Diverse and Aligned Audio-to-Video Generation via Text-to-Video Model Adaptation

Guy Yariv [1 2] **Itai Gat** [3] **Sagie Benaim** [1] **Lior Wolf** [4] **Idan Schwartz** [2 4 *] **Yossi Adi** [1 *]

Abstract

We consider the task of generating diverse and realistic videos guided by natural audio samples from a wide variety of semantic classes. For this task, the videos are required to be aligned both globally and temporally with the input audio: globally, the input audio is semantically associated with the entire output video, and temporally, each segment of the input audio is associated with a corresponding segment of that video. We utilize an existing text-conditioned video generation model and a pre-trained audio encoder model. The proposed method is based on a lightweight adaptor network, which learns to map the audio-based representation to the input representation expected by the text-to-video generation model. As such, it also enables video generation conditioned on text and audio and, for the first time, on both text and audio. We extensively validate our method on three datasets demonstrating significant semantic diversity of audio-video samples. We further propose a novel evaluation metric (AV-Align) to assess the alignment of generated videos with input audio samples. AV-Align is based on detecting and comparing energy peaks in both modalities. Compared to recent state-of-the-art approaches, our method generates videos that are better aligned with the input sound, both for the content and temporal axis. We also show that videos produced by our method present higher visual quality and are more diverse.

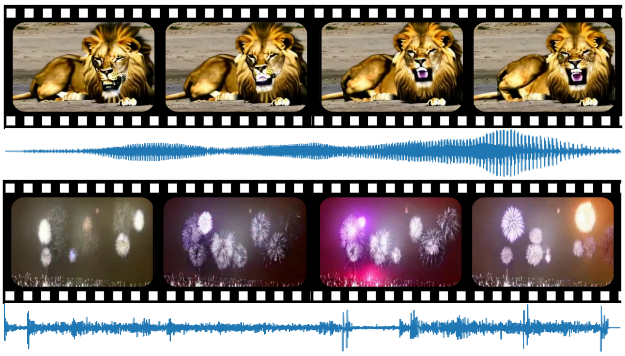

Figure 1. Generated video frames (above) and input audio signal (below the frames) employing our technique. The input to our model is an audio recording from which a representation is extracted. This representation maintains crucial temporal attributes and is then mapped into a text-based latent space representation incorporating both local and global audio context. Subsequently, this latent representation is fed into a pre-trained text-to-video diffusion generative model, ensuring the synchronized generation of video which is closely aligned with the input audio.

1. Introduction

Neural generative models have changed the way we create and consume digital content. From generating high-quality images and videos (Ho et al., 2020; Rombach et al., 2022), speech and audio (Wang et al., 2023a; Sheffer & Adi, 2023; Copet et al., 2023; Kreuk et al., 2022; Hassid et al., 2023), through generating long textual spans (Touvron et al., 2023a;b; Brown et al., 2020), these models have shown impressive results.

In the context of video generation, progress has been more elusive, with recent work making progress in generating short videos conditioned on text (Singer et al., 2022; Ho et al., 2022). Although audio is tightly connected to videos (e.g., providing important cues for motion in a scene), most of the prior work did not consider audio in the generation process. For instance, the action of 'playing drums' or the 'motion of waves' can be distinctively associated with a naturally occurring sound. Moreover, audio is comprised of structural components such as pitch and envelope that provide important cues for the type of scene and motion depicted.

We tackle the problem of generating diverse and realistic

[*]Equal contribution [1]The Hebrew University of Jerusalem [2]NetApp [3]Technion [4]Tel-Aviv University. Correspondence to: Guy Yariv <guy.yariv@mail.huji.ac.il>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

videos guided by natural audio samples. Our generated videos capture diverse and real-life settings from various semantic classes and are aligned globally and temporally with the input audio. Globally, the input audio is semantically associated with the entire output video, and temporally, each segment of the input audio is associated with a corresponding segment of that video. An example generation video can be seen in Figure 1.

Prior work on audio-guided video generation was mainly focused on either global information in the videos (i.e., capturing the semantic class) or specific scenes (e.g., speech). (Mama et al., 2021; Park et al., 2022; Kumar et al., 2020) generate talking heads conditioned on speech, but these are limited to videos of human faces and are conditioned on speech and not natural audio. More closely related to our setting, given an input video and an audio sample, Chatterjee & Cherian (2020) generate a continuation of the video that is aligned with the audio. Our method, however, generates videos from audio-only. Ge et al. (2022) proposed a method for generating aligned videos conditioned on audio. While impressive, generated videos are highly limited in diversity. Other works such as Chen et al. (2017); Hao et al. (2022); Ruan et al. (2023) generate videos that are globally aligned to the semantic class of the input audio sample (e.g., dancing, drums, etc.) but are unable to generate videos in which every segment is temporally aligned to each segment in the input audio sample.

In contrast to the above methods, our approach enables the generation of diverse and realistic videos associated and aligned with the input audio from a wide variety of semantic classes. Our work utilizes a pre-trained text-conditioned video generation engine and converts the input audio to a sequence of pseudo tokens. Given an input audio sample, we first encode it using an audio encoder, producing a latent representation of the audio signal. To capture local-to-global information, we construct the representation considering the i-th segment and neighboring segments. In particular, we use windows of varying sizes and average the embeddings corresponding to audio segments in these windows. Next, to produce the N-th video frame, we divide the audio embedding into N consecutive segments. We then train an adapter network to map these segments to a set of pseudo-tokens. Lastly, to produce the corresponding video, we feed the output of the audio mapping module into the pre-trained text-to-video generation model.

Intuitively, we learn a mapping between the audio representation obtained by the pre-trained audio encoder, to the textual tokens' representation used for conditioning the pre-trained text-to-video model. By that, video conditioning can be extended to audio tokens. To validate our approach, we consider a number of datasets that exhibit a diverse set of videos and input audio samples. We consider the Landscape dataset (Lee et al., 2022), which captures landscape videos. The AudioSet-Drums dataset (Gemmeke et al., 2017) which captures drums videos, and the VGGSound dataset (Chen et al., 2020) which consists of a diverse set of real-world videos from 309 different semantic classes.

We compare our method to state-of-the-art approaches, both in terms of objective evaluation and human study. We evaluate the audio-video alignment as well as video quality and diversity. To capture temporal alignment, we devise a new metric based on detecting energy peaks in both modalities separately and measuring their alignment. Further, we provide an ablation study where we consider alternative approaches to condition the video model.

Our contributions: (i) A state-of-the-art audio-to-video generation model that captures diverse and naturally occurring real-life settings from a wide variety of input videos of different semantic classes; (ii) We present a method that is based on a lightweight adapter, which learns to map audio-based tokens to pseudo-text tokens. As such, it also allows video generation conditioned on text, audio, or both text and audio. As far as we are aware, our method is the first to enable video generation conditioned both on audio and text; and (iii) Our method can generate natural videos aligned with the input sound, both globally and temporally. To validate this, we present a novel evaluation function to measure audio-video alignment. Since, as far as we can ascertain, we are the first to generate diverse and natural videos guided by audio inputs, such an evaluation function is critical to making progress in the field.

2. Related Work

Audio-to-image generation. Text-to-image generation has seen great advances recently, using either autoregressive methods (Ramesh et al., 2021; Gafni et al., 2022; Yu et al., 2022) or diffusion based models (Nichol et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Rombach et al., 2022; Ramesh et al., 2022; Rombach et al., 2022). This inspired a new line of work concerning audio-to-image generation. Żelaszczyk & Mańdziuk (2022); Wan et al. (2019) proposed to generate images based on audio recordings using a GANŻelaszczyk & Mańdziuk (2022). Żelaszczyk & Mańdziuk (2022) present results for generating MNIST digits only and did not generalize to general audio sounds, while Wan et al. (2019) generate images from general audio. In Wav2Clip Wu et al. (2022b), the authors learn a Contrastive Language-Image Pre-Training (CLIP) (Radford et al., 2021) like a model for learning joint representation for audio-image pairs. Later on, such representation can be used to generate images using VQ-GAN (Esser et al., 2021) under the VQ-GAN CLIP (Crowson et al., 2022) framework. The most relevant related work to ours is AudioToken (Yariv et al., 2023), in which the authors learn an audio token while

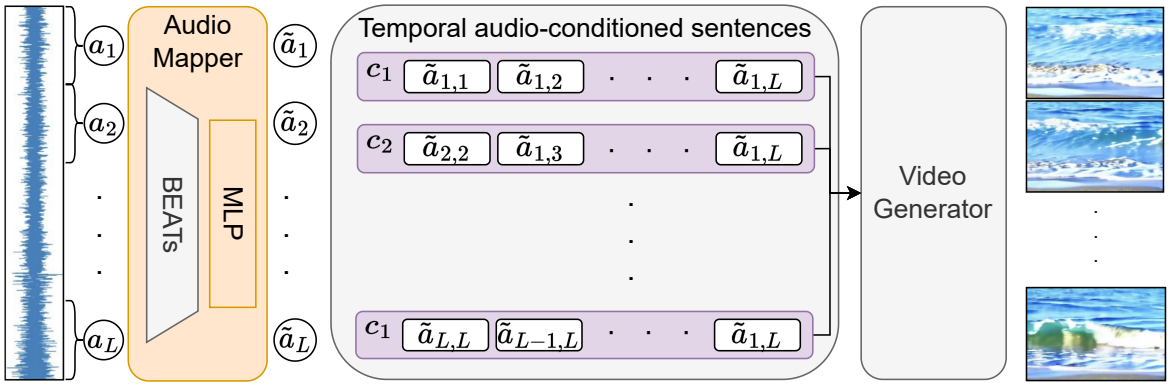

Figure 2. An illustration of the proposed model architecture and method. The input audio is first passed through a pre-trained audio encoder model (BEATs). Then, the resulting representations are fed into a trainable MLP layer, establishing a mapping between audio and text tokens. These text-based representations are then used to condition each frame via a temporal audio-conditioned sequence. This sequence effectively takes into account both local and global audio segments. Furthermore, an attentive token ($\tilde{a}_{\text{atten}}$) is included to learn the identification of significant audio signals using a pooling attention layer. Lastly, the conditioned components are utilized to generate frames through a pre-trained video generator. Notably, optimization is only applied to the MLP within the AudioMapper model and the pooling attention module.

adapting a diffusion-based text-to-image model to generate images using audio inputs.

Text-to-video generation. Early attempts to establish a connection between text and video relied on conditioned retrieval methods (Ali et al., 2022). Later, Wu et al. (2021) introduces the novel integration of 2D VQVAE and sparse attention in a text-to-video generation, facilitating the generation of highly realistic scenes. Wu et al. (2022a) extends GODIVA and presents a unified representation for various generation tasks in a multitask learning scheme. Later on, CogVideo (Hong et al., 2022) is built on top of a frozen text-to-image model by adding additional temporal attention modules. Singer et al. (2022) further improves generation quality following a similar modeling paradigm. Video Diffusion Models (He et al., 2022) uses a space-time factorized U-Net with joint image and video data training. Other approaches, such as Villegas et al. (2022) and Villegas et al. (2022) and (Yu et al., 2023) proposed transformer-based approaches to generate long videos or for multi-task-learning. The most relevant prior work to ours is Wang et al. (2023b), which proposed ModelScope. ModelScope is a latent diffusion-based text-to-video generation model with spatiotemporal blocks. By that, ModelScope enables consistent frame generation and smooth movement transitions.

Audio-to-video generation models can be roughly divided into two: (i) speech-to-video generation (talking heads); and (ii) general audio-to-video. Under the speech-to-video generation, Mama et al. (2021) proposed learning a discrete latent representation of the video signal using VQ-VAE, which will be later modeled via an auto-encoder conditioned on speech spectrogram. Park et al. (2022) generates talking face focusing a piece of phonetic information via *Audio-Lip Memory* module, while (Kumar et al., 2020) proposed a one-shot approach for fast speaker adaptation.

When considering general audio-to-video generation, Chatterjee & Cherian (2020) first proposed a method of generating aligned videos conditioned on both audio and video prompts. Ge et al. (2022) introduced a transformer-based approach for generating videos conditioned on either audio or textual features. Although providing impressive generations, their videos are not diverse and were demonstrated on drum generation only. Chen et al. (2017) suggest using separate frameworks for audio-to-image and image-to-audio generation. Hao et al. (2022) also suggest modeling both audio-to-image and image-to-audio using bidirectional transformers, however, using a unified framework. The authors prove it is better than two separate ones. Lastly, Ruan et al. (2023), follows the same modeling paradigm, however, using latent diffusion models.

3. Method

The proposed method is composed of three main components: (i) an AudioMapper, (ii) multiple audio-conditioned temporal sequences, and (iii) a text-to-video generation module. As our goal in this study is to enrich video generation models using audio inputs, we leverage a pre-trained diffusion-based text-to-video model and augment it with audio conditioning capabilities. A visual description of the proposed method can be seen in Figure 2.

In contrast to converting audio to image, transforming audio to video presents two additional challenges: (i) ensuring

the creation of coherent frames and (ii) synchronization between the audio and video components. For example, consider the scenario of having an audio recording of a dog barking. In the resulting video, it is crucial not only for the dog's appearance to remain consistent across all frames but also for the match between the timing of the barking sound and the dog's motion. In this work, we focus on item (ii) by temporally conditioning the generation of each of the video frames by a contextualized representation of the input audio.

Formally, we are interested in the generation of a video, denoted as $v = (v^{(1)}, \ldots, v^{(L)})$, where $v^{(i)} \in \mathbb{R}^{3\times H\times W}$ is an output frame, driven by a corresponding audio condition $a = (a_1, \ldots, a_R)$, where $a_i \in [-1, 1]$ is an audio sample at a given sampling rate in the time domain. We seek to establish a conditional probabilistic model, $p_\theta(v|a)$, encompassing the entire frame-set, where each frame $v^{(i)}$ is conditioned on a, which denotes the audio condition.

Note that the conditioning of each frame considers the entire audio input but is built differently for each frame. More details can be found in the paragraph on Audio-conditioned temporal sequence.

AudioMapper maps the audio representation obtained from a pre-trained audio encoder to pseudo-tokens compatible with the pre-trained text-to-video model. We denote the output of the AudioMapper as TEMPOTOKENS.

Formally, the model gets as input embedded audio, which originates from a pre-trained audio encoder $h : [-1, 1]^R \to \mathbb{R}^{R'\times H\times d}$, where H is the number of layers the representation is collected from, d is the inner dimension of the encoder, and R' is the segment length that h operates on. To force both audio and video latent representations to have the same dimension, we fix $R' = L$ by employing a pooling layer. Specifically, we use the BEATs model (Chen et al., 2022) as the audio encoder h. Different layers encapsulate a range of specificity levels. Representations derived from BEATs' final layers are strongly tied to class-related attributes, whereas earlier layers encompass low-level audio features (Gat et al., 2022; Adi et al., 2019). We embed an audio segment into a token representation using a non-linear neural network $g : \mathbb{R}^{L\times H\times d} \to \mathbb{R}^{L\times H\times d_t}$:

$$\tilde{a}^{(i)} = g\left(h(a)^{(i)}\right), \tag{1}$$

where $\tilde{a}^{(i)} \in \mathbb{R}^{L\times H\times d_t}$, and d_t is the embedding dimension of the text-conditioned tokens of the video generation process. The network g consists of four sequential linear layers with GELU non-linearity between them. We denote $\tilde{a}^{(i)}$ as TEMPOTOKENS. Subsequently, we generate a temporal conditioning sequence for each video frame using TEMPOTOKENS. We provide a detailed description of the process in the following paragraph.

Audio-conditioned temporal sequence. Next, to better capture the local context around each video frame, we apply an expanding *context window* technique over the obtained TEMPOTOKENS. This approach captures the surrounding sound signals of the i-th frame as follows:

$$c^{(i)} = \left(\tilde{a}_{\max(1,i-j),\min(i+j,K)} \mid j = 2^k\right)_{k=0}^{\log K}, \tag{2}$$

where

$$\tilde{a}_{l,r} = \frac{1}{r-l}\sum_{s=l}^{r}\tilde{a}^{(s)}. \tag{3}$$

This context window expands exponentially with increasing temporal distance from the target position, facilitating consideration of a wider local-to-global audio context range. The exponential expansion effectively balances local and global contexts, encompassing important distant audio components that can provide valuable insights into the audio class and close temporal changes needed for audio-video alignment. Figure 3 visually describes the audio-conditioned temporal sequence. Finally, we consider a context window that encompasses all audio signals. We substitute average operation with a trainable attentive pooling layer (Schwartz et al., 2019). Thus,

$$\tilde{a}_{\text{atten}} = \sum_{u=1}^{L} p(u)\tilde{a}^{(u)}, \tag{4}$$

where $p(u)\geq 0 \ \forall u$ is a probability distribution (i.e., $\sum_{u=1}^{L} p(u) = 1$) over the audio components. The probability distribution takes the form:

$$p(u) \propto \exp\left(\alpha_l\theta_l(u) + \alpha_c\theta_c(u)\right). \tag{5}$$

The local potential is $\theta_l(u) = v_l^\top \text{relu}(V_l a_u)$, and the cross potential between the audio components is:

$$\theta_c(u) = \sum_{i=1}^{L}\left(\left(\frac{W_1\tilde{a}^{(u)}}{||W_1\tilde{a}^{(u)}||}\right)^\top\left(\frac{W_2\tilde{a}^{(i)}}{||W_2\tilde{a}^{(i)}||}\right)\right). \tag{6}$$

The trainable parameters are (i) V_l, W_1, W_2, which re-embed the data to tune the attention, (ii) $v_l \in \mathbb{R}^{(L\cdot H\cdot d_t)\times 1}$ that scores the sound component (iii) α_l, α_c that calibrates the local and cross potentials. The attention mechanism enables learning the significance of the audio components.

Text-to-video. Lastly, we leverage a pre-trained latent diffusion text-to-video model to learn the aforementioned temporal audio tokens, $c = \{c^{(i)}\}_{i=1}^{L}$.

Diffusion models are a family of generative models designed to learn the data distribution $p(x)$. This is done by learning the reverse Markov process of length T. Given a timestamp $t \in [0, 1]$, the denoising function $\epsilon_\theta : \mathbb{R}^d \to \mathbb{R}^d$ learns to predict a clean version of the perturbed x_t from the training

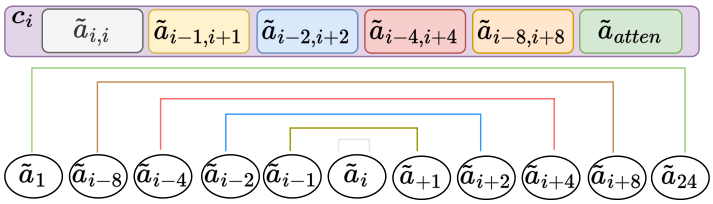

Figure 3. Illustration of the audio-conditioned temporal sequence for the case of 24 audio components. For the i-th frame, the window sizes grow exponentially, considering local audio details to aid in aligning audio and video, as well as the broader global information that enhances the differentiation of video classes. Additionally, we introduce a token that encompasses all audio components and identifies the significant ones through an attention pooling layer ($\tilde{a}_{\text{atten}}$).

distribution. The generative process can be conditioned on a given input, i.e., modeling $p(x|y)$ where y is a condition vector. In that case, the objective function is $\mathcal{L}_{\text{CLDM}} \triangleq$,

$$\mathbb{E}_{(v,a)\sim S, t\sim U(0,1), \epsilon\sim\mathcal{N}(0,I)} \left[\left\| \epsilon - \epsilon_\theta(f\left(v_t, c\right), t) \right\|_2^2 \right], \quad (7)$$

where each video frame, $v^{(i)}$, is conditioned on a dedicated condition vector $c^{(i)}$.

Specifically, in this work, we set ϵ_θ to be a state-of-the-art text-to-video model, ModelScope, which is comprised of a 3D-UNet integrated with a temporal attention layer as outlined in Wang et al. (2023b). ModelScope was trained on ∼10M text-video pairs and ∼2B text-image pairs (Wang et al., 2023b). Notice that the proposed framework is not limited to ModelScope and can be used for any differentiable text-to-video model.

Model optimization. We optimize the AudioMapper and the attentive pooling layer only and backpropagate gradients through ϵ_θ while keeping its parameters unchanged. Optimization minimizes the loss $\mathcal{L}_{\text{CLDM}}$ for reconstructing a frame $v^{(i)}$ conditioned on $c^{(i)}$ (see Equation (7)), with an added weight decay regularization for the encoded TEMPO-TOKENS. Overall, we optimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{CLDM}} + \frac{\lambda_{l_1}}{L} \sum_{u=1}^{\log L} \tilde{a}^{(u)}, \quad (8)$$

where λ_{l_1} is a trade-off hyper-parameter between the loss term and the regularization.

4. Evaluation Metrics

We evaluate our method on three main axes: video quality and diversity, audio-video alignment, and a human study.

Video quality and diversity. We report standard evaluation metrics in the domain of video generation for assessing quality and diversity. We utilize the following metrics: (i) Frechet Video Distance (FVD) metric, which quantifies the visual disparity between feature embeddings extracted from generated and reference videos (Unterthiner et al., 2019) and is used to assess quality and diversity; (ii) Inception Score (IS), which is computed with a trained C3D model (Tran et al. (2015)) on UCF-101 (Soomro et al., 2012) and assesses video quality.

Audio-video alignment. We distinguish between two types of audio-video alignment: (i) Semantic (or global) alignment, in which the semantic class (e.g., playing drums) of the input audio is depicted by the output video (e.g., a video of people playing drums). To this end, we consider the CLIP Similarity (CLIPSIM) metric (Wu et al., 2021), which gauges the alignment between generated video content and its corresponding audio label; (ii) Temporal alignment, in which we consider if each input audio segment is synchronized with its corresponding generated video segment. To measure this type of alignment, we introduce a novel evaluation metric.

The new metric is based on detecting energy peaks in both modalities separately and measuring their alignment. The premise behind this metric is that fast temporal energy changes in the audio signal often correspond to an object movement producing this sound. For instance, consider an audio waveform of fireworks. A successful audio-video temporal alignment would ensure that the video frames portraying the fireworks exhibit a noticeable change synchronously. Conversely, when the video exhibits a significant change, a corresponding peak should be observed in the audio waveform at that precise moment.

Our audio-video alignment metric operates as follows. We first detect candidate alignment points by considering each modality separately. We detect audio peaks using an Onset Detection algorithm (Böck & Widmer, 2013), pinpointing instances of heightened auditory intensity. To detect the changes within the video, we calculate the mean of the Optical Flow (Horn & Schunck, 1981) magnitude for each frame and identify rapid changes over time. Then, for each peak in one modality, we validate whether a pick was also detected in the other modality within a three-frame temporal window and vice-versa.

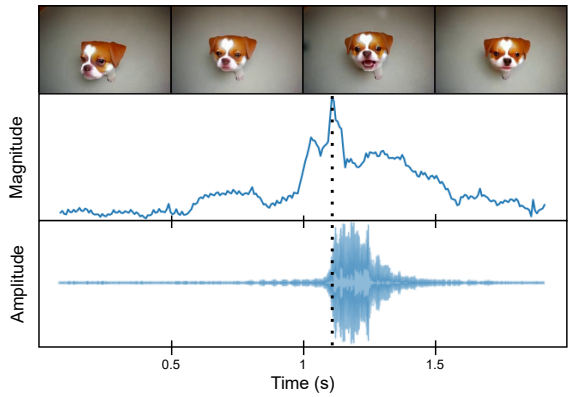

Figure 4. Audio-Video alignment metric illustration. The first row presents four frames from a generated video featuring a dog. The second row depicts the mean magnitude of optical flow for each frame, capturing video changes. The bottom row shows the amplitude of the audio waveform. The vertical line in the middle and the bottom graphs marks the onset of the waveform, while the peak of video change is also indicated.

Finally, we normalize by the number of peaks to derive the alignment score ranging between zero and one. Such a metric reflects the model's proficiency in synchronizing audio and video. More formally, given $\mathcal{A}$ and $\mathcal{V}$, audio and video peaks were obtained from the onset detection algorithms and optical flow, respectively. The alignment score is defined as:

$$\text{AV-Align} = \frac{1}{2|\mathcal{A} \cup \mathcal{V}|} \left(\sum_{a \in \mathcal{A}} \mathbf{1}[a \in \mathcal{V}] + \sum_{v \in \mathcal{V}} \mathbf{1}[v \in \mathcal{A}] \right), \quad (9)$$

where we consider a valid peak if placed within a window of three frames in the other modality. The above metric can be interpreted as the Intersection-over-Union metric.

To facilitate comprehension, Figure 4 illustrates the alignment process visually, depicting audio peaks and corresponding video changes, emphasizing the interplay between the auditory and visual domains.

Human study. We perform Mean Opinion Scores (MOS) experiments considering both quality and audio-video alignment. In this setup, human raters are presented with several short video samples and are instructed to evaluate their quality and alignment on a scale between 1–5 with increments of 1.0. Specifically, we ask raters to evaluate the videos considering overall quality, global alignment to the audio file, and local alignment between the visual and sound of the video files. We evaluate 20 videos per method and enforce ten raters per sample. The full questionnaire we asked the raters can be found in the supplemental material.

Model	FVD (↓)	CLIPSIM (↑)	IS (↑)	AV-Align (↑)
	VGGSound			
ModelScope Text2Vid	**801**	**0.69**	**15.55**	0.27
ModelScope Random	1023	0.47	6.32	0.26
Ours	923	0.57	11.04	**0.35**
	AudioSet-Drums			
TATS	303	0.69	2.10	0.28
Ours	**299**	**0.70**	**2.78**	**0.61**
	Landscape			
MM-Diffusion	922	0.53	2.85	0.41
Ours	**784**	**0.57**	**4.49**	**0.54**

Table 1. Automatic video generation results. We report FVD, CLIPSIM, IS, and Alignment ('align') scores for both the proposed method (Ours) and the baselines. For a fair comparison, we compare our method to TATS (Ge et al., 2022) and to MM-Diffusion (Ruan et al., 2023) using the benchmarks reported by the authors in the original paper.

5. Experimental Setup

Implementation details. The proposed method contains ∼35M trainable parameters. We optimized the model using two A6000 GPUs for 10K iterations. We use AdamW optimizer with learning rate of 1e-05 using constant learning rate scheduler. Each batch comprises 8 videos with 24 frames per video, sampled randomly for one-second granularity. To enhance training efficiency and mitigate memory consumption, we integrated gradient checkpointing into the training process of the 3D U-net architecture. Code and pre-trained models will be publicly available upon acceptance.

Datasets. We utilize the VGGSound dataset (Chen et al., 2020), derived from YouTube videos containing ∼180K clips of 10 seconds duration, annotated across 309 classes. To enhance data quality, we filtered ∼60K videos in which audio-video alignment is weak. During this filtering procedure, we utilized a pre-trained audio classifier to categorize sound events present in each clip. Simultaneously, a pre-trained image classifier was employed to classify the middle frame of every video clip. We then computed the CLIP (Radford et al., 2021) score by comparing the predicted labels from both classifiers. Then, filtering is done by removing videos that do not pass a pre-defined threshold. Our exploration of alternative filtering criteria, focusing on frames with maximum similarity to text labels rather than uniformly choosing the middle frame, reveals minimal differences (approximately 0.01) in CLIPSIM (Wu et al., 2021), IS (Tran et al., 2015), and AV-Align scores, leading us to use the middle frame.

Additionally, to have a fair comparison with prior work, we experimented with two additional datasets. (i) The *Landscape* dataset (Lee et al., 2022), which contains 928 nature

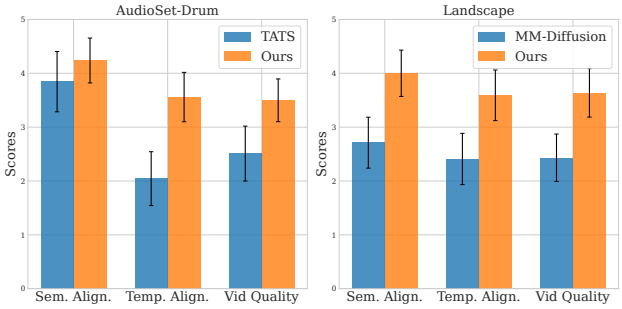

Figure 5. Human study. We consider the MOS score for three metrics: (i). *Semantic alignment*, where we ask users to rate how well the video matches the input audio semantic label, (ii). *Temporal alignment*, where we ask users to rate how well each input audio segment is aligned with the generated video segments, and (iii) *Video quality*, where we ask users to rate the generated video quality. On the LHS, we consider video models trained on AudioSet-Drum, and on the RHS, we consider video models trained on Landscape.

videos divided into 10-second clips, covering nine distinct scenes; (ii) The *AudioSet-Drum* dataset (Gemmeke et al., 2017), contains ∼7k videos of drumming. We used the same split as proposed by Ge et al. (2022), where ∼6k is used as the training set while the rest serves as a test set.

Baselines. We compare the proposed method to previous state-of-the-art models generating videos conditioned on audio inputs. Ge et al. (2022) proposed Time Sensitive Transformer (TATS) model, which projects audio latent embeddings onto video embeddings, enabling cross-modal alignment. Ruan et al. (2023) recently proposed MM-Diffusion, which employs coupled denoising auto-encoders to generate joint audio and video content. Each of the above-mentioned baselines, i.e., TATS and MM-Diffusion, were originally evaluated using different benchmarks, i.e., AudioSet-Drums and Landscape, respectively. For a fair comparison, we evaluate the proposed method using each of the datasets suggested in the original papers.

Moreover, we consider two naive baselines based on text-to-video models. In the first one, we generate videos from text description and retrieve random audio from the training set which corresponds to the same class as the generated video, denoted as *ModelScope Text-To-Video*. For the second one, denoted as *ModelScope Random*, we generate videos unconditionally (i.e., without any specific textual conditions), and match it with a random audio segment. For both baselines, we use the pre-trained publicly available zeroscope-v2 model [1].

[1]we use the zeroscope-v2 576w as can be found in the following link: https://huggingface.co/cerspense/zeroscope_v2_576w

6. Results

We start by presenting results for audio-to-video generation considering both objective metrics presented above and human study. Next, we empirically demonstrate how the proposed method can be used to generate videos conditioned on both text and audio modalities, thus enhancing text-to-video generations. Lastly, we conducted an ablation study to understand better the effect of our audio conditioning technique on generation quality and alignment. Visual results are provided in the supplementary.

6.1. Audio-to-Video Generation

Objective evaluation. As can be seen in Tab. 1, our method outperforms the baselines on all metrics for the AudioSet-Drums and Landscape datasets. Specifically, our method improves the quality of the generated videos (FVD and IS scores) and the audio-video alignment (AV-Align and CLIP-SIM scores). As expected, the gap between the methods is larger when considering the alignment scores.

Notice the alignment scores changed significantly when considering different benchmarks. Sound events can also be produced by objects not seen in the video; this is especially noticeable in the VGGSound benchmark, in which the AV-Align score of the original videos is 0.51.

Next, we compare our method to the original ModelScope model, both text-condition (ModelScope Text2Vid) and unconditionally (ModelScope Random). As we do not modify the model, we consider the text-condition setup as a topline in terms of video quality metrics. Recall the audio in both models is retrieved from our training set, using either the video class for ModelScope Text2Vid or randomly ModelScope Random. As expected, our model outperforms ModelScope Random, considering all metrics. The ModelScope Text2Vid is superior to our model in terms of video quality. However, when considering audio-video alignment, our method is significantly better.

Human study. We present results using a human study considering both video quality and alignment (both semantic and temporal). Results are depicted in Figure 5. As can be seen for both the AudioSet-Drum and Landscape datasets, users found our videos significantly more temporally aligned. Our method improves semantic alignment on both TATS and MM-Diffusion, with a significant gap to MM-Diffusion on the Landscape dataset. Finally, on video quality, users found our videos significantly superior.

6.2. Joint Audio-Text to Video Generation

Combining text and audio to guide generation involves adding text tokens for conditioning. In Tab. 2, we show results using "A video of <class>" for text conditioning

Figure 6. Examples of added text tokens for altering the output video. We show results for fire and flowing water audio.

Cond.	FVD (↓)	CLIPSIM (↑)	IS (↑)	AV-Align (↑)
Text	**801**	**0.69**	**15.55**	0.27
Audio	923	0.57	11.04	0.35
Text+Audio	859	0.58	11.66	**0.36**

Table 2. Results of the proposed method using different modalities as conditioning. As model conditioning, we report results for Text, Audio, and Text+Audio modalities.

Cond.	FVD (↓)	CLIPSIM (↑)	IS (↑)	AV-Align (↑)
vec.	948	0.57	10.12	0.29
win. (1-res.)	998	0.56	9.22	**0.36**
win. (2-res.)	965	0.56	9.87	0.35
win. (3-res.)	972	0.56	10.01	0.34
win. (4-res.)	950	0.56	10.13	0.35
win. (5-res.)	**923**	**0.57**	**11.04**	0.35

Table 3. An ablation study exploring the different audio conditioning. We report FVD, CLIPSIM, IS, and Alignment scores on VGGSound (Chen et al., 2020) considering single-vector conditioning (vec.), time-dependent condition using one window size (win. (1-res.), and different windows of size k (win. (k-res.)).

and "A video of <TemporalAudioAtokens> <class>" for Text+Audio. Combining text and audio conditioning outperforms audio-only in all metrics, especially FVD. Text-only provides the highest video quality but lacks alignment.

In Fig. 6, we present how we merge text tokens to temporal audio tokens, which enables style manipulation. For example, we can depict the sound of a river flowing over the moon using the prompt "on the moon".

6.3. Ablation Study

Recall our method, which consists of using context windows of varying sizes to capture a local-to-global context of the input audio. In Tab. 3, we assess the effect of using different windows of size $K \in \{1, 2, 3, 4\}$ denoted as win. (K-res.). In practice, the window size is determined by $\log K$; we use K for readability. Using only the local context window ($K = 1$) results in a good alignment. As we increase the global context (i.e., increasing K), the video quality is improved while the alignment scores are comparable.

We additionally consider a single audio conditioning vector (vec) by averaging all the audio components. Despite high video quality scores, the absence of local temporal information results in a notably worse AV-Align score.

7. Limitations

Our method, using a pre-trained text-to-video model, involves adapting between text and audio tokens, posing challenges in mapping between their latent representations. Due to hardware limitations, our method generates relatively short video segments with temporal conditioning limited to 24 frames. Additionally, discrepancies can arise between visual and audio modalities, such as a video showing a dog in a car while the audio only features a radio playing. This limitation is not specific to our method but rather a general challenge in the domain.

8. Conclusion

We introduced a state-of-the-art audio-to-video generation model that generates diverse and realistic videos aligned to input audio samples. Leveraging a lightweight adapter for mapping between audio and text representations enables conditioning video generation on both audio and text for the first time. Our expanding context window technique captures local and global context, and we propose the AV-Align metric for assessing temporal alignment.

Future work aims to explore incorporating additional modalities, such as depth, images, or IMU, alongside audio and text for video generation.

References

Adi, Y., Zeghidour, N., Collobert, R., Usunier, N., Liptchinsky, V., and Synnaeve, G. To reverse the gradient or not: An empirical comparison of adversarial and multi-task learning in speech recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3742–3746. IEEE, 2019.

Ali, A., Schwartz, I., Hazan, T., and Wolf, L. Video and text matching with conditioned embeddings. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1565–1574, 2022.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023.

Böck, S. and Widmer, G. Maximum filter vibrato suppression for onset detection. In *DAFx-13*, 2013.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.

Chatterjee, M. and Cherian, A. Sound2sight: Generating visual dynamics from sound and context. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pp. 701–719. Springer, 2020.

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

Chen, L., Srivastava, S., Duan, Z., and Xu, C. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 349–357, 2017.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., and Wei, F. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *arXiv preprint arXiv:2306.05284*, 2023.

Crowson, K., Biderman, S., Kornis, D., Stander, D., Hallahan, E., Castricato, L., and Raff, E. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pp. 88–105. Springer, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022.

Gat, I., Lorberbom, G., Schwartz, I., and Hazan, T. Latent space explanation by intervention. In *AAAI*, 2022.

Ge, S., Hayes, T., Yang, H., Yin, X., Pang, G., Jacobs, D., Huang, J.-B., and Parikh, D. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Hao, W., Guan, H., and Zhang, Z. Vag: A uniform model for cross-modal visual-audio mutual generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Hassid, M., Remez, T., Nguyen, T. A., Gat, I., Conneau, A., Kreuk, F., Copet, J., Defossez, A., Synnaeve, G., Dupoux, E., et al. Textually pretrained speech language models. *arXiv preprint arXiv:2305.13009*, 2023.

He, Y., Yang, T., Zhang, Y., Shan, Y., and Chen, Q. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Horn, B. K. and Schunck, B. G. Determining optical flow. *Artificial Intelligence*, 17

(1):185–203, 1981. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(81)90024-2. URL https://www.sciencedirect.com/science/article/pii/0004370281900242.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

Kumar, N., Goel, S., Narang, A., and Hasan, M. Robust one shot audio to video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 770–771, 2020.

Lee, S. H., Oh, G., Byeon, W., Kim, C., Ryoo, W. J., Yoon, S. H., Cho, H., Bae, J., Kim, J., and Kim, S. Sound-guided semantic video generation. In *European Conference on Computer Vision*, pp. 34–50. Springer, 2022.

Mama, R., Tyndel, M. S., Kadhim, H., Clifford, C., and Thurairatnam, R. Nwt: towards natural audio-to-video generation with representation learning. *arXiv preprint arXiv:2106.04283*, 2021.

Nichol, A. Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804. PMLR, 2022.

Park, S. J., Kim, M., Hong, J., Choi, J., and Ro, Y. M. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2062–2070, 2022.

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

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruan, L., Ma, Y., Yang, H., He, H., Liu, B., Fu, J., Yuan, N. J., Jin, Q., and Guo, B. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Gontijo-Lopes, R., Ayan, B. K., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Schwartz, I., Yu, S., Hazan, T., and Schwing, A. G. Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2039–2048, 2019.

Sheffer, R. and Adi, Y. I hear your true colors: Image guided audio generation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. Learning spatiotemporal features with 3d convolutional networks, 2015.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Towards accurate generative models of video: A new metric challenges, 2019.

Villegas, R., Babaeizadeh, M., Kindermans, P.-J., Moraldo, H., Zhang, H., Saffar, M. T., Castro, S., Kunze, J., and Erhan, D. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Wan, C.-H., Chuang, S.-P., and Lee, H.-Y. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 496–500. IEEE, 2019.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report, 2023b.

Wu, C., Huang, L., Zhang, Q., Li, B., Ji, L., Yang, F., Sapiro, G., and Duan, N. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021.

Wu, C., Liang, J., Ji, L., Yang, F., Fang, Y., Jiang, D., and Duan, N. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European conference on computer vision*, pp. 720–736. Springer, 2022a.

Wu, H.-H., Seetharaman, P., Kumar, K., and Bello, J. P. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022b.

Yariv, G., Gat, I., Wolf, L., Adi, Y., and Schwartz, I. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. *arXiv preprint arXiv:2305.13050*, 2023.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A. G., Yang, M.-H., Hao, Y., Essa, I., et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Żelaszczyk, M. and Mańdziuk, J. Audio-to-image cross-modal generation. In *IJCNN*, 2022.

In this supplementary, we include multimedia results of videos given in the form of an attached webpage (please open index.html attached. All files are local and anonymized). We further explain concepts presented in the main paper and describe the ablations done. The supplementary comprises the following subsections:

1. Video Generation: Explanation of the underlying architecture for video generation
2. Design choices for the audio encoder.

A. Video Generation

We employ ModelScope as the backbone for video generation (Wang et al., 2023b). Using a latent video diffusion model, it generates a video v^{pr} from a text prompt p. Both the training video v^{gt} and generated video v^{pr} exist in the visual space, while the diffusion process and denoising UNet ϵ_θ operate in the latent space, facilitated by VQGAN (Esser et al., 2021)'s encoder $\mathcal{E}$ and decoder $\mathcal{D}$. The latent representation of a training video $v^{gt} = [f_1, \ldots, f_F]$ with F frames is obtained by encoding it with $\mathcal{E}$, resulting in Z_0^{gt}. During training, Z_0^{gt} evolves into Z_T^{gt} through the diffusion process, involving the addition of Gaussian noise $[\epsilon_1^{gt}, \ldots, \epsilon_T^{gt}]$ over T steps. In contrast, during inference, the UNet predicts noise for each step, ultimately generating Z_0^{pr} from an initial random noise Z_T^{pr}. The final video v^{pr} is reconstructed using the VQGAN decoder $\mathcal{D}$.

Text Conditioning: To ensure a robust alignment, it employs a pre-trained CLIP ViT-H/14 (Radford et al., 2021) text encoder to convert the prompt p into the text embedding $c \in \mathbb{R}^{N_p \times d_t}$, where N_p represents the maximum token length of the prompt, and d_t represents the dimension of the token embedding.

In our approach, we replace text tokens with audio-conditioned temporal tokens, enabling unique conditioning for each frame. Specifically, we have $c \in \mathbb{R}^{F \times N_p \times d_t}$.

Cond.	FVD (↓)	CLIPSIM (↑)	IS (↑)	AV-Align (↑)
2-layers MLP	1305	0.57	3.93	0.35
4-layers MLP	**1227**	**0.58**	**4.68**	**0.37**

Table 4. Performance evaluation on a subset of the VGGSound dataset across four distinct classes: playing electric guitar, playing drum kit, dog barking, and fireworks banging.

A.1. 3D Denoising UNet

The denoising UNet ϵ_θ is the central component of the latent video diffusion model, responsible for denoising the latent space from Z_T to Z_0 by predicting step-wise noise. It incorporates textual information from the prompt p through a text embedding c. During training, it minimizes the difference between predicted noise $\epsilon_{\hat{t}}^{pr}$ and ground-truth noise $\epsilon_{\hat{t}}^{gt}$ for each step $\hat{t}$, yielding the loss $\mathcal{L}$:

$$\mathcal{L} = \mathbb{E}_{Z_{\hat{t}}, \epsilon_{\hat{t}}^{gt} \sim \mathcal{N}(0,1), \hat{t}} \left[||\epsilon_{\hat{t}}^{gt} - \epsilon_{\hat{t}}^{pr}||_2^2 \right]. \tag{10}$$

A.1.1. Spatio-temporal Block:

The spatio-temporal block for capturing intricate spatial and temporal dependencies (Blattmann et al., 2023). To enhance video synthesis quality, it leverages spatio-temporal convolutions and attention. The spatiotemporal block integrates spatial convolution, temporal convolution, spatial attention, and temporal attention. Spatio-temporal convolutions involve spatial convolutions using a 3×3 kernel on each frame's $\frac{H}{8} \times \frac{W}{8}$ latent features and temporal convolutions on F frames. The spatio-temporal attention consists of spatial and temporal attention modules. Spatial attention operates on $\frac{HW}{64}$ latent spatial features, while temporal attention processes the temporal dimension with size F. Both attention mechanisms are implemented using the Transformer architecture.

B. Design choices for the audio encoder

In Tab. 4, we conducted an ablation study by varying the number of linear layers in our AudioMapper module. We found that using four layers yielded superior performance compared to employing only two layers. This indicates that a deeper linear architecture contributes to better model performance, likely due to its increased capacity for capturing complex patterns and representations in the data.