

DiVA: Enabling Interactive Digital Life Simulation via Video Models

Cheng Chen^{1,2,3}, Hao Ouyang³, Qiuyu Wang³, Ka Leong Cheng³, Wen Wang³,
Yihao Meng³, Hanlin Wang³, Yixuan Li³, Jiacheng Wei¹, Zhenshan Tan⁴,
Yanhong Zeng³, Yujun Shen³, Guosheng Lin¹, Fayao Liu^{2*}

¹ Nanyang Technological University, Singapore .

² Institute for Infocomm Research, A*STAR, Singapore .

³ Ant Group, China .

⁴ Nanjing University of Information Science and Technology, Nanjing, China .

Abstract

We present DiVA, a deeply interactive digital life simulator pioneering a new paradigm for long-term, open-ended interactive experiences within digital character worlds. DiVA’s architecture pairs a Multimodal Large Language Model (MLLM) as a router with a meticulously designed stacked video pipeline for seamless, multi-turn interactions with action and audio response. To maintain continuity and avoid degradation, we model generation as a three-part coupled system: waiting video, action video, and the transitions between them. These transitions are critically handled by our Anchored Video Continuation (AVC) module, which returns the character to stable states to prevent degradation. By encoding information from the preceding action video segment, AVC ensures smooth transitions, significantly reducing camera jitter and inconsistencies common in current video transition methods. This design also enables complex pose changes (e.g., sitting to standing) typically difficult for audio-driven models. These system designs together ensure high-fidelity identity, coherence, and dynamics for extended experiences. **In this journal extension, we further analyze the necessity of the three-stage design, anchor-state selection, transition naturalness, spatial grounding, and the quality-latency trade-off, and we expand the comparison to additional long-form audio-driven avatar models.** To validate our pipeline design, we comprehensively compare our system against alternatives by replacing our core generation module with mainstream long-video, continuation, and interpolation methods. Results confirm DiVA is markedly superior in maintaining long-term visual quality and realism, validating its effectiveness as a sustainable, interactive simulation.

Keywords: video generation, interactive video, digital life simulation, long video generation, audio-driven avatar, controllable video synthesis

1 Introduction

Pushing video generation [Wan et al \(2025\)](#); [Kong et al \(2025\)](#); [Bai et al \(2025\)](#), from creating short and passive clips to simulating long-term and interactive digital lives, represents an exciting new

frontier for the field. We frame this vision as Digital Life Simulation (DLS): all behaviors, visuals, and dialogue are generated on demand by models, with no predefined rules or handcrafted assets. Users interact through open-ended language and spatial clicks, while the system maintains long-term

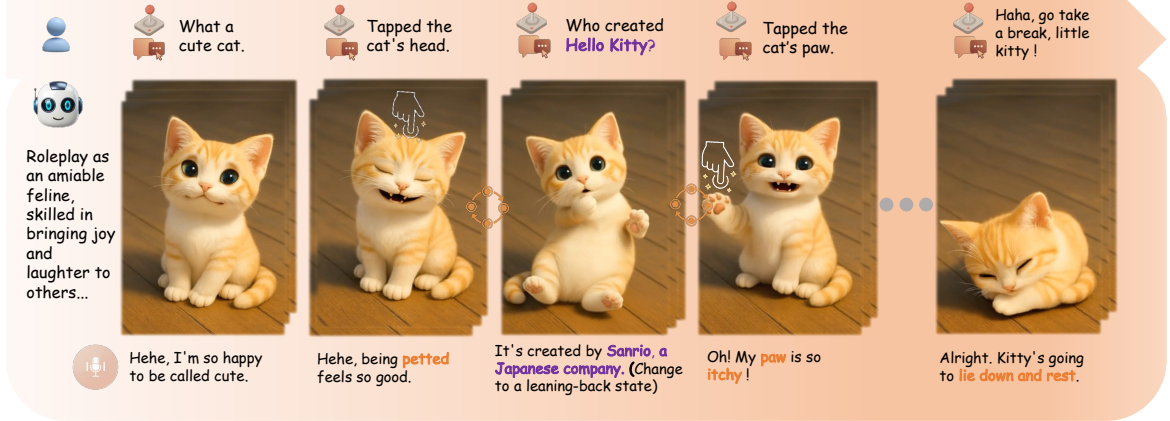


Fig. 1 Our DiVA enables infinite, continuous interaction with undiminished quality (five rounds shown here). Users can define character personality and interact via simultaneous verbal and non-verbal (spatial click) inputs, receiving corresponding realistic actions and voice replies. DiVA demonstrates advanced state transition capabilities (3rd interaction), while also inheriting the MLLM’s strong ability for factual responses (e.g., describing Hello Kitty), enabling limitless simulation within interactive digital worlds.

persona and memory and continuously produces role-centric response shots aligned with user intent.

This idea resonates with recent discussions of infinite games, where experiences evolve without preset boundaries or fixed rules. This concept has been explored by generating image series based on evolving prompts [Li et al \(2024\)](#). Taking this a step further, previous work [Cheng et al \(2025\)](#) lets users interact with characters in a virtual world through language and generates multi-turn states composed of dynamic animation shots, moving toward openness. However, it mainly adopts a third-person perspective of the character and produces isolated, non-contiguous video segments, which makes it hard for users to embody the character and limits conversational, continuous video expression and coherence.

To address these gaps, we present DiVA: a deeply interactive digital life simulator that maintains high dynamics and strong expressivity over long-form, multi-turn interactions while closely tracking user input intent. Specifically, we designed both linguistic (language) and spatial (e.g., spatial clicks) interactions to allow users to engage more immersively with the character. During long-term interaction, the system maintains the character’s high-fidelity visual identity and long-range persona, and continuously renders role-centric response shots. These shots synchronize the character’s spoken dialogue with non-verbal actions (e.g., gestures, expressions) that align with the user’s interaction input.

The recent advent of Multimodal Large Models (MLLMs) [Bai et al \(2023\)](#); [Achiam et al \(2023\)](#) has brought significant advancements to role-playing, achieving high levels of personalization and coherence using in-context learning [Dong et al \(2022\)](#). Motivated by these advancements, the MLLM in DiVA functions as the simulation router. Initially, it is conditioned via in-context exemplars to instantiate the character’s persona and interaction environment. Subsequently, during the multi-turn interaction, it reasons over user input (both linguistic and spatial) and the interaction history to produce structured directives. These directives jointly encode the character’s dialog response and the behavioral prompts for the video generator, directing it to render the precise, synchronized response shots mentioned earlier.

Relying on a single, monolithic video generator presents a significant gap in achieving both high-quality expressivity and long-term stability, while also severely limiting controllability. To address these limitations and achieve naturalness, expressivity, and long-term stability, we design a three-part coupled video generation pipeline. First, waiting video produces subtle idle motion at the start to avoid a frozen avatar feel. This is followed by Action Video, which performs MLLM-guided audio-video alignment to generate gestures and facial expressions tightly coupled to speech.

To handle the critical transitions between these multi-turn interactions, we introduce anchored

video continuation (AVC). We found that simple interpolation from the prior video’s last frame causes camera and motion discontinuity. Our AVC therefore instead conditions on a segment of the preceding action video while targeting on preset, high-quality anchor frames. This mechanism allows the character to reset to a high-quality state after each action, which is essential for suppressing identity drift. Critically, the AVC phase is decoupled from strict audio alignment, which limits its expressive potential. This allows our method to expand the character’s motion range for more creative dynamic effects, such as performing natural pose changes concurrently with camera movement that is difficult to achieve in the tightly-coupled action video phase.

To validate our pipeline design, we conduct a comprehensive comparison. We benchmark our system against alternative approaches by systematically substituting our core video module with mainstream long-video generation, state-of-the-art continuation, and interpolation-based methods. These results demonstrate that in long-duration, multi-turn digital simulations, DiVA is significantly superior in maintaining visual quality, character response accuracy, and high-dynamics behavior. These capabilities validate our design’s effectiveness, providing a robust foundation that realizes an infinite exploration and pushes generative media toward a future of persistent, interactive digital worlds.

Compared with the conference version, this journal extension provides a more complete analysis of the stateful simulation design. We expand the related-work discussion and baseline coverage, detail anchor construction and routing, analyze anchor sensitivity and transition naturalness, quantify spatial-click grounding and the quality–latency trade-off, and discuss boundary cases of the current video backbone.

2 Related Work

2.1 Video Generation Models

Inspired by the scaling laws observed in language models, a surge of works (Polyak et al, 2024; Kong et al, 2024; Blattmann et al, 2023; Wan et al, 2025; Gao et al, 2025b; OpenAI, 2025; Google, 2025; Kuaishou Technology, 2025) has begun to extend the capabilities of video generative models through

scalable data curation strategies and large-scale model training. Benefiting from the extensive training scale, these models exhibit impressive visual quality.

Recent systems further improve the temporal prior through larger diffusion-transformer backbones, stronger visual–text encoders, and multi-stage training. These foundation models provide the high-fidelity appearance and motion prior required by DiVA, but they are primarily optimized for bounded clips. Repeatedly applying a short-video generator does not by itself create a persistent simulation, because each imperfectly generated segment becomes the condition for subsequent turns and local errors can therefore accumulate.

Long-video generation is critical for applications requiring long-term coherence, such as embodied AI and world models. Key strategies to address generation challenges include enhancing robustness by injecting and denoising noise (e.g., Rolling Diffusion (Ruhe et al, ???)); compressing historical frames into a fixed latent (Zhang et al, 2025a; Cai et al, 2025) or combining compression with future frame planning (e.g., FramePack (Zhang and Agrawala, 2025)); and using task-specific pretraining to prevent color drift (e.g., LongCat-Video (Team et al, 2025)). Additionally, domain-specific models exist, such as InfiniteTalk (Yang et al, 2025b), which specializes in generating long talking-body videos. In our experiments, we compare our pipeline against the aforementioned approaches, demonstrating its effectiveness during ultra-long, turn-by-turn video generation.

LongLive adopts causal frame-level autoregression with KV-recache and streaming long tuning to support real-time prompt changes (Yang et al, 2025a), whereas HoloCine models a complete multi-shot scene through shot-localized text control and sparse inter-shot attention (Meng et al, 2026). These methods substantially extend duration and cross-shot coherence, but most still follow a predominantly single-stream generation trajectory: the current generated state is repeatedly reused as future context, and restoration of a clean visual state is handled only indirectly through prompts, memory, or context compression. iMontage provides a complementary example in which the temporal prior of a pretrained video backbone is repurposed for variable-cardinality many-to-many image generation (Fu et al, 2026). This supports a broader principle also adopted by DiVA: strong

pretrained video priors can be reorganized through task-specific structure rather than replaced by a monolithic model trained from scratch.

2.2 Controllable Video Synthesis

Concurrently, a parallel line of research has focused on enhancing the controllability of video generation through various modalities. For instance, some works (Geng et al, 2025; Zhang et al, 2025b; Yin et al, 2023) utilize trajectories to guide the generation process, while others (Bai et al, 2025; Luo et al, 2025; Bahmani et al, 2025) concentrate on the control of cameras. Although these methods enable precise control, it is typically limited to single, short-form video clips and they are not designed for immersive, long-duration interaction.

Recent controllable systems broaden the conditioning interface. WorldCanvas jointly uses reference images, trajectories, and text to specify the identity, timing, visibility, and semantics of promptable world events (Wang et al, 2025); CADCrafter and Sculpt3D introduce structured geometric or camera-aware conditions for more controllable synthesis (Chen et al, 2025, 2024). Human-animation methods additionally condition on audio, poses, end effectors, or reference identities to obtain semantically meaningful character motion (Tian et al, 2025; Kong et al, 2025; Gao et al, 2025a; Meng et al, 2025). These studies demonstrate that additional conditions can substantially improve local control, but sustained interaction further requires the system to decide when the visual state should change and how quality should be restored after many turns.

First-last-frame generation and video continuation provide temporal endpoint control closely related to our transition problem. A target frame specifies the desired terminal appearance, while preceding content constrains short-term continuity. Standard first-last-frame interpolation, however, normally represents the starting state with only one frame. In an interactive setting, the terminal frame alone is an incomplete description of ongoing motion: two clips may end at visually similar frames while having different body momentum, gesture phase, and camera movement. DiVA therefore conditions AVC on a preceding video segment, so that the generated transition behaves as a continuation of the ongoing motion rather than a newly initialized interpolation.

2.3 Digital Life Generation

Talking Avatar Generation. Many pioneering works have focused on endowing AI with a visual embodiment. As speech is a critical medium for character simulation, research has predominantly centered on generating talking faces or talking bodies. The generation of talking avatars can be broadly categorized into two approaches: 3D-based (Chan et al, 2022; Qiu et al, 2025; Qian et al, 2024) and video-based (Kong et al, 2025; Tian et al, 2025; Gao et al, 2025a; Meng et al, 2025). 3D-based methods typically rely on an intermediate 3D representation for animation, a process that often results in lower visual fidelity and naturalism. In contrast, video-based methods leverage end-to-end training with large-scale data curation, which tends to yield superior realism in facial expressions and body dynamics. Despite their strong performance in lip-syncing scenarios, generating avatars with a consistent personality and natural long-term motion still requires an additional process.

Long-form audio-driven systems such as InfiniteTalk, StableAvatar, WanS2V, and Live Avatar extend this paradigm through sparse reference frames, sliding windows, causal adaptation, or streaming inference (Yang et al, 2025b; Tu et al, 2025; Gao et al, 2025a; Huang et al, 2025). X-Streamer further integrates multimodal reasoning and audiovisual generation in a Thinker-Actor architecture for persistent video-call interaction (Xie et al, 2025). These systems are highly relevant to conversational avatars, but their primary objective remains continuous audio-conditioned rendering. Strong audio coupling benefits lip synchronization and local conversational motion, whereas digital life simulation additionally requires deliberate macroscopic state changes, spatial interaction, and explicit long-horizon quality control.

AI Agents in Games and Simulations. There has been a line of exploration in enabling AI to simulate various roles in games, such as competitors, designers, or teammates (Zhu et al, 2021; Pell, 1992). For instance, Unbounded (Li et al, 2024) generates an infinite game world, utilizing an LLM to generate text responses and a pre-trained text-to-image model enhanced with LoRA (Hu et al, 2022) for character-consistent image generation. Building on this, AnimeGamer (Cheng et al,

2025) extends the concept to anime life simulation, employing a video model to generate dynamic animation shots. While both Unbounded and AnimeGamer utilize a Multimodal Large Language Model (MLLM) (Achiam et al, 2023; Bai et al, 2023; Liu et al, 2024) as a router to transform dialogue into character information, their limitations distinguish them from our work. Unbounded is primarily focused on static image generation. AnimeGamer, despite using a video model to simulate character movements, mainly adopts a third-person perspective and produces isolated video segments. This approach makes it hard for users to embody the character and limits interaction-driven, continuous video expression.

Multi-turn visual interaction has also been explored through visual dialog and interactive character systems (Chen et al, 2022). Taken together, existing studies establish language models as effective interaction agents, but leave an important rendering-side question: how can open-ended multimodal intent be converted into continuous, speech-aligned video while retaining both expressive large motion and stable visual identity over many turns?

We view existing long-form character generation as two complementary paradigms. General long-video models provide broad motion, but do not explicitly route interaction states or deliberately restore a vetted visual state. Audio-driven avatar models provide strong speech alignment and local identity preservation, but large pose and camera transitions remain constrained by the audio-coupled generation objective. DiVA addresses this stability-expressiveness trade-off with a decoupled stateful pipeline: Action Video handles speech-aligned responses, AVC performs audio-decoupled macroscopic transitions, and curated anchors provide explicit visual reset states selected by the MLLM. Table 1 summarizes this positioning.

3 Methods

3.1 Task Formulation

As shown in Fig. 2, we focus on the challenging task of digital life simulation in this paper, wherein a digital character is expected to deliver an enduring, consistent, and dynamic immersive interactive simulation guided by both a user-defined personality and ongoing user input. We employ a Multimodal

Large Language Model (MLLM) as a router to process user input and translate it into a corresponding character response for the generative models. Subsequently, a video model is utilized to render the actions of the digital character. To support natural, long-term, and multi-turn action generation, our video generation model is composed of three key modules: a waiting video module, an action video module, and an anchored video continuation module.

To make the stateful interaction process explicit, let $\mathcal{C}$ denote the static context, including the character identity, persona, scene, voice configuration, and a portfolio of anchor states $\mathcal{A} = \{I_k^a\}_{k=1}^K$. At turn t , the user provides multimodal input U_t , which may contain language and a spatial click, while H_{t-1} stores the interaction history. In addition to the generated video, DiVA maintains a semantic state selected from $\mathcal{A}$. This differs from ordinary long-video continuation, where the system state is represented only by the most recently generated frames and therefore inherits their accumulated errors.

3.2 MLLM as Digital Life Simulation Router

The recent advent of MLLMs has brought significant advancements to role-playing, achieving high levels of persona fidelity and temporal coherence. Our empirical evaluations revealed that MLLMs possess potent zero-shot and few-shot capabilities, where an In-Context Learning (ICL) (Dong et al, 2022) strategy achieves strong performance without costly fine-tuning. Therefore, for architectural parsimony and generalization, we adopt the ICL approach in our simulation. In our implementation, we primarily leverage large models such as GPT-5 (Achiam et al, 2023) and Qwen (Bai et al, 2023). We observed that robust performance is achievable with a concise set of ICL exemplars.

In our framework, the MLLM functions as the central semantic orchestrator. Its responsibilities begin prior to the simulation, where it is conditioned with a series of descriptive exemplars defining the interaction context, the character’s emotional range, and its dialogic style. Our framework also supports visual persona definition. Users can provide a single reference image, which is then processed by a generative vision module (Gemini (Gemini Team, 2024) or Qwen (Bai et al, 2023)) to

Table 1 Positioning of representative generation paradigms for multi-turn digital life simulation. “Explicit reset” denotes deliberately returning generation to a high-quality visual state rather than only propagating recent generated frames.

Paradigm	Speech	Macro motion	Routing	Reset	Primary objective
General long-video generation (FramePack, LongLive, LongCat-Video)	Optional	Broad, indirect	Prompt	No	Efficient, coherent long-form continuation
Audio-driven avatars (InfiniteTalk, StableAvatar, WanS2V, Live Avatar)	Strong	Audio-constrained	Implicit	No	Long identity-preserving talking-avatar synthesis
Unified interactive avatar (X-Streamer)	Strong	Conversational	Learned	Identity ref.	Real-time persistent audiovisual interaction
DiVA	Strong	Explicit AVC	MLLM	Anchors	Stateful DLS with long-term quality control

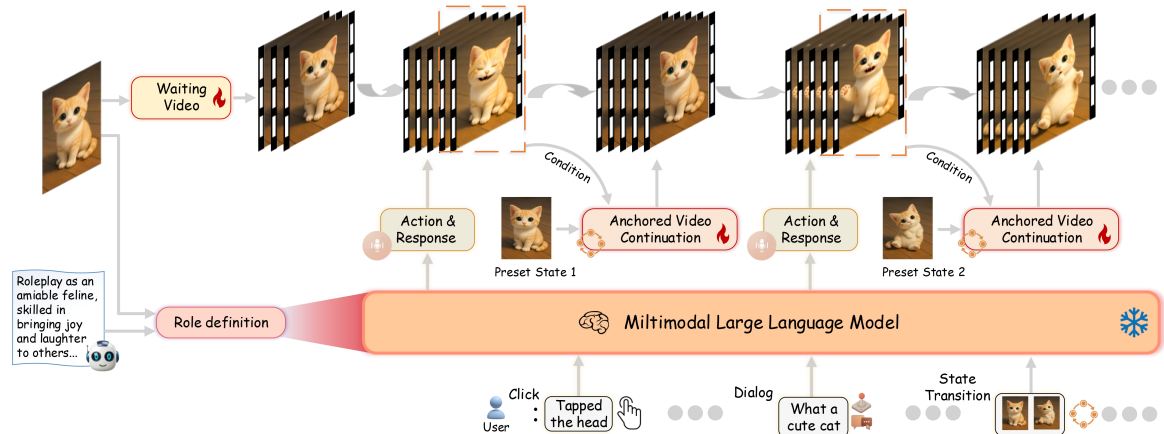


Fig. 2 Given a text prompt defining the character’s role, an MLLM acts as a router, processing user inputs (e.g., clicks and dialogue) and translating them into character responses and behavioral prompts for the audio-text conditional action model. Our video generation component starts with a waiting video, followed by iterative action videos based on the MLLM output. Finally, the anchored video continuation (AVC) module, conditioned on the prior action, targets preset, high-quality anchor frames that represent distinct character states (e.g., sitting or leaning back). Guided by the MLLM state assessment, AVC transitions to the appropriate anchor state. This mechanism smoothly restores the sequence to a stable, high-quality condition and enables expressive transitions between states, significantly expanding the character’s motion range.

synthesize a portfolio of anchor states—visual representations of the character in distinct emotional or physical states using instruction-based editing.

Subsequently, the MLLM manages the turn-by-turn dynamics. Critically, it performs unified reasoning over multimodal user input (linguistic and spatial) and the interaction history. This process yields the character’s response, including the dynamic selection of the appropriate anchor state from the pre-computed visual state-space, which triggers coherent visual transitions. The MLLM output at each turn is a structured semantic output $O_t = (P_t^B, R_t^D)$, jointly encoding the behavioral prompt P_t^B and the dialogic response text R_t^D .

This multi-turn process, governed by the MLLM $\mathcal{M}$, can be formalized as the following

auto-regressive function at each turn t :

$$O_t = \mathcal{M}(U_t, H_{t-1}, \mathcal{C}), \quad (1)$$

where $\mathcal{C}$ is the static context, U_t is the current multimodal user input, H_{t-1} is the interaction history, and $O_t = (P_t^B, R_t^D)$ is the structured output.

Anchor portfolio and state decision.

The anchor states are not hard-coded actions or a fixed script. They form a replaceable and extensible visual portfolio that may contain different poses, emotions, camera distances, or interaction contexts. The MLLM receives the indexed anchor images together with short state descriptions and

follows a persistence prior: it remains in the current state unless a different anchor better matches a meaningful change requested by the interaction. Open-ended dialogue, gestures, expressions, and object interactions remain the responsibility of the action generator; anchors are selected only when a macroscopic visual change or a high-quality reset is useful. In implementation, the structured output therefore additionally includes a state decision k_t , represented as either **KEEP** or **SWITCH TO IMAGE: k**.

Prompt organization.

The ICL prompt contains five functional parts: the character role and simulation goal; the persona, scene, voice, and indexed visual states; a state-management rule that preserves the current state by default; an interaction rule for language and spatial clicks; and generation constraints that keep the spoken reply concise while assigning large pose changes to AVC. The required output separates the **Character Reply**, **Video Prompt**, and state decision. This simple schema keeps the routing process inspectable and prevents the action prompt from being entangled with state-transition control.

For the spatial instruction modality, we explored two grounding mechanisms. A naive approach of providing raw click coordinates proved ineffective, as it requires the MLLM to solve a difficult numerical-to-visual grounding problem. Our second, more robust approach transforms the spatial coordinate task into a visual-referential grounding problem. When a user interaction is detected, we dynamically augment the current video frame with a salient visual marker (e.g., a dilated red dot) at the point of interaction. This augmented frame is then fed to the MLLM. Simultaneously, the ICL prompt is updated to condition the model to interpret this visual marker as the locus of user intent (e.g., the user is clicking the red dot). This method obviates the need for explicit coordinate mapping, instead leveraging the MLLM’s potent visual-contextual reasoning capabilities to infer the user’s intent.

More formally, given the current frame I_t and a click coordinate (x_t, y_t) , we construct

$$\tilde{I}_t = \text{DrawMarker}(I_t; x_t, y_t, r), \quad (2)$$

where r is the marker radius. The MLLM then performs ordinary visual referring on $\tilde{I}_t$, avoiding the

need for fine-grained click-action video supervision.

3.3 Video Model as a Digital Life Rendering Engine

In this section, we elaborate on the architecture of our video generation pipeline, engineered to produce visual results that are immersive, coherent across multiple turns, and highly dynamic. To achieve these goals, our pipeline operates in three phases. Here, we detail the waiting video generation and action video generation phases.

Our simulation commences with the waiting video module, as we posit that a digital life should exhibit natural, subtle idle motions as a baseline state. To this end, we specialize a pre-trained video model by fine-tuning it on a dataset of natural and fluid waiting animations. This module is formulated as an Image-to-Video (I2V) generation task. We employ Low-Rank Adaptation (LoRA) (Hu et al, 2022) for this specialization, optimizing solely on an I2V loss. The use of LoRA obviates the need for extensive parameter retraining, allowing for lightweight and efficient integration.

Upon user-initiated interaction, the system transitions to the action video generation module. In this phase, the video model renders output based on semantic directives generated by the upstream MLLM. To create an immersive experience, the MLLM output is a structured representation containing two key components: a behavioral prompt (a textual description of the character’s actions, expressions, and scene changes) and the dialogic response (the character’s verbal response as text).

Consequently, the model for this stage must support synchronized audio-video generation, for which we leverage an existing open-source audio-text conditional video model (Kong et al, 2025). To satisfy this model’s multimodal input requirements, the MLLM textual dialogic response is seamlessly synthesized into an audio waveform via an instant TTS engine, CosyVoice (Du et al, 2024). This audio, jointly with the MLLM behavioral prompt, forms the complete conditional input used by the video model to generate the final audio-visual output.

Module switching and execution.

The MLLM dynamically governs the switching logic, but the rendering order is fixed. First, Action Video generates the immediate audio-visual

response from the current anchor. Second, AVC transitions from the ending action segment to the selected target anchor. Third, the waiting video associated with that anchor is looped until the next user input arrives. If the router returns **KEEP**, AVC returns to the current anchor; if it returns **SWITCH**, AVC targets the selected state. This explicit sequence makes state changes inspectable and controllable instead of allowing them to emerge implicitly from recent latent history.

3.4 Anchored Video Continuation

Generating long-duration, multi-turn interaction videos of a digital life is hindered by two primary issues. The first challenge is quality degradation, as long-term, turn-by-turn generation leads to declining visual quality and character identity drift due to cumulative errors. The second challenge is limited motion dynamics, which occurs when the action model must simultaneously align video with audio (for lip-sync) and text (for actions). This strict synchronization requirement often forces a compromise in the range of motion, resulting in subdued movements that fail to capture larger, more dynamic actions like pose changes.

To address these issues, we propose an anchored video continuation mechanism. The core idea is to generate a dynamic transition from the end of each action segment to a pre-defined, high-quality anchor state. This design ensures long-term quality by resetting cumulative error, as the character periodically returns to a stable state. At the same time, by intentionally decoupling this transition from strict audio-alignment constraints, the mechanism allows for the generation of large, dynamic motions, significantly enhancing expressiveness.

A naive approach to achieve dynamic transition is to train a frame-based interpolation model to generate a transition video v_{trans} by interpolating between the last frame of the action video I_{end} and a preset anchor frame I_{anchor} . While this method connects the start and end points, we found its results to be suboptimal. Because the interpolation model only considers two static frames, it completely disregards the motion inertia of the character and the camera dynamics from the preceding video. This often results in a transition that feels disjointed and discontinuous from the previous action, as shown in our ablation study.

Therefore, we propose to train a segment-based continuation model. This model takes the final segment of the action video, which we denote as v_{context} , as a condition and continues it by generating a seamless transition to the anchor frame I_{anchor} . This generation process can be formulated as

$$v_{\text{trans}} \sim P(v \mid v_{\text{context}}, I_{\text{anchor}}). \quad (3)$$

We define v_{context} as the last 25% frames from the action video. Because the action videos themselves have varying lengths, the frame count of v_{context} is also variable, which trains the model to handle different context lengths. We inject this condition in two ways: for detail preservation, all frames of v_{context} and the anchor frame I_{anchor} are compressed via VAE into a latent $Z_c \in \mathbb{R}^{c \times t \times h \times w}$ and concatenated with the noise latent z_t ; simultaneously, for global context, we only extract the CLIP features from the first frame of v_{context} and the anchor frame, injecting them into the DiT model via decoupled cross-attention.

Training-pair construction.

We curate AVC training clips that contain meaningful changes in body pose, camera distance, or viewpoint rather than restricting the data to talking-head videos. The collection pipeline consists of multi-source collection, automated filtering, and annotation. We remove low-quality, poorly illuminated, or discontinuous clips using aesthetic and luminance filtering, retain single-shot videos with usable endpoints, and use Qwen2.5-VL to describe character appearance, action, scene content, and camera behavior. The resulting set contains 45k video-text pairs. The first quarter of each variable-length clip is used as the observed context and the final frame as the target anchor, producing diverse transition magnitudes without manually annotated transition labels. The waiting-video subset contains 3k clips with loop-compatible low-amplitude motion, such as breathing, blinking, minor head motion, hair movement, and environmental dynamics.

AVC implementation.

AVC is initialized from the Wan 2.1 14B first-last-frame model and adapted with LoRA. Let E_{vae} denote its spatiotemporal VAE. The context frames and target anchor are encoded into $Z_c =$

$E_{\text{vae}}(v_{\text{context}} \cup I_{\text{anchor}})$ and concatenated channel-wise with the noisy video latent. A temporal mask $M \in \{0, 1\}^{1 \times T \times h \times w}$ marks every supplied context and anchor position:

$$M_j = \begin{cases} 1, & j \in \mathcal{T}_{\text{context}} \cup \mathcal{T}_{\text{anchor}}, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Unlike the original first-last-frame formulation, which preserves only the sequence boundaries, this mask supports an arbitrary conditioning segment. CLIP image features from the first context frame and the target anchor are injected through decoupled cross-attention. Using only the endpoint CLIP features limits memory growth, while the VAE latent retains frame-level appearance and motion over the full context. LoRA is applied to the attention projections and the original diffusion/flow-matching objective is retained.

4 Experiments

In this section, we demonstrate how DiVA addresses key limitations of prior models in supporting interactive digital life simulation via video models.

4.1 Implementation Details

We train our anchored video continuation model on our curated in-house dataset of 45k video-text pairs. All videos are processed at a resolution of 480×832 . For training, the starting video is obtained by taking the first 1/4 of the frames from a video, and the last frame of that video is extracted to serve as the interpolation target. Since the videos in our dataset vary in length, the starting videos used for training therefore have different lengths. This enables our model to support starting video inputs of varying lengths. During inference, we uniformly use 21 frames from the previous interaction video as the starting video.

Our base model is the Wan 2.1 14B FLF (Wan et al, 2025). We train the model using LoRA with a rank of 128. The model is trained for 10k steps with a learning rate of 1×10^{-5} and a linear warmup schedule. The entire training process is conducted on 48 NVIDIA H800 GPUs with a total batch size of 48. For the waiting video module, we also use Wan 2.1 14B as the base model, which allows our

video base models to be shared across both tasks. We train this model using LoRA with a rank of 32. The model is trained for 5k steps on a dataset of 3k videos featuring subtle, live-wallpaper-style motions.

Evaluation protocol.

Our long-horizon benchmark contains 60 multi-turn character simulations spanning diverse human, anime, stylized, and animal identities. Each sequence contains repeated language interactions and explicit state transitions and comprises thousands of generated frames. All compared methods receive the same reference identity, turn prompts, and state-change instructions whenever their interfaces permit. FramePack and InfiniteTalk condition every turn after the first on the final frame of the previous turn. LongCat-Video follows its official continuation protocol, including its restoration and super-resolution stages. Methods without audio conditioning are evaluated for visual quality and state following rather than lip synchronization. Spatial grounding is evaluated on 50 interaction pairs with clicks on semantically distinct regions.

4.2 Comparison

We select the powerful base model FramePack (Zhang and Agrawala, 2025) (an I2V model with anti-drift designs to support long-video generation), InfiniteTalk (Yang et al, 2025b) (a strong open-source audio-driven video generation model), and LongCat-Video (Team et al, 2025) (a large pre-trained video continuation model) as our baselines, to evaluate our model’s capabilities in maintaining persistent expressiveness in long-term interaction and accurate semantic understanding for state transition, respectively. We adopt the same prompts and a consistent turn-by-turn generation process. In each turn, we use the duration recommended by the corresponding method to avoid drift. For FramePack and InfiniteTalk, every turn after the first is conditioned on the final frame of the previous turn. For LongCat-Video, we follow its official guidelines, using the recommended video length from the previous turn for continuation, and its super-resolution and restoration functions are also applied in each turn. Since FramePack and LongCat-Video lack audio conditioning, we generate their results using only video prompts and focus the comparison on visual quality.

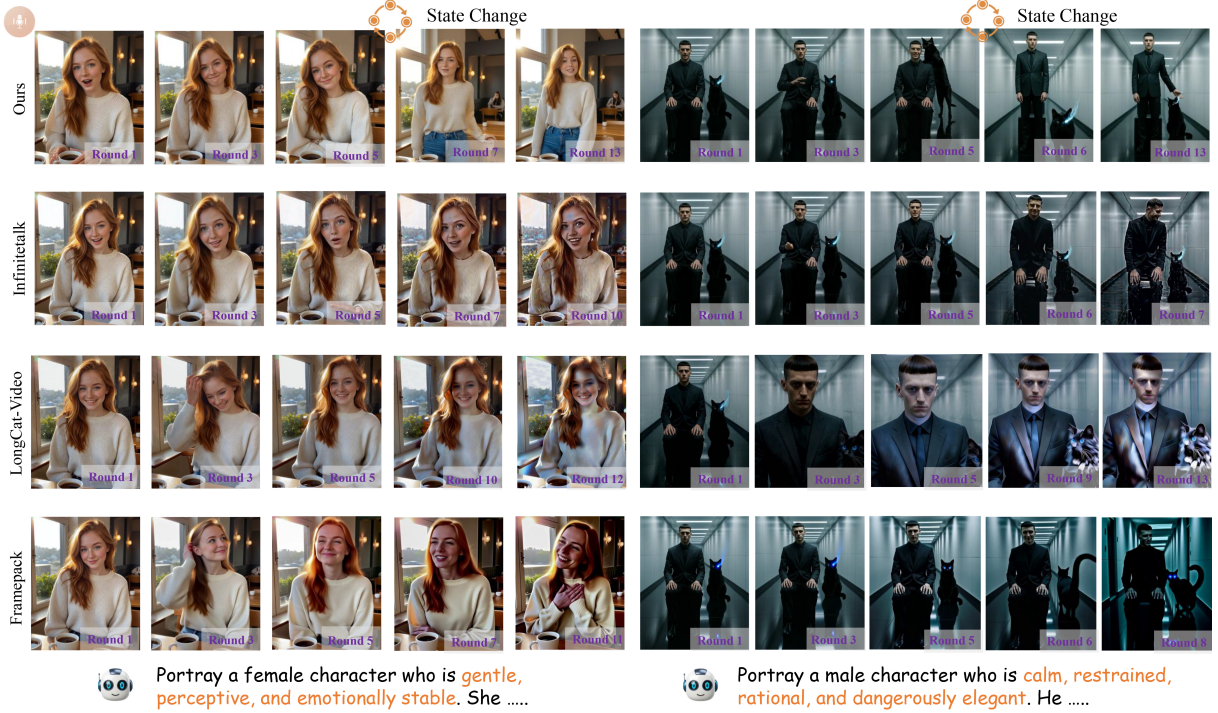


Fig. 3 Qualitative comparison on long-term digital life simulation. Our method successfully generates high-quality, expressive, and drift-free results, accurately performing state transitions (e.g., the 3rd to 4th transition in both examples). In contrast, all baselines suffer from severe drift and fail to execute precise state transitions; for instance, InfiniteTalk’s subject only partially stands up in the second example. Notably, all baselines exhibit severe drift in fewer than 10 interaction rounds (marked in the bottom-right of each image), whereas our method remains stable indefinitely. Best viewed zoomed in.

4.2.1 Qualitative Results

We present a comparison of our results against baselines in Fig. 3. These results were obtained from an interaction process spanning tens of rounds. To be concise, we visualize five of these rounds.

From the results, we can observe the following key characteristics. First, benefiting from our pipeline design, our method leverages a precise anchored video continuation module to accurately control the character’s state during the interaction, thereby achieving natural state transitions. As illustrated in the figure, during the transition from the third interaction scene to the subsequent scenes, only our method successfully enables the character to assume a natural standing posture in both scenes. In contrast, FramePack attempts to stand but only achieves a half-standing state due to the uncontrollable nature of its generative model in the second example. In other examples, the baselines fail to follow the state transition entirely.

Second, our method consistently maintains high-fidelity visual quality and identity preservation throughout long-term interactions. When



Fig. 4 Demonstration of our pipeline’s high-quality continuation capability. In this example, the model initially generates a coffee cup not present in the user-provided scene. Our AVC module seamlessly resolves this by generating a high-quality continuation of the character, plausibly setting the cup down. Subsequently, AVC executes a complex state transition: zooming out as the character stands and simultaneously closes the laptop, enabling more dynamic possibilities for character simulation.

interacting with a character, it is often necessary to engage in ten or more rounds within a single scene. Our experiments reveal that while the baselines

exhibit some anti-drift capabilities, visual degradation in the first few rounds may be subtle to humans; these minor degradations accumulate and are significantly amplified over multiple rounds. This leads to a substantial deterioration in color fidelity and character identity. We observe that the baseline models typically show noticeable degradation around the 5th round, followed by a sharp decline in quality near the 8th round.

These phenomena demonstrate that our video generation pipeline design significantly enhances model controllability. This not only facilitates more accurate state transitions but also ensures that the generation quality remains high and does not degrade, even over extended temporal durations.

To demonstrate the high-quality, consistent continuation capability of our pipeline, we present an example in Fig. 4. In this case, our model successfully handles interactions with newly generated objects (e.g., the character coherently picks up and later smoothly puts down a coffee cup via our AVC module). Furthermore, the model maintains continuity through pose changes. In the second row, for example, the character stands up with coherent camera movement. The model also logically integrates other actions, like closing the computer. This strong video continuation capability is essential for long-term, multi-turn interactions, proving our method’s effectiveness in achieving high-quality, natural generation.

4.2.2 Quantitative Results

To evaluate our pipeline, we generated 60 videos across different scenes and assessed them for both general video quality and click response accuracy. For general quality, we adopted the VBench (Huang et al, 2024) benchmark to measure subject consistency, dynamic degree, and image quality. Recognizing that significant quality differences can emerge between the beginning and end of videos, we also analyzed temporal drifting. Following FramePack (Zhang and Agrawala, 2025), we report quality drift in Table 2, which is defined as the absolute difference in image quality between the first 15% and the last 15% of the frames. We also report lip-sync metrics in the table. For click accuracy, we created 50 diverse interaction pairs using GPT with visually marked images and then manually evaluated whether the model’s generated actions are logical and correct (e.g., ensuring a

click on the “right ear” triggered a corresponding right-sided action).

The results indicate that the pre-trained LongCat-Video model maintains visual quality relatively well compared to other baselines, owing to its foundation in video continuation and the use of super-resolution. However, it tends to generate static scenes with minimal motion, resulting in a low dynamic degree. Although FramePack and InfiniteTalk can mitigate temporal drift to some degree, FramePack exhibits significant subject offsets over time, while its lack of audio-alignment constraints permits greater dynamism compared with other baselines. Additionally, all methods exhibit significant quality drift, while our pipeline achieves a strong balance among quality, motion, and lip-sync accuracy. Specifically, our AVC mechanism enables the model to revert to stable anchor states and dynamically transition between different anchors, which generates more expressive and superior dynamics, yielding strong all-around performance. For click accuracy, we found that using visual markers significantly improves the MLLM performance to a high precision. The MLLM consistently generates logical video prompts and responses, while occasional errors stem from the video model’s fidelity in following these instructions. This straightforward approach validates our pipeline design, proving both its effectiveness and usability.

For completeness, Quality Drift is computed as

$$\Delta M_{\text{drift}}(V) = |M(V_{\text{start}}) - M(V_{\text{end}})|, \quad (5)$$

where V_{start} and V_{end} denote the first and last 15% of the generated sequence. DiVA reduces Quality Drift from 0.0947 for InfiniteTalk, the strongest baseline on this metric, to 0.0341, a relative reduction of approximately 64%. Importantly, this improvement is not obtained by suppressing motion: DiVA also achieves the highest Dynamic Degree of 0.4833.

4.3 Additional Audio-Driven Avatar Baselines

To determine whether the advantage is specific to InfiniteTalk, we further evaluate StableAvatar and WanS2V under the same multi-turn DLS protocol. Table 3 shows that both models accumulate more quality drift than InfiniteTalk and remain

Table 2 Quantitative results. The best result is shown in **bold**. Higher is better except for Quality Drift and Sync-D.

Method	Subject Consis.↑	Dynamic Degree↑	Image Quality↑	Quality Drift↓	Sync-C↑	Sync-D↓	Click Acc.↑
FramePack	0.8979	0.4167	0.5961	0.1293	–	–	–
InfiniteTalk	0.9202	0.3667	0.6143	0.0947	7.21	7.32	–
LongCat-Video	0.8948	0.2833	0.6238	0.1195	–	–	–
DiVA	0.9451	0.4833	0.6481	0.0341	7.53	7.16	0.8200

Table 3 Quality Drift for additional audio-driven avatar baselines. Lower is better.

Method	Quality Drift↓
StableAvatar (Tu et al, 2025)	0.1393
WanS2V (Gao et al, 2025a)	0.1151
InfiniteTalk (Yang et al, 2025b)	0.0947
DiVA	0.0341

substantially behind DiVA. These methods are valuable when the primary requirement is uninterrupted speech-driven avatar rendering with limited state change. DiVA targets a different operating point in which speech-aligned responses are interleaved with explicit large-motion transitions and controlled visual resets.

4.4 Why a Three-Stage Pipeline?

The comparison reveals the stability–expressiveness trade-off that motivates our architecture. General long-video models allow broad motion, but their generated states are repeatedly propagated and therefore drift. Audio-driven avatar models preserve local identity and speech alignment more effectively, but their strong audio coupling restricts macroscopic pose and camera changes. DiVA assigns these conflicting objectives to compatible stages: Waiting Video is optimized for subtle idle naturalness, Action Video for speech alignment and interaction-specific response, and AVC for endpoint control, large state change, and quality restoration. The simultaneous improvement in Dynamic Degree and Quality Drift provides quantitative evidence that the decomposition is not merely an implementation convenience.

The design also exposes the switching logic. The MLLM selects a target state, Action Video completes the spoken response, AVC transitions to the selected anchor, and the corresponding waiting loop persists until the next interaction. State transitions are therefore explicit and inspectable rather than emerging implicitly from a single stream of recent generated frames.

Table 4 User study results on transition naturalness. Preference scores for Transition Fidelity (TF) and Physical Plausibility (PhyP). The best results are in **bold**.

Method	Transition Fidelity↑	Physical Plausibility↑
Wan FLF2V	0.233	0.283
DiVA	0.767	0.717

4.5 Ablation Study

A key component of our design is the anchored video continuation module, which generates the intermediate frames between a previously generated video clip and anchor frames. To validate AVC, we conduct a comparative analysis against the strong Wan FLF2V baseline, with qualitative results presented in Fig. 5. We observe that conventional single-frame interpolation models lack prior knowledge of the character motion and camera dynamics established in the preceding video. Consequently, their interpolated results often suffer from unnatural transitions and camera movements. For instance, in the second example, the character generated by Wan FLF2V suddenly moves out of the frame and executes a rapid, unnatural shift between two distinct actions. In contrast, our model produces a significantly more natural and coherent transition. These temporal artifacts are difficult to capture with static images or general frame-level metrics and are far more pronounced when viewed as videos.

Human evaluation of transition naturalness.

We conducted a user study because camera trajectory smoothness and physical motion continuity are not reliably captured by frame-level metrics. Thirty participants evaluated 60 transition videos spanning 10 character styles and species. Each participant compared DiVA with Wan FLF2V in terms of Transition Fidelity, which assesses camera stability and framing continuity, and Physical Plausibility, which assesses the realism of character and camera motion. As shown in Table 4, DiVA is

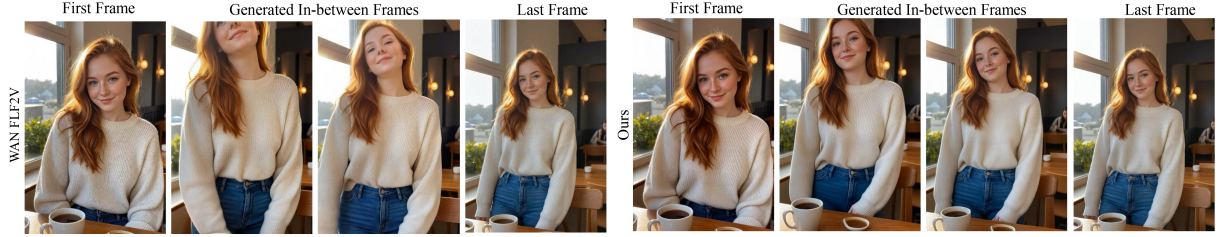


Fig. 5 Comparison of video frame interpolation methods. We compare our method against Wan FL2V. Our approach generates more natural camera movements and character pose transitions, while the base Wan method (relying solely on the last frame) exhibits less harmonious camera work and character motion. Our method effectively avoids these inconsistencies. We recommend viewing the video samples for a clear demonstration of camera and character motion.

preferred in 76.7% of comparisons for Transition Fidelity and 71.7% for Physical Plausibility.

Figure 6 provides a broader qualitative gallery covering photorealistic humans, stylized characters, and anime. The examples further demonstrate that segment conditioning remains effective across appearance domains and substantial viewpoint changes.

Alternative Designs for CLIP Injection. In addition to latent concatenation, we leverage the image encoder of CLIP (Radford et al, 2021) to extract high-level feature representations from the condition images. These representations are subsequently injected into the DiT model via decoupled cross-attention. Let the extracted visual patch features be denoted as $\mathbf{f}_c \in \mathbb{R}^{l \times n \times c}$, where l represents the number of condition images, n is the number of patches per image, and c is the feature dimension. Directly concatenating the features of all condition images along the sequence dimension would result in an excessively long context sequence. Furthermore, adjacent frames in a video segment typically share highly redundant semantic information. To address this, we explore two strategies for processing the CLIP embeddings: (1) *Keyframe Selection*, where we retain only the first frame of the video segment and the stable-state anchor frame; and (2) *Temporal Averaging*, where we compute the mean of the CLIP features across all frames in the video segment to obtain a condensed representation in $\mathbb{R}^{1 \times n \times c}$, which is then concatenated with the feature of the stable-state anchor frame.

A comparison of general quality across 100 generated transition videos reveals negligible performance differences between the two methods. This implies that the global CLIP condition is robust to frame selection (multi-frame versus first-frame), as sufficient information is integrated by

Table 5 Anchor configuration ablation. Historical anchors are generated frames from the preceding interaction history rather than curated visual states.

Configuration	Subject Consis.↑	Dynamic Degree↑	Quality Drift↓
1 Anchor (static)	0.8842	0.4145	0.0911
Historical anchors	0.8633	0.4512	0.1425
5 Curated	0.9451	0.4833	0.0341

the latent condition. We therefore use keyframe selection for efficiency.

4.6 Anchor-State Analysis

We compare three anchor configurations in Table 5. A single static anchor provides a reliable reset but forces diverse interactions back to one pose, reducing dynamic range and causing unnatural spatial warping. Historical anchors use recent generated frames, similar to a single-stream continuation strategy; they preserve more recent motion but also preserve and amplify generated errors. Five curated anchors provide both pose coverage and clean reset points, producing the strongest subject consistency, highest Dynamic Degree, and lowest Quality Drift.

These results show that anchor quality and state coverage are both important. A reset state must be sufficiently clean to remove accumulated error, but the portfolio must also cover enough macroscopic poses to avoid forcing every interaction through one rigid endpoint. The five-anchor configuration provides the best balance between expressive state changes and long-term stability.

4.7 Spatial Click Grounding

Figure 7 compares raw coordinate prompting with our visual-marker strategy. Numerical coordinates require the MLLM to learn an explicit

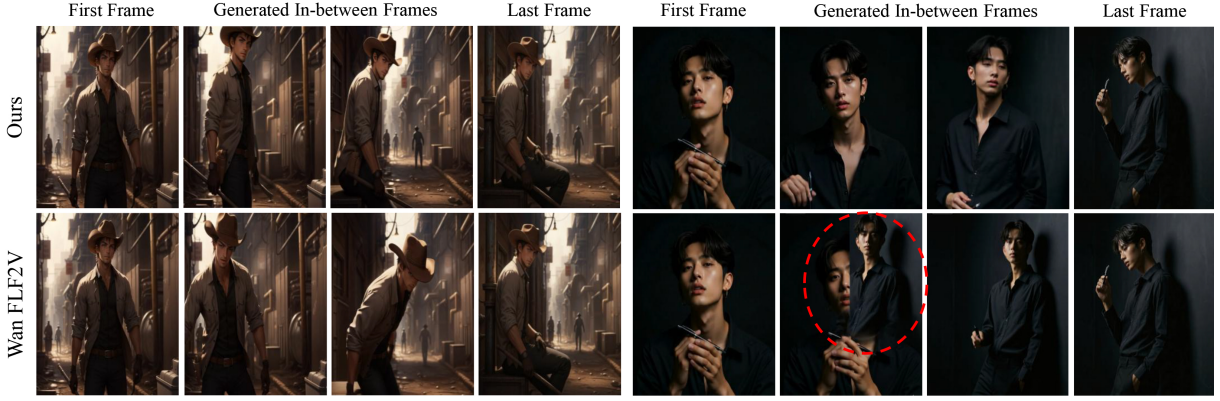


Fig. 6 Qualitative gallery of transition naturalness across diverse character styles. Under substantial pose and camera shifts, Wan FL2V often introduces an abrupt shot change, whereas DiVA maintains continuous motion and camera flow.



Fig. 7 Ablation of spatial click input. Direct coordinate input struggles with precise localization, leading to semantic confusion between adjacent features (e.g., ear versus eye). Visual markers provide explicit spatial grounding for the MLLM.

coordinate-to-pixel correspondence and frequently confuse nearby semantic regions. Rendering the click directly on the current frame turns the problem into visual referring and raises click-response accuracy to 0.82 on 50 interaction pairs. The mechanism requires no additional video-model training and can be applied to arbitrary character categories.

Inference Cost and Speed. Our framework is highly adaptable, offering a flexible trade-off between inference speed and generation quality. By integrating INT8 quantization, feature caching, and step distillation, our primary high-fidelity configuration achieves high-quality motion in four generation steps (approximately 6s per step on an H800 GPU). Importantly, our architecture readily supports causal distillation to enable real-time inference. However, as noted in recent works such as LingBot-World (Team et al, 2026), real-time causal mechanisms inherently compromise complex state transitions, lip-sync accuracy, and movement degree. Therefore, aligned with previous work

(Cheng et al, 2025), we deliberately prioritize high visual fidelity to ensure an immersive simulation. This flexibility establishes our framework as a versatile paradigm for both premium-quality generation and future real-time deployments.

Under the causal configuration evaluated on NVIDIA H800 GPUs, DiVA achieves approximately 1.5s end-to-end response latency and interactive rendering at around 15 FPS. Under the same evaluation setting, FramePack and InfiniteTalk require approximately 28.5s and 24.5s per turn, respectively. X-Streamer reports 25 FPS at 256×256 resolution (Xie et al, 2025), but targets video-call-style portrait streaming. Direct FPS comparison should therefore be interpreted cautiously, because DiVA additionally performs large pose and camera transitions at higher resolution and explicitly routes among action, transition, and waiting states. Figure 8 summarizes the quality-latency trade-off.

4.8 Applications of DiVA

Our model inherently understands multimodal knowledge and can simulate diverse behavioral styles, enabling it to adeptly perform a wide range of tasks. In Fig. 9, we demonstrate several such scenarios. In the first example, the user defines the digital character as a companion. In this role, the character’s actions and linguistic responses adopt a gentle and supportive demeanor, with each reply consistently aligning with the defined companion persona. In the second example, the character is defined as an academic assistant and provided with external documentation to lecture on. In this



Fig. 8 Quality-latency trade-off for DiVA configurations. Causal distillation, feature caching, and quantization enable interactive throughput, while the full configuration prioritizes generation quality.



Fig. 9 Persona-driven interaction. Our method dynamically defines distinct character personas, enabling varied roles with appropriate verbal and non-verbal responses. The examples include a companion persona offering empathetic interaction with fitting gestures and an academic assistant explaining complex PDFs (e.g., “Attention Is All You Need”). The assistant provides factually correct explanations and uses gestures that enhance clarity.

context, the character adopts a more formal posture, and its responses are factually accurate; for instance, it correctly recommends the Annotated Transformer (Rush, 2018) and provides a precise explanation of the attention mechanism. This versatility in role-playing highlights the significant potential of our model for future digital character applications.

4.9 Discussion of Visual Hallucination

While video models achieve high realism in character simulation, they are prone to illogical artifacts akin to clipping in 3D modeling, which we term visual hallucination. As illustrated in Fig. 10, we present two instances. In the first case, the model faithfully renders a rectangular matrix with nodes based on a transformer-related prompt, yet the resulting object lacks physical grounding. In the second case, the character interacts with a key that suddenly appears, although it is absent from the original frame. These instances demonstrate the model’s robust instruction-following. This highlights a trade-off where the model’s creative and imaginative response, often valued by users, takes

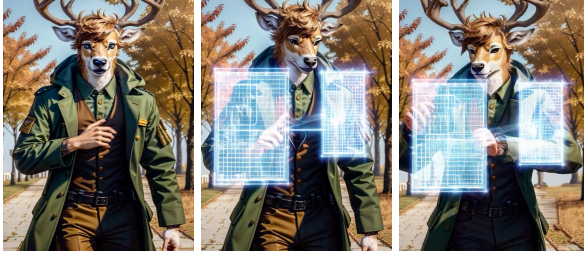
precedence over strict physical realism. Balancing this trade-off remains a direction for future research.

An anchor can limit the persistence of a transient hallucination when AVC generates a plausible removal action before returning to a clean state, as illustrated by the coffee-cup example in Fig. 4. This should not be interpreted as guaranteed semantic correction: AVC controls the transition endpoint and can interrupt local error accumulation, but it does not solve the underlying physical reasoning limitations of the video backbone.

4.10 Limitations and Boundary Cases

DiVA inherits fine-grained generation failures from the underlying video models, especially for hands, object contacts, and physically unusual interactions. When the selected target anchor differs extremely from the current action segment, AVC generally preserves a smooth transition but may become cross-fade-like rather than producing fully plausible intermediate motion. Moreover, the anchor portfolio discretizes macroscopic states; although users may append new anchors without retraining, a small portfolio cannot cover

Because I've learned **Transformer**..
 Prompt : As his fingers move, ethereal lines materialize in the thin air, tracing his gestures to form a **rectangular matrix of nodes and connections**. He ...



Perhaps you've already made it to the doorway. You're just missing **a key**.
 Prompt: The demoness slowly uncrosses her arms, placing one hand on her hip while the other gestures subtly as **if holding an invisible key**. Her...



Fig. 10 Discussion of visual hallucination. The video model’s powerful generative imagination allows it to follow complex prompts, such as visualizing “Transformer” as a square matrix or hallucinating a key. These cases demonstrate advanced prompt-following fidelity, even when the results defy physical common sense.

every possible pose or scene configuration. Future work may combine automatic anchor expansion, stronger physical priors, and online state discovery to enlarge the reachable visual state space while retaining explicit quality control.

5 Conclusion

We introduce DiVA, a novel deep simulator designed for building an open-ended interactive digital character world. This represents a crucial step in pushing video generation beyond passive clips toward long-term, immersive digital life simulation. DiVA leverages an MLLM as a router to dynamically process complex user linguistic and spatial input, enabling a highly persona-driven, open-ended simulation experience. To maintain high dynamics and visual stability during continuous interaction, we designed a three-part coupled video pipeline (consisting of waiting, action, and anchored video continuation). This pipeline, featuring the proposed AVC mechanism, effectively suppresses temporal identity drift while simultaneously unlocking greater dynamic expression. Ultimately, DiVA provides a robust foundation for simulating long-duration, controllable characters, advancing the future of generative interactive worlds.

The expanded analyses further show that the benefit comes from the combination of explicit state routing, segment-conditioned transition generation, and curated visual reset states: removing any of these properties weakens either expressiveness or long-term stability. At the same time, the remaining hand, contact, and extreme-transition

failures indicate that stateful system design complements rather than replaces progress in the underlying video foundation models.

Declarations

Funding

Funding information will be completed before submission.

Competing Interests

The authors declare that they have no known competing financial or non-financial interests that could have influenced the work reported in this paper.

Ethics Approval and Consent to Participate

The user study reports aggregate preferences and does not collect sensitive personal data. The authors will verify whether institutional ethics approval or an exemption statement is required before submission.

Data Availability

The publicly available datasets and pretrained models used in this study are cited in the manuscript. The internally curated training data cannot currently be redistributed because of source-data and institutional restrictions. Evaluation prompts, protocols, and generated results are intended to be released with the project materials, subject to institutional approval.

Code Availability

The authors intend to release inference and evaluation code, model adaptations, and prompt templates subject to institutional approval.

Author Contributions

Author-contribution statements will be completed and verified before submission.

References

- Achiam J, Adler S, Agarwal S, et al (2023) Gpt-4 technical report. arXiv preprint arXiv:230308774
- Bahmani S, Skorokhodov I, Qian G, et al (2025) Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 22875–22889
- Bai J, Bai S, Chu Y, et al (2023) Qwen technical report. arXiv preprint arXiv:230916609
- Bai J, Xia M, Fu X, et al (2025) Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:250311647
- Blattmann A, Dockhorn T, Kulal S, et al (2023) Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:231115127
- Cai S, Yang C, Zhang L, et al (2025) Mixture of contexts for long video generation. arXiv preprint arXiv:250821058
- Chan ER, Lin CZ, Chan MA, et al (2022) Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 16123–16133
- Chen C, Tan Z, Cheng Q, et al (2022) UTC: A unified transformer with inter-task contrastive learning for visual dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 18103–18112
- Chen C, Yang X, Yang F, et al (2024) Sculpt3D: Multi-view consistent text-to-3d generation with sparse 3d prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10228–10237
- Chen C, Wei J, Chen T, et al (2025) CADCrafter: Generating computer-aided design models from unconstrained images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11073–11082
- Cheng J, Ge Y, Ge Y, et al (2025) Animegamer: Infinite anime life simulation with next game state prediction. arXiv preprint arXiv:250401014
- Dong Q, Li L, Dai D, et al (2022) A survey on in-context learning. arXiv preprint arXiv:230100234
- Du Z, Chen Q, Zhang S, et al (2024) Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. arXiv preprint arXiv:240705407
- Fu Z, Zeng X, Lan J, et al (2026) iMontage: Unified, versatile, highly dynamic many-to-many image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16237–16247
- Gao X, Hu L, Hu S, et al (2025a) Wan-s2v: Audio-driven cinematic video generation. arXiv preprint arXiv:250818621
- Gao Y, Guo H, Hoang T, et al (2025b) Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:250609113
- Gemini Team G (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Tech. rep., arXiv preprint arXiv:2403.05530, URL <https://arxiv.org/abs/2403.05530>
- Geng D, Herrmann C, Hur J, et al (2025) Motion prompting: Controlling video generation with motion trajectories. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 1–12
- Google (2025) Veo 3. <https://aistudio.google.com/models/veo-3/>, accessed: 2025-10-17

- Hu EJ, Shen Y, Wallis P, et al (2022) Lora: Low-rank adaptation of large language models. ICLR 1(2):3
- Huang Y, Guo H, Wu F, et al (2025) Live avatar: Streaming real-time audio-driven avatar generation with infinite length. arXiv preprint arXiv:251204677
- Huang Z, He Y, Yu J, et al (2024) Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 21807–21818
- Kong W, Tian Q, Zhang Z, et al (2024) Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:241203603
- Kong Z, Gao F, Zhang Y, et al (2025) Let them talk: Audio-driven multi-person conversational video generation. arXiv preprint arXiv:250522647
- Kuaishou Technology (2025) Kling video generation model. <https://klingai.com/cn/>, accessed: 2025-10-17
- Li J, Li Y, Wadhwa N, et al (2024) Unbounded: A generative infinite game of character life simulation. arXiv preprint arXiv:241018975
- Liu A, Feng B, Xue B, et al (2024) Deepseek-v3 technical report. arXiv preprint arXiv:241219437
- Luo Y, Bai J, Shi X, et al (2025) Camclonemaster: Enabling reference-based camera control for video generation. arXiv preprint arXiv:250603140
- Meng R, Zhang X, Li Y, et al (2025) Echomimicv2: Towards striking, simplified, and semi-body human animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 5489–5498
- Meng Y, Ouyang H, Yu Y, et al (2026) HoloCine: Holistic generation of cinematic multi-shot long video narratives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 461–471
- OpenAI (2025) Sora 2. <https://openai.com/index/sora-2/>, accessed: 2025-10-17
- Pell B (1992) Metagame: A new challenge for games and learning. Tech. rep., University of Cambridge, Computer Laboratory
- Polyak A, Zohar A, Brown A, et al (2024) Movie gen: A cast of media foundation models. arXiv preprint arXiv:241013720
- Qian S, Kirschstein T, Schoneveld L, et al (2024) Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 20299–20309
- Qiu L, Gu X, Li P, et al (2025) Lhm: Large animatable human reconstruction model from a single image in seconds. arXiv preprint arXiv:250310625
- Radford A, Kim JW, Hallacy C, et al (2021) Learning transferable visual models from natural language supervision. In: International conference on machine learning, PmLR, pp 8748–8763
- Ruhe D, Heek J, Salimans T, et al (????) Rolling diffusion models, 2024. URL <https://arxiv.org/abs/240209470>
- Rush A (2018) The annotated transformer. <http://nlp.seas.harvard.edu/2018/04/03/attention.html>
- Team ML, Cai X, Huang Q, et al (2025) Longcat-video technical report. arXiv preprint arXiv:251022200
- Team R, Gao Z, Wang Q, et al (2026) Advancing open-source world models. arXiv preprint arXiv:260120540
- Tian L, Hu S, Wang Q, et al (2025) Emo2: End-effector guided audio-driven avatar video generation. arXiv preprint arXiv:250110687
- Tu S, Pan Y, Huang Y, et al (2025) Stableavatar: Infinite-length audio-driven avatar video generation. arXiv preprint arXiv:250808248
- Wan T, Wang A, Ai B, et al (2025) Wan: Open and advanced large-scale video generative models.

arXiv preprint arXiv:250320314

Wang H, Ouyang H, Wang Q, et al (2025) The world is your canvas: Painting promptable events with reference images, trajectories, and text. arXiv preprint arXiv:251216924

Xie Y, Gu T, Li Z, et al (2025) X-streamer: Unified human world modeling with audiovisual interaction. arXiv preprint arXiv:250921574

Yang S, Huang W, Chu R, et al (2025a) Longlive: Real-time interactive long video generation. arXiv preprint arXiv:250922622

Yang S, Kong Z, Gao F, et al (2025b) Infinitetalk: Audio-driven video generation for sparse-frame video dubbing. arXiv preprint arXiv:250814033

Yin S, Wu C, Liang J, et al (2023) Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:230808089

Zhang L, Agrawala M (2025) Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:250412626

Zhang T, Bi S, Hong Y, et al (2025a) Test-time training done right. arXiv preprint arXiv:250523884

Zhang Z, Liao J, Li M, et al (2025b) Tora: Trajectory-oriented diffusion transformer for video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference, pp 2063–2073

Zhu J, Villareale J, Javvaji N, et al (2021) Player-ai interaction: What neural network games reveal about ai as play. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp 1–17