Engineering Applications of Artificial Intelligence 149 (2025) 110457

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

Semantically complex audio to video generation with audio source separation

Sieun Kim [a], Jaehwan Jeong [a], Sumin In [a], Seung Hyun Lee [a], Seungryong Kim [b], Saerom Kim [a], Wooyeol Baek [a], Sang Ho Yoon [c], Eugenio Culurciello [d], Sangpil Kim [a,*]

[a] Department of Artificial Intelligence, Korea University, Seoul, 02841, Republic of Korea
[b] Kim Jaechul Graduate School of Artificial Intelligence, KAIST, Daejeon 34141, Republic of Korea
[c] Graduate School of Culture Technology, KAIST, Daejeon 34141, Republic of Korea
[d] ECE, Purdue University, 610 Purdue Mall, West Lafayette, 7907, IN, USA

ARTICLE INFO

MSC:
C6260
C6264
C3370P
C1250
C1230
C6130M

Keywords:
Multi-source audio
Video generation
Deep learning
Artificial intelligence

ABSTRACT

Recent advancements in artificial intelligence for audio-to-video generation have shown the ability to generate high-quality videos from audio, particularly by focusing on temporal semantics and magnitude. However, existing works struggle to capture all semantics from audio, as real world audios often consist of mixed sources, making it challenging to generate semantically aligned videos. To solve this problem, we present a novel multi-source audio-to-video generation framework that incorporates decomposed multiple audio sources into video generative models. Specifically, our proposed Attention Mosaic directly maps each decomposed audio feature to the corresponding spatial attention feature. In addition, our condition injection module is helpful for producing more natural contexts with non-audible objects by leveraging the knowledge of existing generative models. Our experiments show that the proposed framework achieves state-of-the-art performance in representing both multi- and single-source audio-to-video generation methods.

1. Introduction

Recent generative models have shown remarkable progress with various modalities, such as text (Rombach et al., 2022; Chen et al., 2023; Oh et al., 2023; Zhao et al., 2023; Blattmann et al., 2023; Liu et al., 2024; Lee et al., 2024), mask (Voleti et al., 2022; Chang et al., 2023; Jain et al., 2024; Ma et al., 2023; Chen et al., 2024a), and audio (Chatterjee and Cherian, 2020; Biner et al., 2024; Lee et al., 2023a; Jeong et al., 2023; Yariv et al., 2024; Zhang et al., 2024; Xing et al., 2024). Text modality is a conventional user-provided condition type, but it still presents challenges in controlling motion dynamics solely based on the text input. Beyond that, audio has inherent features, such as intensity, timbre and volume, which enable the generation of temporally dynamic features with infinite variation. Using these characteristics, recent works (Lee et al., 2022, 2023b, 2022) have demonstrated the superiority of audio conditioning, emphasizing the potential of audio in the generation and manipulation of visual content. Traditionally, audio has been used for generating talking heads (Kumar et al., 2020; Park et al., 2022; Peng et al., 2024), where each frame is generated to synchronize the face and lip movements with the given speech audio. Due to the limitation of handling only specific classes of audio, works have attempted to extend into generating videos from more diverse classes of audio, i.e., general audio-to-video generation. Early works (Chatterjee and Cherian, 2020; Le Moing et al., 2021; Lee et al., 2022) utilized Generative Adversarial Networks (GAN) for generating frames, but they did not consider the temporal semantics of audio. To overcome this limitation, recent works (Lee et al., 2023a; Jeong et al., 2023; Yariv et al., 2024; Zhang et al., 2024) have leveraged time-aware semantic audio information as input condition in diffusion model with the magnitude information of the given audio. However, they do not accurately represent semantically complex audio, especially when sounds from multiple objects overlap. As shown in Fig. 1, since multiple audio sources are mixed in a real video scenario, it is necessary to decompose the mixed audio into individual components and produce visual elements corresponding to the individual audio sources.

In this paper, we propose Maestro, a novel method of multi-source audio-to-video generation, which aims to accurately produce multiple

* Corresponding author.
E-mail addresses: seeeun67@korea.ac.kr (S. Kim), jhwan@korea.ac.kr (J. Jeong), ism0705@korea.ac.kr (S. In), easter3163@korea.ac.kr (S.H. Lee), seungryong.kim@kaist.ac.kr (S. Kim), rlatofha3597@gmail.com (S. Kim), 100wooyeol@gmail.com (W. Baek), sangho@kaist.ac.kr (S.H. Yoon), euge@purdue.edu (E. Culurciello), spk7@korea.ac.kr (S. Kim).

https://doi.org/10.1016/j.engappai.2025.110457
Received 19 September 2024; Received in revised form 30 January 2025; Accepted 26 February 2025
Available online 13 March 2025

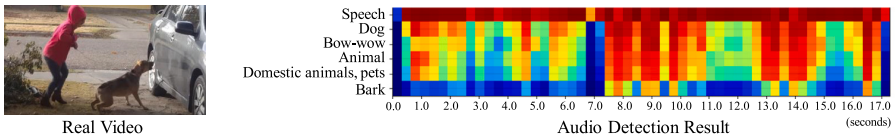

Fig. 1. Audio detection visualization of a 17 s real video. A real video contains a mixture of various audio meanings, which are represented correspondingly in the video.

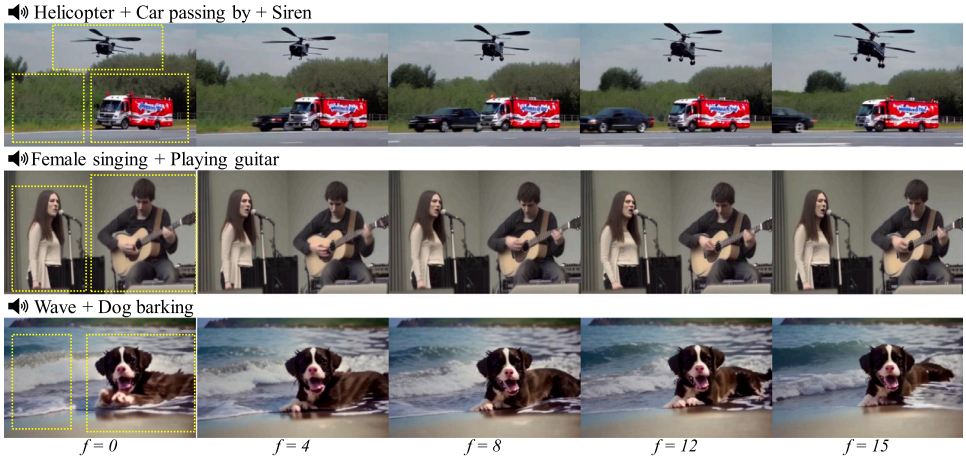

Fig. 2. Examples of our video generation based on multi-source audio inputs. Maestro creates impressive output that fully reflects semantically complex audio and naturally expresses various objects. The yellow boxes indicate the input bounding boxes.

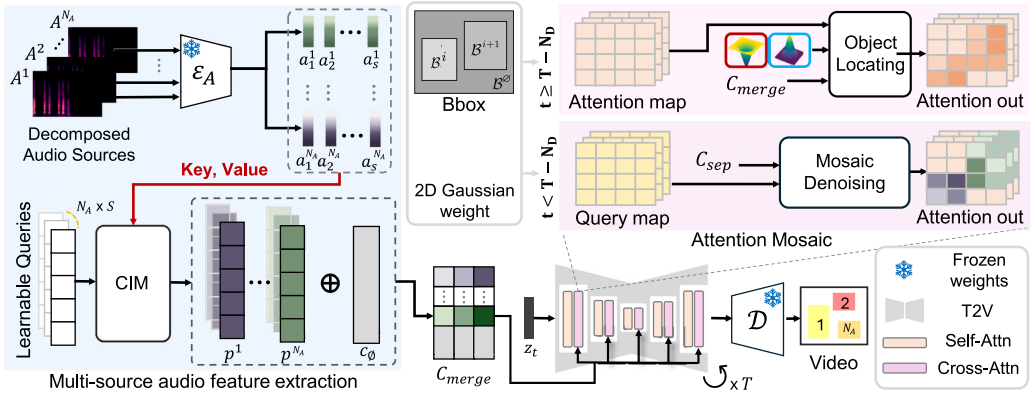

Fig. 3. Overview of Maestro. Our model consists of two main components: (i) Multi-source audio feature extraction, which produces time-dependent audio tokens for each segment of denoised audio by leveraging a pre-trained condition injection module and (ii) Attention Mosaic, directly mapping each audio feature to designated locations in the spatial attention module for generating multiple objects.

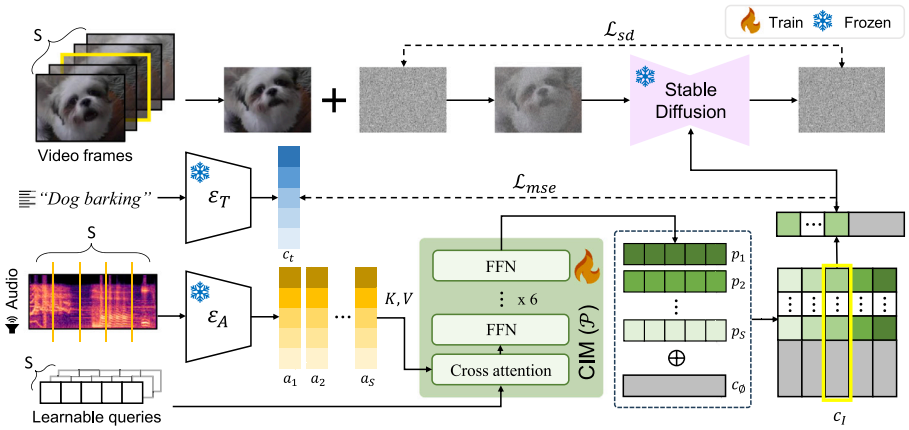

Fig. 4. Condition injection module training process. Maestro generates time-dependent audio tokens with Condition Injection Module (CIM). Audio input is divided into S segments, which are processed through the audio encoder and CIM to be injected into the video generation model. By employing Stable Diffusion, CIM trains the video frame information corresponding to the audio.

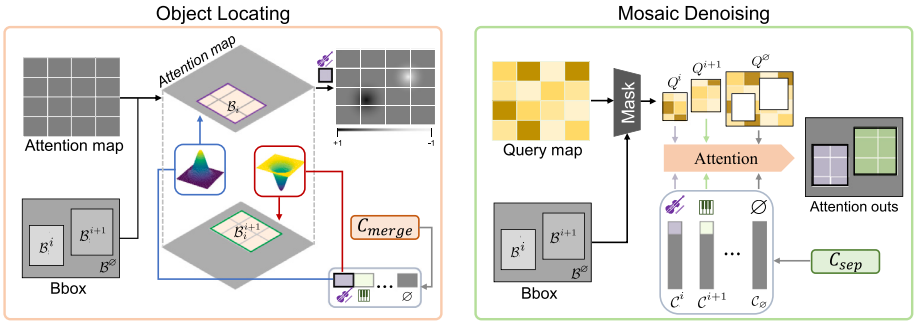

Fig. 5. Attention Mosaic. At timestep t, noise is predicted using either the Object Locating (Left) or Mosaic Denoising (Right) method. In the case of Object Locating, a 2D Gaussian is utilized in the *Attention map* dimension to create objects at the desired locations. Mosaic Denoising, on the other hand, ensures that each condition only attends to the corresponding bounding box B region to generate objects. Each method is applied with its respective edited condition, referred to as C_{merge} (Eq. (4)) and C_{sep} (Eq. (9)).

objects in the input semantically complex audio. Unlike previous approaches that focus on representing the semantics and transitions of single-source audio, Maestro tackles this challenge through Attention Mosaic with multiple decomposed audio sources and introduces an effective condition injection module optimization strategy to attain better contextual expression, even for non-audible objects within the video. Instead of performing a denoising step for each object, we adopt the "Mosaic Denoising" method, which directly maps multiple decomposed audio semantics to the desired space of video using spatial attention modules. Furthermore, to enhance contextual expression, we leverage Stable Diffusion (Rombach et al., 2022) to train the frame information of actual videos corresponding to the audio with our curated VGGSound and Landscape dataset (Lee et al., 2022). For example, in a video where a dog is barking, the audio label might simply state "dog barking". However, by utilizing video frames, we can gather information on whether the dog is barking indoors or outdoors in the yard, providing insights into the surrounding environment and context.

This paper is organized as follows: We first discuss the related work in Section 2. Then our approach is presented in Section 3, including the Condition Injection Module and Attention Mosaic. Experimental evaluation and analysis is conducted in Section 4 and conclusion are drawn in Section 5. In the Appendix, we include experiments on a different dataset, as well as a discussion of limitations, future works, and additional qualitative results. Experimental results show that our method can produce semantically rich video from various multi-source audio. For example, it accurately represents three objects, two people, and effectively captures foreground and background sound objects (see Fig. 2). Moreover, we examine the effectiveness of proposed our methods in ablation studies.

Our main contributions are listed as follows:

- We propose a novel approach that generates video by decomposing audio sources, addressing the limitation of existing audio-to-video generation methods that often fail to capture multiple audio semantics.
- Our Mosaic Denoising aligns multi-source semantics by directly mapping each audio feature to designated locations in the spatial attention module, enhancing the instance-level audio-visual alignment.
- To achieve better contextual expression with non-audible objects, beyond contrastive learning for audio encoder, we also applied pixel-level optimization by leveraging the knowledge of Stable Diffusion.
- Our extensive experimental results show that our proposed pipeline outperforms existing audio-to-video generation models

on various metrics, such as video quality, audio-visual alignment, and user study.

2. Related work

2.1. Text-to-video generation

Text-to-video (T2V) generation models (Ho et al., 2022b,a; Singer et al., 2022; Oh et al., 2023; Guo et al., 2023; Hong et al., 2022; Zhao et al., 2023; Blattmann et al., 2023; Liu et al., 2024; Lee et al., 2024) have made significant advancements in recent years, driven by the progress in diffusion models. Video Diffusion Models (VDM) (Ho et al., 2022b) was the first to apply diffusion models to video generation. It uses a spatio-temporally factorized U-Net architecture, which maintains temporal continuity and spatial consistency while generating videos. VDM requires high computational resources because it learns the entire video using 3D convolutions rather than processing each frame individually. Therefore, Imagen Video (Ho et al., 2022a) reduced computational complexity by processing each frame of the video data individually and improved overall quality based on a cascaded diffusion model.

Other approaches, such as Make-A-Video (Singer et al., 2022) and CogVideo (Hong et al., 2022), extended existing text-to-image (T2I) models to T2V models by adding a temporal layer for the video domain. They follow the fine-tuning process with video datasets while leveraging the knowledge of image generation priors. Subsequently, AnimateDiff (Guo et al., 2023) and VideoCrafter (Chen et al., 2023) improved both temporal consistency and visual quality by integrating temporal attention modules into the existing 2D latent diffusion model. Alternative approaches (Bar-Tal et al., 2024; Ma et al., 2024) proposed transformer-based diffusion models, resulting in notable enhancements in generation quality.

The most related prior work to ours is VideoCrafter2 (Chen et al., 2024b), an extended video generation model from Stable Diffusion (Rombach et al., 2022), capable of generating high-quality videos from either text prompts or image conditions. Our model preserves this desirable image quality per frame and temporal coherence between frames achieved in their work while taking audio as input.

2.2. Audio-to-video generation

In conditional video generation, text condition can convey semantic information well, but it does not effectively express situations that change over time. On the other hand, sound can effectively reflect dynamic changes. Therefore, recent research has been focusing on video generation using sound modalities. Early studies of audio-to-video (A2V) generation primarily focused on the talking face generation task to synchronize facial and lip movements in each frame (Kumar et al., 2020; Park et al., 2022; Peng et al., 2024). This approach has the limitation of being unable to utilize various audio types, such as those from nature or animals.

Subsequent research has focused on generating more diverse videos by utilizing semantic features from the given audio. Prior studies in A2V, such as Sound2Sight (Chatterjee and Cherian, 2020), CCVS (Le Moing et al., 2021), and Lee et al. (2022), used Generative Adversarial Networks (GAN) based approaches to generate the next frame sequences from audio input. However, these approaches focus on the semantic features of audio, largely neglecting its temporal features.

Recent approaches leverage temporal segments and utilize diffusion models to achieve high-quality temporal and visual results. AADiff (Lee et al., 2023a) controlled temporal dynamics by adjusting the weights of the text-image cross-attention map, using the audio magnitude of each frame as a condition. TPoS (Jeong et al., 2023) aligns audio segments with video frames and uses these segments as a condition for T2I models to generate each frame. TempoToken (Yariv et al., 2024) extracts features from audio segments and integrates them into the T2V model using adapters for effective fusion. ASVA (Zhang et al., 2024) used ImageBind (Girdhar et al., 2023) to encode the input audio into semantically aware, time-dependent tokens. Additionally, to improve the generation of audio-synchronized videos, audio cross attention and temporal attention layers were added to a pre-trained image latent diffusion model (LDM) for better audio conditioning and synchronization. However, conditioning semantically complex audio with multiple overlapping classes is still challenging. In this work, we focus on generating multiple objects from multiple decomposed audio inputs while also enhancing the representation of single-source audio.

3. Methods

As shown in Fig. 3, our model has two main components: (i) Multi-source audio feature extraction (see Section 3.1) and (ii) Attention Mosaic (see Section 3.2). Our multi-source audio feature extraction enables the model to express not only the time-dependent semantic information but also contextual information from audio inputs (e.g., given a sound of playing violin, our model represents not only the violin object but also the violinist and their surroundings). Conditioned on these audio features, Attention Mosaic allows for the generation of multiple objects by ensuring that the desired condition is attended to at the designated location during the process of video inference. In this paper, we propose a novel multi-source audio-to-video generation method.

3.1. Multi-source audio feature extraction

3.1.1. Feature extraction

We leverage AudioSep (Liu et al., 2023), a foundational model for open-domain sound separation, to decompose multi-source audio into N_A single-source audio. For $i \in \{1, \dots, N_A\}$, they are converted to Mel-spectrogram $A^i \in \mathbb{R}^{e \times w}$ where e represents the number of melfrequency bins and w is the width of the spectrogram and then divided into several segments to capture temporal conditions. Each time-dependent segment, denoted as $A^i_s \in \mathbb{R}^{e \times \frac{w}{S}}$, is encoded into ImageBind's unified latent space (Girdhar et al., 2023) where $s \in \{1, \dots, S\}$ and $S = \lceil \frac{L}{2} \rceil$, with L being the input audio duration. Specifically, while ImageBind's audio encoder $\mathcal{E}_A(\cdot)$ usually extracts only the classification token for audio embedding, we leverage both the classification token and local patch tokens, i.e. $a^i_s = \mathcal{E}_A(A^i_s)$, to obtain richer audio information from audio segments.

3.1.2. Condition Injection Module (CIM)

To control video generation using audio, we connect the audio encoder $\mathcal{E}_A(\cdot)$ and video diffusion model with a condition injection module (CIM), so that the denoising UNet can interpret the audio information. Inspired by Flamingo (Alayrac et al., 2022), we use a learnable lightweight model $\mathcal{P}(\cdot)$ as our condition injection module. It produces a fixed number of audio outputs using extracted multi-source audio features a^i_s as inputs. We employ the query transformer architecture (Jaegle et al., 2021), which consists of N stacked layers of cross-attention and feed forward networks (FFN). In more details, during inference, $\mathcal{P}(\cdot)$ makes various time-dependent audio embeddings $a^i = (a^i_1, \dots, a^i_S)$ to a constant number of audio tokens $p^i = (p^i_1, \dots, p^i_S)$ using the N_Q learned latent queries where $p^i_s \in \mathbb{R}^{N_Q \times d}$ and $d \equiv 1024$.

3.1.3. Training condition injection module

As shown in Fig. 4, our training method of condition injection module $\mathcal{P}$ is based on a lightweight pre-trained T2I model, not the T2V model. Utilizing an image generation model particularly allows $\mathcal{P}$ to converge in less time and with fewer resources, enhancing better contextual expression of non-audible objects (see Fig. 13). To train better contextual expression, we extract S video frames from the video corresponding to S audio segments, starting from the first frame and then at intervals of n frames, where $n = \lceil \frac{F}{S-1} \rceil$ and F being the

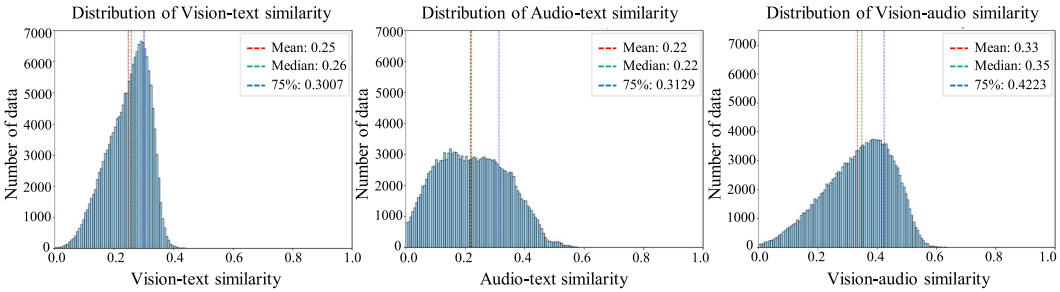

Fig. 6. Histogram visualization of semantic alignment across different modalities. Since semantically misaligned audio-video data pairs can negatively impact the training of the condition injection module, we compute cosine similarity score for Keyframe-Text (Left), Audio-Text (Middle), Keyframe-Audio (Right). We selected only the videos where all three scores were higher than the median.

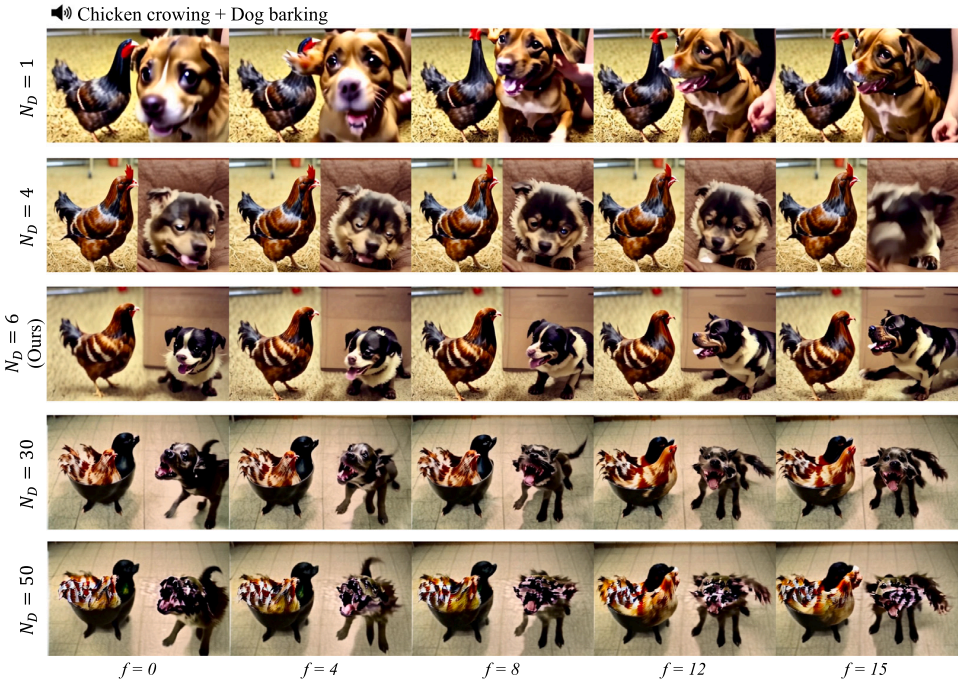

Fig. 7. Hyperparameter N_D. When the N_D value is low, the Object Locating (OL) step is not performed sufficiently, so the location of the object becomes unclear and the frame appears unnatural. On the other hand, when the N_D value increases, the Mosaic Denoising (MD) step will not be performed sufficiently, which will reduce the quality of the object and introduce artifacts. When the N_D value is 6, the OL and MD stages are properly balanced to ensure that objects are clearly positioned and create natural frames, resulting in optimal results.

total number of frames in the input video. Given S video frames and time-dependent audio embeddings $a_1, \ldots, a_S$, our CIM ($\mathcal{P}$) is trained to produce $p_1, \ldots, p_S$, applying denoising loss functions and Mean Squared Error (MSE) loss jointly. Specifically, for denoising loss, we leverage pre-trained Stable Diffusion UNet $\epsilon_{sd}(\cdot)$ (Rombach et al., 2022) and integrated conditioning c_I, which is combined time-dependent audio tokens $p_1, \ldots, p_S$ with unconditional null embeddings $\mathcal{E}_T(c_\varnothing) \in \mathbb{R}^{1\times|W|\times d}$. $c_\varnothing$ represents an unconditioned text prompt and $\mathcal{E}_T(\cdot)$ is pre-trained CLIP-based text encoder.

The denoising loss is thus formulated as:

$$\mathcal{L}_{sd} = \mathbb{E}_{\mathcal{E}(x),c_I,\epsilon\sim\mathcal{N}(0,I),t}\left[\|\epsilon - \epsilon_{sd}(z_t, t, c_I)\|_2^2\right] \tag{1}$$

A extracted frame x, corresponding to a audio token, is encoded into a lower-dimensional latent space $z = \mathcal{E}(x)$ using a pre-trained perceptual auto-encoder $\mathcal{E}(\cdot)$. The corrupted latent representation z_t at time step t is obtained by adding Gaussian noise to the latent representation z.

An MSE loss is defined as the squared norm difference between the audio tokens and the text tokens c_t from dataset's label:

$$\mathcal{L}_{mse} = \|p - c_t\|_2^2, \tag{2}$$

where $p = \mathcal{P}(a)$ and $a = (a_1, \ldots, a_S)$. We update only the parameters of the CIM ($\mathcal{P}$) while freezing Stable Diffusion. By leveraging the spatial prior knowledge from Stable Diffusion, we can effectively learn visual information that corresponds to frames matched with audio segments. In addition, we apply data augmentation to produce audio tokens that contain high-quality audio features. For audio inputs, we apply the SpecAugment (Park et al., 2019), augmenting Mel-spectrogram acoustic

Fig. 8. Qualitative results of loss function optimized parameter λ_1.

features by warping the features and masking blocks of frequency channels. To summarize, we minimize the following loss function $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \lambda_1(\mathcal{L}^A_{sd} + \mathcal{L}^{A'}_{sd}) + (1-\lambda_1)(\mathcal{L}^A_{mse} + \mathcal{L}^{A'}_{mse}), \tag{3}$$

where A' represents augmented audio A.

3.2. Attention mosaic

The part where the input condition affects video generation is the spatial cross-attention module within the denoising UNet of the generative model (Rombach et al., 2022). This module performs cross-attention using the query $Q \in \mathbb{R}^{N_F \times d_h \times d}$ from the self-attention located in the previous stage and the transformed key K, value $V \in \mathbb{R}^{N_F \times |W| \times d}$ from the input condition ($|W| \equiv 77$ and $d \equiv 1024$, which is because our baseline model used CLIP (Radford et al., 2021) as the text embedding model). While this existing module performs well for generation with a single condition, it struggles to represent multiple conditions effectively. To address this limitation, we divide the UNet denoiser into two stages (shown in Fig. 5). The first is *Object Locating*, a method for specifying areas to ensure that multiple objects are accurately placed within a single frame (see Section 3.2.1), and the next is *Mosaic Denoising*, a method for mapping each audio source to a designated area and applying focused conditioning within that area to enhance the fine details of multiple objects (see Section 3.2.2). The former occurs during the early steps $t \in \{T-1, \dots, T-N_D\}$, while the latter takes place in the later steps $t \in \{T-N_D-1, \dots, 0\}$, where T represents the total number of denoising time steps and N_D is a hyperparameter that separates these two processes. The parameter setting is detailed on Section 4 and Fig. 7.

To control the positions of objects in both stages, inspired by Ma et al. (2023) and Jain et al. (2024), we leverage bounding boxes (Bboxes) in the same number as the decomposed audio sources. We present the procedure of Object Locating and Mosaic Denoising in Alg. 1 to facilitate understanding, and the overall procedures are illustrated in Fig. 5.

Table 1
Quantitative results of parameter λ_1.

λ_1	Landscape		VGGSound	
	IA ↑	AV-Align ↑	IA ↑	AV-Align ↑
0.0	0.25	0.45	0.28	0.50
0.5	0.32	0.53	0.45	0.55
0.7	0.34	0.54	0.47	**0.58**
0.9	**0.37**	**0.55**	**0.48**	**0.58**
1.0	0.35	**0.55**	0.45	0.57

Table 2
Quantitative results for Learnable Query Size N_Q.

N_Q	Landscape		VGGSound	
	IA ↑	AV-Align ↑	IA ↑	AV-Align ↑
1	0.2814	0.4788	0.4439	0.5099
5	0.3661	0.5335	0.4797	0.5676
10	**0.3743**	**0.5502**	**0.4831**	0.5825
15	0.3636	0.5489	0.4782	**0.5841**
20	0.3577	0.5407	0.4776	0.5802

3.2.1. Object locating

Given the decomposed audio sources N_A as input, time-dependent audio tokens $p^1, \dots, p^i, \dots, p^{N_A}$ are obtained through pre-trained condition injection module (CIM) where $p^i = (p^i_1, \dots, p^i_S)$ and $S = \lceil \frac{L}{2} \rceil$, with L being the input audio duration. Then, we use the unconditioned text C and repeated audio tokens $\hat{p}^i_s \in \mathbb{R}^{N_S \times N_Q \times d}$ to generate the merged audio condition, where N_S is frame size (which is the total number of frames N_F divided by S) and $s \in \{1, \dots, S\}$, with N_Q representing the size of the learnable queries. The merged condition is defined as:

$$C[s \cdot N_s : (1+s) \cdot N_s, (i-1) \cdot N_Q : i \cdot N_Q] = \hat{p}^i_s, \tag{4}$$

where $i \in \{1, \dots, N_A\}$ and this is denoted as C_{merge}.

In the cross-attention module, the query representation $Q \in \mathbb{R}^{N_F \times d_h \times d}$ is used to compute attention with the representations K,

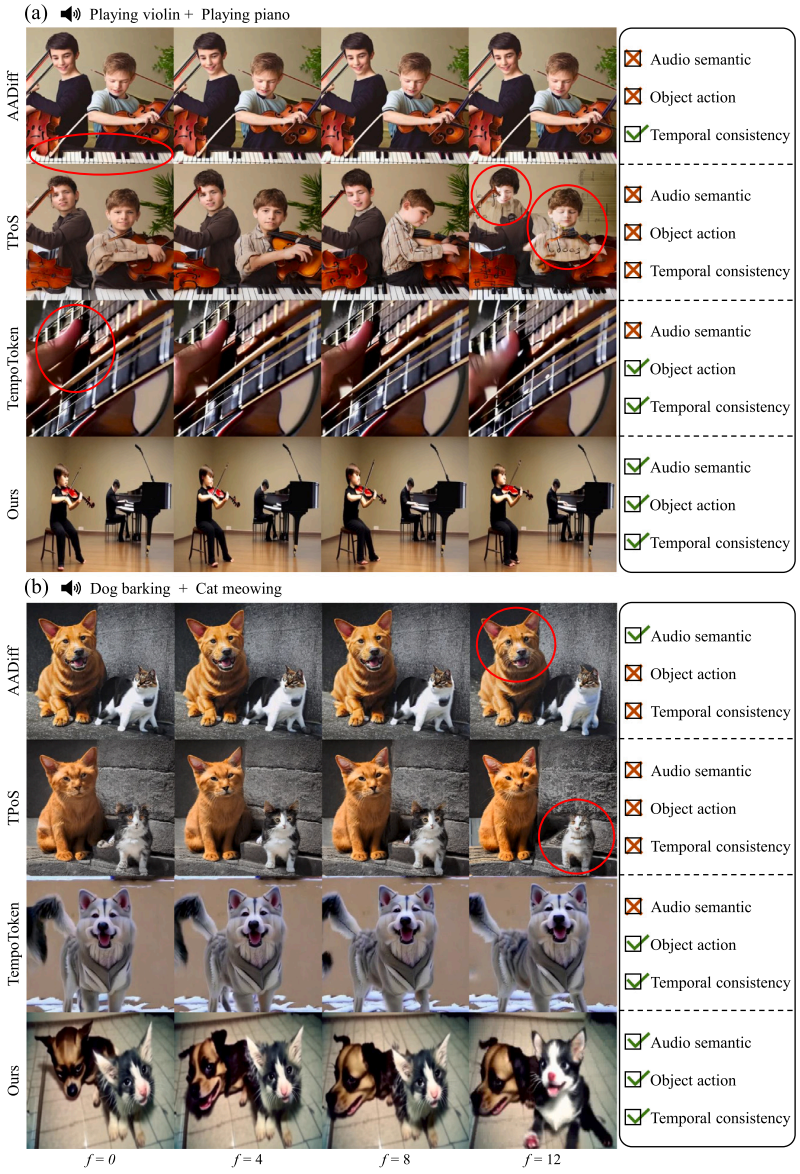

Fig. 9. Qualitative results. Examples of generated video frames (given (a) playing violin + playing piano (b) dog barking + cat meowing) by AADiff, TPoS, TempoToken and ours. "+" indicates overlapping audio inputs.

$V \in \mathbb{R}^{N_F \times |W| \times d}$, derived from the C_{merge}, where $d_h \equiv w \times h$, defined by the spatial resolution width and height within the respective block. The cross attention map $\mathcal{A}$ is then defined as:

$$\mathcal{A} = Softmax(\frac{QK^T}{\sqrt{d}}), \tag{5}$$

where $Softmax(z_k) = \frac{e^{z_k}}{\sum_{t=1}^{K} e^{k_t}}$ for $k = 1, 2, \ldots, K$. Since multi-source audio contains sounds from multiple objects, we use bounding boxes (Bboxes) $\mathcal{B} = \{\mathcal{B}_f^i | i, f \in \mathbb{N}, 1 \le i \le N_A, 1 \le f \le N_F\}$ to specify the location of each object. Given Bbox $\mathcal{B}(x, y)$ created from the resolution d_h, the additional weight function $\mathcal{G} \in \mathbb{R}^{N_F \times w \times h \times |W|}$ that generates the values at given (x, y) using a 2D gaussian function $g(\cdot, \cdot)$ is defined by:

$$\mathcal{G}[f, x, y, (i-1) \cdot N_Q : i \cdot N_Q] = \delta \cdot g(x, y) \tag{6}$$

$$\delta = \begin{cases} +1, & \text{if } (x, y) \in \mathcal{B}_f^i \\ -1, & \text{if } (x, y) \in \mathcal{B}_f^j \end{cases}, \; j = i^c \tag{7}$$

Finally, the (x, y) coordinates of the attention map $\mathcal{A}$ in Eq. (5) are updated by adding $\mathcal{G}$ from Eq. (6), as follows:

$$\mathcal{A}_f^i(x, y) := \mathcal{A}_f^i(x, y) + \mathcal{G}_f^i(x, y), \quad \text{where } (x, y) \in \mathcal{B}_f^i \tag{8}$$

3.2.2. Mosaic denoising

If in the previous step we aimed to position objects precisely at intended locations within the initial latent space $z_T \sim \mathcal{N}(0, I)$, the subsequent step focuses on ensuring that the distributions of located objects are influenced only by their corresponding conditions. The previous approach utilizing attention maps enabled the simultaneous

Algorithm 1 Multi-object Inference

Input: Obtained of previous Self-attention Q, Bounding boxes $\mathcal{B} \in \{B_1, B_2, ..., B_{N_A}\}$, 2D gaussian function $g(\cdot,\cdot)$, Audio embedding $p^i \in \{p^1, p^2, ..., p^{N_A}\} = P_A$, Unconditional null embeddings C, and pre-trained attention linear weight **Linear**$(\cdot)$
Output: Cross-attention value to be input to the next block

// T = Total number of diffusion steps
// N_D = Hyper-parameter dividing inference steps
// N_F = Number of Video Frames
// N_Q = Condition injection module's query size
// (x, y) = Coordinates within the Bboxes $\mathcal{B}$
function MERGE(C, P_A, N_Q)
  for each $i = 1, 2, ..., N_A$ **do**
    Compute C with Eq. (4)
  $K, V \leftarrow$ **Linear**$_{K,V}(C)$
  return K, V
function SEPARATE(C, P_A, N_Q)
  for each $i = 1, 2, ..., N_A$ **do**
    C^i = COPY(C)
    Compute C^i with Eq. (9)
    $K^i, V^i \leftarrow$ **Linear**$_{K,V}(C^i)$
  return $K^{0:N_A}, V^{0:N_A}$

for each timestep $t = T-1, T-2, ..., 0$ **do**
  // **Object Locating**
  if $t \geq T - N_D$ **then**
    $K, V \leftarrow$ MERGE(C, P_A, N_Q)
    $A \leftarrow Softmax(QK^T/\sqrt{d})$
    for each $i = 1, 2, ..., N_A$ **do**
      $s = (i-1) \times N_Q$
      $e = i \times N_Q$
      $A_{s:e}(x, y) = A_{s:e}(x, y) + g(B^i) - g(B^{i^C})$
    Output $\leftarrow A \times V$
  // **Mosaic Denoising**
  else if $t < T - N_D$ **then**
    $K^{0:N_A}, V^{0:N_A} \leftarrow$ SEPARATE(C, P_A, N_Q)
    for each $i = 1, 2, ..., N_A$ **do**
      $\hat{Q}^{N_F \times \hat{d_h} \times d} = \text{MASK}(B^i, Q^{N_F \times d_h \times d})$
      $O^i \leftarrow Softmax(\hat{Q}(K^i)^T/\sqrt{d}) \times V^i$
    Output $\leftarrow$ Concat$(O^1, O^2, ..., O^{N_A})$

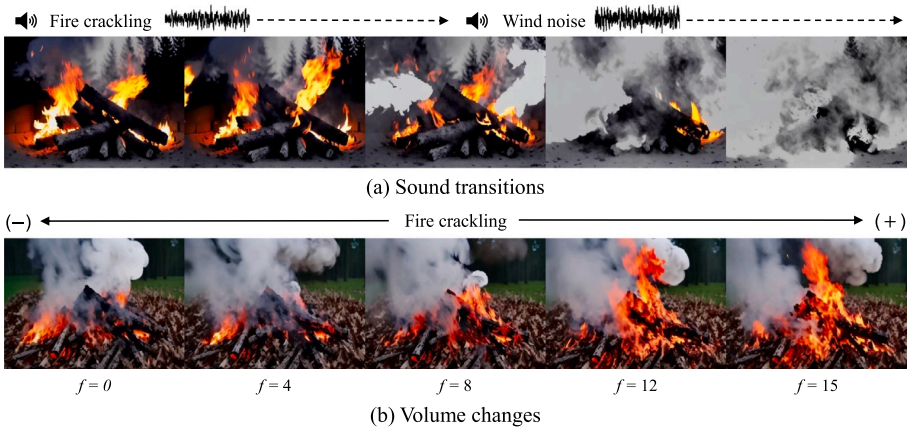

Fig. 10. Result of sound control. (a) Sound transitions (*Fire crackling* → *Wind noise*). We observe a natural process where a fire burns, then diminishes with smoke due to the wind. (b) Volume changes (×0.5 → ×1.5). As we increase the magnitude of sound *Fire crackling*, the fire becomes larger.

generation of multiple objects, but a problem remains where theses generated objects are still incomplete due to being influenced by various condition. This issue arises from the probabilistic nature of adjusting weights before the $Softmax$ stage, where conditions that should focus on a single bounding box area are influenced by other areas due to low probabilities. Therefore, to address this limitation, we propose a method called $MosaicDenoising$ to mask the query map $\hat{Q} \in \mathbb{R}^{N_F \times h \times w \times d}$ (which is a resized form of the query Q) in alignment with the bounding boxes $B^i(x, y)$ before entering the *Attention* operation. Since we need to create individual K and V corresponding to the masked $\hat{Q}^i(x, y)$, we leverage unconditioned text C^i to generate separated audio conditions, which is a different approach from Eq. (4). The separated condition C^i is defined as:

$$C^i[N_s : (1+s) \cdot N_s, : N_Q] = \hat{p}^i_s, \tag{9}$$

where $\hat{p}^i_s \in \mathbb{R}^{N_S \times N_Q \times d}$ represents repeated audio tokens, which is the same as mentioned in Section 3.2.1 and this is denoted as C_{sep}. Specifically, $i \in \{1, \ldots, N_A\}$ and $s \in \{1, \ldots, S\}$ where $S = \lceil \frac{L}{2} \rceil$, with L being the input audio duration.

The K^i and V^i pairs generated from this C^i perform *Attention* with the corresponding $\hat{Q}^i(x, y)$ as follows:

$$\mathcal{O}^i = Softmax(\frac{\hat{Q}^i(x, y){K^i}^T}{\sqrt{d}})V^i \tag{10}$$

$\hat{Q}^i(x, y) \in \mathbb{R}^{N_F \times \hat{d_h} \times d}$ refers to the query corresponding to the $B^i(x, y)$ regions and is as follows: $\hat{d_h} = x \times y$. (The softmax operation is the same as mentioned in Section 3.2.1.) Afterward, outputs $\mathcal{O}^1, \ldots, \mathcal{O}^{N_A}$ are merged back together.

4. Experiments

In Sections 4.1 and 4.2, we present our dataset curation process using the VGGSound dataset (Chen et al., 2020) and provide implementation details. In Section 4.3, we discuss the optimization of the Condition Injection Module, focusing on finding the optimal values for the loss function and the learnable query size. Furthermore, in Sections 4.4 and 4.5, we showcase the qualitative and quantitative results of our model in comparison with baselines. Finally, in Section 4.6, we demonstrate the effectiveness of each component of our proposed method.

4.1. Dataset

Dataset Curation. Existing audio-video datasets, including the VGGSound dataset (Chen et al., 2020), often have "in the wild" videos from YouTube, containing diverse classes. However, many audio-video pairs are temporally or semantically unaligned and have noisy audio, such as out-of-frame audio sources or audio sources with a magnitude of 0. Additionally, there are static videos with very little movement, which can have a negative impact on video processing.

To leverage clean (well-matched) video-audio pairs for training, we curate a dataset by filtering the VGGSound dataset. VGGSound

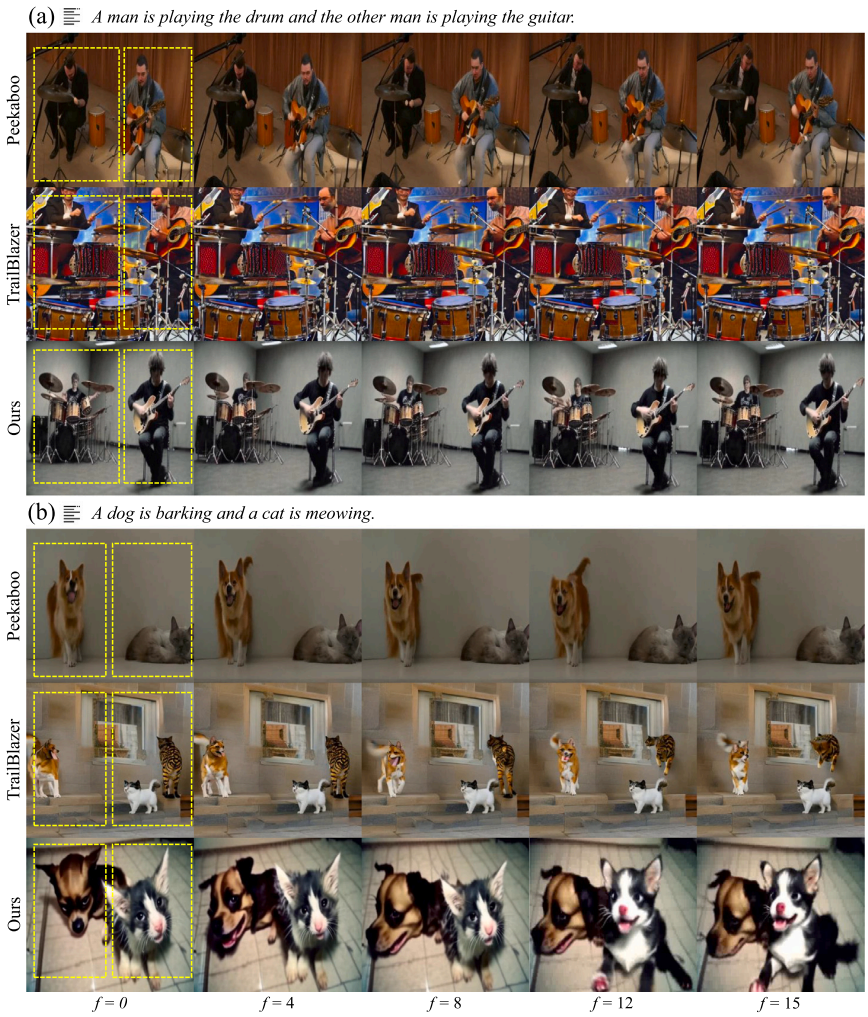

Fig. 11. Qualitative results of our method with text-based baselines. We compare the representation of multiple objects with our method and existing text-to-video generation models. The text at the top refers to input from the Peekaboo and Trailblazer models. Our method leverages overlapping sounds of drum playing and guitar playing as input in (a), and overlapping sounds of dog barking and cat meowing as input in (b). The yellow boxes indicate the input bounding boxes.

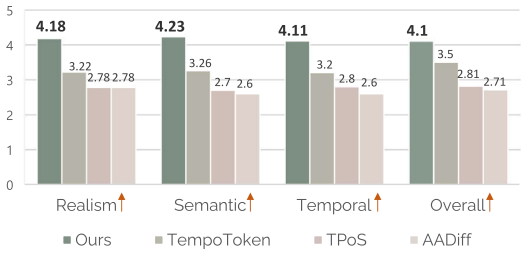

Fig. 12. Human evaluation results.

contains over 200K audio-visual pairs, which consists of 309 classes. For measuring (i) semantic alignment, we compute the alignment scores for Keyframe-Audio, Keyframe-Text, and Audio-Text leveraging ImageBind (Girdhar et al., 2023). Since the pre-trained models (Chen et al., 2024b; Rombach et al., 2022) used in our method employ a clip-based text encoder trained to maximize cosine similarity between image and text embeddings for correct pairs (Radford et al., 2021), we choose cosine similarity as the semantic alignment score. To determine the optimal threshold, We manually evaluate the quality of the top 100 videos before and after thresholding at the 25%, 50%, and 75% boundaries for each pair. For example, in the 25% threshold range of the Keyframe-Audio pair, we observe that the keyframe features only a cat indoors, while the audio contains the louder sound of a car horn outside the window. Finally, we keep the videos where all three values are higher than the median (see Fig. 6). To remove (ii) static videos, we follow the optical flow measurement method of Stable Video diffusion (Blattmann et al., 2023), to obtain dense optical flow maps at 2fps using the OpenCV (Itseez, 2015) implementation of the Farneback algorithm and set threshold to 0.1. Additionally, we exclude videos with an audio magnitude of 0 and with several labels that are difficult to distinguish by audio alone.

For pre-training the condition injection module, we use two audio-visual datasets: curated VGGSound and Landscape (Lee et al., 2022). Curated VGGSound is high-quality dataset containing 40K audio-visual pairs in the wild. The dataset has audio data of various objects, such as people, animals and cars. To produce richer videos, we use the Landscape dataset consisting of 9 natural classes, with 900 clips as the

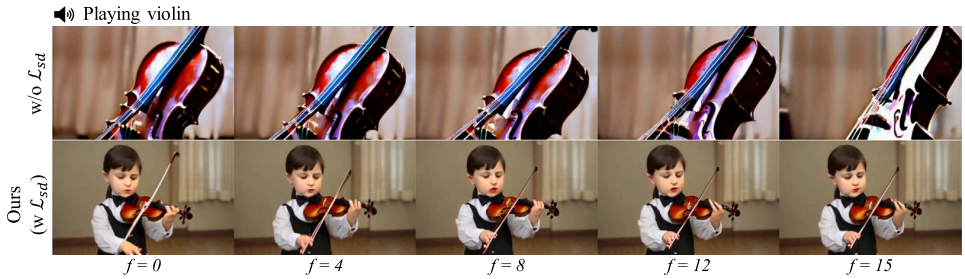

Fig. 13. Effectiveness of denoising loss $\mathcal{L}_{sd}$. Generated video frames using two versions of audio query transformer: one using the denoising loss of Stable Diffusion (second row) and one without (first row) in the pre-training process.

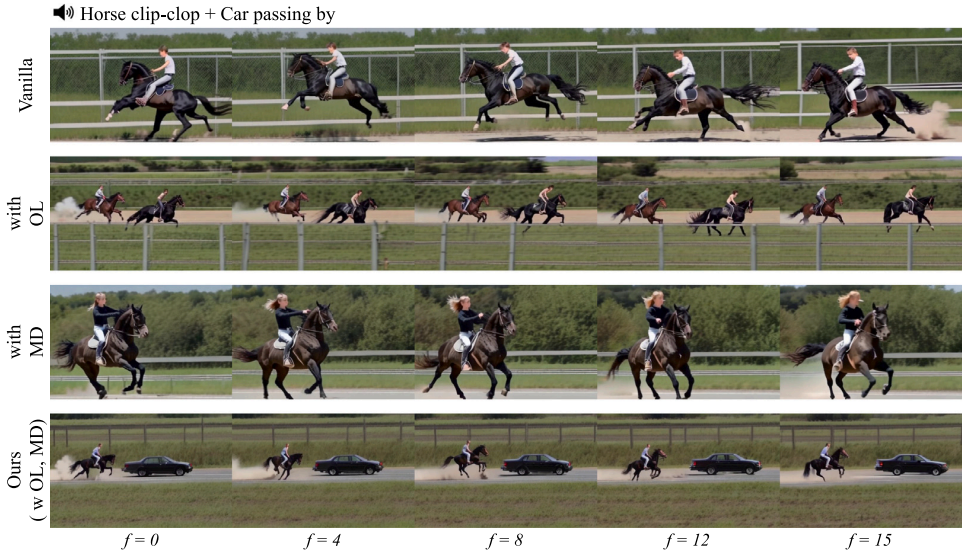

Fig. 14. Effectiveness of Attention Mosaic. Generated video clips with and without proposed our modules, Object Locating (OL) and Mosaic Denoising (MD), in Attention Mosaic.

training and 100 clips as the testing set (Zhang et al., 2024; Yariv et al., 2024).

Effects of Dataset Curation. By lowering Kolmogorov complexity through the reduction of unnecessary data that disrupts the generation of learning patterns (Bolón-Canedo and Remeseiro, 2019; Kabir and Garg, 2023), the curated VGGSound dataset can also potentially improve the accuracy of the Condition Injection Module, which creates time-dependent audio tokens that effectively capture audio information. Furthermore, Stable Video Diffusion (Blattmann et al., 2023) has demonstrated the necessity of a well-curated pre-training dataset for high-quality video generation, incorporating captioning and filtering strategies.

4.2. Implementation details

4.2.1. Baselines

We compared our methods with several state-of-the-art audio-to-video generation methods, AADiff (Lee et al., 2023a), TPoS (Jeong et al., 2023), and TempoToken (Yariv et al., 2024). We re-implemented AADiff and leveraged TPoS and TempoToken's pre-trained checkpoints on Landscape and VGGSound. For TempoToken, we evaluated the approach of using both an audio and a text prompt, referred to as TempoToken (w/ text) in quantitative study. We used the text prompt "a photo of ⟨class label⟩".

Moreover, for evaluating video generation performance using multiple objects, we consider two text-to-video generation models : TrailBlazer (Ma et al., 2023) and Peekaboo (Jain et al., 2024). Both baselines use the pre-trained publicly available zeroscope-v2 model (Wang et al., 2023). We utilized audio class label as a text input and configured identical fixed bounding boxes for all. Specifically, for TrailBlazer, we leveraged Scene Compositing method which synthesizes multiple subjects using the prompt and the stored single subject latents.

4.2.2. Training condition injection module

Condition injection module (CIM) allows the denoising UNet in the T2V (Text-to-Video) model to interpret the audio information and produces time-dependent audio tokens. We leveraged the pre-trained Stable Diffusion V2-1 (Rombach et al., 2022) for $\mathcal{L}_{sd}$ and ImageBind (Girdhar et al., 2023) as audio encoder. The input image size is fixed as 512 × 512 and the input audio duration is fixed as 10 s for training the CIM. We employed the ViT-H-14 architecture of CLIP text encoder (Radford et al., 2021), which is used in our base video model, VideoCrafter2 (Chen et al., 2024b). We utilized the Adam optimizer and set the learning rate to 0.0001. Training was performed on the curated VGGSound dataset for 30,000 iterations, requiring approximately 23 h, and on the Landscape dataset for 15,000 iterations, taking around 3 h. Both were conducted using 4 NVIDIA RTX 4090 GPUs. TPoS (Jeong

Table 3
Quantitative results. Compared with audio-to-video generation baselines on two datasets: VGGSound and Landscape. The underlined values are the second best-performing values, and the bolded values are the first best-performing values.

Model	Inputs		Landscape			VGGSound	
	Text	Audio	FVD ↓	IA ↑	AV-Align ↑	IA ↑	AV-Align ↑
AADiff (Lee et al., 2023a)	✓	✓	2151.8	0.24	0.12	0.30	0.30
TPoS (Jeong et al., 2023)	✓	✓	2314.1	<u>0.26</u>	<u>0.44</u>	0.23	0.43
TempoToken (Yariv et al., 2024)	✓	✓	<u>1912.8</u>	0.22	0.40	<u>0.38</u>	<u>0.52</u>
	✗	✓	2108.8	0.12	<u>0.44</u>	0.10	0.48
Ours	✗	✓	**585.9**	**0.37**	**0.55**	**0.48**	**0.58**

et al., 2023) was trained on the combined Landscape and VGGSound datasets for 26 h using 4 NVIDIA RTX 3090 GPUs. For TempoToken (Yariv et al., 2024), training on the VGGSound dataset took 32 h and training on the Landscape dataset took 5 h, using 2 A6000 GPUs.

4.2.3. Hyper-parameters setting

We conduct experiments on the value of N_D, which distinguishes the time steps T between *Object Locating* (OL) and *Mosaic Denoising* (MD) in Attention Mosaic (Section 3.2). The upper part of Fig. 7 shows that when OL is performed less time steps and MD more, the resulting video appears unnatural, with frames split in half. Conversely, increasing the N_D (which makes OL be performed more steps and MD less) leads to clear object locations and more natural frames, but at the cost of lower object quality, resulting in artifacts such as the merging of a dog and a chicken. The optimal compromise was found at $N_D = 6$, where objects are generated at precise locations corresponding to the Bboxes, resulting in stable quality and natural frames.

4.3. Discussion of condition injection module

4.3.1. Analysis of loss function

As shown in Table 1 and Fig. 8, when only the MSE loss L_{mse} is used ($\lambda_1 = 0.0$), the quality of the generated videos is lower in terms of alignment with the audio. This is because various audio samples from the same class are mapped to a single class label (text), resulting in the loss of diverse audio representation information. Additionally, while the "gun shooting" label correctly generates the gun, it fails to represent the subject shooting the gun. (see Fig. 8 first row) On the other hand, as the ratio of the Denoising loss L_{sd} increases, both qualitative and quantitative results show improvement. As shown in the last row of Table 1 ($\lambda_1 = 1.0$), we observe a slight drop in the IA score. In cases where no text is used, the single-source audio is learned with the entire image information that contains multiple meanings, the alignment accuracy decreases. Therefore, we conclude that incorporating both image and text information is essential for training. Consequently, we set λ_1 to 0.9 (see Fig. 8, fourth row). The effect of the Denoising loss L_{sd} is further explained in the Ablation study in 4.6.

4.3.2. Analysis of learnable query size

In our method, the maximum number of audio inputs is determined by the value of N_Q. According to Eq. (4), the product $N_A \times N_Q$ must be less than 77. For example, when $N_Q = 10$, a maximum of 7 audio inputs can be used. Since our study addresses the task of generating multiple objects from multiple audio inputs, we set the maximum value of N_Q to 20 to ensure at least three audio can be utilized as inputs. Table 2 shows that Image-Audio Semantic alignment (IA) is sufficient with $N_Q = 5$, while Temporal alignment (AV-Align) is sufficient with $N_Q = 10$. When $N_Q = 1$, we observe significantly lower performance, indicating that $N_Q = 1$ is insufficient to capture both audio and frame information effectively. Therefore, we select $N_Q = 10$ as it effectively captures essential information while minimizing computing complexity. Additionally, we experiment with the Animal Kingdom dataset (Ng et al., 2022) and the results can be found in Appendix.

4.4. Qualitative results

4.4.1. Comparison with baselines

Fig. 9 shows the visual comparison of Maestro and several audio-to-video generation baselines. We include video results from AADiff, TPoS, TempoToken, and Maestro with 3 criteria: *Audio semantic, Object action and Temporal consistency*. We evaluate the accurate representation of semantically complex audio (Audio semantic), the exhibition of object actions (Object action), and the consistency of consecutive frames (Temporal consistency). We used the audio inputs of (a) playing violin + playing piano and (b) dog barking + cat meowing, where "+" indicates overlapping audio inputs. Most baselines fail to accurately capture semantically complex audio. AADiff and TPoS encounter difficulties in representing natural object movements and fail to accurately simulate human actions such as violin or piano playing. Furthermore, (b) shows poor object consistency performance. TempoToken effectively represents natural action in videos but struggles to accurately capture overlapped audio. (a) shows that mixed sounds of dog barking and cat meowing are mistakenly interpreted as guitar sounds. However, Maestro excels in accurately representing non-audible objects, such as natural arm movements of musicians playing instruments. In addition, (b) shows that Maestro effectively represents simple meanings like "dog" and "cat", while realistically depicting "barking" and "meowing" with natural mouth movements.

We also show the ability of our model to respond to changes in audio volume and transition, as shown in Fig. 10. (a) shows the outputs that seamlessly represent transitions associated with each audio input. Additionally, (b) shows that Maestro reflects changes in audio volume in the video objects. Therefore, Maestro enables the generation of richer and more diverse videos using only audio input.

4.4.2. Comparison to multi-object generation model with text

We compare the performance of representing multiple objects to existing text-to-video generation models. In Fig. 11, we provide examples of generated video frames by Peekaboo (Jain et al., 2024) and Trailblazer (Ma et al., 2023). Peekaboo fails to perfectly represent objects (see first row), while Trailblazer often exhibits unnatural behavior in multi-object interactions (see second row). However, Maestro depict the fine details of a person playing the drum and guitar, including the naturalness of their actions and alignment with bboxes positions. Moreover, (b) shows that Maestro most accurately represents the meanings of "barking" and "meowing" through natural mouth movements.

4.5. Quantitative results

4.5.1. Comparison with baselines

We quantitatively compare our method to baselines on the Landscape (Lee et al., 2022) and VGGSound (Chen et al., 2020) dataset. We use the following three metrics: (i) Frechet Video Distance (FVD) (Unterthiner et al., 2018) to measure video quality, (ii) Image-Audio (IA) (Girdhar et al., 2023) to measure the average alignment between video frames and audio semantics using CLIP (Radford et al., 2021) features, and (iii) Audio-Visual Alignment (AV-Align) (Yariv et al., 2024) to assess the temporal alignment. AV-Align detects audio peaks using an Onset Detection algorithm (Böck and Widmer, 2013) and

video changes by calculating the mean of the Optical Flow (Horn and Schunck, 1981) magnitude, respectively. To implement these, we finetune Inflated 3D ConvNet with Landscape dataset for FVD and use encoder of ImageBind (Girdhar et al., 2023) for IA. As shown in Table 3, our method outperforms in video quality, audio-visual alignment and semantic alignment score on both Landscape and VGGSound dataset. AADiff and TPoS perform worse than TempoToken in terms of video quality (FVD), but they show strong image-audio alignment scores (IA). In addition, TempoToken is worse IA score than TempoToken (w/ text), but has similar FVD score. This is expected, as TempoToken relies heavily on text conditions to convey semantics.

4.5.2. User study

We conduct a human evaluation study by recruiting 100 participants from Amazon Mechanical Turk (AMT). Participants are shown videos generated by four different audio-to-video generation models: AADiff (Lee et al., 2023a), TPoS (Jeong et al., 2023), TempoToken (Yariv et al., 2024), and ours. Participants evaluate each method considering realism, semantic, temporal consistency, and overall quality over 20 videos generated from 10 single-source audio and 10 multi-source audio. We use a Likert scale ranging from 1 (low quality) to 5 (high quality). As shown in Fig. 12, our proposed method outperforms other baselines across four criteria. More details are provided in Appendix A.

4.6. Ablation studies

To demonstrate the effectiveness of each component of our method, we perform ablation studies on the following elements: (1) Stable Diffusion denoising loss $\mathcal{L}_{sd}$, (2) Attention Mosaic. Specifically, in Attention Mosaic, we focus two methods: object locating and mosaic denoising. The evaluation is conducted with the VGGSound test dataset.

Effectiveness of denoising loss $\mathcal{L}_{sd}$. In the text-to-video model, the text conditions "playing violin". and "A girl is playing violin on the stage". make a big difference in video quality. Therefore, to extract more specific information from audio, we pre-train the condition injection module (CIM) of our model using the knowledge of Stable Diffusion. The effectiveness of this learning method, as shown in Fig. 13, is to better represent the situation than a method using only text MSE, which represents only the simple meaning of the audio. Specifically, the representation of non-audible objects such as the violinist and the background becomes more natural and rich.

Effectiveness of Attention Mosaic. We qualitatively demonstrate the effectiveness of our methods at the inference stage. Specifically, we focus on two tasks: *Object Locating* and *Mosaic Denoising*. The results are illustrated in Fig. 14. In the basic Vanilla model, when two conditions are input, either only one condition is generated or they are merged into a single output. This issue arises because the cross-attention mechanisms corresponding to each condition tend to focus on the same object. OL method addresses this by assigning Bboxes for each condition, effectively separating the objects. While this approach successfully generates multiple objects, it fails to distinctly focus on each condition, resulting in objects that are influenced by other conditions. MD method, on the other hand, proposes a deterministic approach to apply conditions to the respective Bbox regions, rather than a probabilistic one. However, similar to the Vanilla outputs, it does not achieve clear object separation, leading to unstable outcomes. By combining the OL and MD, we are able to generate objects corresponding to each condition while simultaneously applying denoising steps. This ensured that only the appropriate conditions are applied to each region, resulting in high-quality outputs.

5. Conclusions

We propose Maestro, a novel pipeline for multi-source audio-to-video generation, which aims to accurately produce multiple objects within frames by decomposing audio sources. Our Attention Mosaic technique ensures that generated objects are placed precisely where desired, enhancing fine details. Additionally, by applying an effective optimization strategy to the condition injection module, we achieve the generation of better contextual expressions involving non-audible objects. Results from the user study indicate that Maestro significantly produces high-quality video from both multi-source and single source audio, improving audio-visual alignment. We also observe that our method effectively captures basic characteristics such as volume changes and transitions.

CRediT authorship contribution statement

Sieun Kim: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Jaehwan Jeong:** Writing – original draft, Software, Methodology, Investigation. **Sumin In:** Validation, Software, Methodology, Investigation, Data curation. **Seung Hyun Lee:** Writing – original draft, Methodology, Conceptualization. **Seungryong Kim:** Writing – review & editing, Resources. **Saerom Kim:** Validation, Formal analysis, Data curation. **Wooyeol Baek:** Validation, Formal analysis. **Sang Ho Yoon:** Writing – review & editing, Formal analysis. **Eugenio Culurciello:** Writing – review & editing, Formal analysis. **Sangpil Kim:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 39%; International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 30%; Development of technology for dataset copyright of multimodal generative AI model, RS-2024-00333068, 30%), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190079, 1%), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

Appendix A. User study

We execute a human evaluation study to assess the video results based on four criteria: realism, audio alignment, temporal consistency, and overall quality. We utilize a five-point Likert scale where the best video will receive all five points for the following four questions. First, "*How natural and realistic does this video look, considering the consistency between the background and the foreground.*" evaluates the realism of the generated video and the coherence between the objects and the background. Second, "*How well does the video correspond with the audio.*" evaluates the accurate representation of semantically complex audio. Third, "*How smoothly the content of videos changes in response to the given audio.*" evaluates the consistency of multiple objects and the background in the generated video frames. Finally, "*Considering the three questions above, please assign a score for their overall video quality.*" evaluates the overall video quality that best reflects the flow of the audio, considering the three questions mentioned above (see Fig. A.15).

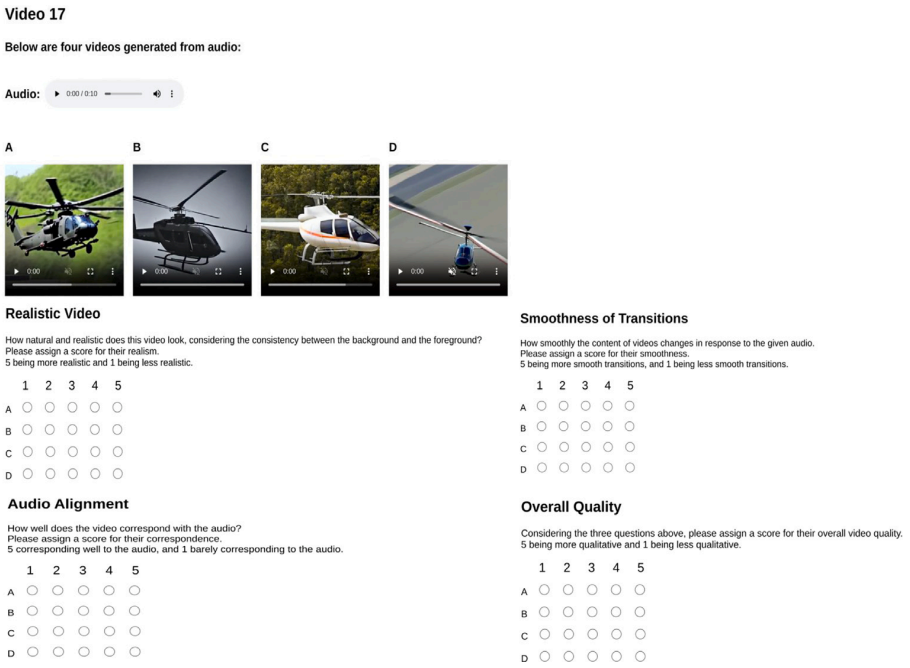

Fig. A.15. Example of human evaluation survey. Our proposed method and audio-to-video generation baselines are labeled A, B, C, and D. In each question, participants are presented with an audio input and the outputs generated by the four models. Then, participants evaluate the quality of each model's output based on four evaluation criteria.

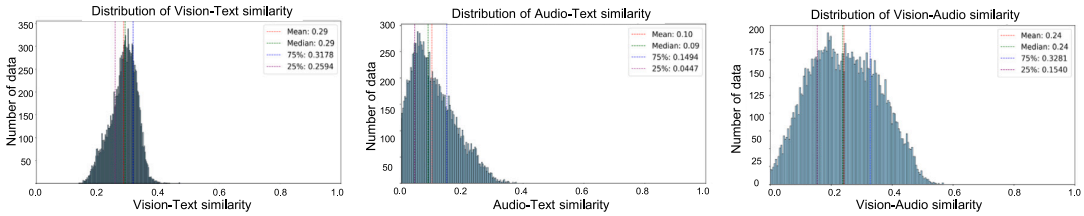

Fig. B.16. Histogram visualization of semantic alignment across different modalities.

Appendix B. Animal Kingdom dataset

We conduct experiments using the Animal Kingdom dataset (Ng et al., 2022) to generate a broader range of object categories. The Animal Kingdom dataset is a large and diverse dataset for animal behavior understanding, consisting of high-quality video and audio. It consists of six key animal classes (e.g., reptiles, birds, fishes, etc.), which we use as text labels. The dataset contains 4,301 videos, with a total duration of 50 h. After curating the dataset, we evenly allocate the inputs across four NVIDIA GeForce RTX 4090 GPUs and train the Condition Injection Module for 15k iterations, taking 4 h to complete.

B.1. Dataset curation

The Animal Kingdom dataset contains a significant amount of narration audio, which we consider noise. Therefore, we first used AudioSep (Liu et al., 2023) to remove the narration audio and split all the videos into 10-s segments for training. To leverage clean (well-matched) video-audio pairs for training, we also curate the dataset similarly to the VGGSound dataset (Chen et al., 2020) (see Fig. B.16). We set the metric thresholds after checking the quality of the videos around the 25%, 50%, and 75% thresholds similar to VGGSound. As a result, we experiment with a total of 2500 videos, setting the training set to 2000 videos and the test set to 500 videos.

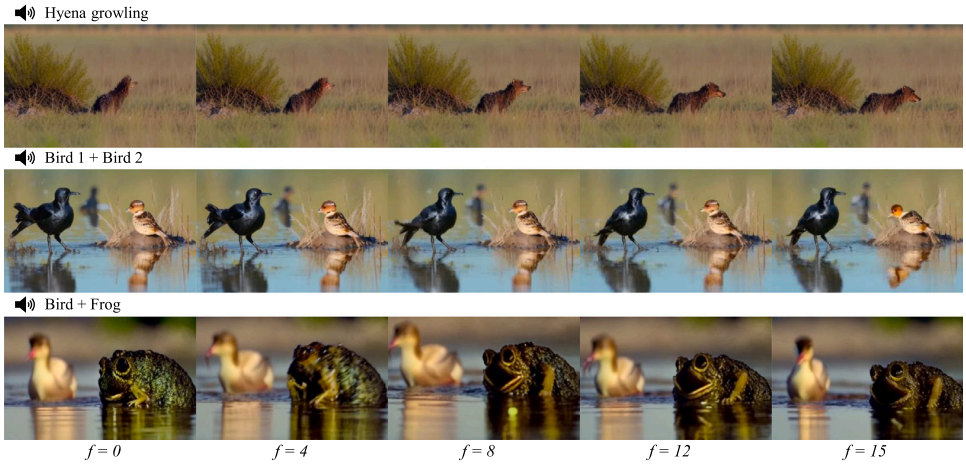

Fig. B.17. Qualitative results of our method on the Animal Kingdom dataset.

Table B.4
Quantitative results for learnable query size N_Q.

N_Q	Animal kingdom	
	IA ↑	AV-Align ↑
1	0.3324	0.4109
5	0.3781	0.4309
10	**0.3816**	**0.4344**
15	0.3776	0.4299
20	0.3796	0.4273

B.2. Learnable query size

Table B.4 shows the Animal Kingdom dataset (Ng et al., 2022) also exhibits poor performance at $N_Q = 1$, similar to VGGSound (Chen et al., 2020) and Landscape (Lee et al., 2022). To achieve good performance, we need at least $N_Q = 5$, and the best performance is observed at $N_Q = 10$.

B.3. Qualitative results

As shown in Fig. B.17, we provide generated videos leveraging sounds as input: (a) hyena growling, (b) overlapping sounds of two birds, and (c) overlapping sounds of a bird and a frog.

Appendix C. Limitations and future works

While showing impressive potential, our proposed Maestro still has some limitations. First, our model has difficulty handling complex soundscapes with multiple overlapping audio sources. Our Condition Injection Module is trained using single-source audio paired with corresponding video data, and during inference, decomposed audio sources are mapped into the video frame space. However, even state-of-the-art audio separation models struggle to achieve perfect single-source decomposition in complex soundscapes, making it difficult for our model to use single-source audio during both training and inference. Additionally, our training approach relies more on image information, which helps the model better recognize foreground objects with richer and clearer cues, allowing it to naturally excel at generating objects that align with the foreground sound.

Second, the range of generable foreground objects and backgrounds is limited. The availability of single-source audio datasets that are semantically well-aligned with videos remains highly constrained. To address this limitation, we use the VGGSound (Chen et al., 2020), Landscape (Lee et al., 2022), and Animal Kingdom (Ng et al., 2022) datasets to generate as many diverse scenarios as possible. The release of high-quality single-source video datasets with diverse classes would further enable the generation of more varied and higher quality videos.

Third, as Maestro relies on a pre-trained T2V diffusion model, the generation quality of Maestro is constrained by the capabilities and limitations of the pre-trained T2V model. Furthermore, the length of the resulting video clips is constrained by the output of the pre-trained model. VideoCrafter is designed to generate videos with 16 frames by default, but it is also capable of generating videos with 32 and 48 frames. When using text as input, the 16 frames video produces the best results, which is why we conducted our research based on 16 frames. As shown in Fig. C.18, three videos effectively capture the semantic information of the audio. Specifically, the 16-frame video (see Fig. C.18(a)) shows barking and roaring well. However, the quality of the 32-frame and 48-frame videos is slightly lower. (see Fig. C.18(b),(c)). While our current focus is on effectively representing multi-object scenarios, further research is needed to extend this capability by handling audio in long videos, in addition to representing multiple objects.

Appendix D. Qualitative results

In this section, we provide more qualitative results of our method with baselines. Specifically, Fig. D.19 represents the qualitative comparison with the state-of-the-art Audio-to-Video models (Lee et al., 2023a; Jeong et al., 2023; Yariv et al., 2024). In Figs. D.20 and D.21, we showcase the multi-source audio to video generation results of overlapping one background sound source with various foreground sound sources. In Figs. D.22 and D.23, we showcase the results of overlapping four different sound sources, each including one and two background sounds, respectively. Finally, we visualize the generated videos conditioning on single-source audio (see Fig. D.24).

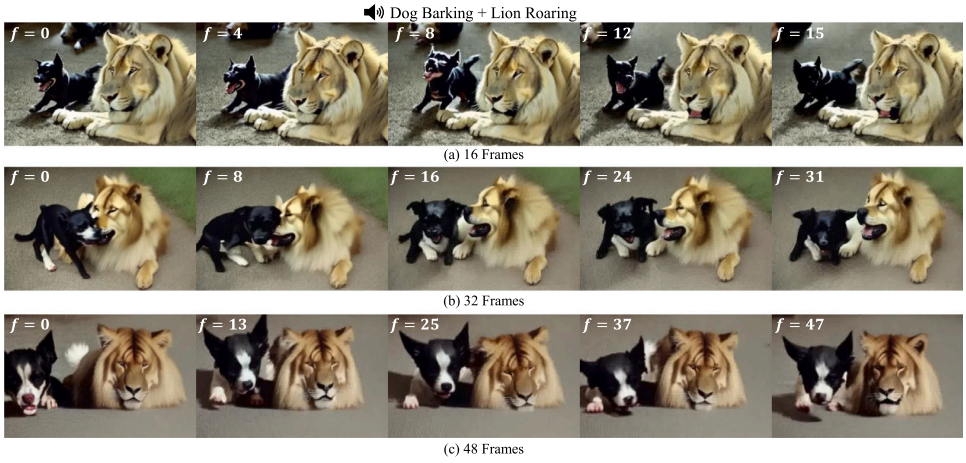

Fig. C.18. Qualitative results from 16, 32, and 48 frames with two audio sources.

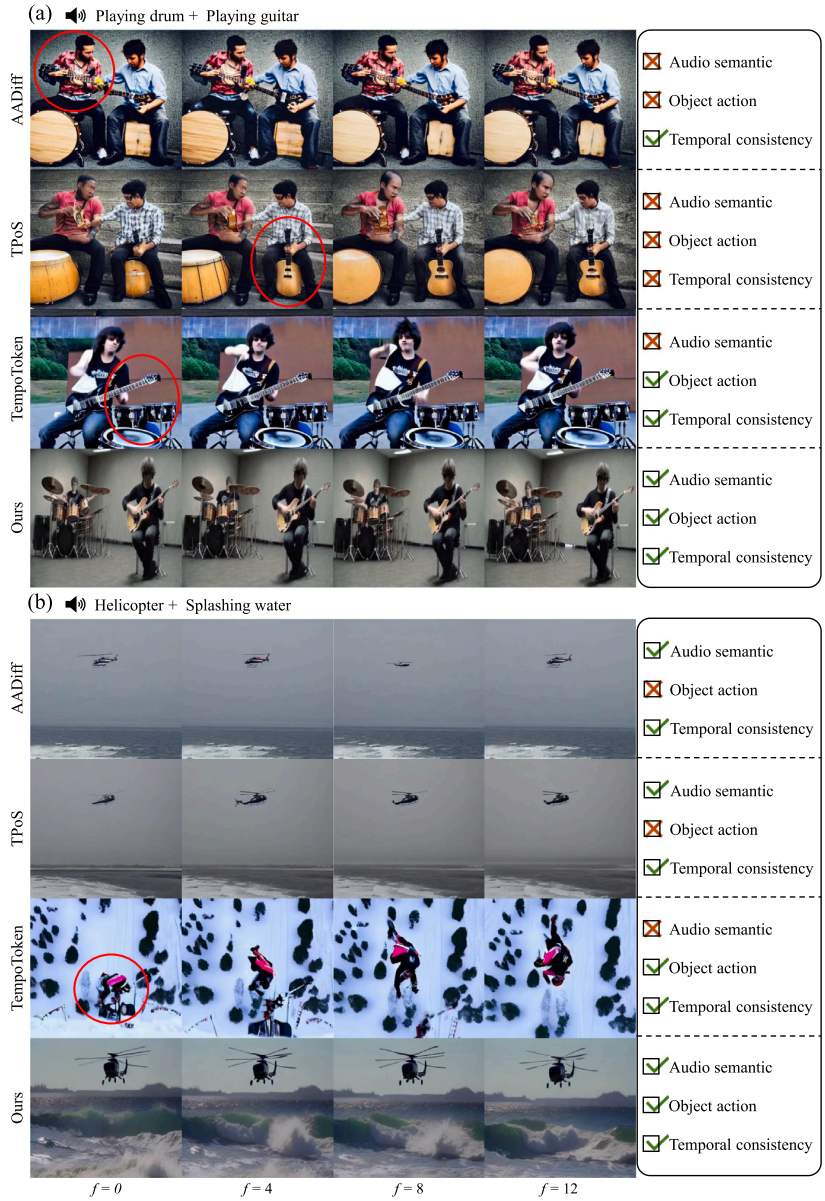

Fig. D.19. Qualitative comparison results. Examples of generated video frames (given (a) playing drum + playing guitar (b) helicopter + splashing water) by AADiff, TPoS, TempoToken and ours. "+" indicates overlapping audio inputs.

Fig. D.20. Qualitative results from multi-source audio.

Fig. D.21. Qualitative results from multi-source audio.

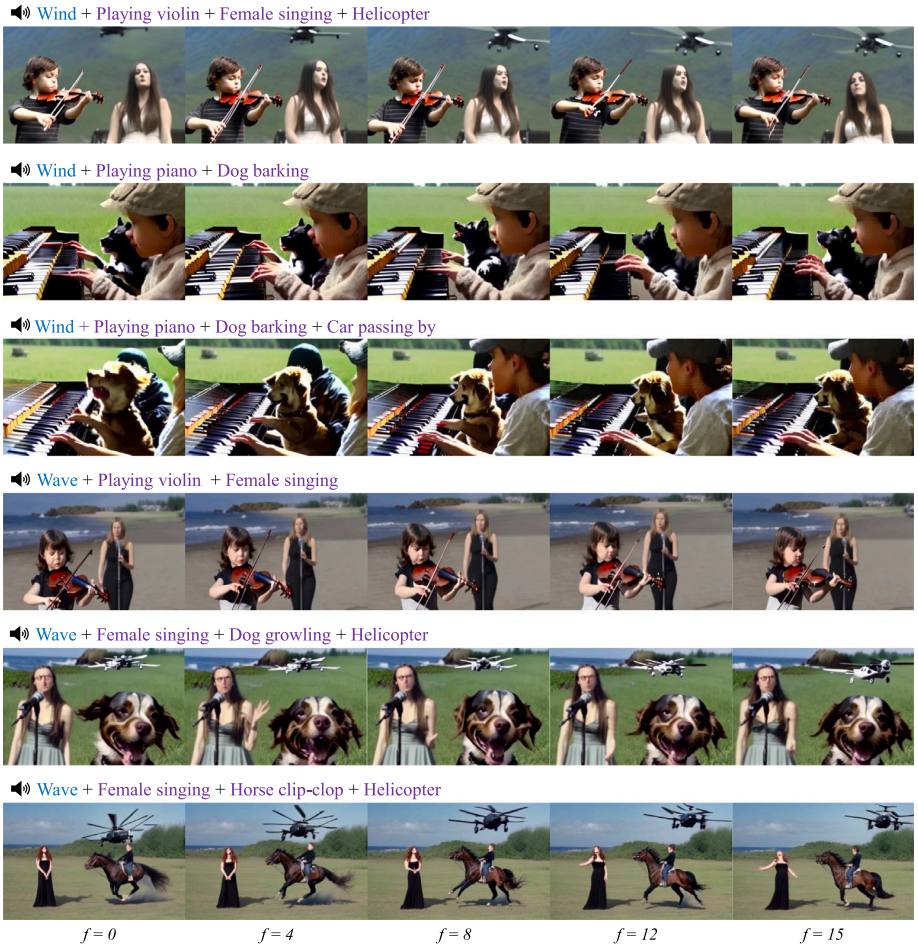

Fig. D.22. Qualitative results from multi-source audio.

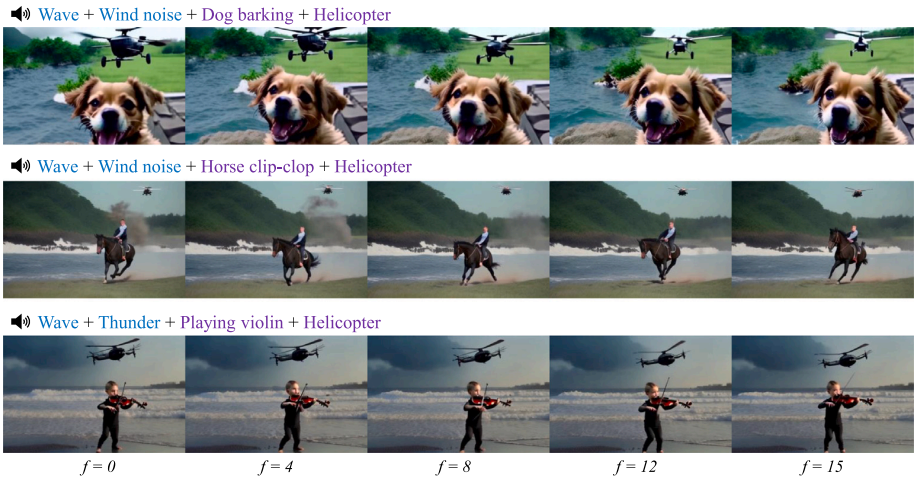

Fig. D.23. Qualitative results from multi-source audio.

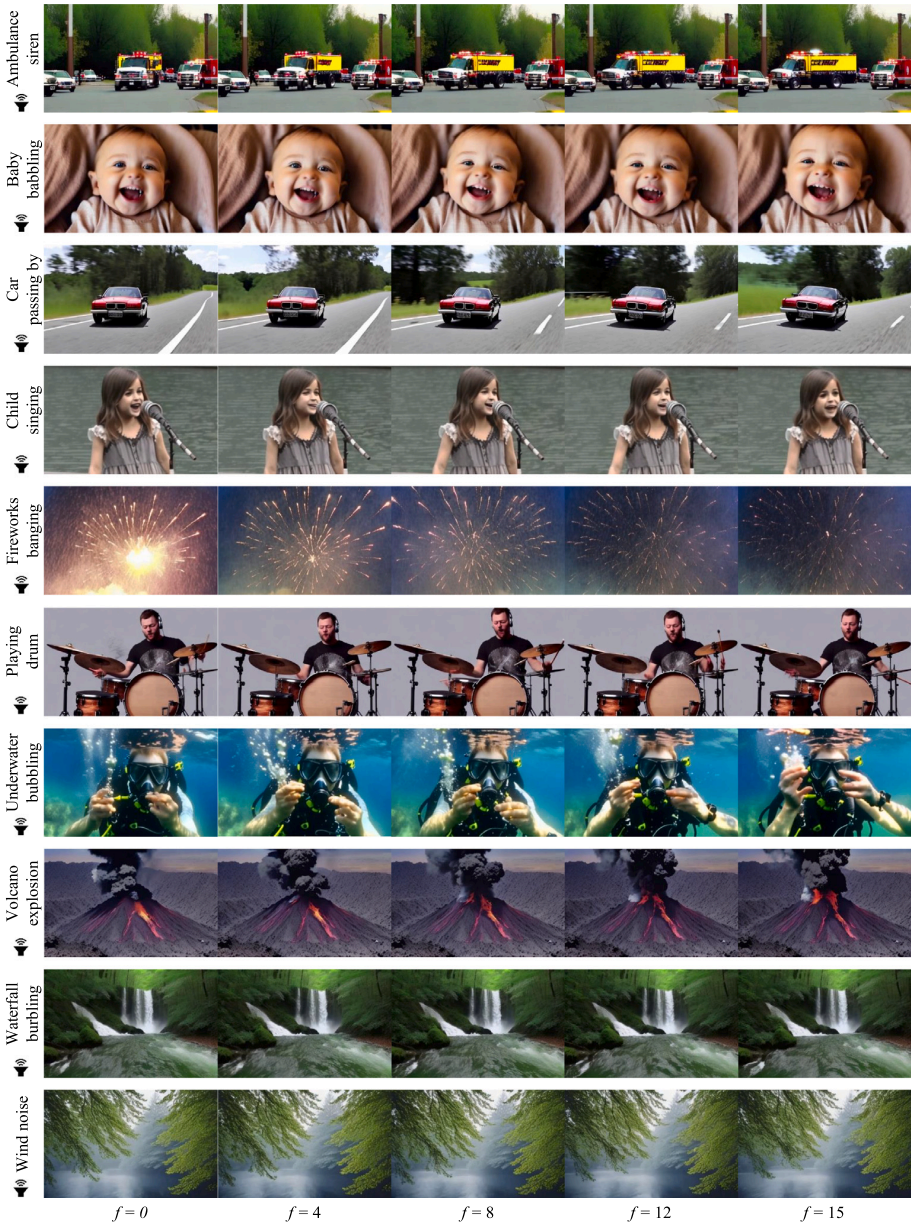

Fig. D.24. Qualitative results from single-source audio.

Data availability

The authors do not have permission to share data.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. 35, 23716–23736.

Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Li, Y., Michaeli, T., et al., 2024. Lumiere: A space-time diffusion model for video generation. arXiv preprint arXiv:2401.12945.

Biner, B.C., Sofian, F.M., Karakaş, U.B., Ceylan, D., Erdem, E., Erdem, A., 2024. SonicDiffusion: Audio-driven image generation and editing with pretrained diffusion models. arXiv preprint arXiv:2405.00878.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al., 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127.

Böck, S., Widmer, G., 2013. Maximum filter vibrato suppression for onset detection. In: Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013). Vol. 7, Citeseer, p. 4.

Bolón-Canedo, V., Remeseiro, B., 2019. Feature selection in image analysis: a survey. Artif. Intell. Rev. 53, 2905–2931, URL: https://api.semanticscholar.org/CorpusID:199511374.

Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.-H., Murphy, K., Freeman, W.T., Rubinstein, M., et al., 2023. Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704.

Chatterjee, M., Cherian, A., 2020. 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. Springer, pp. 701–719.

Chen, C., Shu, J., Chen, L., He, G., Wang, C., Li, Y., 2024a. Motion-zero: Zero-shot moving object control framework for diffusion-based video generation. arXiv preprint arXiv:2401.10150.

Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al., 2023. Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512.

Chen, H., Xie, W., Vedaldi, A., Zisserman, A., 2020. Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 721–725.

Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y., 2024b. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I., 2023. Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B., 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725.

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

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J., 2022b. Video diffusion models. Adv. Neural Inf. Process. Syst. 35, 8633–8646.

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

Horn, B.K., Schunck, B.G., 1981. Determining optical flow. Artificial Intelligence (ISSN: 0004-3702) 17 (1), 185–203. http://dx.doi.org/10.1016/0004-3702(81)90024-2, URL: https://www.sciencedirect.com/science/article/pii/0004370281900242.

Itseez, 2015. Open source computer vision library. https://github.com/itseez/opencv.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J., 2021. Perceiver: General perception with iterative attention. In: International Conference on Machine Learning. PMLR, pp. 4651–4664.

Jain, Y., Nasery, A., Vineet, V., Behl, H., 2024. Peekaboo: Interactive video generation via masked-diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8079–8088.

Jeong, Y., Ryoo, W., Lee, S., Seo, D., Byeon, W., Kim, S., Kim, J., 2023. The power of sound (tpos): Audio reactive video generation with stable diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7822–7832.

Kabir, H., Garg, N., 2023. Machine learning enabled orthogonal camera goniometry for accurate and robust contact angle measurements. Sci. Rep. 13, URL: https://api.semanticscholar.org/CorpusID:256277662.

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

Le Moing, G., Ponce, J., Schmid, C., 2021. Ccvs: Context-aware controllable video synthesis. Adv. Neural Inf. Process. Syst. 34, 14042–14055.

Lee, S.H., Kim, S., Yoo, I., Yang, F., Cho, D., Kim, Y., Chang, H., Kim, J., Kim, S., 2023b. Soundini: Sound-guided diffusion for natural video editing. arXiv preprint arXiv:2304.06818.

Lee, S., Kong, C., Jeon, D., Kwak, N., 2023a. AADiff: Audio-aligned video synthesis with text-to-image diffusion. arXiv preprint arXiv:2305.04001.

Lee, S.H., Li, Y., Ke, J., Yoo, I., Zhang, H., Yu, J., Wang, Q., Deng, F., Entis, G., He, J., et al., 2024. Parrot: Pareto-optimal multi-reward reinforcement learning framework for text-to-image generation. arXiv preprint arXiv:2401.05675.

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

Liu, X., Kong, Q., Zhao, Y., Liu, H., Yuan, Y., Liu, Y., Xia, R., Wang, Y., Plumbley, M.D., Wang, W., 2023. Separate anything you describe. arXiv preprint arXiv:2308.05037.

Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al., 2024. Sora: A review on background, technology, limitations, and opportunities of large vision models. arXiv preprint arXiv:2402.17177.

Ma, W.-D.K., Lewis, J., Kleijn, W.B., 2023. TrailBlazer: Trajectory control for diffusion-based video generation. arXiv preprint arXiv:2401.00896.

Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.-F., Chen, C., Qiao, Y., 2024. Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048.

Ng, X.L., Ong, K.E., Zheng, Q., Ni, Y., Yeo, S.Y., Liu, J., 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 19023–19034.

Oh, G., Jeong, J., Kim, S., Byeon, W., Kim, J., Kim, S., Kwon, H., Kim, S., 2023. MTVG: Multi-text video generation with text-to-video models. arXiv preprint arXiv:2312.04086.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

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

Peng, Z., Hu, W., Shi, Y., Zhu, X., Zhang, X., Zhao, H., He, J., Liu, H., Fan, Z., 2024. Synctalk: The devil is in the synchronization for talking head synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 666–676.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695.

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

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S., 2018. Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717.

Voleti, V., Jolicoeur-Martineau, A., Pal, C., 2022. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Adv. Neural Inf. Process. Syst. 35, 23371–23385.

Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S., 2023. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571.

Xing, Y., He, Y., Tian, Z., Wang, X., Chen, Q., 2024. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7151–7161.

Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., Adi, Y., 2024. Diverse and aligned audio-to-video generation via text-to-video model adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38, pp. 6639–6647.

Zhang, L., Mo, S., Zhang, Y., Morgado, P., 2024. Audio-synchronized visual animation. arXiv preprint arXiv:2403.05659.

Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J., 2023. Controlvideo: Adding conditional control for one shot text-to-video editing. arXiv preprint arXiv:2305.17098.