Novelty Assessment Report

Paper: LongLive: Real-time Interactive Long Video Generation
PDF URL: https://openreview.net/pdf?id=nCAODkpsPJ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

We present LongLive, a frame-level autoregressive (AR) framework for real-time and interactive long video generation. Long video generation presents challenges in both efficiency and quality. Diffusion and Diffusion-Forcing models can produce high-quality videos but suffer from low efficiency due to bidirectional attention. Causal attention AR models support KV caching for faster inference but often degrade in quality on long videos due to memory challenges during long-video training. In addition, beyond static prompt-based generation, interactive capabilities, such as streaming prompt inputs, are critical for dynamic content creation, enabling users to guide narratives in real time. This interactive requirement significantly increases the complexity, especially in ensuring visual consistency and semantic coherence during prompt transitions. To address these challenges, LongLive adopts a causal, frame-level AR design that integrates a KV-recache mechanism that refreshes cached states with the new prompt for smooth, adherent switches streaming long tuning to enable long video training and to align training and inference (train-long–test-long); and short window attention paired with a frame-level attention sink, preserving long-range consistency while enabling faster generation. With these key designs, LongLive fine-tunes a 1.3B-parameter short-clip model to minute-long generation in just 32 GPU-days. At inference, LongLive sustains 20.7 FPS on a single NVIDIA H100, achieves strong performance on VBench in both short- and long-video settings. LongLive supports up to 240-second videos on a single H100 GPU. With FP8 quantization, LongLive boosts inference to 24.8 FPS with marginal quality loss.

Core Task Landscape

This paper addresses: **Real-Time Interactive Long Video Generation**

A total of **50 papers** were analyzed and organized into a taxonomy with **29 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Streaming Autoregressive Generation Architectures**
- **Real-Time Diffusion-Based Interactive Generation**
- **Interactive World Models and Game Engines**
- **Multimodal Interactive Avatar and Digital Human Synthesis**
- **Domain-Specific Real-Time Interactive Video Applications**
- **Supporting Techniques for Real-Time Interactive Video**
- **Surveys, Benchmarks, and Foundational Frameworks**
- **Specialized Real-Time Synthesis and Rendering Techniques**

Complete Taxonomy Tree

- Real-Time Interactive Long Video Generation Survey Taxonomy
- Streaming Autoregressive Generation Architectures
  - Frame-Level Autoregressive Models with Memory Mechanisms ★ (3 papers)
  - [0] LongLive: Real-time Interactive Long Video Generation (Anon et al., 2026) View paper
  - [7] VideoSSM: Autoregressive Long Video Generation with Hybrid State-Space Memory (Yifei Yu, 2025) View paper
  - [49] RELIC: Interactive Video World Model with Long-Horizon Memory (Yicong Hong, 2025) View paper
  - Chunk-Based and Block-Diffusion Autoregressive Models (2 papers)
  - [14] Inferix: A Block-Diffusion based Next-Generation Inference Engine for World Simulation (Inferix Team, 2025) View paper
  - [16] MAGI-1: Autoregressive Video Generation at Scale (Sand. ai, 2025) View paper
  - Autoregressive Models with Adversarial or Distillation Training (2 papers)
  - [9] Rolling forcing: Autoregressive long video diffusion in real time (Liu, 2025) View paper
  - [20] Autoregressive Adversarial Post-Training for Real-Time Interactive Video Generation (Lin, 2025) View paper
- Real-Time Diffusion-Based Interactive Generation
  - Flow Matching and Streaming Diffusion Frameworks (2 papers)
  - [2] Streamdit: Real-time streaming text-to-video generation (Kodaira Akio, 2025) View paper
  - [35] StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation (Tianrui Feng, 2025) View paper
  - Motion-Conditioned Real-Time Diffusion Models (1 papers)
  - [4] MotionStream: Real-Time Video Generation with Interactive Motion Controls (Shin, 2025) View paper
  - Distributed Inference and Pipeline Parallelism for Diffusion (1 papers)
  - [19] Live Avatar: Streaming Real-time Audio-Driven Avatar Generation with Infinite Length (Yubo Huang, 2025) View paper
- Interactive World Models and Game Engines
  - Neural Game Engines with Action Conditioning (4 papers)
  - [1] Yan: Foundational interactive video generation (Ye, 2025) View paper
  - [3] Diffusion models are real-time game engines (Valevski, 2024) View paper

    - [10] Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition (Li JiaQi, 2025) View paper
    - [36] The Matrix: Infinite-Horizon World Generation with Real-Time Moving Control (Feng Rui-li, 2024) View paper
    - Action-Conditioned World Models for Long-Horizon Planning (2 papers)
    - [17] Matrix-Game 2.0: An Open-Source, Real-Time, and Streaming Interactive World Model (Peng Chunli, 2025) View paper
    - [28] Learning World Models for Interactive Video Generation (Hu Xun, 2025) View paper
    - Scene Reconstruction and Physics-Based Interactive Environments (1 papers)
    - [15] Video2Game: Real-time, Interactive, Realistic and Browser-Compatible Environment from a Single Video (Hongchi Xia, 2024) View paper
- Multimodal Interactive Avatar and Digital Human Synthesis
    - Audio-Driven Real-Time Portrait Video Generation (2 papers)
    - [13] LLIA - Enabling Low-Latency Interactive Avatars: Real-Time Audio-Driven Portrait Video Generation with Diffusion Models (Yu, 2025) View paper
    - [26] ChatAnyone: Stylized Real-time Portrait Video Generation with Hierarchical Motion Diffusion Model (Qi, 2025) View paper
    - Multimodal Interactive Digital Human Frameworks (2 papers)
    - [8] MIDAS: Multimodal Interactive Digital-humAn Synthesis via Real-time Autoregressive Video Generation (Chen Ming, 2025) View paper
    - [11] X-Streamer: Unified Human World Modeling with Audiovisual Interaction (Xie You, 2025) View paper
    - Autoregressive Interactive Head and Body Motion Generation (1 papers)
    - [42] ARIG: Autoregressive Interactive Head Generation for Real-time Conversations (Guo Ying, 2025) View paper
- Domain-Specific Real-Time Interactive Video Applications
    - Traffic Scene and Vehicular Network Applications (1 papers)
    - [6] AIGC-Driven Real-Time Interactive 4-D Traffic Scene Generation in Vehicular Networks (Xiaolong Li, 2025) View paper
    - Embodied AI and Robotics with Multimodal Reasoning (3 papers)
    - [23] Robovqa: Multimodal long-horizon reasoning for robotics (Pierre Sermanet, 2024) View paper
    - [27] Proactive Assistant Dialogue Generation from Streaming Egocentric Videos (Zhang, 2025) View paper
    - [43] Vinci: A real-time embodied smart assistant based on egocentric vision-language model (Huang Yi-fei, 2024) View paper
    - Extended Reality and Immersive Video Streaming (4 papers)
    - [12] VARFVV: View-Adaptive Real-Time Interactive Free-View Video Streaming With Edge Computing (Qiang Hu, 2025) View paper
    - [37] Ray tracing-based construction of 3D background model for real-time stereoscopic rendering of live immersive video (Youngwook Kim, 2024) View paper
    - [38] From 2D to 3D video conferencing: modular RGB-D capture and reconstruction for interactive natural user representations in immersive extended reality (XR) communication (Simon N. B. Gunkel, 2023) View paper
    - [50] Live4D: A Real-time Capture System for Streamable Volumetric Video (Yifeng Zhou, 2023) View paper
- Supporting Techniques for Real-Time Interactive Video
    - Real-Time Rendering and Compression for Neural Representations (2 papers)
    - [32] Dimensionality Reduction for the Real-Time Light-Field View Synthesis of Kernel-Based Models (Martijn Courteaux, 2024) View paper
    - [48] BakedAvatar: Baking Neural Fields for Real-Time Head Avatar Synthesis (Hao-Bin Duan, 2023) View paper
    - Real-Time Motion Control and Character Animation (2 papers)
    - [22] DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control (Zhao, 2024) View paper
    - [24] Interactive Character Control with Auto-Regressive Motion Diffusion Models (Yi Shi, 2023) View paper
    - Streaming Video Understanding and Online Processing (2 papers)
    - [25] VideoLLM-online: Online Video Large Language Model for Streaming Video (Joya Chen, 2024) View paper
    - [30] Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions (Zhang Pan, 2024) View paper
    - Dynamic Texture Synthesis and Real-Time Procedural Generation (1 papers)
    - [39] DyNCA: Real-Time Dynamic Texture Synthesis Using Neural Cellular Automata (Ehsan Pajouheshgar, 2022) View paper
- Surveys, Benchmarks, and Foundational Frameworks
    - Surveys and Taxonomies of Interactive Generative Video (2 papers)
    - [5] A survey of interactive generative video (Yu Jiwen, 2025) View paper
    - [41] From Masks to Worlds: A Hitchhiker's Guide to World Models (Bai Jin-bin, 2025) View paper
    - Reactive Closed-Loop Benchmarks and Evaluation Systems (1 papers)
    - [40] Bench2Drive-R: Turning Real World Data into Reactive Closed-Loop Autonomous Driving Benchmark by Generative Model (YOU Junqi, 2024) View paper
    - Predictive Control and Real-Time Behavior Synthesis (1 papers)
    - [18] Predictive Sampling: Real-time Behaviour Synthesis with MuJoCo (Howell, 2022) View paper
- Specialized Real-Time Synthesis and Rendering Techniques
    - Photo-Realistic Avatar Reconstruction and Animation (2 papers)
    - [44] StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video (Lizhen Wang, 2023) View paper
    - [46] Animatable Virtual Humans: Learning Pose-Dependent Human Representations in UV Space for Interactive Performance Synthesis (Wieland Morgenstern, 2023) View paper
    - Real-Time Interactive Projections and Extended Reality Installations (2 papers)
    - [31] Image Synthesis from a Collection of Depth Enhanced Panoramas: Creating Interactive Extended Reality Experiences from Static Images (Thomas Marrinan, 2024) View paper
    - [33] Island Design Camps - Interactive Video Projections as Extended Realities (Bert Bongers, 2023) View paper
    - Real-Time Tactile and Embodied Perception Systems (1 papers)
    - [34] GelSLAM: A Real-time, High-Fidelity, and Robust 3D Tactile SLAM System (Huang, 2025) View paper
    - Minute-Length High-Resolution Generation with Linear Complexity (1 papers)
    - [47] LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity (Hongjie Wang, 2025) View paper
    - Hierarchical World Models and Unified Control Frameworks (1 papers)

- [29] MinD: Unified Visual Imagination and Control via Hierarchical World Models (X Chi, 2025) View paper
- Social Interactive Human Video Synthesis (1 papers)
- [45] Social interactive human video synthesis (Dumebi Okwechime, 2010) View paper
- Personalized Short Video Content Generation (1 papers)
- [21] Application of multimodal generation model in short video content personalized generation (Minghui Yang, 2025) View paper

Narrative

Core task: real-time interactive long video generation. The field encompasses diverse architectural paradigms and application domains, organized into eight main branches. Streaming Autoregressive Generation Architectures focus on frame-by-frame synthesis with memory mechanisms to maintain temporal coherence, as seen in works like VideoSSM[7] and RELIC[49]. Real-Time Diffusion-Based Interactive Generation explores efficient diffusion sampling strategies such as StreamDiffusionV2[35] and Streamdit[2] that enable low-latency synthesis. Interactive World Models and Game Engines, exemplified by Diffusion Game Engines[3] and Hunyuan GameCraft[10], build controllable environments for interactive simulation. Multimodal Interactive Avatar and Digital Human Synthesis addresses real-time character animation and conversational agents, while Domain-Specific Real-Time Interactive Video Applications target specialized use cases like autonomous driving scenarios (AIGC Traffic Scene[6]) and robotics. Supporting Techniques provide foundational methods for compression, scheduling, and optimization, and Surveys, Benchmarks, and Foundational Frameworks (Interactive Generative Video Survey[5]) offer structured evaluations. Specialized Real-Time Synthesis and Rendering Techniques handle rendering pipelines and view synthesis for immersive experiences.

Several active lines of work reveal key trade-offs between generation quality, latency, and controllability. Autoregressive models with memory mechanisms balance long-range coherence against computational overhead, while diffusion-based approaches trade sampling steps for real-time responsiveness. Interactive world models emphasize user control and physical plausibility, often at the cost of visual fidelity compared to purely generative methods. LongLive[0] sits within the Streaming Autoregressive Generation branch, specifically among Frame-Level Autoregressive Models with Memory Mechanisms, alongside VideoSSM[7] and RELIC[49]. While VideoSSM[7] leverages state-space models for efficient temporal modeling and RELIC[49] emphasizes retrieval-augmented context, LongLive[0] appears to prioritize extended temporal consistency across very long sequences, addressing the challenge of maintaining coherent narratives and visual stability over extended interactive sessions without catastrophic drift.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. VideoSSM: Autoregressive Long Video Generation with Hybrid State-Space Memory

Authors: Yifei Yu, Xiaoshan Wu, Xinting Hu, Tao Hu, Yangtian Sun, et al. (11 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Autoregressive (AR) diffusion enables streaming, interactive long-video generation by producing frames causally, yet maintaining coherence over minute-scale horizons remains challenging due to accumulated errors, motion drift, and content repetition. We approach this problem from a memory perspective, treating video synthesis as a recurrent dynamical process that requires coordinated short- and long-term context. We propose VideoSSM, a Long Video Model that unifies AR diffusion with a hybrid sta...

Relationship Analysis

Both papers belong to the Frame-Level Autoregressive Models with Memory Mechanisms category, employing causal frame-by-frame generation with memory strategies for long-horizon video coherence. They overlap in addressing real-time interactive long video generation through KV-cache management, streaming training strategies (streaming long tuning vs. self-rollout), and attention mechanisms (short window + frame sink vs. sliding window + SSM). The key difference is that LongLive uses a static frame-level attention sink with KV-recache for prompt switching, while VideoSSM introduces a dynamic hybrid state-space memory (SSM-based global compressed memory) that continuously evolves to avoid content repetition and frozen patterns.

---

2. RELIC: Interactive Video World Model with Long-Horizon Memory

Authors: Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

A truly interactive world model requires three key ingredients: real-time long-horizon streaming, consistent spatial memory, and precise user control. However, most existing approaches address only one of these aspects in isolation, as achieving all three simultaneously is highly challenging-for example, long-term memory mechanisms often degrade real-time performance. In this work, we present RELIC, a unified framework that tackles these three challenges altogether. Given a single image and a te...

Relationship Analysis

Both papers belong to the Frame-Level Autoregressive Models with Memory Mechanisms category, employing causal frame-by-frame generation with KV-cache mechanisms for long-horizon video generation. They overlap in using streaming autoregressive architectures with memory mechanisms (KV-cache, attention sinks) to maintain temporal coherence during real-time generation. However, LongLive focuses on interactive prompt switching via KV-recache and streaming long tuning for text-driven generation, while RELIC emphasizes camera-pose-aware memory compression and spatial consistency for action-controlled world modeling with explicit 3D scene retrieval capabilities.

Contributions Analysis

Overall novelty summary. LongLive proposes a frame-level autoregressive framework for real-time interactive long video generation, combining KV-recache for prompt switching, streaming long tuning for extended temporal coherence, and short window attention with attention sinks. The paper resides in the 'Frame-Level Autoregressive Models with Memory Mechanisms' leaf, which contains only three papers total (including LongLive itself). This is a relatively sparse research direction within the broader taxonomy of fifty papers across twenty-nine leaf nodes, suggesting the work targets a specific niche at the intersection of causal autoregressive generation and long-horizon interactive synthesis.

The taxonomy reveals that LongLive's leaf sits within the 'Streaming Autoregressive Generation Architectures' branch, which also includes chunk-based and adversarial autoregressive methods. Neighboring branches explore real-time diffusion frameworks (e.g., flow matching, pipeline parallelism) and interactive world models with action conditioning. While diffusion-based methods prioritize visual quality through bidirectional attention, LongLive's causal design trades some modeling capacity for KV-caching efficiency. The sibling papers VideoSSM and RELIC address similar memory challenges but through state-space models and retrieval augmentation respectively, whereas LongLive emphasizes recache mechanisms and streaming tuning for interactive prompt transitions.

Among twenty-three candidates examined, the contribution-level analysis shows varied prior work overlap. The KV-recache mechanism for interactive prompt switching examined seven candidates with no clear refutations, suggesting relative novelty in this specific interactive control paradigm. The streaming long tuning strategy examined ten candidates and found one refutable match, indicating some existing work on long-sequence training alignment. The short window attention with frame-level attention sink examined six

candidates and identified one refutable prior, suggesting that attention sink techniques for autoregressive video models have been explored previously, though possibly in different architectural contexts or application domains.

Based on the limited search scope of twenty-three semantically similar candidates, LongLive appears to combine several existing techniques (attention sinks, streaming training) with a novel interactive control mechanism (KV-recache). The work's position in a sparse taxonomy leaf and the absence of refutations for the KV-recache contribution suggest some originality in the interactive prompt-switching aspect. However, the analysis does not cover the full breadth of autoregressive video generation literature, and the two refuted contributions indicate that components of the approach build on established methods for long-sequence modeling and attention optimization.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: KV-recache mechanism for interactive prompt switching

Description: The authors introduce a KV-recache technique that refreshes cached key-value states at prompt boundaries by recomputing them using previously generated frames and the new prompt. This enables smooth visual transitions while maintaining semantic alignment with the new prompt during interactive video generation.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. MemFlow: Flowing Adaptive Memory for Consistent and Efficient Long Video Narratives

URL: View paper

Brief Assessment

MemFlow[53] focuses on memory retrieval and activation strategies for long-term consistency, not on refreshing cached states at prompt boundaries. The candidate's memory mechanism operates differently from the KV-recache technique described in the original paper.

---

2. FilmWeaver: Weaving Consistent Multi-Shot Videos with Cache-Guided Autoregressive Diffusion

URL: View paper

Brief Assessment

FilmWeaver[51] addresses multi-shot video generation with a dual-level cache (shot cache and temporal cache) for maintaining consistency across different shots in narrative videos, not interactive prompt switching during continuous generation. The technical focus and application domain differ fundamentally from the original paper's KV-recache for real-time prompt transitions.

---

3. Playing For You: Text Prompt-guided Joint Audio-visual Generation for Narrating Faces using Multi-entangled Latent Space

URL: View paper

Brief Assessment

Playing For You[56] focuses on joint audio-visual generation for talking faces using multi-entangled latent spaces between audio and video modalities. It does not address KV-cache mechanisms for prompt switching in video generation.

---

4. Contextual Knowledge Infusion via Iterative Semantic Tracing for Visionâ□□Language Understanding

URL: View paper

Brief Assessment

Contextual Knowledge Infusion[54] focuses on vision-language understanding tasks, not video generation with prompt switching. The candidate's 'dynamic key-value memory module' serves a different purpose than the original's KV-recache for interactive video generation.

---

5. EgoLCD: Egocentric Video Generation with Long Context Diffusion

URL: View paper

Brief Assessment

EgoLCD[55] focuses on long-term sparse KV caching for egocentric video generation with semantic retrieval, not interactive prompt switching. The candidate's memory mechanism retrieves historical segments based on semantic similarity for temporal consistency, whereas the original paper's KV-recache specifically refreshes cached states at prompt boundaries to enable smooth transitions during interactive generation with multiple user prompts.

---

6. SneakPeek: Future-Guided Instructional Streaming Video Generation

URL: View paper

Brief Assessment

SneakPeek[52] focuses on future-guided streaming video generation with dual-region KV caching for past frames and future keyframes, not on refreshing cached states at prompt boundaries for interactive prompt switching as in the original paper.

---

7. MotionStream: Real-Time Video Generation with Interactive Motion Controls

URL: View paper

Brief Assessment

MotionStream[4] focuses on motion-conditioned video generation with sliding-window causal attention and attention sinks for infinite-length streaming, not on prompt switching or refreshing cached states with new prompts.

Contribution 2: Streaming long tuning strategy

Description: The authors propose a train-long-test-long training procedure that exposes the model to extended self-generated sequences during training. This approach aligns training with inference conditions by iteratively generating short clips conditioned on previously cached states, mitigating error accumulation and quality degradation in long videos.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Macro-from-micro planning for high-quality and parallelized autoregressive long video generation

URL: View paper

Brief Assessment

Macro-from-micro Planning[65] focuses on hierarchical planning (micro and macro planning) for long video generation, not on train-long-test-long training procedures. The candidate does not address streaming long tuning or iterative generation with cached states during training.

---

2. DeepVerse: 4D Autoregressive Video Generation as a World Model

URL: View paper

Brief Assessment

DeepVerse[68] focuses on 4D autoregressive video generation with geometric constraints for world modeling, not on train-long-test-long strategies for general video generation consistency. The candidate does not address the streaming long tuning approach described in the original paper.

---

3. Autoregressive Adversarial Post-Training for Real-Time Interactive Video Generation

URL: View paper

Brief Assessment

Autoregressive Adversarial Post-Training[20] focuses on adversarial post-training for single-step generation with student-forcing, not on the train-long-test-long strategy of iteratively generating short clips conditioned on cached states during training.

---

4. Loong: Generating Minute-level Long Videos with Autoregressive Language Models

URL: View paper

Prior Art Analysis

Loong[66] demonstrates that similar train-long-test-long strategies for autoregressive video generation were proposed prior to the original paper. Both papers address the same fundamental problem: the train-short-test-long mismatch in autoregressive video models leads to quality degradation on long videos. Loong[66] proposes 'progressive short-to-long training' that exposes the model to increasingly longer sequences during training, which directly parallels the original paper's 'streaming long tuning' approach of training on extended self-generated sequences. Both methods aim to align training with inference conditions to mitigate error accumulation in long video generation.

Evidence

Evidence 1 - **Rationale**: Both papers propose training strategies that progressively expose the model to longer sequences during training to address the train-test mismatch in autoregressive video generation. The original paper's 'train-long-test-long strategy' and Loong[66]'s 'progressive short-to-long training' serve the same purpose of aligning training with inference conditions. - **Original**: To address this mismatch, we propose atrain-long-test-long strategy. During training, the model synthesizes long sequences by conditioning on its own imperfect predictions, with supervision applied throughout the entire rollout. This exposes the model to extended, self-generated, and progressively d... - **Candidate**: We propose progressive short-to-long training with a loss re-weighting scheme to mitigate the loss imbalance problem for long video training.

Evidence 2 - **Rationale**: Both papers explicitly address error accumulation as a key problem in long video generation and propose training strategies to mitigate it by exposing models to their own predictions during training. - **Original**: This exposes the model to extended, self-generated, and progressively degraded frames already in training, aligning training with inference, mitigating error accumulation to improve fidelity and consistency. - **Candidate**: We further investigate inference strategies, including video token re-encoding and sampling strategies, to diminish error accumulation during inference.

---

5. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models

URL: View paper

Brief Assessment

Fast Autoregressive Video[64] focuses on distilling bidirectional diffusion models into autoregressive models for streaming generation, but does not describe a train-long-test-long strategy that exposes models to extended self-generated sequences during training as proposed in the original paper.

---

6. Streamingt2v: Consistent, dynamic, and extendable long video generation from text

URL: View paper

Brief Assessment

Streamingt2v[63] focuses on autoregressive video generation using short-term (CAM) and long-term (APM) memory modules for chunk transitions, not on train-long-test-long training procedures that expose models to extended self-generated sequences during training.

---

7. MAGI-1: Autoregressive Video Generation at Scale

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

---

8. Videoauteur: Towards long narrative video generation

URL: View paper

Brief Assessment

Videoauteur[67] focuses on long narrative video generation through interleaved auto-regressive modeling with actions, captions, and keyframes in the cooking domain. It does not address the train-long-test-long strategy for autoregressive video generation or the streaming long tuning approach described in the original paper.

---

9. Learning World Models for Interactive Video Generation

URL: View paper

Brief Assessment

Learning World Models[28] focuses on memory mechanisms and retrieval-augmented generation for interactive video world models in game environments, not on train-long-test-long strategies for autoregressive video generation consistency.

---

10. Progressive autoregressive video diffusion models

URL: View paper

Brief Assessment

Progressive Autoregressive Video[62] focuses on progressive noise scheduling for autoregressive video generation, not on train-long-test-long training procedures that expose models to extended self-generated sequences during training.

Contribution 3: Short window attention with frame-level attention sink

Description: The authors introduce a combination of local short-window attention and a frame-level attention sink (frame sink) that maintains persistent global anchor tokens. This design reduces computational cost while preserving long-range temporal consistency in video generation.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Sliding Window Attention Training for Efficient Large Language Models

URL: View paper

Brief Assessment

Sliding Window Attention[57] focuses on efficient long-context handling in transformer-based LLMs for text processing, not video generation. The candidate addresses attention sink in softmax operations for language models, while the original paper applies short window attention with frame-level attention sinks specifically for maintaining temporal consistency in video generation.

---

2. Efficient Vocal Source Separation Through Windowed Sink Attention

URL: View paper

Brief Assessment

Windowed Sink Attention[60] focuses on vocal source separation in audio processing, not video generation. The technical domain and application are fundamentally different from the original paper's video generation framework.

---

3. Streamingvlm: Real-time understanding for infinite video streams

URL: View paper

Brief Assessment

Streamingvlm[58] focuses on vision-language models for infinite video stream understanding with attention sinks for text tokens, not frame-level autoregressive video generation. The candidate's attention sink mechanism is designed for cross-modal streaming inference in VLMs, while the original paper addresses temporal consistency in video generation with frame-level AR models.

---

4. StreamingDialogue: Prolonged Dialogue Learning via Long Context Compression with Minimal Losses

URL: View paper

Brief Assessment

StreamingDialogue[59] focuses on dialogue compression using end-of-utterance tokens as attention sinks in text-based LLMs, not video generation with frame-level attention sinks for temporal consistency.

---

5. MotionStream: Real-Time Video Generation with Interactive Motion Controls

URL: View paper

Prior Art Analysis

MotionStream[4] demonstrates that combining sliding-window causal attention with attention sinks for long-range temporal consistency in video generation was already established prior to the original paper. Both papers use attention sinks to maintain persistent global anchor tokens while employing short/local attention windows to reduce computational cost. MotionStream[4] explicitly describes incorporating 'self-rollout with attention sinks and kv cache rolling during training' to enable 'constant-speed generation of arbitrarily long videos,' which directly parallels the original paper's claim of using short window attention with frame-level attention sink for efficient long video generation.

Evidence

Evidence 1 - **Rationale**: Both papers describe combining short/sliding-window attention with attention sinks as a key technical contribution for efficient video generation, demonstrating that this combination was already known in MotionStream[4]. - **Original**: we introduce short window attentioncombined with a frame-level attention sink(abbreviated as frame sink), which significantly accelerates inference while preserving performance. - **Candidate**: A key to our approach is introducing carefully designed sliding-window causal attention, combined with attention sinks.

Evidence 2 - **Rationale**: MotionStream[4] explicitly describes using attention sinks with a fixed context window for long video generation, which directly corresponds to the original paper's frame sink mechanism that maintains persistent global anchors while using short windows. - **Original**: we introduce short window attentioncombined with a frame-level attention sink(abbreviated as frame sink), which significantly accelerates inference while preserving performance. - **Candidate**: By incorporating self-rollout with attention sinks and kv cache rolling during training, we properly simulate inference-time extrapolations with a fixed context window, enabling constant-speed generation of arbitrarily long videos.

---

6. Reward Forcing: Efficient Streaming Video Generation with Rewarded Distribution Matching Distillation

URL: View paper

Brief Assessment

Reward Forcing[61] uses sliding window attention with attention sink tokens, but focuses on EMA-based sink token updates to prevent over-attention to initial frames, rather than the specific combination of short-window attention with frame-level attention sink for long-range temporal consistency as proposed in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. MemFlow: Flowing Adaptive Memory for Consistent and Efficient Long Video Narratives

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] LongLive: Real-time Interactive Long Video Generation View paper
- [1] Yan: Foundational interactive video generation View paper
- [2] Streamdit: Real-time streaming text-to-video generation View paper

- [3] Diffusion models are real-time game engines View paper
- [4] MotionStream: Real-Time Video Generation with Interactive Motion Controls View paper
- [5] A survey of interactive generative video View paper
- [6] AIGC-Driven Real-Time Interactive 4-D Traffic Scene Generation in Vehicular Networks View paper
- [7] VideoSSM: Autoregressive Long Video Generation with Hybrid State-Space Memory View paper
- [8] MIDAS: Multimodal Interactive Digital-humAn Synthesis via Real-time Autoregressive Video Generation View paper
- [9] Rolling forcing: Autoregressive long video diffusion in real time View paper
- [10] Hunyuan-GameCraft: High-dynamic Interactive Game Video Generation with Hybrid History Condition View paper
- [11] X-Streamer: Unified Human World Modeling with Audiovisual Interaction View paper
- [12] VARFVV: View-Adaptive Real-Time Interactive Free-View Video Streaming With Edge Computing View paper
- [13] LLIA - Enabling Low-Latency Interactive Avatars: Real-Time Audio-Driven Portrait Video Generation with Diffusion Models View paper
- [14] Inferix: A Block-Diffusion based Next-Generation Inference Engine for World Simulation View paper
- [15] Video2Game: Real-time, Interactive, Realistic and Browser-Compatible Environment from a Single Video View paper
- [16] MAGI-1: Autoregressive Video Generation at Scale View paper
- [17] Matrix-Game 2.0: An Open-Source, Real-Time, and Streaming Interactive World Model View paper
- [18] Predictive Sampling: Real-time Behaviour Synthesis with MuJoCo View paper
- [19] Live Avatar: Streaming Real-time Audio-Driven Avatar Generation with Infinite Length View paper
- [20] Autoregressive Adversarial Post-Training for Real-Time Interactive Video Generation View paper
- [21] Application of multimodal generation model in short video content personalized generation View paper
- [22] DartControl: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control View paper
- [23] Robovqa: Multimodal long-horizon reasoning for robotics View paper
- [24] Interactive Character Control with Auto-Regressive Motion Diffusion Models View paper
- [25] VideoLLM-online: Online Video Large Language Model for Streaming Video View paper
- [26] ChatAnyone: Stylized Real-time Portrait Video Generation with Hierarchical Motion Diffusion Model View paper
- [27] Proactive Assistant Dialogue Generation from Streaming Egocentric Videos View paper
- [28] Learning World Models for Interactive Video Generation View paper
- [29] MinD: Unified Visual Imagination and Control via Hierarchical World Models View paper
- [30] Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions View paper
- [31] Image Synthesis from a Collection of Depth Enhanced Panoramas: Creating Interactive Extended Reality Experiences from Static Images View paper
- [32] Dimensionality Reduction for the Real-Time Light-Field View Synthesis of Kernel-Based Models View paper
- [33] Island Design Camps - Interactive Video Projections as Extended Realities View paper
- [34] GelSLAM: A Real-time, High-Fidelity, and Robust 3D Tactile SLAM System View paper
- [35] StreamDiffusionV2: A Streaming System for Dynamic and Interactive Video Generation View paper
- [36] The Matrix: Infinite-Horizon World Generation with Real-Time Moving Control View paper
- [37] Ray tracing-based construction of 3D background model for real-time stereoscopic rendering of live immersive video View paper
- [38] From 2D to 3D video conferencing: modular RGB-D capture and reconstruction for interactive natural user representations in immersive extended reality (XR) communication View paper
- [39] DyNCA: Real-Time Dynamic Texture Synthesis Using Neural Cellular Automata View paper
- [40] Bench2Drive-R: Turning Real World Data into Reactive Closed-Loop Autonomous Driving Benchmark by Generative Model View paper
- [41] From Masks to Worlds: A Hitchhiker's Guide to World Models View paper
- [42] ARIG: Autoregressive Interactive Head Generation for Real-time Conversations View paper
- [43] Vinci: A real-time embodied smart assistant based on egocentric vision-language model View paper
- [44] StyleAvatar: Real-time Photo-realistic Portrait Avatar from a Single Video View paper
- [45] Social interactive human video synthesis View paper
- [46] Animatable Virtual Humans: Learning Pose-Dependent Human Representations in UV Space for Interactive Performance Synthesis View paper
- [47] LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity View paper
- [48] BakedAvatar: Baking Neural Fields for Real-Time Head Avatar Synthesis View paper
- [49] RELIC: Interactive Video World Model with Long-Horizon Memory View paper
- [50] Live4D: A Real-time Capture System for Streamable Volumetric Video View paper
- [51] FilmWeaver: Weaving Consistent Multi-Shot Videos with Cache-Guided Autoregressive Diffusion View paper
- [52] SneakPeek: Future-Guided Instructional Streaming Video Generation View paper
- [53] MemFlow: Flowing Adaptive Memory for Consistent and Efficient Long Video Narratives View paper
- [54] Contextual Knowledge Infusion via Iterative Semantic Tracing for Visionâ□□Language Understanding View paper
- [55] EgoLCD: Egocentric Video Generation with Long Context Diffusion View paper
- [56] Playing For You: Text Prompt-guided Joint Audio-visual Generation for Narrating Faces using Multi-entangled Latent Space View paper
- [57] Sliding Window Attention Training for Efficient Large Language Models View paper
- [58] Streamingvlm: Real-time understanding for infinite video streams View paper
- [59] StreamingDialogue: Prolonged Dialogue Learning via Long Context Compression with Minimal Losses View paper
- [60] Efficient Vocal Source Separation Through Windowed Sink Attention View paper
- [61] Reward Forcing: Efficient Streaming Video Generation with Rewarded Distribution Matching Distillation View paper
- [62] Progressive autoregressive video diffusion models View paper
- [63] Streamingt2v: Consistent, dynamic, and extendable long video generation from text View paper
- [64] From Slow Bidirectional to Fast Autoregressive Video Diffusion Models View paper
- [65] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation View paper
- [66] Loong: Generating Minute-level Long Videos with Autoregressive Language Models View paper
- [67] Videoauteur: Towards long narrative video generation View paper
- [68] DeepVerse: 4D Autoregressive Video Generation as a World Model View paper