YingVideo-MV: Agent-Based Music-Driven Video Generation

AI Lab Team
Giant Network AI Lab
Shanghai, China

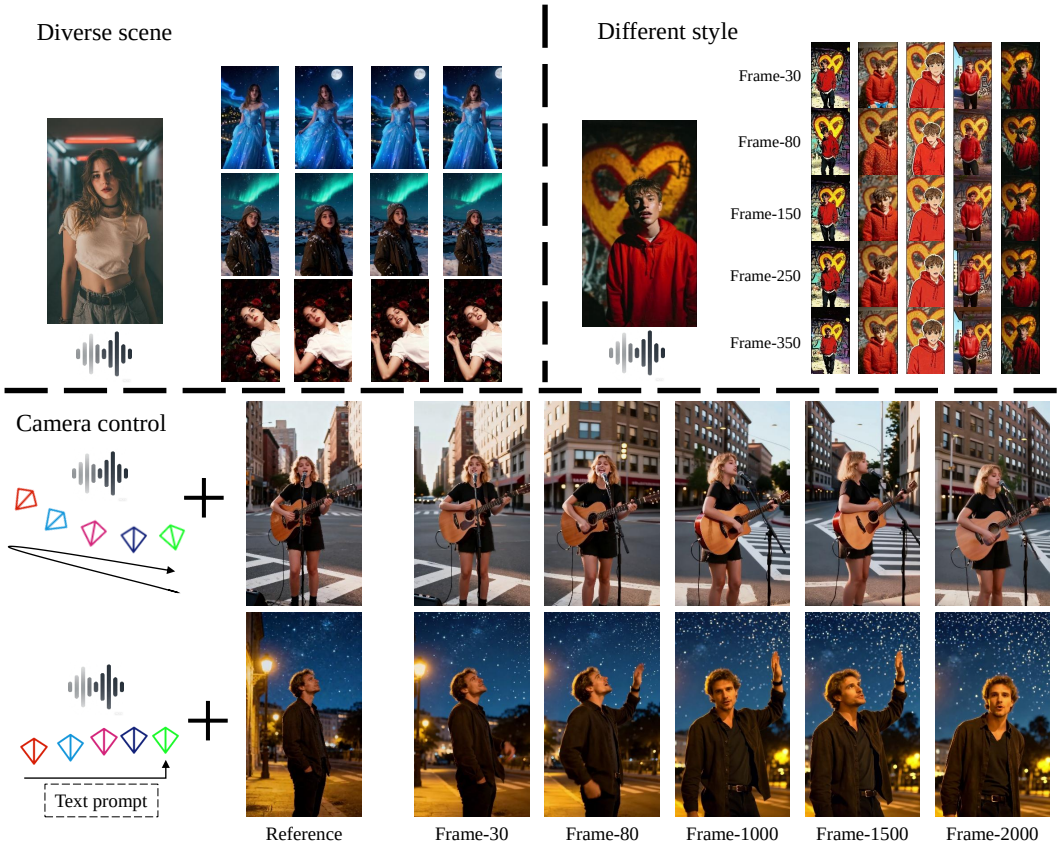

Figure 1: Conditioned on a portrait image, text, and audio input, YingVideo-MV can generate and edit portraits with strong identity consistency, expressive facial features, natural body dynamics, and camera movement. The results demonstrate vivid emotions, rich camera movements, and precise lip-syncing, while also generating animation and cartoon styles.

Abstract

While diffusion model for audio-driven avatar video generation have achieved notable process in synthesizing long sequences with natural audio-visual synchronization and identity consistency, the generation of music-performance videos with camera motions remains largely unexplored. We present YingVideo-MV, the first cascaded framework for music-driven long-video generation. Our approach integrates audio semantic analysis, an interpretable shot planning module (MV-

Director), temporal-aware diffusion Transformer architectures, and long-sequence consistency modeling to enable automatic synthesis of high-quality music performance videos from audio signals. We construct a large-scale Music-in-the-Wild Dataset by collecting web data to support the achievement of diverse, high-quality results. Observing that existing long-video generation methods lack explicit camera motion control, we introduce a camera adapter module that embeds camera poses into latent noise. To enhance continuity between clips during long-sequence inference, we further propose a time-aware dynamic window range strategy that adaptively adjust denoising ranges based on audio embedding. Comprehensive benchmark tests demonstrate that YingVideo-MV achieves outstanding performance in generating coherent and expressive music videos, and enables precise audio-motion-camera synchronization.

1 Introduction

Music-performing avatars have demonstrated significant research and application value across diverse visual media creation domains, including cinema, music videos (MV), vlogs, and advertisements. These models Yang et al. (2025a); Ding et al. (2025); Tu et al. (2025) achieve temporally coherent facial expressions, lip movements Fei et al. (2025), and body poses Wang et al. (2025b) through joint modeling of multi-modal inputs (images, speech, and text), enabling dynamic visualization of musical semantics and emotions. As a form of digital performance, singing-capable virtual beings can convey emotional intent with high fidelity, opening new possibilities for immersive music content expression. Recently, Video Diffusion Transformers (DiTs) Chen et al. (2025b); Meng et al. (2025); Jiang et al. (2024a); Tian et al. (2024); Wei et al. (2025) have emerged as a unified generative paradigm, widely appilied to synthesize highly expressive visual content conditioned on multi-modal signals (e.g., images, speech, and text prompts). Using their powerful spatio-temporal modeling and cross-modal integration capabilities, piror works Wang et al. (2024a); Yang et al. (2025b); Qiu et al. (2025); Zheng et al. (2024); Lin et al. (2025); Jiang et al. (2025) have made a significant process in precise facial expression and lip-sync alignment, natural body motion generation, and large-scale data scalability.

However, substantial challenges remain when extending tasks to the complex domain of music performance video generation. First, effective cinematographic language and visual narrative design are critical in music performance. Camera movement patterns, depth-of-field transitions, and compositional rhythms directly influence audience immersion and emotional delivery. Yet, existing models Yang et al. (2025a); Tu et al. (2025); Chen et al. (2025b); Cui et al. (2025b); Hu et al. (2025); Ren et al. (2025); Kim et al. (2025) often lack systematic modeling of cinematographic principles and scene composition, resulting in monotonous framing, rigid motion artifacts, and poor rhythmic coordination. Current music-driven video generation methods predominantly rely on single-viewpoint or static-scenario configurations, with limited capacity to model multi-camera switching and spatial depth perception—factors crucial for artistic expression and photorealism. Second, cross-modal temporal and rhythmic alignment remains a key bottleneck. Music-conditioned video content must precisely synchronize camera motions, performance rhythms, and musical elements (rhythm, melody, emotion) across temporal dimensions. Ideal systems should not only "understand" audio signals and "interpret" textual semantics but also grasp latent emotional intent and narrative logic to generate empathetic, contextually coherent visuals. Moreover, exisiting methods Gu et al. (2025); Gan et al. (2025) primarily employ frame-wise sequential prediction, which suffers from content driff and temporal coherence degradation in long-sequence generation. This leads to progressive distortion in identity preservation, pose stability, and expression consistency, ultimately limiting the capacity for high-quality, extended-duration music performance synthesis.

To address these challenges, we introduce YingVideo-MV, a cascaded music-driven video generation framework that integrates audio analysis with a temporal-aware diffusion Transformer architecture. This framework enables high-quality talking portrait video generation from music signals while establishing a large-scale Music-in-the-Wild Dataset (MusicMV-Dataset) containing diverse performances. Inspired by the unified perception-action capabilities of intelligent agents Team et al. (2023); Xu et al. (2025); Ding et al. (2025); Hong et al. (2025), we design an MV Director Module that synthesizes multi-modal inputs into structured shot list information. This shot list encapsulates key elements including initial frame composition, rhythmic cues, and character motion patterns, ensuring alignment between generated content and the intended narrative logic and musical aesthetics. The

framework operates through a two-stage pipeline. First, a shot list video is generated from the input music, which then conditions the parallel generation of multiple sub-clips using a music-driven video model. These sub-clips are temporally aligned and visually fused to produce a complete, coherent music performance video with emotional consistency. At the core of our music-dirven video model lies a temporal-aware diffusion Transformer that simultaneously handles lip-sync alignment, facial expression generation, and camera motion synthesis, achieving natural coordination among performer actions, musical rhythm, and camera dynamic. To tackle long-sequence coherence challenges, we propose a dynamic window inference technique that enables seamless transitions and visual consistency in extended video generation by smoothly propagating visual states across adjacent segments. Meanwhile, the incorporation of the DPO training strategy during the optimization process leads to enhanced lip synchronization and improved visual quality.

We summarise our key contributions as follows:

- **MV-Director Framework with Unified instruction Planning.** We propose a global planning module that integrates multi-modal inputs (text prompts, camera trajectories, initial frames, and music segments) into unified semantic instructions. This mechanism elevates music-driven video generation from low-level cue tracking to deep comprehension of musical cinematographic language with explicit control over narrative logic and aesthetic composition.
- **Cascaded Portrait Video Synthesis Pipeline.** We design a two-stage generation framework where first stage establishes high-level semantic and shot planning via MV-Director, while the second stage achieves local dynamics through a temporal-aware diffusion Transformer. This cascaded architecture effectively balances global consistency with local expressiveness, enabling long-sequence video generation with coherent camera movements and smooth transitions, significantly enhancing narrative fluency and musical expressiveness.
- **High-fidelity Generation across diverse scenarios.** YingVideo-MV generates high-fidelity, temporally conherent portrait videos across diverse scenarios. It achieves precise lip-sync alignment, rich facial expression variations, and camera movements synchronized with musical rhythms. The framework demonstrates superior generalization and artistic controllability, providing a robust foundation for multi-modal content generation applications.

Our method produces professional-quality music videos with camera movements perfectly synchronized with the music beat, and highly consistent character portrayals and lip-syncing. YingVideo-MV provides a practical technological approach to automated, high-quality music video creation.

2 Related Work

To enable the synchronized generation of music-driven videos with controllable camera trajectories, we explore three research paradigms: audio-driven video generation, controllable camera poses, and joint audio-visual generation of camera trajectories.

2.1 Audio-driven Video Generation

This dominat paradigm typically employs videos diffusion models for visual content synthesis. Early attempts Blattmann et al. (2023); Chen et al. (2023b); Zeng et al. (2024) utilized U-Net architectures to demonstrate feasibility, but these models suffered from limited capacity to generate high-fidelity frames with temporal coherence. The advent of Diffusion Transformers Peebles, Xie (2023) marked a pivotal advancement, with WAN Wan et al. (2025) leveraging their scalable architecture to process spatio-temporal segments at scale, significantly enhancing video coherence and generation quality. Despite these advancements in achieving precise lip synchronization Chen et al. (2025a); Cui et al. (2025b); Lin et al. (2025); Peng et al. (2024); Yariv et al. (2024), two critical limitations persist: (1) weak semantic alignment between audio-visual modalities, and (2) the fixed camera perspetive constraint that struggles with occlusion handling and view-dependent deformations under dynamic camera movements. To address these challenges, we introduce an audio-conditioned camera trajectory generation method for rapidly synthesizing animations of camera motion.

2.2 Controllable Camera Poses

Precise camera motion control is critical for music MV generation. Early approaches employ fine-tuning techniques such as LoRAs Hu et al. (2022) in AnimateDiff Guo et al. (2023) to manage specific motion types. However, these approaches offer only limited precision. Recent methods like MotionCtrl Wang et al. (2024b) and its successors Kuang et al. (2024); Xu et al. (2024); He et al. (2024a) directly condition video generation on extrinsic camera parameters, mapping sparse brush strokes to Gaussian representations. 4D scene generation approaches Watson et al. (2024); Wu et al. (2025); Sun et al. (2024) provide inherent camera control through spatio-temporal field modeling, though their synthesis quality currently lags behind specialised video generation models. Motion-I2V Shi et al. (2024) and MOFA Niu et al. (2024) introduce two-stage pipelines that first predict motion from strokes then generate videos conditioned on the predicted motion, but this requires maintaining two independent models, increasing system complexity. Recently, TORA Zhang et al. (2025b) achieved state-of-the-art results by leveraging a DiT backbone with camera-aware positional encodings, while Go-with-the-Flow Burgert et al. (2025) explores dense trajectory control through optical flow warping mechanisms. However, the above methods still cause camera jumps in the generated video. Therefore, we propose to incorporate camera coding into noise latents over time steps to improve the camera controllability of the generated video.

2.3 Joint Audio-visual Generation of Camera Trajectories

While prior approaches focus on deploying video diffusion models trained on millions of in-the-wild videos for controllable generation, a parallel line of research emphasizes capturing high-dimensional volumentric representations through complex data pipelines to enable fine-grained control over video synthesis. By using synchronized multi-camera setups, these methods reconstruct 3D/4D representations (e.g., meshes Cagniart et al. (2010); Beeler et al. (2011); Fyffe et al. (2011), NeRFs Lombardi et al. (2019); Işık et al. (2023), or Gaussian splatting He et al. (2024b); Jiang et al. (2024b); Luiten et al. (2024)) for camera trajectory planning, LeviTor Wang et al. (2025a) clusters segmentation masks into sparse points enhanced with depth information, but the lack of correspondence modeling and U-Net-based architecture limitations constrain its performance. FlexTraj Zhang et al. (2025c) introduces a multi-granularity, pose-agnostic trajectory control framework for image-to-video generation, enabling flexible camera motion specification without explicit 3D reconstruction. Moreover, all prior approaches assume trajectories are aligned with the first frame, which restricts their applicability. To address these limitation, we propose YingVideo-MV, a cascaded video generation framework, which integrates camera trajectory and audio into DiT backbone.

3 Method

The work aims to develop a framework for photorealistic portrait animation generation based on music audio supplemented with auxliary multimodal inputs (including text, static images, and video references). The proposed framework would generate temporally coherent output sequences that maintain the target object's visual identity by the conditional inputs under dynamic camera perspectives, while simultaneously achieving precise alignment with both musical rhythm patterns and speech-related articulatory features. Key animation characteristics of generated video include head movements, gestural dynamics, facial expression variations, and lip-reading consistency synchronised with audio phonemes.

3.1 Preliminaries

Flow Matching. Flow Matching Lipman et al. (2022) learns a vector field via conditional flow matching, then transforms the initial noise distribution into target samples through forward ODE integration along stochastic path, entirely bypassing explicit density estimation or inverse transformation procedures. Recent advances Tan et al. (2025); Esser et al. (2024); Wan et al. (2025); Fei et al. (2025) have demonstrated significant performance gains by operating in the latent space of pre-trained autoencoders. Unlike conventional text-to-video models that rely solely on textual conditioning, our approach integrates multi-modal conditioning information comprising driving audio sequences(A), static portrait images(I), reference video clips(V), and associated textual prompts(T). Crucially, during training, our framework learns to transform random noised camera pose vector sequences

($\mathcal{C}$) into structured and stable dynamic camera trajectory sequences through a decicated denoising progress,

$$L_{mse} = \mathbb{E}_{z_0, z_1, t\sim[0,1]}[||v_t - u_\theta(z_t, t, T, I, V, A, \mathcal{C})||_2^2], \tag{1}$$

where u_θ is a trainable denoising net. z_1 and z_0 notes the latent embedding of the training sample and the initialized noise sampled from the Gaussian distribution $\mathcal{N}(0, 1)$. z_t is the training sample constructed using a linear interpolation. Velocity $\mathrm{v}_t = dz_t/dt = z_1 - z_0$ serves as the regression target for the model.

Video Diffusion Transformers. Diffusion Transformers represents a class of generative models built upon the Transformer architecture Peebles, Xie (2023), demonstrating superior performance in video synthesis tasks through the implementation of full spatio-temporal attention across three dimensions. The architecture paradigm enables explicit modeling of long-range dependencies in both spatial and temporal domains, achieving state-of-the-art results in photorealistic video generation.

Low-rank adaptation (LoRA). Low-Rank Adaptation (LoRA) Hu et al. (2022) is a parameter-efficient fine-tuning strategy. Instead of updating all parameters during fine-tuning, LoRA injects a pair of low-rank matrices into the existing weight layers and restricts optimization to these additional parameters. By freezing the original weights and training only a small number of low-rank components, LoRA effectively mitigates catastrophic forgetting Kirkpatrick et al. (2017) while greatly reducing computational overhead. More specifically, the inserted low-rank matrices act as a residual update to the pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$. The resulting adapted weight can be formulated as follows:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W} = \mathbf{W} + AB^T, \tag{2}$$

where $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{n\times r}$ denote the two low-rank matrices, and r represents the rank hyperparameter controlling the adaptation capacity. In practical implementations, LoRA modules are typically integrated only into the attention layers of transformer architectures, further decreasing both fine-tuning time and GPU memory consumption.

Camera Representation. Following prior works Chen et al. (2023a); Kant et al. (2024); He et al. (2024a); Xu et al. (2024), we employ the Plücker embedding Sitzmann et al. (2021) to represent camera poses, as it offers both a strong geometric interpretation and fine-grained per-pixel camera encoding.Specifically, given the camera extrinsic matrix $\mathbf{E} = [\mathbf{R}; \mathbf{t}] \in \mathbb{R}^{3\times4}$, where $\mathbf{R}$ and $\mathbf{t}$ denote the rotation and translation components respectively, and the intrinsic matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$, we derive the Plücker embedding for each image pixel (u, v) as $\mathbf{p} = (\mathbf{o} \times \mathbf{d}', \mathbf{d}')$.Here, $\mathbf{o}$ represents the camera center in world coordinates, the symbol "$\times$" denotes the cross product, and the ray direction from the camera origin toward the pixel is computed as $\mathbf{d} = \mathbf{R}\mathbf{K}^{-1}[u, v, 1]^T + \mathbf{t}$. We then normalize $\mathbf{d}$ to obtain $\mathbf{d}'$. Finally, the Plücker embedding for frame i is expressed as $\mathbf{P}_i \in \mathbb{R}^{6\times H\times W}$, where h and w correspond to the spatial resolution of the associated visual tokens.

3.2 Cascaded Generation Pipeline

As shown in Figure 2, we formulate the task of multimodal music video (MV), creation as a sequential decision-making problem. The MV-Director agent operates in an environment defined by a user-specified high-level goal G — such as narrative intent, visual style , emotional tone, or character design — and a toolbox T containing various multimodal operations, including music analysis, scene planning, image/video generation, editing, and refinement modules. The agent aims to synthesize a sequence of actions $A = (a_1, a_2, \ldots, a_N)$ that transforms the initial state s_0 (consisting of music, user prompts, and optional reference images or videos) into a final state s_N that satisfies the MV creation goal G. The core challenge is one fold — designing a rich and accurate toolbox T that covers the full spectrum of MV production needs.

Toolbox T. The toolbox T includes audio separation, audio understanding, trajectory generation, image generation, video generation, and video editing. These tools allow the agent to analyze music, design scenes, synthesize visuals, and iteratively refine the video.

Music source segmentation. To segment raw music into semantically meaningful clips, we iteratively identify local maxima in the onset strength of beats as potential boundaries, ensuring that the resulting

Figure 2: Illustration of YingVideo-MV's cascaded generation Pipeline. Our framework integrates multimodal inputs (music, text, and images) to enable segmented generation of music-performing portrait videos under the guidance of a global planning module. The planning agent strategically invokes specialized tools according to sub-task requirements, ultimately generating three core outputs conditioned on initial-frame specifications: (1) photorealistic music-performing portrait images, (2) coherent dynamic camera trajectories, and (3) synchronized audio sequences aligned with visual performance cues.

average segment duration approximates one musical bar:

$$\Delta_{bar} = 4\Delta_{beat} = 4 \cdot \frac{60}{\text{bpm}}. \tag{3}$$

The segmentation satisfies two principles: (1) According to the *cut-to-the-beat* rule, secne boundaries should conincide with strong beats; (2) The duration of MV scenes typically correlates with the bar length Prétet et al. (2021).

Music understanding. Inspired by multimodal large language models (MLLMs) Bai et al. (2025b); Hong et al. (2025); Qi et al. (2025), we unify music-related evidence into a shared semantic space to provide high-level control signals for global MV planning. Specifically, we employ Qwen 2.5-Omni Xu et al. (2025) to extract both the transcription and the emotional attributes from each music segment, enabling the system to generate scene-level script content that matches the musical style, mood, and dynamics.

Trajectory generation. Camera trajectory design is critical for MV production, as it directly influences visual storytelling, shot composition and emotional pacing. Existing methods often rely on geometric heuristics or learning-based approaches that lack textual alignment or fine-grained control. Following recent advances in cinematography modeling (e.g., GenDoP Zhang et al. (2025a)), our trajectory generator leverages scene content and script-level guidance to produce smooth, context-aware camera movements. This module generates expressive and musically aligned camera paths, supporting shot framing, depth transitions, and narrative continuity throughout the MV.

Image generation, **video generation**, and **video editing**. This module handles the creation and refinement of visual assets for assets for each scene. Image generation is supported by open-source models such as Flux and SDXL, or closed-source APIs like MidJourney for text-to-image synthesis and reference-guided consistency. Video generation converts images and trajectories into temporally conherent clips, leveraging the S2V model for smooth, music-aligned sequences. Finally, video editing performs post-processing, including video and audio segment concatenation, subtitle addition,

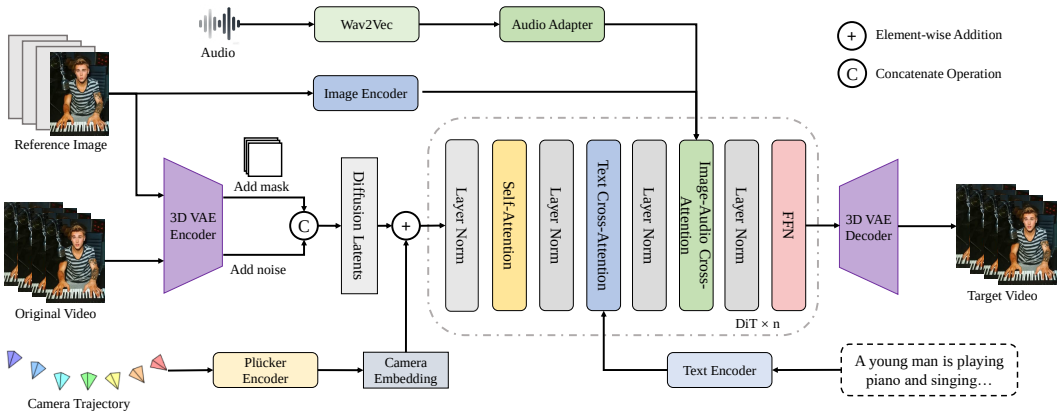

Figure 3: Illustration of Video Generation Model Architecture. Embeddings from the image and text encoders are injected into each block of the DiT. Given audio input, we leverage Wav2Vec to extract audio embeddings, while the camera trajectory is encoded and incorporated into the diffusion latent. To model the joint audio–latent representation, audio embeddings are fed into an audio adapter, the outputs of which are injected into the DiT via cross-attention.

and overall stylistic refinement ensuring narrative and visual consistency across the entire music video.

3.3 Model Architecture

As illustrated in Figure 3, the S2V module of YingVideo-MV builds upon the widely adopted WAN 2.2 Wan et al. (2025) framework and follows established research paradigms. Audio inputs are first processed through Wav2Vec to extract audio embeddings, which are subsequently refined using StableAvatar Tu et al. (2025)'s audio adapter to mitigate potential distribution mismatches. These enhanced embeddings are then fed into the denoising DiT pipeline. Reference images are processed via two parallel pathways: (1) Temporal axis padding with zero-filled frame is followed by latent encoding through a frozen 3D VAE encoder. The resulting latent codes are concatenated along the channel dimension with compressed video frames and binary masks (1 for the first frame, 0 otherwise); (2) Image embeddings are generated via a CLIP image encoder and injected into every image-audio cross-attention block of the denoising DiT to regulate visual appearance. During inference, original input video frames are replaced with the sum of random noise and camera pose encodings, as detailed in Sec. 3.4. A dynamic weighted silding window denoising strategy is introduced to enhance video smoothness in long-sequence generation by fusing latent information, as detailed in Sec. 3.5. Additionally, Direct Preference Optimization (DPO) is applied to align the generated portraits with human aesthetic and perceptual preferences, as detailed in Sec. 3.6

3.4 Camera Controlled Video Generation

After obtaining the Plücker embedding Sitzmann et al. (2021) $\mathbf{P}_i$ that encodes the camera pose of the i-th frame, we represent the complete camera trajectory of a video clip as a sequence of Plücker embeddings $\mathbf{P} \in \mathbb{R}^{L\times 6\times H\times W}$, where L denotes the video clip length. To inject camera information into the video generation backbone and enable explicit camera control, inspired by previous works He et al. (2024a); Bai et al. (2025a), we employ an adapter module to project the camera embeddings so that their tensor shape matches that of the noisy latent. The projected embeddings are then fused with the latent via element-wise addition before being fed into the DiT model. This design provides a simple yet effective mechanism for conditioning video generation on camera motion. The camera adapter architecture is composed of a sequence of `PixelUnshuffle`, `Conv2d`, and `ResidualBlock` layers.

3.5 Timestep-aware Dynamic Window Range Strategy

For long-video generation strategies, traditional approaches first denoise entirely one video clip latent, then moving on to the next clip. To maintain temporal coherence, the last few frames of the previous

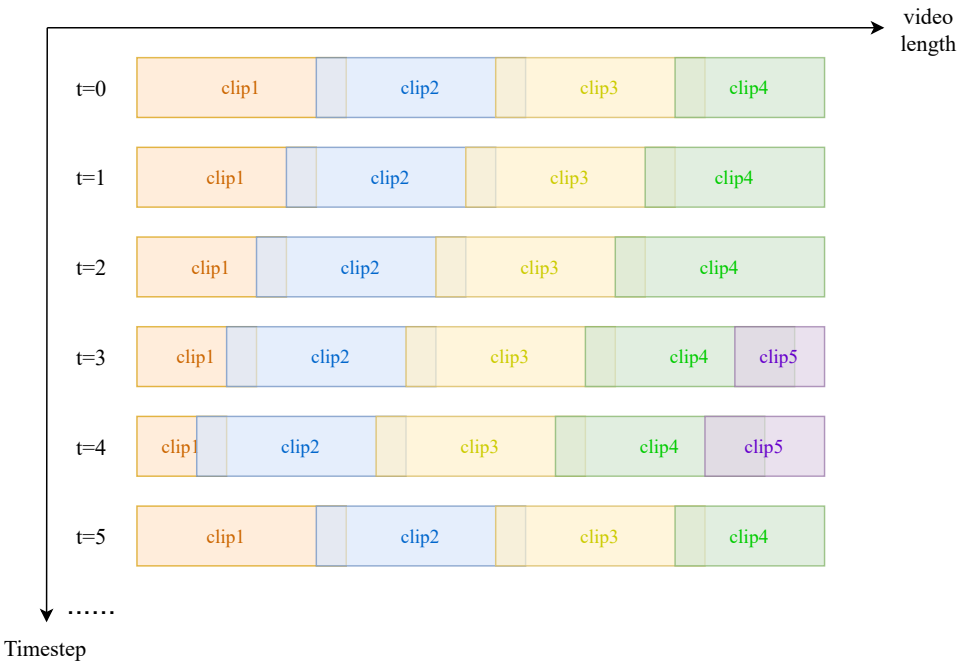

Figure 4: Timestep-aware dynamic window range strategy. From top to bottom, each row represents a denoising process at one timestep. Within each row, each clip of different color represents the segmentation of the long video. There are overlapping areas between each clip. At t=3, the last clip expands its overlap with the preceding clip to satisfy the minimum clip-length constraint. At t=5, the starting offset is reset because the offset accumulated in the previous timestep has reached its maximum allowable value.

clip are used as motion frames to provide conditioning for the subsequent clip. Recent works (e.g., Sonic Ji et al. (2025) and Stable Avatar Tu et al. (2025)) adopt a different long-video generation strategy: They both firstly process the entire sequence of latent clips at each diffusion timestep, then move to next timestep. Additionally, Sonic shifts the starting position at every timestep to enlarge the contextual receptive field of each frame, but its clips have no overlap. In contrast, Stable Avatar restarts denoising from the beginning at every timestep without shifting the starting position, but its clips have overlap.

Unlike Sonic , we do not employ a rolling strategy; that is, the initial frames and the final frames are not merged into a single clip for denoising, as these two portions of the video may have undergone substantial visual changes and are therefore unsuitable for self-attention. And unlike Stable Avatar, our method does shift the starting position at each timestep.

Furthermore, our Timestep-aware dynamic window range strategy pays special attention to the first and last clips. After several steps of shifting, the number of frames in the first clip may become very small. We observe that denoising within such a small window degrades generation quality, so we introduce a constraint: once the first-clip length shrinks to a threshold, the next timestep resets to zero starting offset like timestep 0. Similarly, the last clip may also become too short. Our solution is to extend it forward until it reaches the minimum clip length requirement, which increases the number of overlapping frames between the last and the second-last clips.

As illustrated in Figure 4 and Algorithm 1, our Timestep-aware dynamic window range strategy consists of two nested loops. The outer loop is inverse diffusion process, while the inner loop is sliding window process that model predicts for the each clip on the audio conditions. Within the inner loop, our strategy divides the video into different clips, the position and length of each clip dynamically change at each time step.

Algorithm 1 Timestep-aware dynamic window range strategy

Ensure: Video generator $\mathrm{G}(\cdot)$ with window length f, audio $c_a^{[0,l]}$, camera $c_c^{[0,l]}$, text c_t, image c_i, steps T.

Require: Final denoised latent $z_0^{[0,l]}$.

Initialize noisy latent $z_T^{[0,l]}$, shift offset α, shift step p, max offset m, min clip length n, overlap length o
for $t = T, \ldots, 1$ **do**
  if $\alpha > m$ **then**
    $\alpha \leftarrow 0$ ▷ reached max offset, reset shift
  end if
  $s \leftarrow -\alpha$
  $e \leftarrow s + f$
  $s \leftarrow \max(0, s)$ ▷ ensure no rolling
  while $e < l$ **do**
    $z_{t-1}^{[s,e]} \leftarrow \mathrm{G}(z_t^{[s,e]}, c_a^{[s,e]}, c_c^{[s,e]}, c_t, c_i, t)$
    if $e < l$ **then**
      $s \leftarrow e - o$
      if $s + f < l$ **then**
        $e \leftarrow s + f$
      else
        $e \leftarrow l$
        if $e - s < n$ **then**
          $s \leftarrow e - n$ ▷ ensure minimum clip length
        end if
      end if
    end if
  end while
  $\alpha \leftarrow \alpha + p$ ▷ accumulate shift
end for

3.6 Direct Preference Optimization

Direct Preference Optimization (DPO) formalizes the alignment of the model with human preferences as a policy optimization task based on pairwise preference data. Let the preference dataset be denoted as $D = \{(x, y_w, y_l)\}$, where y_w represents the preferred sample and y_l represents the less-preferred one. For each training sample, we randomly select four video clips and compute three evaluation metrics for each clip: the Sync-C score Li et al. (2024), the hand-quality reward score (Hand-Specific Reward Cui et al. (2025a)), and the VideoReward score Liu et al. (2025). After aggregating these three indicators through weighted scoring, the segment with the highest composite score is designated as y_w, while the one with the lowest score is designated as y_l.

The DPO objective aims to maximize the likelihood of preferred outputs while regularizing the deviation from a reference policy π_{ref}. Formally, it is defined as:

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(x,y^w,y^l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y^w \mid x\right)}{\pi_{\mathrm{ref}}\left(y^w \mid x\right)} - \beta\log\frac{\pi_\theta\left(y^l \mid x\right)}{\pi_{\mathrm{ref}}\left(y^l \mid x\right)}\right)\right], \tag{4}$$

where $\sigma(\cdot)$denotes the sigmoid function, and β is a temperature coefficient controlling the strength of regularization towards π_{ref}.

Following the Flow-DPO loss proposed in VideoReward [Liu et al., 2025],the DPO loss in our training process can be reformulated as:

$$\begin{aligned}\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(y^w,y^l,t)\sim\mathcal{D}}\Bigg[\log\sigma\Bigg(-\frac{\beta_t}{2}\Big(&\left\|\boldsymbol{v}^w - \boldsymbol{v}_\theta\left(\boldsymbol{x}_t^w, t\right)\right\|^2 - \left\|\boldsymbol{v}^w - \boldsymbol{v}_{\mathrm{ref}}\left(\boldsymbol{x}_t^w, t\right)\right\|^2 \\ &- \left(\left\|\boldsymbol{v}^l - \boldsymbol{v}_\theta\left(\boldsymbol{x}_t^l, t\right)\right\|^2 - \left\|\boldsymbol{v}^l - \boldsymbol{v}_{\mathrm{ref}}\left(\boldsymbol{x}_t^l, t\right)\right\|^2\right)\Big)\Bigg)\Bigg]\end{aligned} \tag{5}$$

Here, v_{ref} refers to the velocity field of the reference model, initialized from the diffusion model, while v^w and v^l correspond to the velocity fields obtained from the preferred sample y^w and the dispreferred sample y^l. During training, gradients from L_{DPO} adjust the denoising direction to favor high-reward regions in the trajectory space while preserving the temporal dynamics of the reference policy. This joint optimization enables preference-aware generation without compromising the inherent stability of pretrained diffusion models.

4 Experiments

Implementation Details. We adopt the Wan2.1-I2V-14B architecture as the baseline video diffusion model for our experiments, which was trained using a constant learning rate of 1×10^{-5}. And the MV-Director is powered by a finetuned Qwen-Omni model. The proposed framework was trained on 64 NVIDIA A800-80G GPUs with mixed-precision acceleration. For stage-1 training, we curated a general-purpose video dataset containing approximately 1,500 hours of single-person facial/body performances, with individual clips averaging 10 seconds in duration. To enhance musical performance expressiveness in stage-2 training, we further incorporated 400 hours of domain-specific music performance videos featuring synchconized audio-visual recording of professional singers and virtual avatars.

Test Datasets and Evaluation Metrics. Following established benchmarks, we evaluate our methods performance across diverse scenarios using three datasets (HDTF, CelebV-HQ, and EMTD) and a a camera motion dataset (MultiCamVideo). Performance is quantified through complementary automated metrics and human evaluation. For objective assessment, we employ the following: Fréchet Inception Distance (FID) to measure per-frame visual quality; Fréchet Video Distance (FVD) for temporal consistency evaluation; SyncNet-based Sync-C (confidence score) and Sync-D (lip distance) to qualify lip-sync accuracy; Cosine Similarity (CSIM) scoring identity preservation; and rotation error (RotErr) / translation error (TransErr) He et al. (2024a) to assess camera motion precision. To capture perceptual nuaces beyond automated metrics, we conduct user studies where 20 participants rate 15 generated videos from five dimensions; gesture synchronization with audio prosody, body motion alignment with speech rhythm, lip-sync accuracy, identity consistency, and overall naturalness, yielding 300 feedback samples for comparative analysis.

4.1 Experiment Results

Qualitative Results. Figure 5 presents audio-driven long-video generation results under various camera controls. Our method demonstrates superior performance in identity perservation, motion clarity, and expression control. For instance, in the first two rows, our method exhibits both structural coherence and seamless transitions during secne expansion under camera zoom-out, effectively maintaining spatial consistency while generating plausible out-of-frame content. In the third row, even under camera zoom-in conditions, YingVideo-MV preserbes character identity fidelity and fine-grained scene details (e.g., background textures and lighting consistency). Notably, the generated sequences balance cinematographic precision with visual naturalness across all evaluated camera motion patterns, including complex transitions between dynamic and static shots.

Quantitative Results. As shown in Table 1, we compare our method against camera motion-aware models (CameraCtrl He et al. (2024a), Uni3C Cao et al. (2025)) and long-sequence generation frameworks (StableAvatar Tu et al. (2025), InfiniteTalk Yang et al. (2025a)). Due to the absence of camera motion capabilities in StableAvatar and InfiniteTalk, their performance on camera-related metrics cannot be evaluated. Our method demonstrates significantly better visual quality and identity preservation, outperforming InfiniteTalk's sparse frame audio-driven method (which achieves good FID and FVD scores through local lip region optimization, while our method suffers from reduced FID and FVD scores due to camera movement). While our method does not achieve the minimal rotation error, it surpasses other baselines in translation error, demonstrating more natural camera trajectory estimation. The balanced lip-sync accuracy and camera motion generation indicates a harmonious integration of multimodal temporal alignment and cinematographic control, representing a significant advancement in music-driven video synthesis.

User Study. To further validate our method's effectiveness, we conducted a subjective evaluation on our internal dataset, which is shown in Table 2. Participants rated four key dimensions using a 5-point with 0.5 increments (1=worst, 5=best): (1) smoothness and coherence of camera motions,

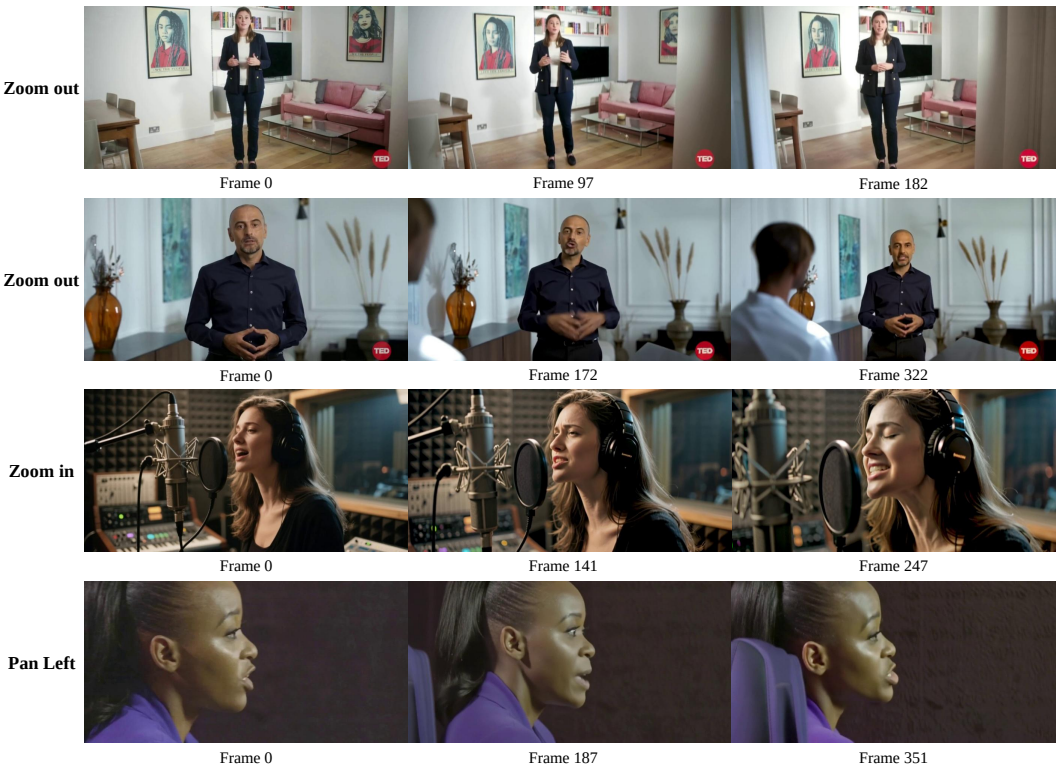

Figure 5: **Visualization of Camera Movment.** This figure illustrates the audio-driven high performance of our framework with synchronized camera motions. The generated sequences demonstrate precise alignment between body movements and camera motion.

Method	Function		Metrics						
	Camera movement	Long video generation	RotErr ↓	TransErr ↓	FID ↓	FVD ↓	CSIM ↑	Sync-C ↑	Sync-D ↓
Stable Avatar	✗	✓	-	-	38.14	375	0.635	5.15	9.49
InfiniteTalk	✗	✓	-	-	**27.14**	**132.54**	<u>0.744</u>	<u>5.61</u>	<u>9.18</u>
CameraCtrl	✓	✗	<u>1.18</u>	9.02	36.72	360.3	0.498	4.35	11.29
Uni3C	✓	✗	**1.13**	<u>7.26</u>	31.75	253.76	0.574	4.66	10.03
Ours	✓	✓	1.22	**4.85**	<u>30.36</u>	<u>193.68</u>	**0.753**	**6.07**	**8.67**

Table 1: **Quantitative comparison with other methods.** This table summarizes common camera motion and long-sequence video generation methods, with our approach uniquely combining both capabilities. Quantitative metrics demonstrate superior performance where **Bold** indicates the best result and <u>Underline</u> denotes the second-best across all evaluated benchmarks.

(2) lip-sync accuracy with musical rhythms, (3) naturalness of character movements, and (4) overall video quality. The results demonstrated statistically significant superiority of our method across all dimensions, with average scores of 4.3±0.6 for camera motion, 4.5±0.5 for lip-sync, 4.2±0.5 for motion naturalness, and 4.4±0.6 for overall quality. Inter-rater reliability analysis showed substantial agreement, confirming the robustness of subjective assessments. Qualitative feedback highlighted our framework's ability to generate cinematographically coherent sequences with natural audio-visual synchronization, particularly in complex scenarios involving dynamic camera transitions and rapid lip movements.

4.2 Ablation Study

We conduct ablation experiments to quantitatively evaluate the impact of reward feedback mechanisms and temporal coherence strategies. Specifically, we train and test models with and without DPO optimization, as well as with and without the dynamic window inference strategy. As shown

Method	Smoothness and coherence of camera motions	Lip-sync accuracy with musical rhythms	Naturalness of character movements	Overall video quality
Stable Avatar	1.3±0.2	3.9±0.4	3.7±0.3	3.9±0.2
InfiniteTalk	1.4±0.4	4.4±0.4	**4.3±0.3**	4.4±0.4
CameraCtrl	3.8±0.3	3.7±0.1	3.4±0.2	3.3±0.3
Uni3C	4.0±0.1	4.0±0.7	3.6±0.2	3.7±0.3
Ours	**4.3±0.6**	**4.5±0.5**	4.2±0.5	**4.4±0.6**

Table 2: **User Study.** The metrics demonstrate superior performance where **Bold** indicates the best result and Underline denotes the second-best across all evaluated benchmarks.

in the table 3, the introduction of DPO leads to improvements across all metrics, demonstrating enhanced video quality, lip-sync accuracy, and identity preservation. Similarly, the dynamic window strategy achieves 6.3% improvement in temporal smoothness (FVD↓). These results indicate that our combination of DPO-driven preference alignment and dynamic window-based temporal coherence not only improves technical fidelity and synchronization but also enhances expressiveness, producing outputs that better align with human perceptual preferences.

	FID ↓	FVD ↓	CSIM ↑	Sync-C ↑	Sync-D ↓
Ours	**30.36**	**193.68**	**0.753**	**6.07**	**8.67**
Ours(w/o DPO)	35.02	203.71	0.728	5.88	8.92
Ours(w/o TDW)	35.63	205.88	0.731	5.79	9.03

Table 3: **Quantitative results of ablation study**. 'TDW' means timestep-aware dynamic window range strategy.

5 Limitation and Future Work

Our framework currently faces challenges in generating animations for non-human entities when provided with reference images exhibiting significant morphological and structural differences from human subjects. While our method achieves high performance on human avatars, the geometric and textural complexities of fantastical creatures (e.g., multi-limbed beings, non-biological structures) exceed the current model's capacity for novel shape synthesis. A promising direction involves integrating an auxiliary reference-aware network to explicitly capture semantic details through hierarchical feature adaptation and cross-domain knowledge transfer. We also observe limitations in modeling complex interpersonal dynamics - extending our framework to support multi-character interactive music videos (MC-MV) remains a key objective for future exploration. This would require advancing inter-person spatial reasoning and behavioral coordination modeling while maintaining strict audio-visual synchronization across multiple agents.

6 Conclusion

In this work, we introduced YingVideo-MV, a cascaded video generation framework that unifies multi-modal inputs with long-sequence music video synthesis. Our two-stage pipeline first employs MV-Director for global shot planning, followed by DiT-based clip-wise generation of high-resolution video details. A dynamic window inference optimization module further refines inter-clip visual transitions to address the critical challenges of cinematic language deficiency and temporal inconsistency in music video generation. Combined with meticulously curated datasets and practical training/inference strategies, our framework achieves faithful global semantic alignment while preserving fine-grained audio-visual details. Experimental results demonstrate YingVideo-MV's ability to generate videos with precise lip synchronization, identity consistency, and controllable camera dynamics. Superiority over baseline methods is further validated through human preference-based metrics. We believe this work provides practical solutions for automated, high-quality music content creation and establishes a novel paradigm for future multi-modal video generation systems.

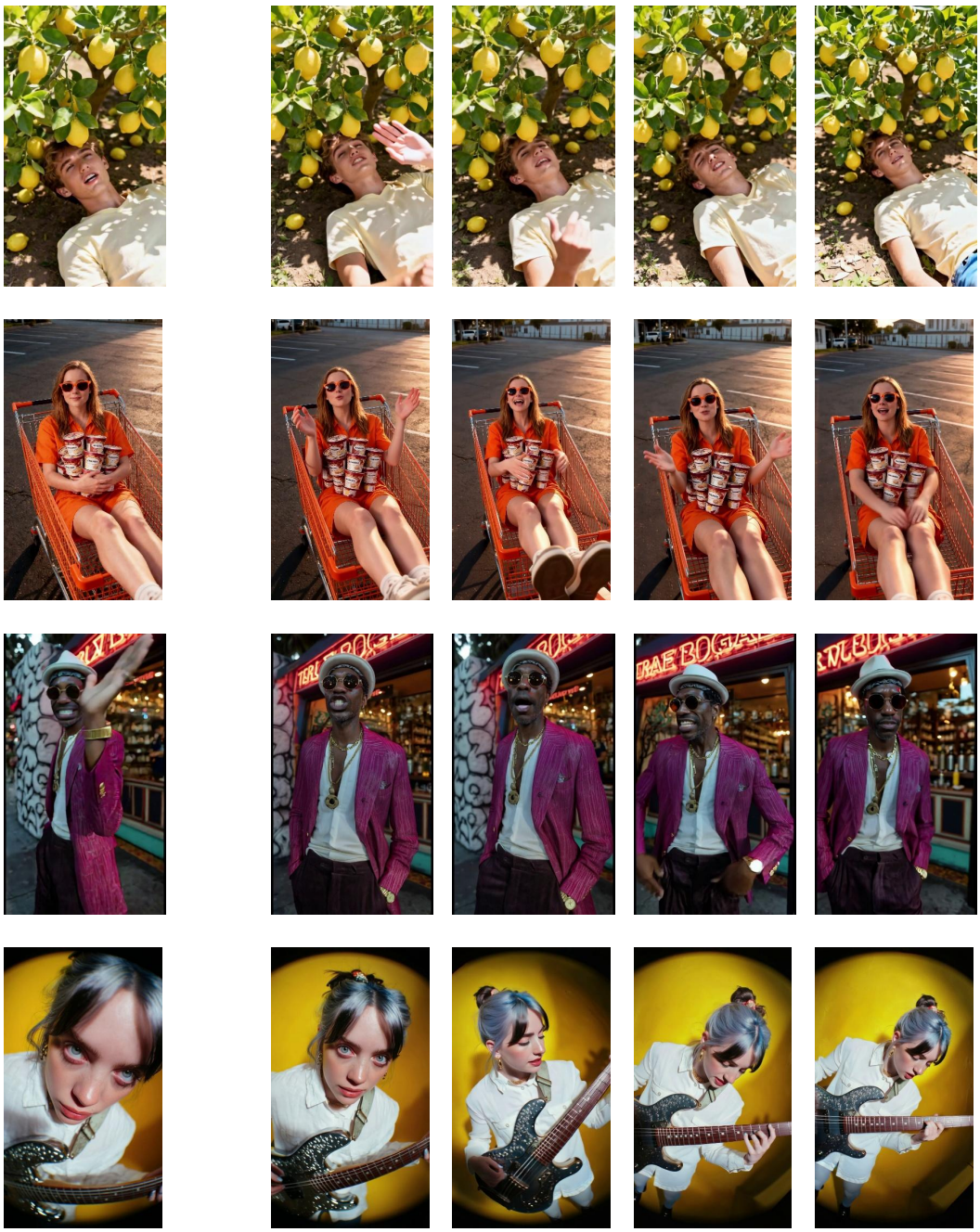

Figure 6: **More generated results of YingVideo-MV**. The table shows the video results generated by YingVideo-MV using various camera motion combinations.

7 Author

Jiahui Chen, Weida Wang, Runhua Shi, Huan Yang, Chaofan Ding.

References

Bai Jianhong, Xia Menghan, Fu Xiao, Wang Xintao, Mu Lianrui, Cao Jinwen, Liu Zuozhu, Hu Haoji, Bai Xiang, Wan Pengfei, others . Recammaster: Camera-controlled generative rendering from a single video // arXiv preprint arXiv:2503.11647. 2025a.

Bai Shuai, Chen Keqin, Liu Xuejing, Wang Jialin, Ge Wenbin, Song Sibo, Dang Kai, Wang Peng, Wang Shijie, Tang Jun, others . Qwen2. 5-vl technical report // arXiv preprint arXiv:2502.13923. 2025b.

Beeler Thabo, Hahn Fabian, Bradley Derek, Bickel Bernd, Beardsley Paul A, Gotsman Craig, Sumner Robert W, Gross Markus H. High-quality passive facial performance capture using anchor frames. // ACM Trans. Graph. 2011. 30, 4. 75.

Blattmann Andreas, Dockhorn Tim, Kulal Sumith, Mendelevitch Daniel, Kilian Maciej, Lorenz Dominik, Levi Yam, English Zion, Voleti Vikram, Letts Adam, others . Stable video diffusion: Scaling latent video diffusion models to large datasets // arXiv preprint arXiv:2311.15127. 2023.

Burgert Ryan, Xu Yuancheng, Xian Wenqi, Pilarski Oliver, Clausen Pascal, He Mingming, Ma Li, Deng Yitong, Li Lingxiao, Mousavi Mohsen, others . Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. 13–23.

Cagniart Cedric, Boyer Edmond, Ilic Slobodan. Probabilistic deformable surface tracking from multiple videos // European conference on computer vision. 2010. 326–339.

Cao Chenjie, Zhou Jingkai, Li Shikai, Liang Jingyun, Yu Chaohui, Wang Fan, Xue Xiangyang, Fu Yanwei. Uni3C: Unifying Precisely 3D-Enhanced Camera and Human Motion Controls for Video Generation // arXiv preprint arXiv:2504.14899. 2025.

Chen Eric Ming, Holalkere Sidhanth, Yan Ruyu, Zhang Kai, Davis Abe. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023a. 23242–23251.

Chen Haoxin, Xia Menghan, He Yingqing, Zhang Yong, Cun Xiaodong, Yang Shaoshu, Xing Jinbo, Liu Yaofang, Chen Qifeng, Wang Xintao, others . Videocrafter1: Open diffusion models for high-quality video generation // arXiv preprint arXiv:2310.19512. 2023b.

Chen Yi, Liang Sen, Zhou Zixiang, Huang Ziyao, Ma Yifeng, Tang Junshu, Lin Qin, Zhou Yuan, Lu Qinglin. HunyuanVideo-Avatar: High-Fidelity Audio-Driven Human Animation for Multiple Characters // arXiv preprint arXiv:2505.20156. 2025a.

Chen Zhiyuan, Cao Jiajiong, Chen Zhiquan, Li Yuming, Ma Chenguang. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions // Proceedings of the AAAI Conference on Artificial Intelligence. 39, 3. 2025b. 2403–2410.

Cui Jiahao, Chen Yan, Xu Mingwang, Shang Hanlin, Chen Yuxuan, Zhan Yun, Dong Zilong, Yao Yao, Wang Jingdong, Zhu Siyu. Hallo4: High-Fidelity Dynamic Portrait Animation via Direct Preference Optimization and Temporal Motion Modulation // arXiv preprint arXiv:2505.23525. 2025a.

Cui Jiahao, Li Hui, Zhan Yun, Shang Hanlin, Cheng Kaihui, Ma Yuqi, Mu Shan, Zhou Hang, Wang Jingdong, Zhu Siyu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025b. 21086–21095.

Ding Yikang, Liu Jiwen, Zhang Wenyuan, Wang Zekun, Hu Wentao, Cui Liyuan, Lao Mingming, Shao Yingchao, Liu Hui, Li Xiaohan, others . Kling-avatar: Grounding multimodal instructions for cascaded long-duration avatar animation synthesis // arXiv preprint arXiv:2509.09595. 2025.

Esser Patrick, Kulal Sumith, Blattmann Andreas, Entezari Rahim, Müller Jonas, Saini Harry, Levi Yam, Lorenz Dominik, Sauer Axel, Boesel Frederic, others . Scaling rectified flow transformers for high-resolution image synthesis // Forty-first international conference on machine learning. 2024.

Fei Zhengcong, Jiang Hao, Qiu Di, Gu Baoxuan, Zhang Youqiang, Wang Jiahua, Bai Jialin, Li Debang, Fan Mingyuan, Chen Guibin, others . SkyReels-Audio: Omni Audio-Conditioned Talking Portraits in Video Diffusion Transformers // arXiv preprint arXiv:2506.00830. 2025.

Fyffe Graham, Hawkins Tim, Watts Chris, Ma Wan-Chun, Debevec Paul. Comprehensive facial performance capture // Computer Graphics Forum. 30, 2. 2011. 425–434.

Gan Qijun, Yang Ruizi, Zhu Jianke, Xue Shaofei, Hoi Steven. OmniAvatar: Efficient Audio-Driven Avatar Video Generation with Adaptive Body Animation // arXiv preprint arXiv:2506.18866. 2025.

Gu Yuchao, Mao Weijia, Shou Mike Zheng. Long-context autoregressive video modeling with next-frame prediction // arXiv preprint arXiv:2503.19325. 2025.

Guo Yuwei, Yang Ceyuan, Rao Anyi, Liang Zhengyang, Wang Yaohui, Qiao Yu, Agrawala Maneesh, Lin Dahua, Dai Bo. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning // arXiv preprint arXiv:2307.04725. 2023.

He Hao, Xu Yinghao, Guo Yuwei, Wetzstein Gordon, Dai Bo, Li Hongsheng, Yang Ceyuan. Cameractrl: Enabling camera control for text-to-video generation // arXiv preprint arXiv:2404.02101. 2024a.

He Mingming, Clausen Pascal, Taşel Ahmet Levent, Ma Li, Pilarski Oliver, Xian Wenqi, Rikker Laszlo, Yu Xueming, Burgert Ryan, Yu Ning, others . Diffrelight: Diffusion-based facial performance relighting // SIGGRAPH Asia 2024 Conference Papers. 2024b. 1–12.

Hong Wenyi, Yu Wenmeng, Gu Xiaotao, Wang Guo, Gan Guobing, Tang Haomiao, Cheng Jiale, Qi Ji, Ji Junhui, Pan Lihang, others . GLM-4.1 V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning // arXiv preprint arXiv:2507.01006. 2025.

Hu Edward J, Shen Yelong, Wallis Phillip, Allen-Zhu Zeyuan, Li Yuanzhi, Wang Shean, Wang Lu, Chen Weizhu, others . Lora: Low-rank adaptation of large language models. // ICLR. 2022. 1, 2. 3.

Hu Li, Wang Guangyuan, Shen Zhen, Gao Xin, Meng Dechao, Zhuo Lian, Zhang Peng, Zhang Bang, Bo Liefeng. Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance // arXiv preprint arXiv:2502.06145. 2025.

Işık Mustafa, Rünz Martin, Georgopoulos Markos, Khakhulin Taras, Starck Jonathan, Agapito Lourdes, Nießner Matthias. Humanrf: High-fidelity neural radiance fields for humans in motion // ACM Transactions on Graphics (TOG). 2023. 42, 4. 1–12.

Ji Xiaozhong, Hu Xiaobin, Xu Zhihong, Zhu Junwei, Lin Chuming, He Qingdong, Zhang Jiangning, Luo Donghao, Chen Yi, Lin Qin, others . Sonic: Shifting focus to global audio perception in portrait animation // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. 193–203.

Jiang Jianwen, Liang Chao, Yang Jiaqi, Lin Gaojie, Zhong Tianyun, Zheng Yanbo. Loopy: Taming audio-driven portrait avatar with long-term motion dependency // arXiv preprint arXiv:2409.02634. 2024a.

Jiang Jianwen, Zeng Weihong, Zheng Zerong, Yang Jiaqi, Liang Chao, Liao Wang, Liang Han, Zhang Yuan, Gao Mingyuan. Omnihuman-1.5: Instilling an active mind in avatars via cognitive simulation // arXiv preprint arXiv:2508.19209. 2025.

Jiang Yuheng, Shen Zhehao, Wang Penghao, Su Zhuo, Hong Yu, Zhang Yingliang, Yu Jingyi, Xu Lan. Hifi4g: High-fidelity human performance rendering via compact gaussian splatting // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024b. 19734–19745.

Kant Yash, Wu Ziyi, Vasilkovsky Michael, Qian Guocheng, Ren Jian, Guler Riza Alp, Ghanem Bernard, Tulyakov Sergey, Gilitschenski Igor, Siarohin Aliaksandr. SPAD : Spatially Aware Multiview Diffusers. 2024.

Kim Geonung, Han Janghyeok, Cho Sunghyun. VideoFrom3D: 3D Scene Video Generation via Complementary Image and Video Diffusion Models // SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25). Hong Kong, Hong Kong: ACM, 2025. 1–11.

Kirkpatrick James, Pascanu Razvan, Rabinowitz Neil, Veness Joel, Desjardins Guillaume, Rusu Andrei A, Milan Kieran, Quan John, Ramalho Tiago, Grabska-Barwinska Agnieszka, others . Overcoming catastrophic forgetting in neural networks // Proceedings of the national academy of sciences. 2017. 114, 13. 3521–3526.

Kuang Zhengfei, Cai Shengqu, He Hao, Xu Yinghao, Li Hongsheng, Guibas Leonidas J, Wetzstein Gordon. Collaborative video diffusion: Consistent multi-video generation with camera control // Advances in Neural Information Processing Systems. 2024. 37. 16240–16271.

Li Chunyu, Zhang Chao, Xu Weikai, Lin Jingyu, Xie Jinghui, Feng Weiguo, Peng Bingyue, Chen Cunjian, Xing Weiwei. LatentSync: Taming Audio-Conditioned Latent Diffusion Models for Lip Sync with SyncNet Supervision // arXiv preprint arXiv:2412.09262. 2024.

Lin Gaojie, Jiang Jianwen, Yang Jiaqi, Zheng Zerong, Liang Chao, Zhang Yuan, Liu Jingtuo. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025. 13847–13858.

Lipman Yaron, Chen Ricky TQ, Ben-Hamu Heli, Nickel Maximilian, Le Matt. Flow matching for generative modeling // arXiv preprint arXiv:2210.02747. 2022.

Liu Jie, Liu Gongye, Liang Jiajun, Yuan Ziyang, Liu Xiaokun, Zheng Mingwu, Wu Xiele, Wang Qiulin, Xia Menghan, Wang Xintao, others . Improving video generation with human feedback // arXiv preprint arXiv:2501.13918. 2025.

Lombardi Stephen, Simon Tomas, Saragih Jason, Schwartz Gabriel, Lehrmann Andreas, Sheikh Yaser. Neural volumes: Learning dynamic renderable volumes from images // arXiv preprint arXiv:1906.07751. 2019.

Luiten Jonathon, Kopanas Georgios, Leibe Bastian, Ramanan Deva. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis // 2024 International Conference on 3D Vision (3DV). 2024. 800–809.

Meng Rang, Zhang Xingyu, Li Yuming, Ma Chenguang. Echomimicv2: Towards striking, simplified, and semi-body human animation // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. 5489–5498.

Niu Muyao, Cun Xiaodong, Wang Xintao, Zhang Yong, Shan Ying, Zheng Yinqiang. Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model // European Conference on Computer Vision. 2024. 111–128.

Peebles William, Xie Saining. Scalable diffusion models with transformers // Proceedings of the IEEE/CVF international conference on computer vision. 2023. 4195–4205.

Peng Ziqiao, Hu Wentao, Shi Yue, Zhu Xiangyu, Zhang Xiaomei, Zhao Hao, He Jun, Liu Hongyan, Fan Zhaoxin. Synctalk: The devil is in the synchronization for talking head synthesis // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. 666–676.

Prétet Laure, Richard Gaël, Peeters Geoffroy. Is there a "language of music-video clips" ? A qualitative and quantitative study // CoRR. 2021. abs/2108.00970.

Qi Ji, Yao Yuan, Bai Yushi, Xu Bin, Li Juanzi, Liu Zhiyuan, Chua Tat-Seng. An LMM for Efficient Video Understanding via Reinforced Compression of Video Cubes // arXiv preprint arXiv:2504.15270. 2025.

Qiu Di, Fei Zhengcong, Wang Rui, Bai Jialin, Yu Changqian, Fan Mingyuan, Chen Guibin, Wen Xiang. Skyreels-a1: Expressive portrait animation in video diffusion transformers // arXiv preprint arXiv:2502.10841. 2025.

Ren Xuanchi, Shen Tianchang, Huang Jiahui, Ling Huan, Lu Yifan, Nimier-David Merlin, Müller Thomas, Keller Alexander, Fidler Sanja, Gao Jun. GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025.

Shi Xiaoyu, Huang Zhaoyang, Wang Fu-Yun, Bian Weikang, Li Dasong, Zhang Yi, Zhang Manyuan, Cheung Ka Chun, See Simon, Qin Hongwei, others . Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling // ACM SIGGRAPH 2024 Conference Papers. 2024. 1–11.

Sitzmann Vincent, Rezchikov Semon, Freeman Bill, Tenenbaum Josh, Durand Fredo. Light field networks: Neural scene representations with single-evaluation rendering // Advances in Neural Information Processing Systems. 2021. 34. 19313–19325.

Sun Wenqiang, Chen Shuo, Liu Fangfu, Chen Zilong, Duan Yueqi, Zhang Jun, Wang Yikai. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion // arXiv preprint arXiv:2411.04928. 2024.

Tan Zhenxiong, Liu Songhua, Yang Xingyi, Xue Qiaochu, Wang Xinchao. Ominicontrol: Minimal and universal control for diffusion transformer // Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025. 14940–14950.

Team Gemini, Anil Rohan, Borgeaud Sebastian, Alayrac Jean-Baptiste, Yu Jiahui, Soricut Radu, Schalkwyk Johan, Dai Andrew M, Hauth Anja, Millican Katie, others . Gemini: a family of highly capable multimodal models // arXiv preprint arXiv:2312.11805. 2023.

Tian Linrui, Wang Qi, Zhang Bang, Bo Liefeng. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions // European Conference on Computer Vision. 2024. 244–260.

Tu Shuyuan, Pan Yueming, Huang Yinming, Han Xintong, Xing Zhen, Dai Qi, Luo Chong, Wu Zuxuan, Jiang Yu-Gang. Stableavatar: Infinite-length audio-driven avatar video generation // arXiv preprint arXiv:2508.08248. 2025.

Wan Team, Wang Ang, Ai Baole, Wen Bin, Mao Chaojie, Xie Chen-Wei, Chen Di, Yu Feiwu, Zhao Haiming, Yang Jianxiao, others . Wan: Open and advanced large-scale video generative models // arXiv preprint arXiv:2503.20314. 2025.

Wang Cong, Tian Kuan, Zhang Jun, Guan Yonghang, Luo Feng, Shen Fei, Jiang Zhiwei, Gu Qing, Han Xiao, Yang Wei. V-express: Conditional dropout for progressive training of portrait video generation // arXiv preprint arXiv:2406.02511. 2024a.

Wang Hanlin, Ouyang Hao, Wang Qiuyu, Wang Wen, Cheng Ka Leong, Chen Qifeng, Shen Yujun, Wang Limin. Levitor: 3d trajectory oriented image-to-video synthesis // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025a. 12490–12500.

Wang Mengchao, Wang Qiang, Jiang Fan, Fan Yaqi, Zhang Yunpeng, Qi Yonggang, Zhao Kun, Xu Mu. Fantasytalking: Realistic talking portrait generation via coherent motion synthesis // Proceedings of the 33rd ACM International Conference on Multimedia. 2025b. 9891–9900.

Wang Zhouxia, Yuan Ziyang, Wang Xintao, Li Yaowei, Chen Tianshui, Xia Menghan, Luo Ping, Shan Ying. Motionctrl: A unified and flexible motion controller for video generation // ACM SIGGRAPH 2024 Conference Papers. 2024b. 1–11.

Watson Daniel, Saxena Saurabh, Li Lala, Tagliasacchi Andrea, Fleet David J. Controlling space and time with diffusion models // arXiv preprint arXiv:2407.07860. 2024.

Wei Cong, Sun Bo, Ma Haoyu, Hou Ji, Juefei-Xu Felix, He Zecheng, Dai Xiaoliang, Zhang Luxin, Li Kunpeng, Hou Tingbo, others . Mocha: Towards movie-grade talking character synthesis // arXiv preprint arXiv:2503.23307. 2025.

Wu Rundi, Gao Ruiqi, Poole Ben, Trevithick Alex, Zheng Changxi, Barron Jonathan T, Holynski Aleksander. Cat4d: Create anything in 4d with multi-view video diffusion models // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025. 26057–26068.

Xu Dejia, Nie Weili, Liu Chao, Liu Sifei, Kautz Jan, Wang Zhangyang, Vahdat Arash. Camco: Camera-controllable 3d-consistent image-to-video generation // arXiv preprint arXiv:2406.02509. 2024.

Xu Jin, Guo Zhifang, He Jinzheng, Hu Hangrui, He Ting, Bai Shuai, Chen Keqin, Wang Jialin, Fan Yang, Dang Kai, others . Qwen2. 5-omni technical report // arXiv preprint arXiv:2503.20215. 2025.

Yang Shaoshu, Kong Zhe, Gao Feng, Cheng Meng, Liu Xiangyu, Zhang Yong, Kang Zhuoliang, Luo Wenhan, Cai Xunliang, He Ran, others . InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing // arXiv preprint arXiv:2508.14033. 2025a.

Yang Shurong, Li Huadong, Wu Juhao, Jing Minhao, Li Linze, Ji Renhe, Liang Jiajun, Fan Haoqiang, Wang Jin. Megactor-sigma: Unlocking flexible mixed-modal control in portrait animation with diffusion transformer // Proceedings of the AAAI Conference on Artificial Intelligence. 39, 9. 2025b. 9256–9264.

Yariv Guy, Gat Itai, Benaim Sagie, Wolf Lior, Schwartz Idan, Adi Yossi. Diverse and aligned audio-to-video generation via text-to-video model adaptation // Proceedings of the AAAI Conference on Artificial Intelligence. 38, 7. 2024. 6639–6647.

Zeng Yan, Wei Guoqiang, Zheng Jiani, Zou Jiaxin, Wei Yang, Zhang Yuchen, Li Hang. Make pixels dance: High-dynamic video generation // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. 8850–8860.

Zhang Mengchen, Wu Tong, Tan Jing, Liu Ziwei, Wetzstein Gordon, Lin Dahua. GenDoP: Auto-regressive Camera Trajectory Generation as a Director of Photography // arXiv preprint arXiv:2504.07083. 2025a.

Zhang Zhenghao, Liao Junchao, Li Menghao, Dai Zuozhuo, Qiu Bingxue, Zhu Siyu, Qin Long, Wang Weizhi. Tora: Trajectory-oriented diffusion transformer for video generation // Proceedings of the Computer Vision and Pattern Recognition Conference. 2025b. 2063–2073.

Zhang Zhiyuan, Wang Can, Chen Dongdong, Liao Jing. FlexTraj: Image-to-Video Generation with Flexible Point Trajectory Control // arXiv preprint arXiv:2510.08527. 2025c.

Zheng Longtao, Zhang Yifan, Guo Hanzhong, Pan Jiachun, Tan Zhenxiong, Lu Jiahao, Tang Chuanxin, An Bo, Yan Shuicheng. Memo: Memory-guided diffusion for expressive talking video generation // arXiv preprint arXiv:2412.04448. 2024.