Audio-Sync Video Generation with Multi-Stream Temporal Control

Shuchen Weng[1†] **Haojie Zheng**[1,2†] **Zheng Chang**[3] **Si Li**[3] **Boxin Shi**[4,5‡] **Xinlong Wang**[1‡]
[1]Beijing Academy of Artificial Intelligence
[2]School of Software and Microelectronics, Peking University
[3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[4]State Key Lab of Multimedia Info. Processing, School of Computer Science, Peking University
[5]Nat'l Eng. Research Ctr. of Visual Tech., School of Computer Science, Peking University
{scweng, wangxinlong}@baai.ac.cn, suimu@stu.pku.edu.cn
{zhengchang98,lisi}@bupt.edu.cn, shiboxin@pku.edu.cn

Abstract

Audio is inherently temporal and closely synchronized with the visual world, making it a naturally aligned and expressive control signal for controllable video generation (*e.g.*, movies). Beyond control, directly translating audio into video is essential for understanding and visualizing rich audio narratives (*e.g.*, Podcasts or historical recordings). However, existing approaches fall short in generating high-quality videos with precise audio-visual synchronization, especially across diverse and complex audio types. In this work, we introduce MTV, a versatile framework for audio-sync video generation. MTV explicitly separates audios into speech, effects, and music tracks, enabling disentangled control over lip motion, event timing, and visual mood, respectively—resulting in fine-grained and semantically aligned video generation. To support the framework, we additionally present DEMIX, a dataset comprising high-quality cinematic videos and demixed audio tracks. DEMIX is structured into five overlapped subsets, enabling scalable multi-stage training for diverse generation scenarios. Extensive experiments demonstrate that MTV achieves state-of-the-art performance across six standard metrics spanning video quality, text-video consistency, and audio-video alignment. Project page: https://hjzheng.net/projects/MTV/.

1 Introduction

Audio is a fundamental medium in daily life, crucial for both information delivery (*e.g.*, communication, notifications, and education) and immersive experiences (*e.g.*, enhancing the impact of film visuals). Despite the prevalence of audio-centric platforms (*e.g.*, Podcasts), content presented solely through audio lacks the visual dimension needed to fully convey the richness of events. Since audio is naturally temporal and inherently synchronized with the visual world, researchers [1–3] have devoted considerable attention to translating audios into corresponding videos to enhance audience understanding of rich audio narratives (*e.g.*, historical recordings).

Despite great progress, existing methods face practical limitations in generating high-fidelity cinematic videos with precise synchronization (*e.g.*, pouring water into the transparent cup), primarily due to: *(i)* **Under-specified audio-visual mapping.** Current approaches handle a wide spectrum of audios and map them to various target scenes (*e.g.*, landscapes [4], dancing [5], music performances [6]). This broad representation scope potentially leads to ambiguous mappings lacking specificity between audio and visual features. *(ii)* **Inaccurate temporal alignment.** Existing methods primarily focus on

[†] Equal contributions. [‡] Corresponding authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

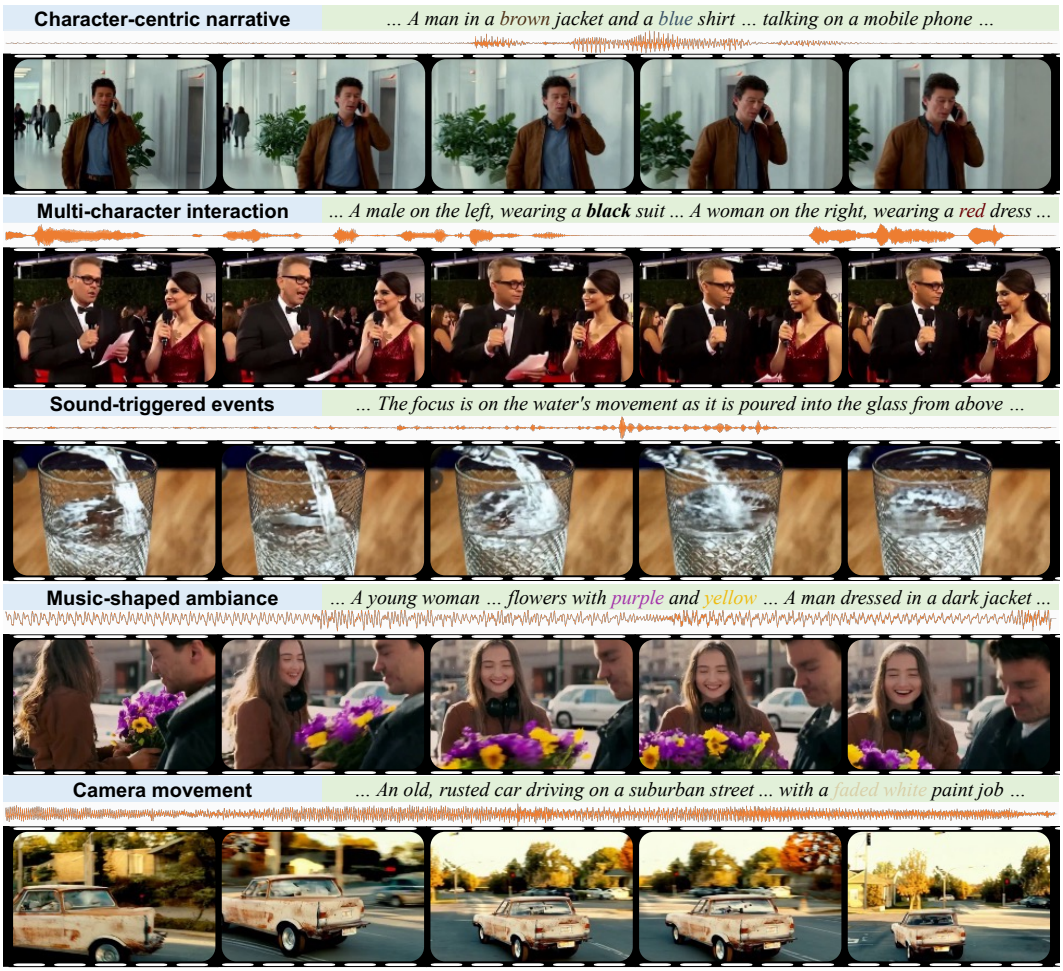

Figure 1: MTV demonstrates versatile audio-sync video generation capabilities following user-provided text descriptions specifying scenes and subjects. Capabilities shown include producing videos centered on targeted characters (1st and 2nd rows) while triggering events with sound effects (3rd row), generating visual mood with accompanying music (4th row), and adaptively handling camera movement (5th row). We present these generated videos in the supplementary materials.

building scene-level semantic consistency (*e.g.*, translating engine sound to a car-centered video), struggling with accurate timing correspondence between individual audio events and their visual features (*e.g.*, speech [7], motion [8], and visual mood [9]).

In this paper, we propose the **MTV** framework, enabling **M**ulti-stream **T**emporal control for audio-sync **V**ideo generation to overcome aforementioned issues, with versatile capabilities across scenarios illustrated in Fig. 1. Instead of attempting a direct mapping from composite audios, we explicitly separate audios into distinct controlling tracks (*i.e.*, speech, effects, and music), inspired by CDX'23[1]. To provide sufficient high-quality video clips with demixed audio tracks, we contribute a large-scale DEMIX dataset with tailored data processing, including 392K video clips with 1.2K hours. These tracks enable the model to precisely control lip motion, event timing, and visual mood, resolving the ambiguous mapping. To further incorporate rich visual semantics beyond direct audio cues, we leverage features (*e.g.*, subject gesture, scene appearance, camera movement) initially derived from a pretrained text-to-video model [10], and subsequently finetuned using video clips from the DEMIX dataset. To enable the progressive extension of learned high-level video semantic features stage-by-stage, this dataset is structured into five overlapped subsets. A multi-stage training strategy is introduced to learn concrete and localized controls (*e.g.*, lip motion) towards more abstract and global influences (*e.g.*, visual mood), leading to clear audio-visual relationships.

[1] `https://www.aicrowd.com/challenges/sound-demixing-challenge-2023`

To achieve accurate temporal alignment, we propose the Multi-Stream Temporal ControlNet (MST-ControlNet) within the MTV framework. The interval stream is designed for specific feature synchronization, which extracts features from the speech and effects tracks. It employs interval interaction blocks to understand each track individually and construct their interplay, maintaining the coherence with inferred semantic features. After that, interval feature injection module inserts features of each track into corresponding time intervals to drive lip motion and event timing. Since visual mood typically covers the entire video clip, the holistic stream is designed for overall aesthetic presentation, which extracts features from the music track using the holistic context encoder. These features then serve as style embeddings, applied uniformly to all frames through global style injection, controlling the visual mood.

We summarize our contributions as follows:

- We present MTV, a versatile audio-sync video generation framework by demixing audio inputs, achieving precise audio-visual mapping and accurate temporal alignment.
- We introduce an audio-sync video generation dataset structured into five overlapped subsets, presenting the multi-stage training strategy for learning audio-visual relationships.
- We propose the multi-stream temporal ControlNet to distinctively process demixed audio tracks and precisely control lip motion, event timing, and visual mood, respectively.

2 Related Works

2.1 Video Diffusion Model

The field of video generation has made significant progress with the adoption of diffusion models. Early approaches [11–13] extend the dynamic modeling capabilities of pretrained text-to-image diffusion models [14] by incorporating temporal layers (*e.g.*, 3D convolutions [15] and temporal attention [16]). However, these methods face inherent challenges in capturing long-range spatial-temporal dependencies due to the convolutional architectures of their backbone (*e.g.*, UNet [17]). To overcome this limitation, Sora report [18] presents the potential of the diffusion transformer (DiT) [19] architecture, prompting a shift towards integrating 3D VAE [20] for spatial-temporal compression and scaling up to train the entire DiT-based model. Further improvement has been achieved by recent foundation models through adaptive layernorm modules [10], progressive scaling [21, 22], and post-training techniques [23]. These advancements in text-to-video models provide a strong foundation and powerful generative priors that could potentially be leveraged for related cross-modal tasks, such as high-quality audio-sync video generation.

2.2 Audio-driven Image Animation

Audio-driven image animation aims to generate dynamic visuals from a static image, synchronized with user-provided audios. Several previous works animate general objects or scenes while maintaining audio-visual consistency. Sound2Sight [24] and CCVS [25] leverage the context of preceding frames to achieve audio-driven subsequent frames generation. TPOS [26] uses audios with variable temporal semantics and amplitude to guide the denoising process. ASVA [27] incorporates a temporal audio control module for effective audio synchronization. Other works concentrate on audio-driven human animation. Talking head [7, 28–30] focus on animating human face images to produce lip motion that synchronize with the speech. Recent works extend animation beyond the head to include half-body movements [31] and introduce pose control for full-body animation [32]. Another specific application is music-to-dance [33, 34], which generates human dance according to the beat of the music. Despite the audio-visual synchronization of these methods, their reliance on static images restricts models' capability to generate dynamic scenes required for cinematic videos.

2.3 Audio-sync Video Generation

Audio-sync video generation does not require additional images for reference, offering the potential for free scene creation. Early works are designed based on VQGAN [35] and StyleGAN [36], achieving audio control through multi-modal autoregressive transformers [2] and style code alignment [4, 37]. Recently, following the success of diffusion models demonstrating effectiveness in general video generation, researchers have turned their attention. Highlighting the benefit of multi-modal

Table 1: Comparison of DEMIX dataset and previous datasets.

Method	Year	Modality		Scene			Audio component				Specifications	
		Text	Audio	People	Objects	Cinematic	Speech	Effects	Music	Demix	Clips	Hours
UCF-101 [38]	2012	–	✓	✓	–	–	–	✓	✓	–	13K	27
HIMV-200K [39]	2017	–	✓	✓	✓	✓	–	–	✓	–	200K	–
AudioSet [40]	2017	–	✓	✓	✓	–	✓	✓	✓	–	2.1M	5.8K
VoxCeleb2 [41]	2018	–	✓	✓	–	–	✓	–	–	–	150K	2.4K
VGGSound [42]	2020	–	✓	✓	✓	–	✓	✓	✓	–	200K	550
WebVid-10M [43]	2021	✓	–	✓	✓	–	–	–	–	–	10.7M	52K
Landscape [4]	2022	–	✓	–	✓	–	–	✓	–	–	9K	26
InternVid [44]	2024	✓	✓	✓	✓	–	✓	✓	✓	–	7.1M	760K
Ours (DEMIX)	2025	✓	✓	✓	✓	✓	✓	✓	✓	✓	392K	1.2K

conditions, TA2V [6] demonstrates that conditioning on both text descriptions and audio inputs significantly enhances the quality of generated videos. To achieve audio-visual alignment at both global and temporal levels, TempoTokens [1] designs a lightweight adapter for text-to-video generation model. Introducing a unified diffusion architecture, MM-Diffusion [5] enables both joint audio-video generation and zero-shot audio-sync video generation. Leveraging diffusion-based latent aligners for open-domain audio-visual generation, Xing *et al.*[3] achieve the audio-sync video editing and open-domain content creation. Although great progress has been made, audio-sync video generation still faces under-specific audio-visual mapping and inaccurate temporal alignment. Therefore, achieving cinematic quality remains challenging.

3 Dataset

We introduce the DEMIX dataset, tailored for training demixed audio-sync video generation models.

Data source. The training data is sourced from three aspects: *(i)* 65 hours of talking head videos from CelebV-HQ [45]; *(ii)* 4,923 hours of cinematic videos from MovieBench [46] (69h), Condensed Movies [47] (1,270h), and Short-Films 20K [48] (3,584h); and *(iii)* 8,903 hours film-related videos from YouTube. All collected videos include their accompanying audio tracks.

Video filtering. Following previous video generation models [10, 12, 49], we use PySceneDetect [50] to segment video into single-shot clips. Audiobox-aesthetics [51] is further used to assess the quality of accompanying audio, removing clips with low scores. For the left video clips, we annotate each one with text descriptions using LLaVA-Video [52].

Demixing filtering. To improve audio demixing reliability, we employ a dual-demixing comparison strategy, comparing demixing outputs from MVSEP [53] (speech, effects, music) and Spleeter [54] (speech, others). After that, we calculate the L1 distance between the speech tracks. Next, the 'others' track from Spleeter is conditionally compared: to the effects track from MVSEP if music is silent (below -45dB), and to the music track if effects are silent. Clips are discarded only if high L1 distances are found on any of the comparable pairs.

Voice-over filtering. To build clear audio-visual relationships for cinematic videos, we first detect whether people are present in the videos using YOLO [55]. Next, we perform speaker diarization for the accompanying audio using Scribe [56] to identify active speaker segments and count the number of speakers. After that, we detect the active speaker from videos for each frame using TalkNet [57]. As a result, we can discard clips where speech occurs in the audio but the video analysis detects neither a visible person nor an active speaker in the corresponding frames.

Subset division. To facilitate multi-stage training for versatile audio-sync video generation models, the filtered DEMIX data is structured into five overlapped subsets. The basic face subset comprises all talking head videos. The remaining cinematic and film-related videos are then categorized to form the other subsets: assignment to single character or multiple characters depends on the annotated human count, while assignment to sound event or visual mood occurs if the respective effects or music track is non-silent.

Data statistics. After data collection and filtering, our DEMIX dataset includes 18K basic face, 54K single character, 39K multiple characters, 166K sound event, and 195K visual mood data, tailored for

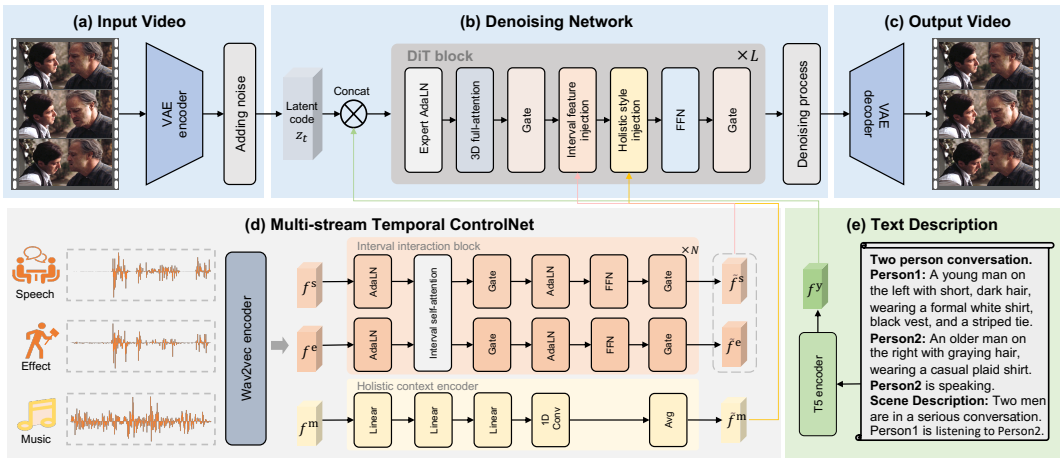

Figure 2: The pipeline of our MTV framework. (a-c) MTV is built on a pretrained text-to-video model [10] that provides strong generative priors for synthesizing diverse visual scenarios. (d) Explicitly separated audio tracks (*i.e.*, speech, effects, music) are fed into our proposed multi-stream temporal ControlNet to ensure synchronization for lip motion, event timing, and visual mood. (e) The MTV framework is trained on our contributed DEMIX dataset with five overlapped subsets and tailored text structures, enabling a multi-stage training strategy for audio-sync video generation.

cinematic videos, totaling non-overlapped 392K clips with 1.2K hours, accompanied by demixed audio tracks[2]. For comprehensive evaluation, we hold out 1K video clips from the dataset to form the testing set. We provide an additional comparison with existing audio-related datasets [4, 38–44] in Tab. 1, highlighting that ours is tailored for versatile audio-sync video generation using demixed audio tracks, while robustly covering scenarios with people, objects, and cinematic visuals.

4 Method

This section begins with an overview of our MTV framework for audio-sync video generation (Sec. 4.1). Next, we detail the Multi-stream Temporal ControlNet (MST-ControlNet), including the interval stream for specific feature synchronization, and the holistic stream for overall aesthetic presentation (Sec. 4.2). Finally, we present the multi-stage training strategy for effectively learning audio-visual relationships (Sec. 4.3).

4.1 Overview

MTV generates audio-sync videos based on user-provided text descriptions y (specifying the scenes and subjects) and demixed audio tracks $a = \{a^{\mathrm{s}}, a^{\mathrm{e}}, a^{\mathrm{m}}\}$ (representing speech, effects, and music) to respectively drive the lip motion, event timing, and visual mood. The pipeline is illustrated in Fig. 2.

Video compression. As presented in Fig. 2 (a), MTV is equipped with a pretrained spatio-temporal variational autoencoder (VAE) encoder $\mathcal{E}$ to map video clips x into latent code $z_0 = \mathcal{E}(x)$. After that, its corresponding VAE decoder $\mathcal{D}$ is used to reconstruct video clips from the latent code $x = \mathcal{D}(z_0)$.

Denoising network. As presented in Fig. 2 (b), we concatenate the text embeddings f^{y} and noised latent code z_t before feeding them into the network to ensure the video-text correspondence. The expert Adaptive LayerNorm (AdaLN) [10] then independently processes text and video features within this unified sequence. Next, 3D full-attention is used to interact semantics of text embeddings with corresponding video features. After being extracted by MST-ControlNet, audio cues are integrated via the interval feature injection and holistic style injection mechanisms. Finally, a feed-forward network (FFN) is used to refine the resulting video features.

Denoising process. As presented in Fig. 2 (c), MTV finally generates audio-sync videos by iteratively denoising latent codes. During training, at each time step $t \in \{0, \ldots, T\}$, Gaussian noise $\epsilon_t \sim$

[2]Dataset samples are visualized in the supplementary materials.

$\mathcal{N}(0,1)$ is added to the clean latent code z_0 to produce a noised latent code $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$. A diffusion transformer ϵ_θ is trained to predict the noise ϵ_t, given the noised latent code z_t, demixed audio tracks a, denoising time step t, and text descriptions y. The diffusion transformer is trained by minimizing the loss:

$$\mathcal{L}_{\text{dm}} = \mathbb{E}_{t, z_0, \epsilon_t \sim \mathcal{N}(0,1)} \left[\|\epsilon_t - \epsilon_\theta(z_t, a, t, y)\|^2 \right]. \tag{1}$$

For inference, we iteratively denoise a randomly sampled noise $z_T \sim \mathcal{N}(0,1)$ to obtain the latent code z_0' to generate video clips with the VAE decoder $x' = \mathcal{D}(z_0')$.

4.2 Multi-stream Temporal ControlNet

After explicitly separating audios into speech, effects, and music tracks, we propose the MST-ControlNet to achieve accurate temporal alignment by respectively controlling lip motion, event timing, and visual mood. As presented in Fig. 2 (d), the architecture consists of an audio encoding module followed by two specialized streams.

Audio encoding. Given demixed audio tracks $a = \{a^{\text{s}}, a^{\text{e}}, a^{\text{m}}\}$, we initially extract their corresponding features $\{f^{\text{s}}, f^{\text{e}}, f^{\text{m}}\}$ from the demixed tracks using wav2vec [58]. After that, speech and effect features are fed into the interval stream for specific feature synchronization. Instead, music features are fed into the holistic stream for overall aesthetic presentation.

Interval stream. We design the interval stream to interval-wise control the lip motion and event timing. Specifically, we separately process speech features f^{s} and effect features f^{e} with a stack of linear layers and concatenate them before feeding them into N interval interaction blocks. Within each block, these features are processed independently (via AdaLN, Gate, and FFN) to refine per-track understanding. To model their interplay at each time interval i, the corresponding speech features f_i^{s} and effects features f_i^{e} are jointly processed by a self-attention $[\tilde{f}_i^{\text{s}}, \tilde{f}_i^{\text{e}}] = \text{SelfAttn}([f_i^{\text{s}}, f_i^{\text{e}}])$. This interaction also maintains the coherence with inferred semantic features. Finally, interacted speech features $\tilde{f}^{\text{s}}$ and effects features $\tilde{f}^{\text{e}}$ are integrated into their corresponding time intervals via the interval feature injection mechanism:

$$h_i^{\text{s}} = \text{CrossAttn}(h_i, \tilde{f}_i^{\text{s}}), \quad h_i^{\text{e}} = \text{CrossAttn}(h_i, \tilde{f}_i^{\text{e}}), \tag{2}$$

where h_i represents the video latent code at i-th interval. $\text{CrossAttn}(\cdot,\cdot)$ means a cross-attention, where the latent code serves as the query and the audio features as the key and value. Let M be the number of intervals, the resulting latent code is then updated as $h' = \{h_i^{\text{s}} + h_i^{\text{e}}\}_{i=1}^{M}$.

Holistic stream. The holistic stream is designed to control the visual mood for the entire video clip. Specifically, we process the music features f^{m} through a holistic context encoder, comprising three linear layers and a 1D convolutional layer to extract features representing the visual mood. Since the environmental ambiance typically covers the entire video clip, an average pooling is applied to merge all the intervals and transform them into holistic music features $\tilde{f}^{\text{m}}$. Next, these features are regarded as style embeddings. By independently transforming these features into scale factor $\gamma^{\text{m}} = \text{Linear}(\tilde{f}^{\text{m}})$ and shift factor $\beta^{\text{m}} = \text{Linear}(\tilde{f}^{\text{m}})$, we modulate the video latent code h' uniformly across all intervals via the holistic style injection:

$$h^{\text{m}} = h' \odot (\gamma^{\text{m}} + 1) + \beta^{\text{m}}, \tag{3}$$

where h^{m} is the modulated latent code, fed into the denoising network to refine video features.

4.3 Multi-stage training strategy

As the dataset is structured as five overlapped subsets, we introduce the multi-stage training strategy to progressively scale up the model stage-by-stage.

Text structure. As presented in Fig. 2 (e), we create a template to structure text descriptions, enabling our MTV framework to be compatible with these distinct training subsets. Specifically, this template begins with a sentence indicating the number of participants (*e.g.*, "Two person conversation"), based on Scribe [56] speaker counts. It then consists of subsequent entries for each individual, starting with a unique identifier (*e.g.*, *Person1*, *Person2*) followed by their respective appearance description. Following these individual entries, an explicit identifier for the currently active speaker is specified. Finally, a sentence provides an overall description of the scene. Notably, when there is no active speaker in the video, only the overall description will be provided.

Table 2: Quantitative experiment results of comparison and ablation. ↑ (↓) means higher (lower) is better. Throughout the paper, best performances are highlighted in **bold**.

Method	FVD ↓	Temp-C (%) ↑	Text-C (%) ↑	Audio-C (%) ↑	Sync-C ↑	Sync-D ↓
Comparison with state-of-the-art methods						
MM-Diffusion [5]	879.77	94.15	15.61	5.43	1.53	11.21
TempoTokens [1]	795.88	93.13	24.68	6.71	1.45	10.48
Xing *et al.* [3]	805.23	93.30	24.51	7.30	1.55	10.50
Ours (MTV)	**626.06**	**95.40**	**26.55**	**26.22**	**3.17**	**9.43**
Ablation study						
W/o SE	667.81	95.30	26.49	24.68	2.46	9.55
W/o SI	626.46	94.84	25.50	19.64	2.53	9.76
W/o TB	698.36	95.14	26.37	24.50	2.31	9.78

Training schedule. We train the model from concrete and localized controls towards more abstract and global influences. Initially, we train the model to learn lip motion using the basic face subset. It then learns human pose, scene appearance, and camera movement on the single character subset. To handle scenarios with multiple speakers, we subsequently train the model on the multiple characters subset. Following this, our training focus shifts to event timing and extending subject understanding from humans to objects using the sound event subset. Finally, we train the model on the environmental ambiance subset to improve its representation of visual mood.

Training details. We initialize our spatial-temporal VAE and DiT backbone with pretrained weights from CogVideoX [10] and train our model to generate audio-sync videos at a 480×720 resolution. For each stage, we train our model for 40K steps on 24 NVIDIA A800 GPUs using the Adam-based optimizer [59] with a learning rate of 1×10^{-5}, where MST-ControlNet and attention layers of the backbone are trainable. For inference, our model requires 280s to generate a 49-frame audio-sync video on a NVIDIA A100 GPU.

5 Experiments

5.1 Comparison with state-of-the-art methods

As audio-sync video generation is an emerging task, the relevant comparison methods are still developing. We compare our method with three recent state-of-the-art approaches in our DEMIX dataset. For TempoTokens [1] and Xing *et al.* [3], we evaluate them using both text descriptions and corresponding audios as their original configuration. Since MM-Diffusion [5] can only support audio inputs and its training focuses on specific landscape and dancing, we finetune it to ensure a fair comparison. 50 videos are randomly selected from the testing set for evaluation.

Quantitative comparisons. As presented in Tab. 2, we quantitatively evaluate performance across three main aspects: *(i)* Visual quality is assessed using Frechét Video Distance (FVD) [60]. *(ii)* Temporal consistency (Temp-C) is measured by calculating similarity between consecutive frames using CLIP [61]. *(iii)* We examine text-video alignment via Text Consistency (Text-C) [62], audio-video alignment using Audio Consistency (Audio-C) [63], and specifically lip motion synchronization with Sync-C and Sync-D [64]. As a result, our framework outperforms state-of-the-art methods across all six quantitative metrics. These metric details are provided in the supplementary materials.

Qualitative comparisons. As presented in Fig. 3, qualitative comparisons with state-of-the-art methods [1, 3, 5] highlight the advantages of our framework. For instance, even after finetuning MM-Diffusion [5] for over 320K steps using the official code on 8 NVIDIA A100 GPUs, it still struggles with generating cinematic videos. TempoTokens [1] struggles to generate cinematic videos for complex text-specified scenarios, resulting in unrealistic human expressions (Fig. 3 left). Xing *et al.* [3] find it difficult to effectively achieve audio synchronization for specific event timing, leading to incorrect rendering of human gestures for guitar performance (Fig. 3 right). In contrast, our MTV framework faithfully generates audio-sync videos with cinematic quality.

Figure 3: Visual comparison results with state-of-the-art methods for audio-sync video generation.

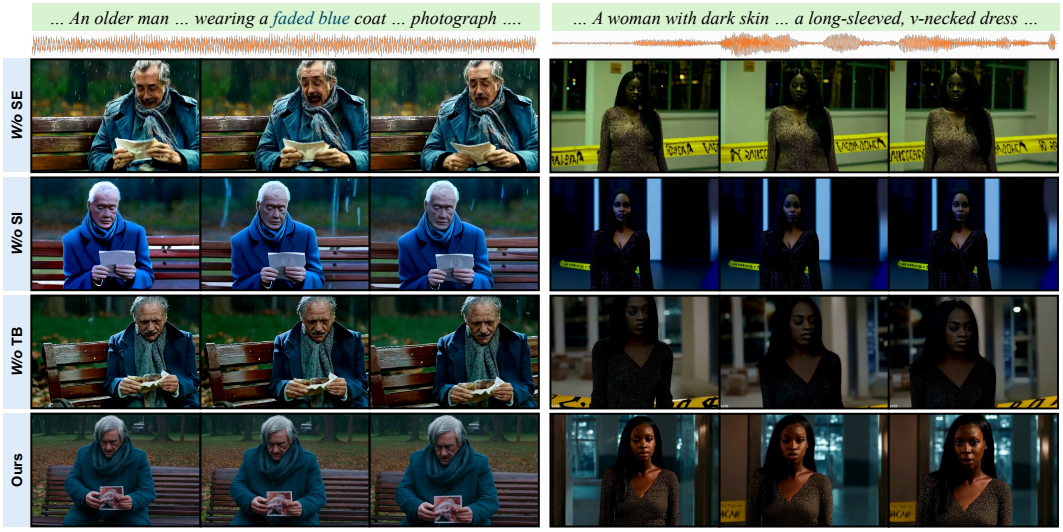

Figure 4: Ablation study results of different MST-ControlNet variants.

5.2 Ablation Study

To evaluate the effectiveness of key components within MST-ControlNet, we conduct ablation studies against three baseline configurations, as shown in Fig. 4 and Tab. 2.

W/o SE (Separate Extraction). We extract all features from demixed audio tracks using interval interaction blocks. This prevents music features from shaping the overall aesthetic presentation, leading to reduced visual mood (Fig. 4 left, degraded FVD and Temp-C).

W/o SI (Separate Injection). We extract features from demixed audio tracks by their respective encoders. These features are then concatenated and injected into the denoising network via a shared cross-attention. This reduces conditional consistency (Fig. 4 left, decreased Text-C and Audio-C).

W/o TB (Training Backbone). We freeze all weights of DiT backbone and only train our proposed MST-ControlNet to preserve more generative priors. This impairs the specific feature synchronization, especially the lip motion synchronization (Fig. 4 right, reduced Sync-C and Sync-D).

Table 3: User study results. Ours (MTV) clearly produces a higher score than state-of-the-art methods.

Subjective criteria	MM-Diffusion [5]	TempoTokens [1]	Xing *et al.* [3]	Ours (MTV)
Semantic consistency	0.96%	13.60%	11.28%	**74.16%**
Motion fluency	0.64%	8.96%	12.56%	**77.84%**
Overall preference	0.72%	12.00%	12.40%	**74.88%**

Table 4: Quantitative experiment results with alternative pre-trained components.

Method	FVD ↓	Temp-C (%) ↑	Text-C (%) ↑	Audio-C (%) ↑	Sync-C ↑	Sync-D ↓
CogVideoX+Wav2Vec	626.06	95.40	26.55	26.22	**3.17**	**9.43**
CogVideoX+Beats	598.53	95.91	26.25	25.28	3.02	9.52
Wan14B+Wav2Vec	**353.61**	**96.36**	**27.23**	**26.49**	3.08	9.56

5.3 User Study

To better evaluate our method from a human perception perspective, we conduct three subjective user study experiments in Tab. 3. We present videos generated by our method and all baselines to participants and ask them to choose the best one based on the following criteria: *(i)* **Semantic consistency.** How well the video content aligns with the text description. *(ii)* **Motion fluency.** The realism and temporal coherence of the motion. *(iii)* **Overall preference.** How good the holistic quality of the video is. For each study, we randomly select 50 text descriptions from the test set, and the evaluations are conducted by 25 volunteers. The table below shows the percentage of times each method is chosen as the winner. Our method is consistently favored by human observers and has achieved the highest scores across all three subjective criteria.

5.4 Analysis of Pre-trained Components

We evaluate the robustness of our proposed method by integrating it with alternative pre-trained components. Specifically, we test replacing the audio encoder (Wav2Vec/BEATs) and the video backbone (CogVideoX/Wan14B) in Tab. 4.

BEATs. Since Wav2Vec [58] is a common setting for speech encoding (*e.g.*, Hallo3 [7]), this baseline only replaces it with BEATs [65] for both the effects and music tracks. As shown in Tab. 4, this baseline achieves comparable (or slightly better) video-related metrics (*i.e.*, FVD and Temp-C) but shows a slight degradation on audio-related metrics (*i.e.*, Audio-C, Sync-C, and Sync-D), suggesting that our current choice of Wav2Vec [58] is a robust and effective one for this task.

Wan14B. Since Wan14B [21] shares a similar DiT-based structure with CogVideoX [10], we can integrate our proposed MST-ControlNet into it without architectural changes. Specifically, our interval feature injection and holistic style injection modules are added after each text cross-attention layer. The quantitative results below show this baseline achieves better performance on video- and text-related metrics (*i.e.*, FVD, Temp-C, and Text-C) due to the stronger capabilities of the Wan14B [21], while achieving comparable performance on all audio-related metrics (*i.e.*, Audio-C, Sync-C, and Sync-D).

5.5 Application

As presented in Fig. 5, our model support four typical scenarios: *(i)* By integrating text-to-video generative priors and learned audio-visual synchronized capabilities, our model can create vivid virtual characters. *(ii)* Given user-provided images and taking them as arbitrary keyframes, our model can drive the image according to the given audios. *(iii)* Although our model generates video segments of 49 frames, it can achieve long video generation by using the generated frame to initialize the next segment. *(iv)* Following training-free approaches [66], our model can generate scene transitions guided by providing time-varying text descriptions.

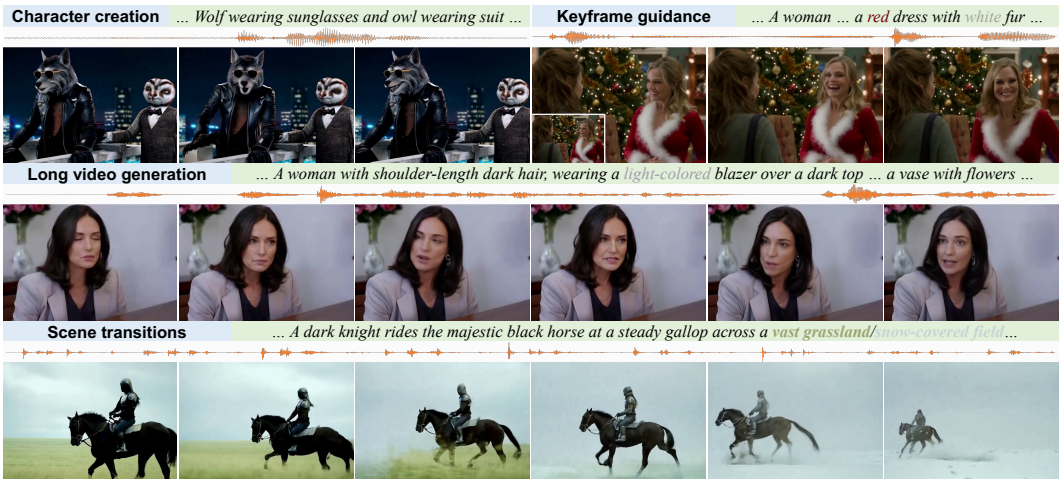

Figure 5: Examples of versatile application scenarios for our proposed MTV framework.

Figure 6: Examples of controllability study for text descriptions and demixed audios.

5.6 Controllability

As shown in Fig. 6, leveraging control from both text descriptions and the three demixed audio tracks (*i.e.*, speech, effects, music), our model can offer controllability across following four key aspects: *(i)* Modifying the text descriptions while keeping all audio tracks fixed allows the visual scene appearance to be edited without affecting the audio synchronization. *(ii)* Given a demixed speech track, the model enables precise control over the synchronized lip motion of the generated character. *(iii)* Similarly, with a demixed effects track, the model accurately synchronizes event timing with the sound effects. *(iv)* By changing the demixed music track, the model creates different visual moods for the generated video.

6 Conclusion

In this work, we presented MTV, a versatile framework for audio-sync video generation. MTV leverages generative priors from pretrained text-to-video models [10] and is trained on our contributed DEMIX dataset that provides sufficient cinematic videos with demixed audio tracks. Equipped with our proposed MST-ControlNet, MTV is able to independently control lip motion, event timing, and visual mood. Combined with a multi-stage training strategy for effective learning of complex audio-visual relationships, MTV achieves state-of-the-art performance across six evaluation metrics.

Limitation. Although our approach demonstrates the potential of using demixed audio tracks for precise video control, it is fundamentally limited by the scope of categories provided by upstream audio demixing techniques [53, 54]. We believe the capabilities of audio-sync video generation methods will further progress with advancements in audio demixing methods.

Acknowledgement. This work is supported by National Natural Science Foundation of China (Grant No. 62136001). We thank all the insightful reviewers for the helpful suggestions, and the colleagues at Beijing Academy of Artificial Intelligence for their support throughout this project.

7 Appendix

7.1 Task Differences

To further present the advantages of our MTV framework, we clarify the distinctions between our approach and other audio-relevant tasks. We discuss relevant methods published before the date of this paper submission (May 15th, 2025).

7.1.1 Audible Video Generation

Audio-sync video generation. Our method belongs to the topic of audio-sync video generation, which receives user-provided audios as inputs, offering the potential for free scene creation with optional text descriptions. With the recent advancements in video models, our comparisons focus on very recent methods (*e.g.*, TempoTokens [1] and Xing *et al.* [3]). Since TATS [2] does not provide the custom audio processing, and other methods [5, 6] are tailored for specific visual categories (*e.g.*, landscapes), we select the one with the higher citations [5] for finetuning to general scenarios. Among these, our method is the first to leverage demixed audio tracks for multi-stream control, achieving state-of-the-art performance across six metrics.

Video-audio joint generation. Different from our audio-sync video generation, video-audio joint generation task aims to generate videos with accompanying audios based on user-provided instructions, where most methods in this area are proposed recently [3, 5, 67–69]. Since the audio is co-generated with the video from a shared input (*e.g.*, text descriptions) that typically lacks explicit temporal control signals, *users typically have limited direct control over the precise event timing within the generated video.* While MM-Diffusion [5] discusses training-free strategies to adapt such joint generation methods for audio-sync video generation, our comparison results in Sec. 5.1 indicate that this adaptation approach still has room for improvement.

Summary. Both video-audio joint generation and audio-sync video generation belong to the audible video generation task. Although video-audio joint generation offers advantages in directly producing audible videos, *audio itself is inherently temporal and closely synchronized with the visual world, making it an ideal control signal for precise temporal guidance.* This makes it highly suitable for controllable video generation, unlocking potential applications (*e.g.*, bringing historical recordings to visual life and creating rich visual narratives for podcasts).

7.1.2 Audible Image Animation

Audio-driven image animation. Audio-driven image animation aims to generate dynamic visuals from a static image, synchronized with user-provided audios. Most of these methods [4, 24–27] handle general objects and scenarios but still struggle with specific feature synchronization (*e.g.*, for speech and events). Animating talking humans is another sub-topic, which requires a human image to be driven mainly by speech. While most methods in this area [70, 28–30] only focus on the head and facial expressions, a few recent methods [31, 32] extend to half-body or full-body generation. Compared to audio-sync video generation, while the reference image required by these methods allows for animating pre-defined subjects, this reliance may also limit the creation freedom for diverse and dynamic video generation.

Discussion with talking human methods. Since code for both CyberHost [31] and OminiHuman-1 [32] is unavailable, we additionally compare our method with SadTalker [71] and Hallo3 [7]. Since both SadTalker [71] and Hallo3 [7] can only animate the frontal face of a single person, it is infeasible to make a comprehensive evaluation even on our single character subset (as videos for single character also contain many frames without a clear frontal face). Consequently, we provide qualitative comparisons in Fig. 7. These results show that our method effectively demonstrates realistic human gestures and reasonable camera movement. In contrast, Hallo3 [7] presents a more static video (*e.g.*, less gesture and stable background), while SadTalker [71] only modifies the face and pastes the remaining regions directly from the source image. Notably, since both SadTalker [71] and Hallo3 [7] require an additional reference image, we take the reference image as the first frame to leverage our model's keyframe guidance capability for a fair comparison.

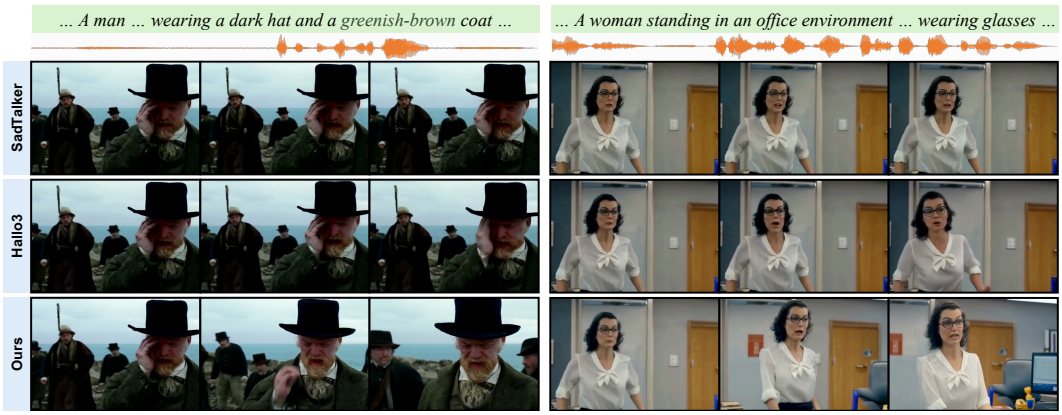

Figure 7: Visual comparison results with state-of-the-art methods for talking human.

7.2 Analysis of MST-ControlNet Depth

As presented in Sec. 4.2, we feed features into N interval interaction blocks within the MST-ControlNet for the interval-wise control. To investigate the impact of this hyperparameter N, we evaluate variants with different depths. The quantitative results presented in Tab. 5 show that increasing N consistently improves the overall visual quality (FVD) and temporal consistency (Temp-C). However, lip motion synchronization metrics demonstrate that they are improved until $N = 4$ before declining. Text-video (Text-C) and audio-video (Audio-C) consistency remain largely stable across different values of N. This suggests a potential trade-off between general video quality and specific lip motion synchronization when varying the depth of interval interaction blocks. Considering this trade-off, we choose $N = 4$ as the setting for our main reported results.

Table 5: Quantitative experiment results of comparison and ablation. ↑ (↓) means higher (lower) is better. Throughout the paper, best performances are highlighted in **bold**.

Method	FVD ↓	Temp-C (%) ↑	Text-C (%) ↑	Audio-C (%) ↑	Sync-C ↑	Sync-D ↓
$N = 1$	677.51	94.94	26.49	**26.32**	2.85	9.47
$N = 4$	626.06	95.40	**26.55**	26.22	**3.17**	**9.43**
$N = 8$	570.62	96.09	26.46	26.26	2.74	9.55
$N = 16$	**485.84**	**97.02**	26.44	26.25	2.42	9.45

7.3 Metrics Details

As described in Sec. 5.1, we adopt six metrics to quantitatively evaluate performance. We present their details below: *(i)* Frechét Video Distance (FVD) [60] is used to assess the video quality by computing the distance between feature distributions from real videos and generated videos. *(ii)* The Temporal consistency (Temp-C) is measured by calculating the cosine similarity between consecutive frame embeddings from the CLIP image encoder [61]. *(iii)* Text consistency (Text-C) is evaluated by cosine similarity between text descriptions and generated videos using VideoCLIP-XL [62]. *(iv)* Audio consistency (Audio-C) is evaluated by cosine similarity between input audios and generated videos using ImageBind [63]. *(v)* Sync-C and Sync-D [64] are common metrics used to evaluate lip motion synchronization.

Notably, AV-Align [1] is another potential metric for evaluating audio-video alignment. This metric detects energy peaks in audio [72] and motion peaks in video [73], respectively. It then validates whether a peak detected in one modality is also detected in the other within a three-frame temporal window, and vice versa. Although this metric is intuitive and reasonable, it seems unsuitable for evaluating the cinematic videos that our MTV framework focuses on. As shown in Tab. 6, real videos unexpectedly achieve the lowest score with this metric. As a result, we only report this metric in the supplementary materials.

Table 6: AV-Align scores for comparison methods. The higher scores are considered better in theory.

Method	MM-Diffusion [5]	TempoTokens [1]	Xing *et al.* [3]	Ours (MTV)	Real videos
AV-Align (%)	33.60	33.66	32.49	25.21	23.19

7.4 Dataset Details

Our dataset processing pipeline is illustrated in Fig. 8, with full processing details provided in Sec. 3 of the main paper. Additional dataset samples from our five subsets (*i.e.*, basic face, single character, multiple characters, sound event, and visual mood) are provided in an *anonymous* GitHub link [3]. Each sample includes a video with its corresponding demixed audio tracks, serving to clearly illustrate the concept of 'audio demixing'.

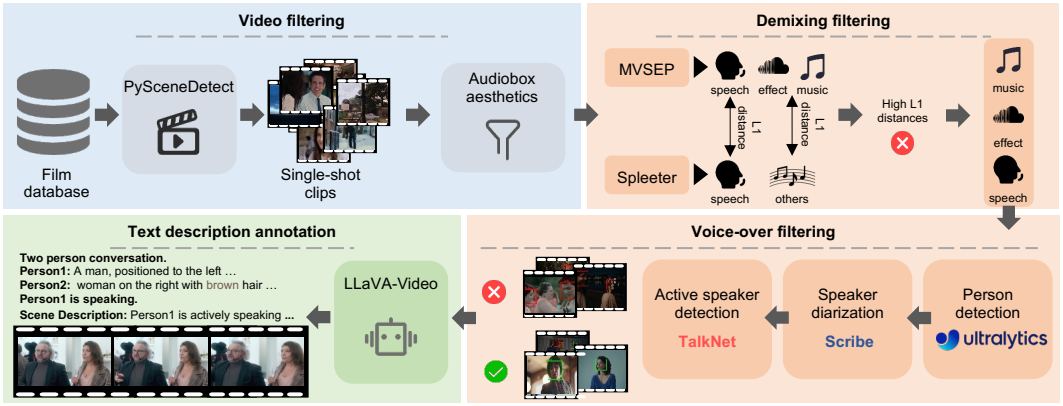

Figure 8: Dataset processing pipeline for our DEMIX dataset.

7.5 Evaluation on Additional Datasets

To demonstrate the generalization capabilities of our MTV framework, we follow TempoTokens [1] to conduct additional experiments on both the Landscape [74] and AudioSet-Drum [75] datasets. To make a fair comparison, we fine-tune all baseline methods (MM-Diffusion [5], TempoTokens [1], and Xing*et al.* [3]) on our DEMIX dataset using their official training schedules, and evaluate them on these separate datasets. As shown in Tab. 7, our MTV framework still achieves significantly better performance. Since neither dataset includes human talking, the lip synchronization metrics (*i.e.*, Sync-C and Sync-D) are not applicable for this evaluation.

Table 7: Quantitative comparison results in Landscape and AudioSet-Drum datasets.

Method	Landscape				Audio-Drum			
	FVD ↓	Temp-C ↑	Text-C ↑	Audio-C ↑	FVD ↓	Temp-C ↑	Text-C ↑	Audio-C ↑
MM-Diffusion [5]	807.65	94.74	14.66	16.59	1520.09	94.59	14.90	14.11
TempoTokens [1]	797.33	94.67	21.73	18.86	1512.97	94.28	23.18	15.59
Xing *et al.* [3]	838.03	94.71	21.04	18.70	1589.46	94.49	23.73	17.84
Ours (MTV)	**697.51**	**96.98**	**25.35**	**23.37**	**1511.53**	**97.50**	**25.62**	**39.61**

7.6 Organization of Supplementary Video

We provide a supplementary video to dynamically showcase our audio-sync video generation results. The video is structured as follows: *(i)* **Versatile capabilities across five scenarios.** We demonstrate

[3] `https://anonymous.4open.science/w/MTV-F4C4/`

five generation scenarios to show our capabilities in character-centric narrative, multi-character interaction, sound-triggered events, music-shaped ambiance, and camera movement. *(ii)* **Application across four typical scenarios.** We present four application scenarios for character creation, keyframe guidance, long video generation, and scene transitions. *(iii)* **Controllability across four key aspects.** We showcase four aspects to control the generated results, including appearance, lip motion, event timing, and visual mood. *(iv)* **Comparison with state-of-the-art methods.** We compare with relevant audio-sync video generation methods [1, 3, 5] to demonstrate our superior performance. *(v)* **Ablation study.** We present the ablation study results to demonstrate the effectiveness of our proposed modules. *(vi)* **Discussion with talking human methods.** We illustrate the task difference with talking human methods [71, 7], where our method animates humans with more realistic human gestures and reasonable camera movement. For a fair comparison with these reference-based methods, we take the reference image as the first keyframe to leverage our model's keyframe guidance capability.

References

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

[2] S. Ge, T. Hayes, H. Yang, X. Yin, G. Pang, D. Jacobs, J.-B. Huang, and D. Parikh, "Long video generation with time-agnostic vqgan and time-sensitive transformer," in *ECCV*, 2022.

[3] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen, "Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners," in *CVPR*, 2024.

[4] S. H. Lee, G. Oh, W. Byeon, J. Bae, C. Kim, W. J. Ryoo, S. H. Yoon, J. Kim, and S. Kim, "Sound-guided semantic video generation," in *ECCV*, 2022.

[5] L. Ruan, Y. Ma, H. Yang, H. He, B. Liu, J. Fu, N. J. Yuan, Q. Jin, and B. Guo, "MM-Diffusion: Learning multi-modal diffusion models for joint audio and video generation," in *CVPR*, 2023.

[6] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li, "TA2V: Text-audio guided video generation," *TMM*, 2024.

[7] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu, "Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer," in *CVPR*, 2025.

[8] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *ICCV*, 2021.

[9] B. M.-K. Ng, S. R. Sudhoff, H. Li, J. Kamphuis, T. Nadolsky, Y. Chen, K. Y.-J. Yun, and Y.-H. Lu, "Visualize music using generative arts," in *CAI*, 2024.

[10] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, *et al.*, "CogVideox: Text-to-video diffusion models with an expert transformer," in *ICLR*, 2025.

[11] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *ICCV*, 2023.

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

[13] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, "Latent video diffusion models for high-fidelity long video generation," *arXiv preprint arXiv:2211.13221*, 2022.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[15] T. Soo Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *CVPR workshops*, 2017.

[16] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?," in *ICML*, 2021.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

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

[19] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023.

[20] L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, V. Birodkar, A. Gupta, X. Gu, A. G. Hauptmann, B. Gong, M.-H. Yang, I. Essa, D. A. Ross, and L. Jiang, "Language model beats diffusion – tokenizer is key to visual generation," in *ICLR*, 2024.

[21] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, *et al.*, "Wan: Open and advanced large-scale video generative models," *arXiv preprint arXiv:2503.20314*, 2025.

[22] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, *et al.*, "Hunyuanvideo: A systematic framework for large video generative models," *arXiv preprint arXiv:2412.03603*, 2024.

[23] G. Ma, H. Huang, K. Yan, L. Chen, N. Duan, S. Yin, C. Wan, R. Ming, X. Song, X. Chen, *et al.*, "Step-Video-T2V technical report: The practice, challenges, and future of video foundation model," *arXiv preprint arXiv:2502.10248*, 2025.

[24] M. Chatterjee and A. Cherian, "Sound2Sight: Generating visual dynamics from sound and context," in *ECCV*, 2020.

[25] G. Le Moing, J. Ponce, and C. Schmid, "CCVS: Context-aware controllable video synthesis," in *NeurIPS*, 2021.

[26] Y. Jeong, W. Ryoo, S. Lee, D. Seo, W. Byeon, S. Kim, and J. Kim, "The power of sound (TPoS): Audio reactive video generation with stable diffusion," in *ICCV*, 2023.

[27] L. Zhang, S. Mo, Y. Zhang, and P. Morgado, "Audio-synchronized visual animation," in *ECCV*, 2024.

[28] J. Jiang, C. Liang, J. Yang, G. Lin, T. Zhong, and Y. Zheng, "Loopy: Taming audio-driven portrait avatar with long-term motion dependency," in *ICLR*, 2025.

[29] H. Wei, Z. Yang, and Z. Wang, "AniPortrait: Audio-driven synthesis of photorealistic portrait animation," *arXiv preprint arXiv:2403.17694*, 2024.

[30] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023.

[31] G. Lin, J. Jiang, C. Liang, T. Zhong, J. Yang, and Y. Zheng, "CyberHost: A one-stage diffusion framework for audio-driven talking body generation," in *ICLR*, 2025.

[32] G. Lin, J. Jiang, J. Yang, Z. Zheng, and C. Liang, "OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models," *arXiv preprint arXiv:2502.01061*, 2025.

[33] W. Xuanchen, W. Heng, L. Dongnan, and W. Cai, "Dance any beat: Blending beats with visuals in dance video generation," in *WACV*, 2025.

[34] Z. Chen, H. Xu, G. Song, Y. Xie, C. Zhang, X. Chen, C. Wang, D. Chang, and L. Luo, "X-dancer: Expressive music to human dance video generation," *arXiv preprint arXiv:2502.17414*, 2025.

[35] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *CVPR*, 2021.

[36] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019.

[37] D. Jeong, S. Doh, and T. Kwon, "Träumerai: Dreaming music with stylegan," *arXiv preprint arXiv:2102.04680*, 2021.

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

[39] S. Hong, W. Im, and H. S. Yang, "Content-based video-music retrieval using soft intra-modal structure constraint," *arXiv preprint arXiv:1704.06761*, 2017.

[40] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[41] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[42] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VggSound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[43] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *ICCV*, 2021.

[44] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Li, G. Chen, X. Chen, Y. Wang, *et al.*, "InternVid: A large-scale video-text dataset for multimodal understanding and generation," in *ICLR*, 2024.

[45] H. Zhu, W. Wu, W. Zhu, L. Jiang, S. Tang, L. Zhang, Z. Liu, and C. C. Loy, "CelebV-HQ: A large-scale video facial attributes dataset," in *ECCV*, 2022.

[46] W. Wu, M. Liu, Z. Zhu, X. Xia, H. Feng, W. Wang, K. Q. Lin, C. Shen, and M. Z. Shou, "MovieBench: A hierarchical movie level dataset for long video generation," *arXiv preprint arXiv:2411.15262*, 2024.

[47] M. Bain, A. Nagrani, A. Brown, and A. Zisserman, "Condensed movies: Story based retrieval with contextual embeddings," in *ACCV*, 2020.

[48] R. Ghermi, X. Wang, V. Kalogeiton, and I. Laptev, "Short film dataset (SFD): A benchmark for story-level video understanding," *arXiv preprint arXiv:2406.10221*, 2024.

[49] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "VideoCrafter2: Overcoming data limitations for high-quality video diffusion models," in *CVPR*, 2024.

[50] B. Castellano, "Video cut detection and analysis tool." https://github.com/Breakthrough/PySceneDetect.

[51] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, *et al.*, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025.

[52] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li, "Video instruction tuning with synthetic data," *arXiv preprint arXiv:2410.02713*, 2024.

[53] R. Solovyev, "Cinematic sound demixing." https://github.com/ZFTurbo/MVSEP-CDX23-Cinematic-Sound-Demixing.

[54] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, 2020.

[55] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO." https://github.com/ultralytics/ultralytics.

[56] Elevenlabs, "Meet scribe." https://elevenlabs.io/blog/meet-scribe.

[57] S. Beliaev and B. Ginsburg, "TalkNet 2: Non-autoregressive depth-wise separable convolutional model for speech synthesis with explicit pitch and duration prediction," *arXiv preprint arXiv:2104.08189*, 2021.

[58] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *NeurIPS*, 2020.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

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

[62] J. Wang, C. Wang, K. Huang, J. Huang, and L. Jin, "VideoCLIP-XL: Advancing long description understanding for video clip models," *arXiv preprint arXiv:2410.00741*, 2024.

[63] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *CVPR*, 2023.

[64] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *ACCV Workshops*, 2017.

[65] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *ICML*, 2023.

[66] M. Cai, X. Cun, X. Li, W. Liu, Z. Zhang, Y. Zhang, Y. Shan, and X. Yue, "DitCtrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation," *arXiv preprint arXiv:2412.18597*, 2024.

[67] Y. Mao, X. Shen, J. Zhang, Z. Qin, J. Zhou, M. Xiang, Y. Zhong, and Y. Dai, "TAVGBench: Benchmarking text to audible-video generation," in *ACM Multimedia*, 2024.

[68] A. Hayakawa, M. Ishii, T. Shibuya, and Y. Mitsufuji, "MMDisco: Multi-modal discriminator-guided cooperative diffusion for joint audio and video generation," in *ICLR*, 2025.

[69] K. Wang, S. Deng, J. Shi, D. Hatzinakos, and Y. Tian, "AV-DiT: Efficient audio-visual diffusion transformer for joint audio and video generation," *arXiv preprint arXiv:2406.07686*, 2024.

[70] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang, "Hallo2: Long-duration and high-resolution audio-driven portrait image animation," in *ICLR*, 2025.

[71] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang, "SadTalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation," in *CVPR*, 2023.

[72] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," Citeseer.

[73] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, 1981.

[74] S. H. Lee, G. Oh, W. Byeon, C. Kim, W. J. Ryoo, S. H. Yoon, H. Cho, J. Bae, J. Kim, and S. Kim, "Sound-guided semantic video generation," in *ECCV*, 2022.

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