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, generating sound solely from text poses challenges in ensuring comprehensive scene depiction and temporal alignment. Meanwhile, video-to-audio generation limits the flexibility to prioritize sound synthesis for specific objects within the scene. 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. Our demo is available at `https://rewas-tv2a.github.io/`.

1 Introduction

Generative models have developed dramatically, making content creation easier for people, including images, videos, and audio, based on text. Especially, text-to-video generation models such as Make-a-Video [34] and Sora [1] 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 realistic sound for the given video. 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 [4, 6],

*Works done during an internship at NAVER AI Lab.

†Corresponding author: lee.j@navercorp.com

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

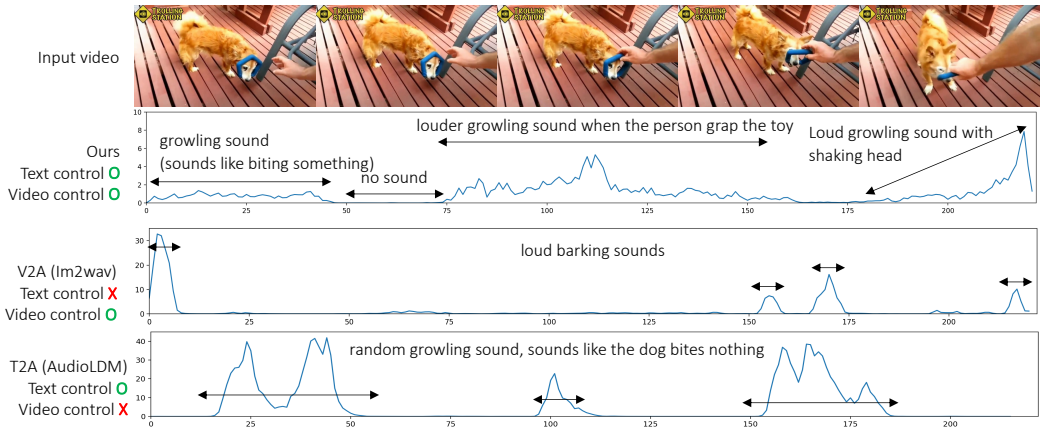


Figure 1: An example of audio generation requiring both text and video control. The text instruction "dog growling" is used for the text control. The video-to-audio (V2A) [33] or text-to-audio (T2A) [25] generation methods cannot understand the detailed semantics from texts (the dog is growling, not barking) or video (the dog is biting something, and the alignment), respectively.

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 [27, 43, 18, 33]. 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 [16, 17, 25, 26, 9]. 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 [25]. Recently, a few studies [40, 12, 3] tried to control the pre-trained AudioLDM more precisely based on ControlNet [46]. 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 [42] and Seeing&Hearing [43] 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) [42], or lacking timestamp-wise control [43]. 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 [46], 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 [20, 12]. 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 [6, 43, 27, 18, 33] on two video-audio aligned datasets, VGGSound [2] and GreatestHits [28]. In the experiments,

Table 1: Comparison of audio generation methods: Can it make a general sound? Can it take text or visual control? and the training efficiency.

Method	General Text sound?	Visual control?	W/o text control?	Efficient mapping?	training?
Sound effect (SFX) generation [4, 6]	✘	✘	✔†	✔	✘
Video-to-audio (V2A) [27, 18, 33]	✔‡	✘	✔	✔	✘
Text-to-audio (T2A) [16, 17, 25, 26, 9]	✔	✔	✘	✔	✘
T2A + Control [40, 12, 3]	✔	✔	✘	✔	✔
Video-to-text & T2A [42, 43]	✔	✔	✔†⋆	✘	✔
ReWaS (ours)	✔	✔	✔	✔	✔

† 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)

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 [6]. As shown in the qualitative study, ReWaS can capture the challenging "short transition" of the video when the boarder jumps into the air, and no skateboarding sound appears in the video.

2 Related work

2.1 Text-to-audio generation

Early work for audio generation was built upon GANs [24, 5], normalizing flows [21], and VAEs [38]. Recently, several studies using diffusion models have shown promising progress on a broad range of acoustic domains. DiffSound [44] employs a diffusion-based token decoder for the first time to transfer text features into mel-spectrogram tokens. Make-An-Audio [17], AudioLDM [25], AudioLDM2 [26], Tango [9] and Make-An-Audio2 [16] are well-founded in latent diffusion model (LDM) [32], 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 [46], 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, MusicControlNet [40] showed control over melody, dynamics, and rhythm, while Guo et al. [12] built a FusionNet between each layer of the U-Net, enabling the fusion of control embeddings for temporal order, pitch, and energy controls. On the other hand, T-Foley [3] introduces Block-FiLM, which generates foley sounds guided by temporal events such as vocal sounds. They have demonstrated that incorporating control signals into the audio generative models provides 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 [2] and AudioSet [8]. Given a set of video features, SpecVQGAN [18] learns a transformer to sample quantized representations (*i.e.*, codebook)

based on visual features to decode spectrogram. Im2wav [33] utilizes rich semantic representations obtained from a pre-trained CLIP [30] as sequential visual conditioning for an audio language model, and applies CFG [15] to steer the generation process. Recently, diffusion-based models have shown the stunning ability to generate high-quality audio [27, 43]. DiffFoley [27] 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 [43] is another diffusion-based model that optimizes the generation process using ImageBind [10] which learns joint embedding space for six modalities (image, text, audio, depth, thermal, and IMU). For V2A purpose, Seeing&hearing utilizes the text-to-audio diffusion model, AudioLDM [25], and aligns its latent space with the video embeddings extracted from the ImageBind Video Encoder during the reverse diffusion process. 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 [4, 42] have focused on creating simplistic SFX (*e.g.*, stick hits) using datasets like CountixAV [47] and GreatestHits [28], which provide fewer classes but more precisely temporal aligned data. CondFoleyGen [6] trains a Transformer to autoregressively predict a sequence of audio codes for a spectrogram VQGAN, conditioned on the given audiovisual example. Syncfusion [4] predicts a discrete onset label that denotes the beginning of a sound for repetitive actions. Recent SonicVisionLM [42] employs a large language model to utilize text as an intermediate product that facilitates user interaction for personalized sound generation. They freeze Tango [9] 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 [25] which generates a latent of mel-spectrogram z computed by VAE [22]. 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 $f_{\text{audio}}(\cdot)$ [41]. Here, the model is conditioned by $\mathbf{E}_a$ using classifier free guidance (CFG) [15].

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 [23].

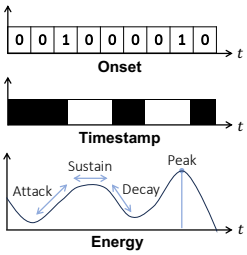


Figure 2: Limitation of timestamp annotations.

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.

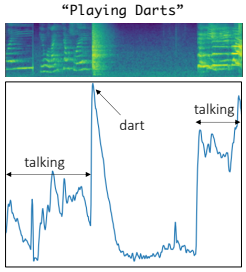


Figure 3: Energy can imply multiple semantics.

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

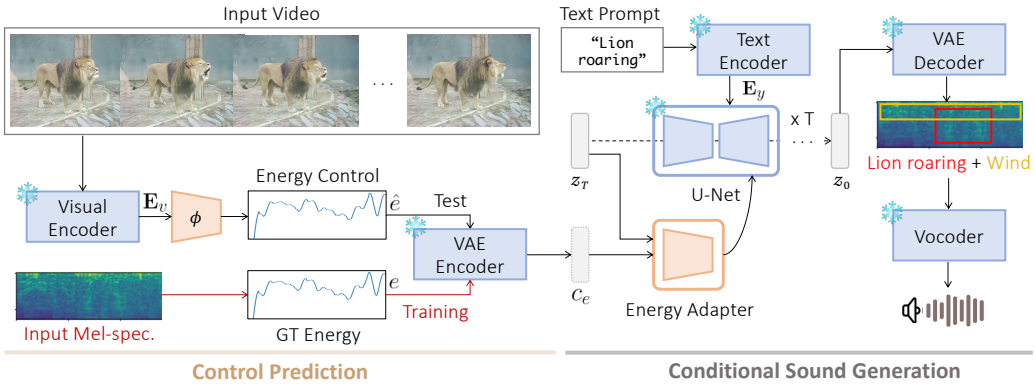


Figure 4: Overall architecture of ReWaS. Our model predicts energy control from a given video, and generates sound with text prompt and control condition. Red lines are used in training only, and replaced to the video-to-energy estimator ϕ in test time.

sound by estimating the onset [4], or audio timestamp [42] 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 [31]. 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 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. Previous approach [27] 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 [20, 36]. 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 previous works [31] and [13], 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 [19] video encoder. We empirically observe that the image encoder (*e.g.*,

CLIP [30]) 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. [31], we calculate the energy from the mel-spectrogram by averaging the frequency bins and further 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 [25]. 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 [39] 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 the framework of ControlNet [46]. 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 [27, 43], our adapter takes the benefit of robust fine-tuning speed (*e.g.*, [27] 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 [7]. 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 [35] 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 [23] as explained in Section 3.1.

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 [2]. 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 [28] test set including the videos of hitting a drumstick with materials. Since Greatest Hits samples have a distinct audio property compared to the other audio samples, we fine-tune ReWaS on the Greatest Hits training samples.

Table 2: Performance comparison on VGGSound [2] with reproduced five seconds audio samples. "Energy" and "TP" denote energy MAE and number of the trainable parameters.

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**

Baselines. We compare ReWaS against open-source V2A generation approaches in priority, SpecVQ-GAN [18], Im2wav [33] and Diff-Foley [27], which are trained on the VGGSound and AudioSet datasets. Furthermore, we compare Seeing&Hearing [43], 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 [6] as a main baseline, which is trained on the Greatest Hits dataset.

Evaluation metrics. Following the implementation of AudioLDM, we employ Fréchet distance (FD) [14], Fréchet audio distance (FAD) [37], and the mean of KL divergence (MKL) [18]. We also measure the alignment between the generated audio and sound categories with CLAP score [17] in VGGSound. However above metrics are limited to evaluating audio-visual temporal alignment, so we employ AV-align [45, 43] 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 [6]. 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) [12] of the energy signals from real and generated sounds for the first time in the sound generation task conditioned on video. Although these evaluation metrics can evaluate different properties of the generated audio, most of them measure the difference between the generated audio and the "ground truth" audio corresponding to the original video. However, one video can sound differently (*e.g.*, human's voice can vary); existing quantitative evaluation metrics have challenges in measuring whether the generated audio is truly suited to the given video. To tackle the issue, we 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.

Table 3: Performance comparison on Greatest Hits [28]. We use material types as text prompts, while CondFoleyGen uses both reference audio and video as inputs.

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

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

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 backbone (our approach) and (3) Text and the estimated energy from the video with Make-An-Audio backbone.

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	0.4170	0.1149
T & Est. E from V	Make-An-Audio	13.89	10.91	2.93	0.4237	0.1368

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.

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

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**

We use three evaluation criteria: audio quality, relevance between audio and video, and temporal alignment. Detailed user instructions are in the Appendix. 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 criterion, and each annotator evaluates five generated samples for each method (*i.e.*, each annotator evaluates 20 audios). For 23 locally hired annotators, we ask them to evaluate 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 difficult or 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. (*e.g.*, "rally car swiftly navigates a turn on the racetrack.") More results can be found in the Appendix and demo page.

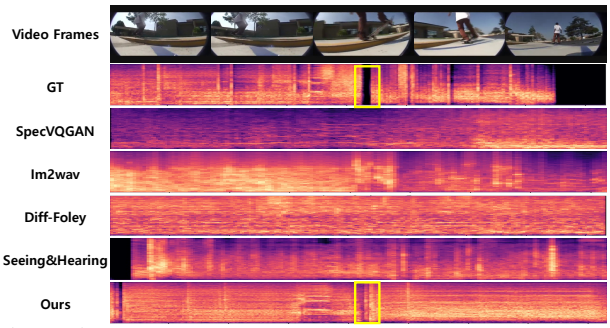
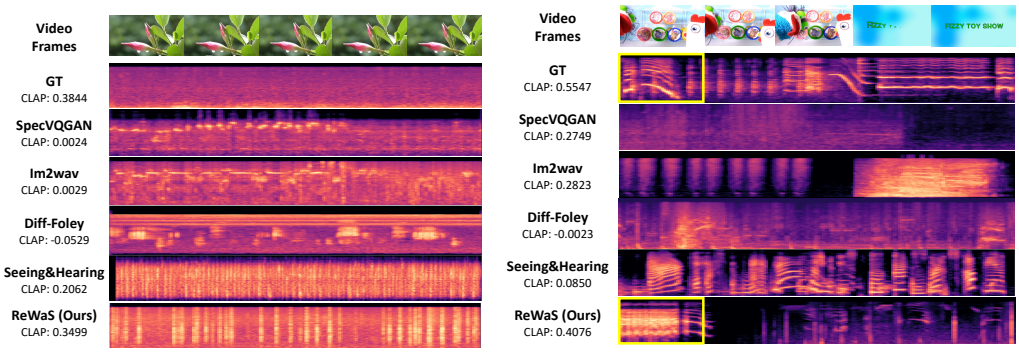


Figure 5: Qualitative comparison on VGGSound. Surprisingly, when the skateboarder jumps, only ReWaS succeded in detecting short transition (yellow box). Text prompt in is "skateboarding".

5.3 Discussion

The impact of the quality of the energy control. To verify the robustness of the energy prediction module, we compare the control by our video-to-energy prediction module and the energy directly extracted from the ground truth audio. 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) It supports the idea that energy information is highly related to visual information, and is easy to estimate solely using video.



(a) (b)

Figure 6: Effectiveness of text prompt. Videos in the real world are sometimes noisy. For example, when videos (a) are hard to distinguish the semantics or (b) contain redundant frames, text prompts used in ReWaS calibrate the results. Text prompt in (a) is "raining", and (b) is "chicken clucking".

We also compare the qualitative results of estimated energy controls with ground truth energy in the Appendix.

T2A Framework. We replace the AudioLDM-M [25] backbone with Make-An-Audio [17], which has fewer parameters than AudioLDM to validate the flexibility of our approach. Details can be found in the Appendix. Table 5 shows the results of the two backbones on VGGSound. Interestingly, ReWaS built upon Make-An-Audio achieves comparable performance to its AudioLDM-M counterpart (See the second row and third row of Table 5). The results demonstrate the robustness of our framework to the choice of the backbone model, and explain our framework has the potential to develop further with advanced T2A models.

General text prompts. To examine the capability of ReWaS with more general text prompts, we generate audio samples with generative videos by KLING[3]. 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. We include more samples in the Appendix.

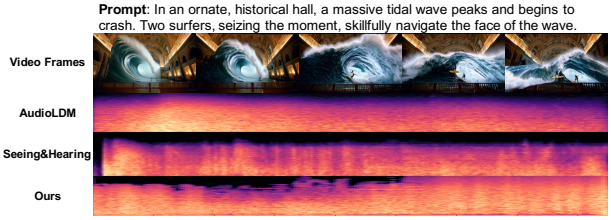


Figure 7: Audio generation with general text prompts.

Effectiveness of visual condition. In the Appendix, we show examples when energy control complementing temporal information. While AudioLDM suffers from inferior temporal alignment and limited sound generation that is mentioned in text prompts but not generated, ReWaS not only generates video-related sounds hidden in the text but also aligns the sound with the frames. This demonstrates the effectiveness of visual condition by ReWaS.

Limitation Although our approach successfully leverages the text and video control simultaneously, our method shares the limitation of AudioLDM, namely, hallucination in generated samples. For example, for a given "basketball bounce" video, ReWaS generates a squeaking sound, even if the player is standing still. This problem might be mitigated if we can use a better AudioLDM model.

6 Conclusion

This paper proposes ReWaS, a novel video-and-text-to-sound generation framework. Our main contribution is that audio structural condition, namely energy, is inferred from video to efficiently and effectively input 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.

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

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024.

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020.

[3] Yoonjin Chung, Junwon Lee, and Juhan Nam. T-foley: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis. In *ICASSP*, 2024.

[4] Marco Comunità, Riccardo F Gramaccioni, Emilian Postolache, Emanuele Rodolà, Danilo Comminiello, and Joshua D Reiss. Syncfusion: Multimodal onset-synchronized video-to-audio foley synthesis. In *ICASSP*, 2024.

[5] Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment. In *AAAI*, 2018.

[6] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *CVPR*, 2023.

[7] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 2018.

[8] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017.

[9] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction tuned llm and latent diffusion model. *arXiv preprint arXiv:2304.13731*, 2023.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

[11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[12] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *AAAI*, 2024.

[13] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18153–18161, 2024.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

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

[16] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023.

[17] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *ICML*, 2023.

[18] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

[19] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. *arXiv preprint arXiv:2401.16423*, 2024.

[20] Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim, and Jinkyu Kim. The power of sound (tpos): Audio reactive video generation with stable diffusion. In *ICCV*, 2023.

[21] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In *NeurIPS*, 2020.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[23] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *NeurIPS*, 2020.

[24] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *ICLR*, 2023.

[25] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

[26] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[27] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2024.

[28] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *CVPR*, 2016.

[29] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[31] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*, 2020.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[33] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *ICASSP*, 2023.

[34] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2022.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020.

[36] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *CVPR*, 2023.

[37] Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of fr\'echet audio distance with human perception of environmental audio is embedding dependant. *arXiv preprint arXiv:2403.17508*, 2024.

[38] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017.

[39] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 2020.

[40] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *arXiv preprint arXiv:2311.07069*, 2023.

[41] 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 *ICASSP*, 2023.

[42] Zhifeng Xie, Shengye Yu, Mengtian Li, Qile He, Chaofeng Chen, and Yu-Gang Jiang. Sonicvisionlm: Playing sound with vision language models. *arXiv preprint arXiv:2401.04394*, 2024.

[43] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *CVPR*, 2024.

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

[45] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In *AAAI*, 2024.

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

[47] Yunhua Zhang, Ling Shao, and Cees GM Snoek. Repetitive activity counting by sight and sound. In *CVPR*, 2021.

A Appendix

A.1 Data Preprocessing

During training, we randomly extract 5-second segments from the VGGSound dataset [2] and 2-second segments from the Greatest Hits dataset. However, during the testing phase, we extract video clips ranging from 2 to 7 seconds in duration for the VGGSound dataset, and from 0 to 2 seconds for the Greatest Hits dataset [28]. Video frames are uniformly sampled at 25 fps. Since ReWaS generates audio based on 5-second videos, we duplicated frames from the Greatest Hits dataset to match the length of these 5-second videos. Subsequently, we trimmed the generated audio to a duration of 2 seconds.

For comparison with baselines, SpecVQGAN [18], Diff-Foley [27] and Seeing&Hearing [43] (10s, 8s, and 10s, respectively) for the test videos. Then, we extract the 5-second clip corresponding to the same video frames used in our method. Since Im2wav [33] is designed to generate sound with a fixed length of 4 seconds, we first generate the initial 4 seconds and extend it by generating an additional 1 second, resulting in a 5-second audio clip.

A.2 Feature Extraction

Video features. We employ SynchFormer [19] trained on VGGSound [2] for the sparse synchronized setting as a video encoder. The video encoder employed in SynchFormer is based on Motionformer [29] pre-trained on Something-Something v2 [11], and fine-tuned on VGGSound and AudioSet [8]. Therefore, the video encoder is strong enough to encode motion dynamics and semantics. We freeze the parameters in the video encoder, and solely train a projection module to estimate energy control. We extract a video feature in the short video clip (0.64 sec). Thus we use a total of 112 length visual embeddings for a 5s video. We note that, for a fair comparison, RGB frames are only used in all methods including ReWaS.

Audio features. Audios of all videos used in our experiments are resampled to 16kHz sampling rate. We follow the default setting of AudioLDM to compute the mel-spectrogram. Specifically, we use 64-bin mel-spectrograms with 1024 window length. While f_{min} and f_{max} are 0 and 8000 respectively, the hop size is 160 and the FFT size is 1024.

A.3 Architecture and training details

Test Dataset. For our experiments, we leverage a subset of 160k videos from VGGSound [2] due to the availability of public videos at the time of training. We split the train data list into training and validation subsets following SpecVQGAN [18].

Energy signal. To encode a video feature into 1-dimensional energy, a projection module ϕ consists of a linear layer, two transformer blocks, and MLPs consisting of four FC layers. We use 768 hidden dimensions for the first linear layer and transformer blocks, and the four FC layers' output dimensions are 128, 64, 16, and 1. The total parameter of ϕ is 22M. We choose AudioLDM-M[4], and the number of training parameters for fine-tuning AudioLDM [25] with our energy adapter is 182M. ReWaS is optimized by AdamW and the learning rate is fixed to 3e-5 during training. We train ReWaS with 4 V100 GPUs for 33 hours on VGGSound, and 1 hour on Greatest Hits [28] respectively.

Details of Make-An-Audio Backbone Framework. The video encoder used in Make-An-Audio [17] is re-trained to predict the appropriate energy scale of mel-spectogram, which is configured with 80 frequency bins and a hop size of 256 samples, different from the AudioLDM-M configuration. Make-An-Audio is notable for its parameter efficiency, requiring significantly fewer parameters than AudioLDM. This reduction in model complexity translates to substantially shorter training times, with the entire model converging in less than one day.

A.4 User study

Figure A.1, Figure A.2, and Figure A.3 show the user instructions used in our human evaluation. Before launching Amazon MTurk (AMT), we first conducted an in-lab study with 23 participants;

[4]weights in `https://github.com/haoheliu/AudioLDM`

each participant evaluated 20 audio samples for each method and each criterion, namely, they evaluated 240 ($20 \times 4 \times 3$) generated audio samples. Based on the observation from the in-lab study, we have set the compensation level for each HIT to \$0.45 so that a worker can earn \$15 per hour. At the same time, we observed that a number of participants had trouble keeping focus on the evaluation with 240 samples (each sample takes five seconds). To prevent the low-quality responses from MTurk annotators, we split each evaluation Human Intelligence Task (HIT) on a smaller scale. Each AMT annotator evaluates five audio samples for each method and one additional ground truth audio to prevent random guessing. We published 50 HITs for each criterion, and 150 responses were collected. Finally, we observe that many AMT annotators consistently score high for all questions (*e.g.*, 4 or 5). To ignore noisy responses, we omit responses having an average score larger than 4.0 for 21 questions. 55 responses were omitted after this filtering process.

Instruction 1

How natural is this audio recording?

Please focus on examining the audio quality and naturalness (noise, timbre, sound clarity, and high-frequency details).

1. Listen to the sample (Click **Play** button to listen audio samples)
2. Select an option
    - Excellent: 5 (Completely natural audio)
    - Good: 4 (Mostly natural audio)
    - Fair: 3 (Equally natural and unnatural audio)
    - Poor: 2 (Mostly unnatural audio)
    - Bad: 1 (Completely unnatural audio)

Figure A.1: User instruction for audio quality (naturalness) test.

Instruction 2

How much is the sound related to the object or material in video?

Please focus on examining the relevance between video and audio, not considering the quality and temporal alignment (*i.e.* sound timing).

1. Watch the sample (Click **Play** button to watch video samples)
2. Select an option
    - Excellent: 5 (Completely relevant audio)
    - Good: 4 (Mostly relevant audio)
    - Fair: 3 (Equally relevant and irrelevant audio)
    - Poor: 2 (Mostly irrelevant audio)
    - Bad: 1 (Completely irrelevant audio)

Figure A.2: User instruction for video-audio relevance test

Instruction 3

How much is the sound temporally aligned to the movements of objects or material in video?

Please focus on examining the temporal alignment between video and audio, not considering audio quality and naturalness.

1. Watch the sample (Click **Play** button to watch video samples)
2. Select an option
    - Excellent: 5 (Completely aligned audio)
    - Good: 4 (Mostly aligned audio)
    - Fair: 3 (Equally aligned and non-aligned audio)
    - Poor: 2 (Mostly non-aligned audio)
    - Bad: 1 (Completely non-aligned audio)

Figure A.3: User instruction for temporal alignment test.

A.5 More qualitative results

Energy controls from Videos. We illustrate estimated energy from video in Figure A.4. The results show the correlation between our energy control generated from video and GT energy obtained from reference audio.

Effectiveness of the text prompt. As shown in Figure A.5, if there are redundant frames, ReWaS can only successfully calibrate the semantics with textual prompt but also it generates “silent" audio sounds when there is a scene change. In contrast, other baseline models such as SpecVQGAN [18], Im2wav [33] and Diff-Foley [27] fail to produce the corresponding sound (e.g., alarm clock ringing) due to misaligned visual and sound contexts, often generating unintended sounds or remaining silent when they should produce sound. Although Seeing&Hearing [43] can produce corresponding sounds, it fails to generate “silent" audio when there is a change in visual scenes. This suggests that baseline models may either resort to generating random sounds when faced with a scene change due to misaligned visual and sound contexts, or they produce sound when they should remain silent, ignoring the visual context.

Effectiveness of visual control. Figure A.6 and Figure A.7 are examples when energy signal serving additional temporal information. As shown in Figure A.6, when a person talking and playing a dart game in an input video, the original AudioLDM [25] generates only the sound of talking, ignoring ‘dart’ prompt. Additionally, aligning generated sound with video is challenging in AudioLDM. In comparison, ReWaS not only generates both the sound of talking and dart but also aligns the sound with the frames. Figure A.7 presents another example. Unlike AudioLDM, which repeatedly generates the same spray and car engine sounds, ReWaS accurately captures the spray sound at the right moment thanks to the visual control without additional text prompt ‘spray’. Furthermore, the result demonstrates the limitation of T2A methods for automatic Foley synthesis, because they cannot watch a video. This demonstrates the effectiveness of visual control by ReWaS.

Subtle Visual Movements. Figure A.8 and Figure A.9 demonstrate the effectiveness of ReWaS in aligning sound with corresponding frames, achieving temporal alignment by accurately capturing small object movements, such as lip synchronization. As illustrated in Figure A.8, the the intensity of growling sound increases as the lion opens its mouth. In another example of Figure A.9, ReWaS also produces temporally synchronized sounds with mouth movements, underscoring its overall effectiveness.

General text prompt Figure A.10 provides an example that evaluates the capability of ReWaS using more general text prompts. We generate audio samples with another generated video from

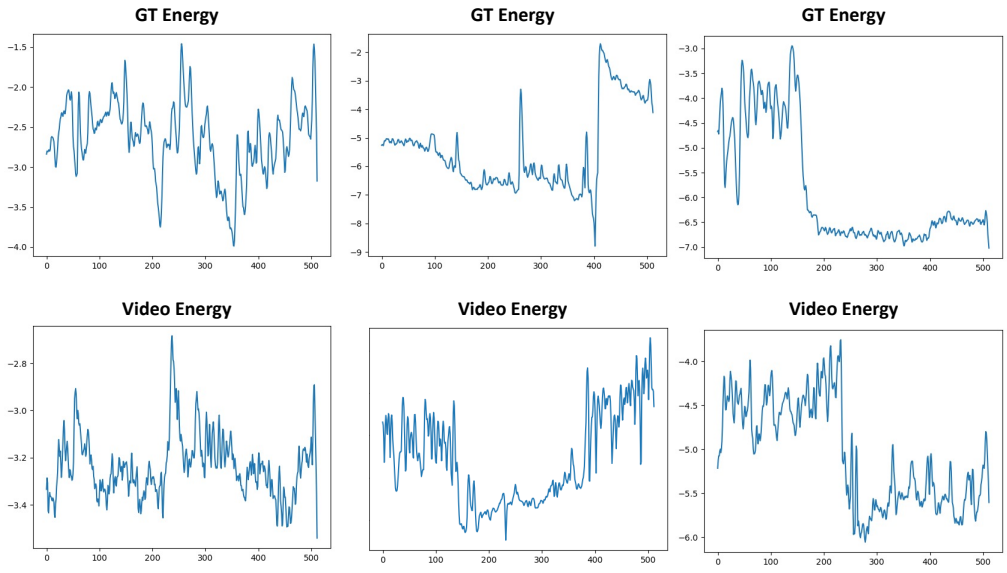


Figure A.4: Examples of energy controls from input videos.

KLING[5]. Our method is the only one that captures the increasing intensity of the sound as the onions are cut from the edge to the center. Both the T2A model, AudioLDM, and the V2T&V2A model, Seeing&Hearing, can generate corresponding sounds, but they lack visual temporal alignment in the generated results.

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

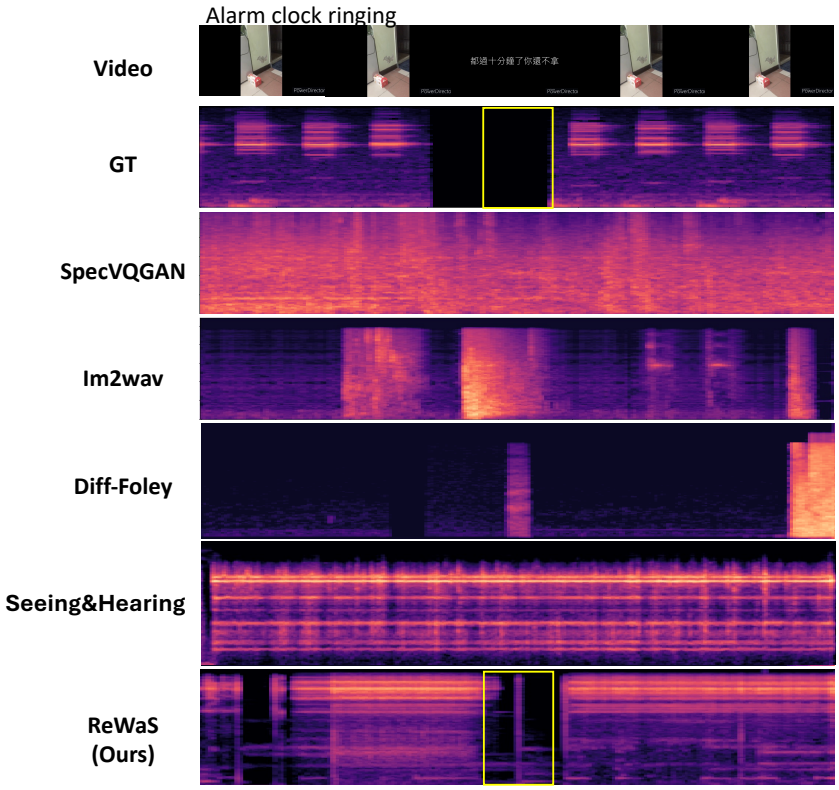


Figure A.5: Example of audio sound from misaligned visual input. ReWaS can make the desired sound and make the silent moment like ground-truth sound.

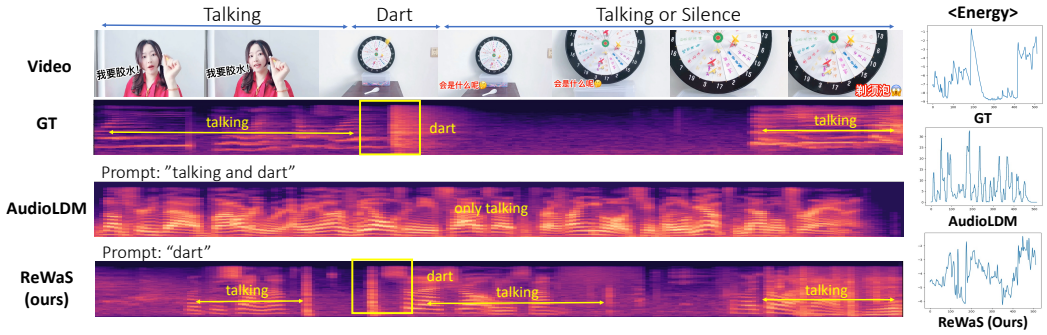


Figure A.6: Effectiveness of video input. In ReWaS, energy control from video input transfers additional temporal information.

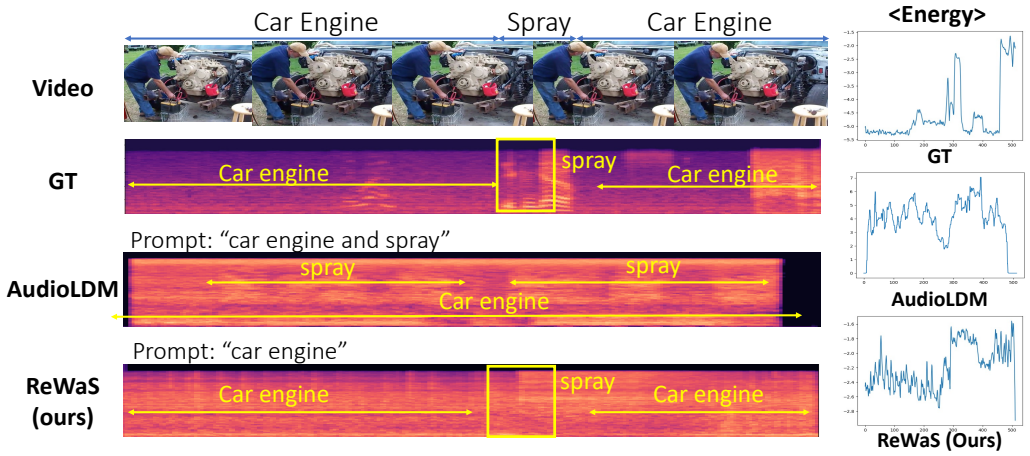


Figure A.7: Additional example of effectiveness of video input.

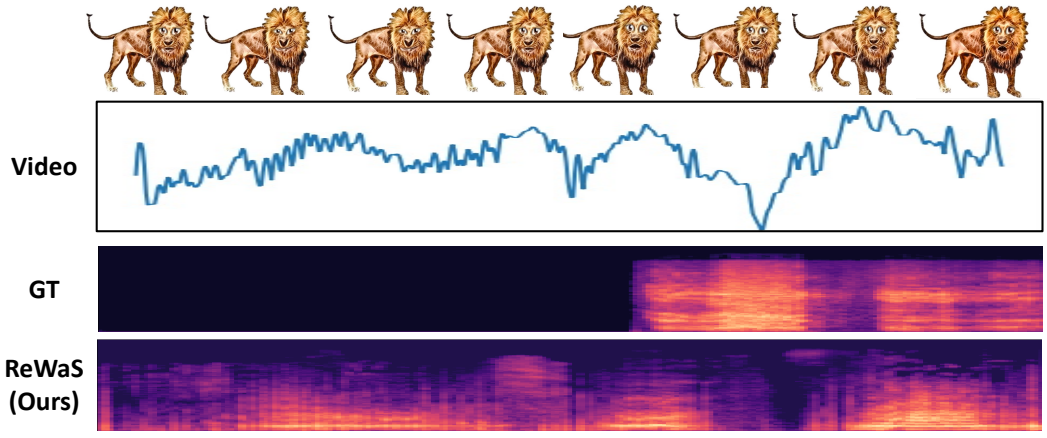


Figure A.8: Example of audio with improved synchronization, capturing small movements (e.g., a lion’s lip synchronization).

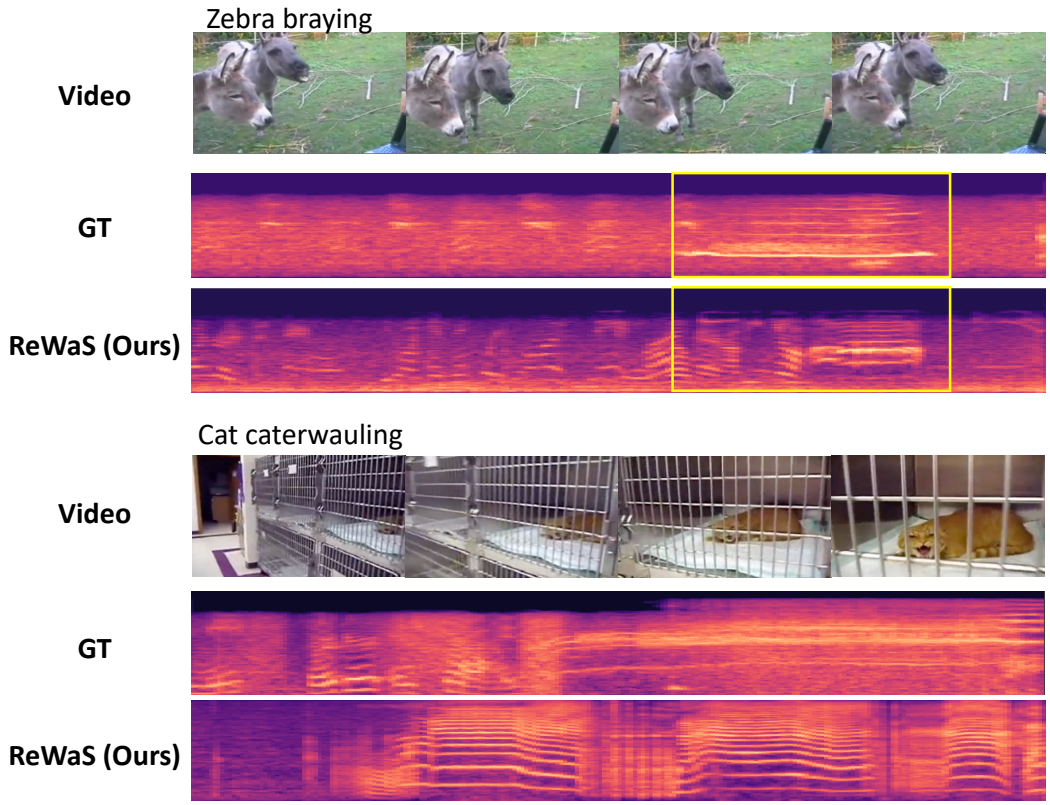


Figure A.9: Examples of generated audio sounds demonstrating the capability of temporal synchronization.

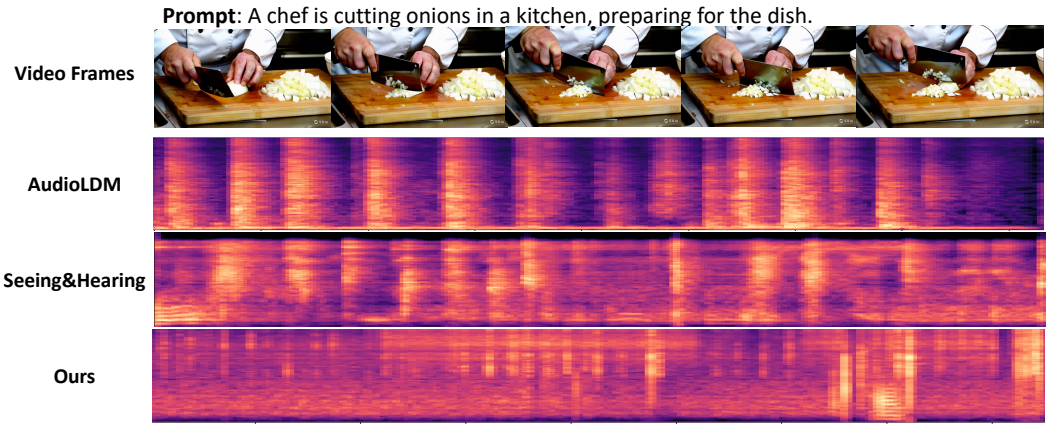


Figure A.10: Example of general user prompt.