aligning the latent space using ImageBind. For a fair comparison, we take the following steps: We first generate the full-length audio by each method, and use a common 5-second clip for evaluation. In the temporal alignment evaluation, we consider CondFoleyGen (Du et al. 2023) as a main baseline, which is trained on the Greatest Hits dataset.

Evaluation metrics. Following the implementation of AudioLDM, we employ Fréchet distance (FD) (Heusel et al. 2017), Fréchet audio distance (FAD) (Tailleur et al. 2024), and the mean of KL divergence (MKL) (Iashin and Rahtu 2021). We also measure the alignment between the generated audio and sound categories with CLAP score (Huang et al. 2023b). However above metrics are limited to evaluating audio-visual temporal alignment, so we employ AV-align (Yariv et al. 2024) based on detecting energy peaks in audio-visual modalities. In the Greatest Hits experiment, we report onset accuracy (Acc) and average precision (AP), following the evaluation protocol introduced by CondFoleyGen. The onset of sound events is a discrete signal obtained by the thresholding of the amplitude gradient. Therefore, relatively quiet sound effects (*e.g.*, scratching leather, touching the leaves) or natural sounds can be excluded from the evaluation. To address this issue, we report the mean absolute error (MAE) (Guo et al. 2024a) of the energy signals from real and generated sounds for the first time in the sound generation task conditioned on video. We also conduct a user study to evaluate the quality and temporal alignment of the generated audio samples.

5.2 Results

Quantitative results. Table 2 shows the quantitative comparisons on the VGGSound. We note that category classes are used as text prompts in the VGGSound. We train 22M parameters for video projection to audio conditional control, and 182M parameters for fine-tuning the AudioLDM with our energy adapter. Since Seeing&Hearing is an optimization-based generation method, we did not report the training parameters. However, they consume twice the time for inference than ReWaS. Our ReWaS achieve the best performance on FD, FAD, energy MAE, and AV-align, showing competitive results in terms of MKL and CLAP scores. Especially, while we use only a quarter of training parameters compared to Diff-Foley, our method outperforms Diff-Foley on all metrics. CLAP scores illustrate the importance of text prompts for semantic alignment. Seeing&Hearing outperforms ReWaS in terms of MKL and CLAP score. However, we argue that Seeing&Hearing is heavily dependent on text prompt, since our method outperforms in terms of MAE and AV-align scores by a large margin. This achieved MAE score result by ReWaS also demonstrates the accuracy of our control prediction module, and generated outputs from ReWaS are most temporally closer to the real audio content.

In addition, we evaluate how the generated audio and the condition video are temporally aligned on Greatest Hits. The dataset distribution of Greatest Hits highly differs from the general audio samples; hence, we fine-tune ReWaS on the Greatest Hits training samples. Table 3 shows the results. ReWaS achieves the best AP and MAE, although ReWaS is not specially designed for Foley like CondFoleyGen.

User study. The quantitative results are limited to measuring how the generated audio sounds realistic and aligned to the given video. To complement it, we conduct a human evaluation study to assess the subjective quality of the generated audio concerning the input video. We ask the human evaluators to evaluate the quality of the audio samples generated by SpecVQGAN, Im2wav, Diff-Foley, and ReWaS. Since Seeing&Hearing shows vulnerable performance in audio-visual alignment, we exclude it from the user study.

We evaluate three criteria: audio quality, relevance between audio and video, and temporal alignment. We use a five-point Likert scale to measure mean opinion score (MOS), where an ideal video with its ideal audio receives a rating of 5 across all criteria. We recruit human annotators via two separate channels: Amazon Mechanical Turk (AMT) and local hiring. We recruit 50 AMT annotators for each cri-

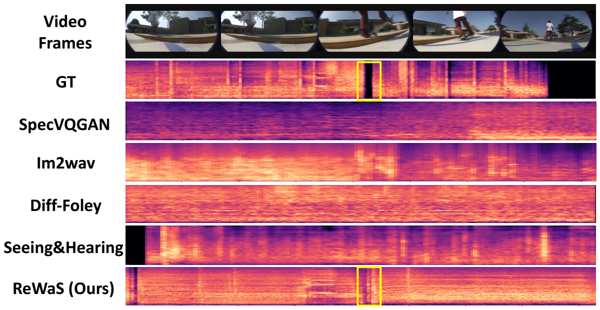

Figure 5: Qualitative comparison on VGGSound. Surprisingly, when the skateboarder jumps, only ReWaS succeded in detecting short transition (yellow box). Text prompt in is "skateboarding".

Control	Backbone	FD↓	FAD↓	KL↓	CLAP↑	MAE ↓
T & GT E from A	AudioLDM-M	13.93	2.65	2.15	0.4497	0.1195
T & Est. E from V	AudioLDM-M	15.24	2.16	2.78	0.4353	0.1149
T & Est. E from V	Make-An-Audio	13.89	10.91	2.93	0.4237	0.1368

Table 5: Impact of the energy control's quality on VGGSound. (1) Text and the ground-truth audio energy with AudioLDM backbone (upper bound), (2) Text and the estimated energy from the video with AudioLDM (our approach) and (3) with Make-An-Audio.

terion, and each annotator evaluates five generated samples for each method (*i.e.*, each annotator evaluates 20 audios). Locally hired 23 annotators evaluated 20 generated samples for each method and criterion. Surprisingly, ReWaS achieves the best in all categories with large margins as shown in Table 4. This subjective result is consistent with our quantitative and qualitative findings, further validating the effectiveness of ReWaS in generating high-quality, relevant, and temporally synchronized audio for the given video.

Qualitative results. Figure 5 shows qualitative results in baselines and ReWaS. Given the skateboarding video, SpecVQGAN and Diff-Foley fail to generate the sound of skate wheels rolling on the floor. Although Im2wav generates that sound, it cannot capture a short transition. We also demonstrate the effectiveness of the text prompt in Figure 6 with CLAP similarity, when redundant frames exist. In this case, V2A methods also struggle to generate corresponding sound. However, ReWaS can effectively calibrate the semantics by user text prompt. Note that the prompt can also be longer and more general if desired by users.

5.3 Discussion

Robustness. Table 5 demonstrates that although we use the estimated energy, the quality of the generated audio is very similar to the audio samples controlled by the ground truth audio energy. (See the first row and second row of Table 5) Interestingly, ReWaS built upon Make-An-Audio achieves comparable performance to its AudioLDM-M counterpart, demonstrating the robustness of our framework (See the second row and third row of Table 5).

General text prompts. To examine the capability of ReWaS

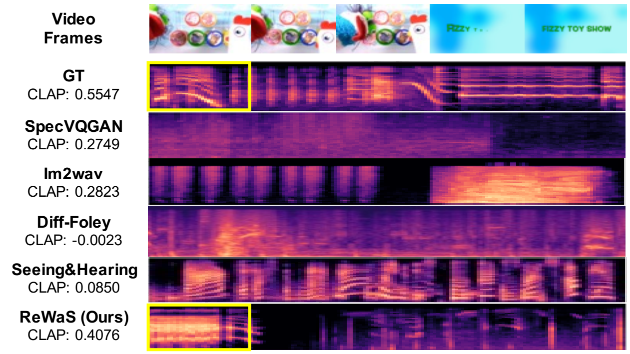

Figure 6: Effectiveness of text prompt. When videos contain both semantic and redundant frames, text prompts used in ReWaS calibrate the results. Text prompt in is "chicken clucking"

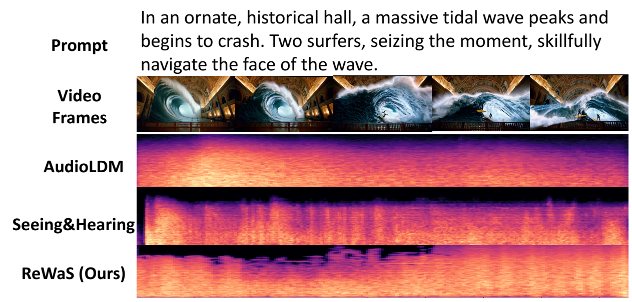

Figure 7: Audio generation with general text prompts.

with more general text prompts, we generate audio samples with generative videos by KLING[1]. As shown in Figure 7, only our method can capture the visual information of the wave, namely, the sound getting louder as the wave crashes.

6 Conclusion

This paper proposes ReWaS, a novel video-and-text-to-audio generation framework. Our key idea lies in inferring audio structural condition, namely energy, from video to efficiently and effectively input the visual condition to the robust T2A model. Therefore, ReWaS can generate complex sounds in the real world without the need for a difficult control design. Quantitative results on VGGSound and Greatest Hits datasets, subjective human study, and qualitative results consistently support that ReWaS can generate natural, temporally well-aligned, and relevant audio for the given video by employing text and video as control.

Acknowledgements

The NAVER Smart Machine Learning (NSML) platform (Kim et al. 2018) had been used for experiments.

[1]https://kling.kuaishou.com/en

References

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP*.

Chung, Y.; Lee, J.; and Nam, J. 2024. T-Foley: A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis. In *ICASSP*.

Comunità, M.; Gramaccioni, R. F.; Postolache, E.; Rodolà, E.; Comminiello, D.; and Reiss, J. D. 2024. Syncfusion: Multimodal Onset-Synchronized Video-to-Audio Foley Synthesis. In *ICASSP*.

Dong, H.-W.; Hsiao, W.-Y.; Yang, L.-C.; and Yang, Y.-H. 2018. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI*.

Du, Y.; Chen, Z.; Salamon, J.; Russell, B.; and Owens, A. 2023. Conditional Generation of Audio From Video via Foley Analogies. In *CVPR*.

Elfwing, S.; Uchibe, E.; and Doya, K. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*.

Ghosal, D.; Majumder, N.; Mehrish, A.; and Poria, S. 2023. Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model. arXiv:2304.13731.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *CVPR*.

Guo, Z.; Mao, J.; Tao, R.; Yan, L.; Ouchi, K.; Liu, H.; and Wang, X. 2024a. Audio Generation with Multiple Conditional Diffusion Model. In *AAAI*.

Guo, Z.; Mao, J.; Tao, R.; Yan, L.; Ouchi, K.; Liu, H.; and Wang, X. 2024b. Audio Generation with Multiple Conditional Diffusion Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 18153–18161.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Ho, J.; and Salimans, T. 2022. Classifier-Free Diffusion Guidance. arXiv:2207.12598.

Huang, J.; Ren, Y.; Huang, R.; Yang, D.; Ye, Z.; Zhang, C.; Liu, J.; Yin, X.; Ma, Z.; and Zhao, Z. 2023a. Make-An-Audio 2: Temporal-Enhanced Text-to-Audio Generation. arXiv:2305.18474.

Huang, R.; Huang, J.; Yang, D.; Ren, Y.; Liu, L.; Li, M.; Ye, Z.; Liu, J.; Yin, X.; and Zhao, Z. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *ICML*.

Iashin, V.; and Rahtu, E. 2021. Taming Visually Guided Sound Generation. arXiv:2110.08791.

Iashin, V.; Xie, W.; Rahtu, E.; and Zisserman, A. 2024. Synchformer: Efficient Synchronization From Sparse Cues. In *ICASSP*.

Jeong, Y.; Ryoo, W.; Lee, S.; Seo, D.; Byeon, W.; Kim, S.; and Kim, J. 2023. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *ICCV*.

Kim, H.; Kim, M.; Seo, D.; Kim, J.; Park, H.; Park, S.; Jo, H.; Kim, K.; Yang, Y.; Kim, Y.; Sung, N.; and Ha, J.-W. 2018. NSML: Meet the MLaaS platform with a real-world case study. arXiv:1810.09957.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *NeurIPS*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*.

Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; and Adi, Y. 2023. AudioGen: Textually Guided Audio Generation. In *ICLR*.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. arXiv:2301.12503.

Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; and Plumbley, M. D. 2024. AudioLDM 2: Learning Holistic Audio Generation with Self-supervised Pretraining. arXiv:2308.05734.

Luo, S.; Yan, C.; Hu, C.; and Zhao, H. 2024. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*.

Owens, A.; Isola, P.; McDermott, J.; Torralba, A.; Adelson, E. H.; and Freeman, W. T. 2016. Visually indicated sounds. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *ICLR*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Sheffer, R.; and Adi, Y. 2023. I hear your true colors: Image guided audio generation. In *ICASSP*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *ICLR*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Sung-Bin, K.; Senocak, A.; Ha, H.; Owens, A.; and Oh, T.-H. 2023. Sound to visual scene generation by audio-to-visual latent alignment. In *CVPR*.

Tailleur, M.; Lee, J.; Lagrange, M.; Choi, K.; Heller, L. M.; Imoto, K.; and Okamoto, Y. 2024. Correlation of Fréchet Audio Distance With Human Perception of Environmental Audio Is Embedding Dependant. arXiv:2403.17508.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. In *NeurIPS*.

Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*.

Wu, S.-L.; Donahue, C.; Watanabe, S.; and Bryan, N. J. 2023. Music ControlNet: Multiple Time-varying Controls for Music Generation. arXiv:2311.07069.

Wu*, Y.; Chen*, K.; Zhang*, T.; Hui*, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP*.

Xie, Z.; Yu, S.; He, Q.; and Li, M. 2024. SonicVisionLM: Playing Sound with Vision Language Models. arXiv:2401.04394.

Xing, Y.; He, Y.; Tian, Z.; Wang, X.; and Chen, Q. 2024. Seeing and Hearing: Open-domain Visual-Audio Generation with Diffusion Latent Aligners. In *CVPR*.

Yang, D.; Yu, J.; Wang, H.; Wang, W.; Weng, C.; Zou, Y.; and Yu, D. 2023. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Yariv, G.; Gat, I.; Benaim, S.; Wolf, L.; Schwartz, I.; and Adi, Y. 2024. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.

Zhang, Y.; Shao, L.; and Snoek, C. G. 2021. Repetitive activity counting by sight and sound. In *CVPR*.