Novelty Assessment Report

Paper: SteinsGate: Adding Causality to Diffusions for Long Video Generation via Path Integral

PDF URL: https://openreview.net/pdf?id=8WS5nDWIWE

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-08

Abstract

Video generation has advanced rapidly, but current models remain limited to short clips, far from the length and complexity of real-world narratives. Long video generation is thus both important and challenging. Existing approaches either attempt to extend the modeling length of video diffusion models directly or merge short clips via shared frames. However, due to the lack of temporal causality modeling for video data, they achieve only limited extensions, suffer from discontinuous or even contradictory actions, and fail to support flexible and fine-grained temporal control. Thus, we propose Instruct-Video-Continuation (InstructVC), combining Temporal Action Binding for fine-grained temporal control and Causal Video Continuation for natural long-term simulation. Temporal Action Binding decomposes complex long videos by temporal causality into scene descriptions and action sequences with predicted durations, while Causal Video Continuation autoregressively generates coherent video narratives from the text story. We further introduce SteinsGate, an inference-time instance of InstructVC that uses an MLLM for Temporal Action Binding and Video Path Integral to enforce causality between actions, converting a pre-trained TI2V diffusion model into an autoregressive video continuation model. Benchmark results demonstrate the advantages of SteinsGate and InstructVC in achieving accurate temporal control and generating natural, smooth multi-action long videos.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Multi-Action Long Video Generation with Temporal Causality**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Causal Video Generation and Continuation**
- **Multi-Event and Multi-Action Video Synthesis**
- **Long-Horizon Robotic Manipulation and Planning**
- **Temporal Reasoning and Video Understanding**
- **Long-Context and Memory-Augmented Video Models**
- **Representation Learning for Long Videos**
- **Domain-Specific Long Video Applications**
- **Survey and Methodological Foundations**

Complete Taxonomy Tree

- Multi-Action Long Video Generation with Temporal Causality Survey Taxonomy
- Causal Video Generation and Continuation
  - Causal Diffusion-Based Video Generation ★ (2 papers)
  - [0] SteinsGate: Adding Causality to Diffusions for Long Video Generation via Path Integral (Anon et al., 2026) View paper
  - [7] Video prediction by modeling videos as continuous multi-dimensional processes (Gaurav Shrivastava, 2024) View paper
  - Autoregressive Video Continuation (3 papers)
  - [16] Infinity-RoPE: Action-Controllable Infinite Video Generation Emerges From Autoregressive Self-Rollout (Hidir Yesiltepe, 2025) View paper
  - [45] Knot Forcing: Taming Autoregressive Video Diffusion Models for Real-time Infinite Interactive Portrait Animation (Steven Xiao, 2025) View paper
  - [48] Inference-based GAN Video Generation (Jingbo Yang, 2025) View paper
- Multi-Event and Multi-Action Video Synthesis
  - Sequential Action Video Generation (2 papers)
  - [4] Mind the Time: Temporally-Controlled Multi-Event Video Generation (Ziyi Wu, 2025) View paper
  - [6] Mavin: Multi-action video generation with diffusion models via transition video infilling (Zhang Bo-wen, 2024) View paper
  - Multi-Text Conditioned Long Video Generation (2 papers)
  - [1] Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising (Wang, 2023) View paper
  - [46] AlcheMinT: Fine-grained Temporal Control for Multi-Reference Consistent Video Generation (Sharath Girish, 2025) View paper
  - Narrative and Instructional Video Generation (2 papers)
  - [35] VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention (Zheng Mingzhe, 2025) View paper
  - [41] SneakPeek: Future-Guided Instructional Streaming Video Generation (Cheeun Hong, 2025) View paper
- Long-Horizon Robotic Manipulation and Planning
  - Vision-Language-Action Models for Long-Horizon Tasks (5 papers)

  - [18] HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models (Minghui Lin, 2025) View paper
  - [22] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers (Wang Yating, 2025) View paper
  - [23] LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks (Yang Yi, 2025) View paper
  - [36] Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation (Ding, 2025) View paper
  - [40] SeqVLA: Sequential Task Execution for Long-Horizon Manipulation with Completion-Aware Vision-Language-Action Model (Yang Ran, 2025) View paper
  - Video-Based World Models and Predictive Planning (5 papers)
  - [9] Vid2World: Crafting Video Diffusion Models to Interactive World Models (Wu Jialong, 2025) View paper
  - [10] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation (Yang Liu-di, 2025) View paper
  - [20] Vavim and vavam: Autonomous driving through video generative modeling (Bartoccioni, 2025) View paper
  - [49] BEYOND SINGLE-STEP: MULTI-FRAME ACTION-CONDITIONED VIDEO GENERATION FOR REINFORCE-MENT LEARNING ENVIRONMENTS (Z Li, 2025) View paper
  - [50] Video Language Planning (Du, 2023) View paper
  - Skill Chaining and Hierarchical Task Decomposition (2 papers)
  - [2] Long-horizon visual imitation learning via plan and code reflection (Chen Quan, 2025) View paper
  - [27] Generative skill chaining: Long-horizon skill planning with diffusion models (Mishra, 2023) View paper
  - Embodied Video Generation for Manipulation (2 papers)
  - [42] FantasyHSI: Video-Generation-Centric 4D Human Synthesis In Any Scene through A Graph-based Multi-Agent Framework (Wang Qiang, 2025) View paper
  - [44] MIND-V: Hierarchical Video Generation for Long-Horizon Robotic Manipulation with RL-based Physical Alignment (Ruicheng Zhang, 2025) View paper
- Temporal Reasoning and Video Understanding
  - Video Reasoning and Scene Graph Generation (3 papers)
  - [3] Video-of-thought: Step-by-step video reasoning from perception to cognition (Fei Hao, 2024) View paper
  - [17] End-to-end video scene graph generation with temporal propagation transformer (Yong Zhang, 2023) View paper
  - [19] STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training (Haiyi Qiu, 2025) View paper
  - Long-Term Action Anticipation (5 papers)
  - [24] Object-centric Video Representation for Long-term Action Anticipation (Ce Zhang, 2024) View paper
  - [26] Adamsformer for spatial action localization in the future (Hyung-Gun Chi, 2023) View paper
  - [30] Rethinking learning approaches for long-term action anticipation (Megha Nawhal, 2022) View paper
  - [31] Action anticipation using pairwise human-object interactions and transformers (Debaditya Roy, 2021) View paper
  - [33] Video + CLIP Baseline for Ego4D Long-term Action Anticipation (Das, 2022) View paper
  - Chain-of-Thought Video Reasoning (2 papers)
  - [15] Vchain: Chain-of-visual-thought for reasoning in video generation (Huang Ziqi, 2025) View paper
  - [29] Plan-X: Instruct Video Generation via Semantic Planning (Lun Huang, 2025) View paper
- Long-Context and Memory-Augmented Video Models
  - Sparse Attention and Context Routing (2 papers)
  - [8] Mixture of Contexts for Long Video Generation (Cai, 2025) View paper
  - [38] Long Context Tuning for Video Generation (Guo, 2025) View paper
  - Dual-Speed and Episodic Memory Learning (1 papers)
  - [5] Slowfast-vgen: Slow-fast learning for action-driven long video generation (Hong, 2024) View paper
- Representation Learning for Long Videos
  - Frame-Wise and Contrastive Representation Learning (1 papers)
  - [32] Frame-wise Action Representations for Long Videos via Sequence Contrastive Learning (Minghao Chen, 2022) View paper
  - Object-Centric and Human-Object Interaction Representations (2 papers)
  - [37] Learning a Generative Model for Multiâ□□Step Humanâ□□Object Interactions from Videos (He Wang, 2019) View paper
  - [47] Keystate-Driven Long-Term Generation of Bimanual Object Manipulation Sequences (Haziq Razali, 2025) View paper
- Domain-Specific Long Video Applications
  - Temporal Forecasting and Prediction (3 papers)
  - [25] Improving Tropical Cyclone Forecasting With Video Diffusion Models (Ren Zhibo, 2025) View paper
  - [28] Predictive autonomy for UAV remote sensing: A survey of video prediction (Zhan Chen, 2025) View paper
  - [34] : pseudo-image sequence evolution-based 3D pose prediction (X Liu, 2022) View paper
  - Long-Horizon Motion and Pose Synthesis (2 papers)
  - [21] Causal Motion Tokenizer for Streaming Motion Generation (B Jiang, 2025) View paper
  - [43] Glocalnet: Class-aware long-term human motion synthesis (Neeraj Battan, 2021) View paper
  - Specialized Robotic and Assistive Applications (3 papers)
  - [11] LAVA: Long-horizon Visual Action based Food Acquisition (Amisha Bhaskar, 2024) View paper
  - [12] Generative World-Model Planning for Long-Horizon User Preference Evolution and Responsible Personalization (Zare, 2025) View paper
  - [13] Robovqa: Multimodal long-horizon reasoning for robotics (Pierre Sermanet, 2024) View paper
- Survey and Methodological Foundations (2 papers)
  - [14] Conditional Video Generation Guided by Multimodal Inputs: A Comprehensive Survey (Niu Kai, 2024) View paper
  - [39] Towards Efficient Video Understanding and Generation: Free Training Signals to Faster Inference (Kahatapitiya, 2025) View paper

Narrative

Core task: Multi-action long video generation with temporal causality. The field addresses the challenge of synthesizing extended video sequences that unfold coherent, causally linked events over time. The taxonomy reveals several complementary branches: Causal Video Generation and Continuation focuses on diffusion-based and autoregressive methods that explicitly model temporal dependencies; Multi-Event and Multi-Action Video Synthesis explores compositional approaches for chaining diverse actions; Long-Horizon Robotic

Manipulation and Planning emphasizes embodied agents executing multi-step tasks; Temporal Reasoning and Video Understanding targets anticipation and comprehension of future states; Long-Context and Memory-Augmented Video Models develop architectures that scale to extended sequences; Representation Learning for Long Videos investigates efficient encodings; Domain-Specific Long Video Applications tackle specialized settings like weather forecasting or surveillance; and Survey and Methodological Foundations provide overarching perspectives. Works such as Gen-L-Video[1] and Video-of-Thought[3] illustrate early efforts to bridge reasoning and generation, while methods like Mind the Time[4] and Slowfast-vgen[5] highlight the importance of temporal structure.

A particularly active line of research centers on causal diffusion-based generation, where models learn to propagate temporal dependencies through latent dynamics or explicit causal masking. SteinsGate[0] sits squarely within this branch, emphasizing causal mechanisms for multi-action sequences, and shares conceptual ground with Continuous Multi-Dimensional[7], which also explores continuous temporal modeling. In contrast, works like VideoGen-of-Thought[35] and Plan Code Reflection[2] adopt more symbolic or planning-driven strategies, trading off end-to-end learning for interpretability and compositional control. Meanwhile, robotic manipulation studies such as Long-VLA[36] and LoHoVLA[23] prioritize action-conditioned prediction in embodied settings, raising questions about how causal video models can transfer to interactive environments. The interplay between diffusion-based synthesis, memory-augmented architectures, and domain-specific constraints remains an open frontier, with SteinsGate[0] contributing a causal lens that complements the broader landscape of long-horizon video generation.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Video prediction by modeling videos as continuous multi-dimensional processes

Authors: Gaurav Shrivastava, Abhinav Shrivastava | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ This multi-step diffusion process has been instrumental in â¦ approach videos as a series of images, generating separate â¦ scenario is posited as the causative factor for the observed â¦

Relationship Analysis

Both papers belong to the Causal Diffusion-Based Video Generation category, employing diffusion models to generate temporally coherent videos with causal mechanisms. While SteinsGate focuses on multi-action long video generation through explicit temporal action binding and autoregressive continuation using Video Path Integral to enforce causality between action segments, the candidate paper (CVP) treats videos as continuous multi-dimensional processes with interpolation between consecutive frames, primarily addressing video prediction tasks rather than multi-action narrative generation. The key distinction is that SteinsGate targets action-rich long video synthesis with explicit action-duration control, whereas CVP focuses on frame-level prediction through continuous interpolation without explicit action decomposition.

Contributions Analysis

Overall novelty summary. The paper proposes InstructVC, a framework combining Temporal Action Binding and Causal Video Continuation to generate long, multi-action videos with explicit temporal causality. It resides in the 'Causal Diffusion-Based Video Generation' leaf, which contains only two papers including this one. This leaf sits within the broader 'Causal Video Generation and Continuation' branch, indicating a relatively sparse but emerging research direction focused on diffusion models that explicitly model temporal dependencies rather than treating video generation as a purely spatial-temporal extension problem.

The taxonomy reveals that neighboring leaves address related but distinct challenges: 'Autoregressive Video Continuation' (three papers) explores chunk-based streaming approaches, while 'Sequential Action Video Generation' and 'Multi-Text Conditioned Long Video Generation' focus on compositional control without necessarily enforcing causal structure. The 'Long-Horizon Robotic Manipulation' branch (twelve papers across four leaves) emphasizes embodied action prediction, suggesting that causal video modeling intersects with but remains distinct from interactive planning domains. The paper's emphasis on temporal causality and action binding differentiates it from purely compositional or memory-augmented approaches found in other branches.

Among thirty candidates examined, none clearly refute the three core contributions: the InstructVC framework (ten candidates, zero refutable), the SteinsGate inference method (ten candidates, zero refutable), and the Video Path Integral technique (ten candidates, zero refutable). The single sibling paper in the same leaf addresses continuous temporal modeling but does not appear to overlap with the specific combination of action binding, causal continuation, and path integral guidance. This limited search scope suggests that within the examined literature, the integration of MLLM-driven temporal decomposition with causal diffusion appears relatively unexplored, though the analysis does not claim exhaustive coverage.

Given the sparse population of the causal diffusion leaf and the absence of refuting candidates among thirty examined papers, the work appears to occupy a distinct position within long video generation. However, the limited search scale and the broader taxonomy structure—showing active research in autoregressive continuation, robotic world models, and compositional synthesis—indicate that the novelty assessment is provisional. A more comprehensive search across the fifty-paper taxonomy and beyond would be needed to fully contextualize the contributions against the wider landscape of temporal reasoning and multi-action video synthesis.

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

Contribution 1: Instruct-Video-Continuation (InstructVC) framework

Description: A two-stage framework for multi-action long video generation that decomposes complex videos into scene descriptions and action sequences with predicted durations (Temporal Action Binding), then autoregressively generates coherent video narratives from the text story (Causal Video Continuation).

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

URL: View paper

Brief Assessment

FilmWeaver[73] focuses on multi-shot video generation with inter-shot consistency through cache mechanisms, while InstructVC addresses multi-action long video generation with temporal causality modeling through action binding and autoregressive continuation. The technical approaches and problem formulations differ fundamentally.

2. Videotetris: Towards compositional text-to-video generation

URL: View paper

Brief Assessment

VideoTetris[78] focuses on compositional video generation with multiple objects and spatial-temporal composition, not on temporal action binding with predicted durations or autoregressive continuation based on causal action sequences.

3. Long-context autoregressive video modeling with next-frame prediction

URL: View paper

Brief Assessment

Long-context Autoregressive[74] focuses on frame-level autoregressive modeling with next-frame prediction for long videos, using asymmetric patchify kernels to reduce token redundancy. It does not address multi-action temporal binding or decomposition of complex videos into scene descriptions and action sequences with predicted durations, which are core to InstructVC's two-stage framework.

4. Ar-diffusion: Asynchronous video generation with auto-regressive diffusion

URL: View paper

Brief Assessment

AR-Diffusion[71] focuses on asynchronous auto-regressive diffusion for video generation with frame-level timestep control, not on temporal action binding with predicted durations or multi-action narrative decomposition as in InstructVC.

5. MAGI-1: Autoregressive Video Generation at Scale

URL: View paper

Brief Assessment

MAGI-1[77] focuses on autoregressive chunk-based video generation with per-chunk noise scheduling, not on temporal action binding with predicted durations or the two-stage decomposition (scene descriptions + action sequences) that characterizes InstructVC.

6. Progressive autoregressive video diffusion models

URL: View paper

Brief Assessment

Progressive Autoregressive[69] focuses on progressive noise scheduling for autoregressive video generation without temporal action binding or explicit action-duration decomposition. The candidate does not address multi-action temporal planning with predicted durations.

7. Autoregressive Video Generation without Vector Quantization

URL: View paper

Brief Assessment

Autoregressive Without VQ[76] focuses on non-quantized autoregressive video generation with temporal frame-by-frame and spatial set-by-set prediction, without vector quantization. It does not address temporal action binding with predicted durations or multi-action decomposition as in InstructVC.

8. Diffusion forcing: Next-token prediction meets full-sequence diffusion

URL: View paper

Brief Assessment

Diffusion Forcing[66] focuses on a general training paradigm for sequence modeling with independent per-token noise levels, not specifically on multi-action long video generation with temporal action binding and autoregressive continuation as described in the original paper.

9. Learning Real-World Action-Video Dynamics with Heterogeneous Masked Autoregression

URL: View paper

Brief Assessment

Heterogeneous Masked Autoregression[75] focuses on learning action-video dynamics across heterogeneous robotic embodiments using masked autoregression for real-time simulation and policy evaluation. It does not address the specific two-stage framework of temporal action binding with duration prediction followed by causal video continuation for multi-action long video generation that characterizes InstructVC.

10. Seine: Short-to-long video diffusion model for generative transition and prediction

URL: View paper

Brief Assessment

Seine[72] focuses on generative transition between different scenes and autoregressive prediction for long videos, but does not propose temporal action binding with predicted durations or the two-stage framework (Temporal Action Binding + Causal Video Continuation) that characterizes InstructVC. Seine's approach uses random-mask video diffusion for scene transitions rather than decomposing videos into action sequences with explicit duration control.

Contribution 2: SteinsGate inference-time method

Description: A plug-and-play inference-time implementation that combines a Multi-modal Large Language Model for temporal action binding with a novel Video Path Integral technique to convert pre-trained text-and-image-to-video diffusion models into autoregressive video continuation models without additional training.

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. Test-Time Temporal Sampling for Efficient MLLM Video Understanding

URL: View paper

Brief Assessment

Test-Time Temporal Sampling[59] focuses on efficient video token sampling for MLLMs through multi-subsequence inference, not on converting diffusion models into autoregressive video continuation models using path integral techniques.

2. Emu3: Next-token prediction is all you need

URL: View paper

Brief Assessment

Emu3[51] focuses on next-token prediction for multimodal generation (images, videos, text) using a unified transformer architecture, not on inference-time autoregressive video continuation from pre-trained diffusion models. The technical approaches are fundamentally different: Emu3 trains from scratch on discrete tokens, while SteinsGate converts existing text-and-image-to-video diffusion models without additional training.

3. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception

URL: View paper

Brief Assessment

VideoChat-r1.5[53] focuses on iterative visual perception for multimodal reasoning in MLLMs, not on converting text-and-image-to-video diffusion models into autoregressive video continuation models. The candidate addresses test-time scaling through progressive attention refinement, while the original contribution specifically targets video generation continuation without additional training.

4. Generative AI for Text-to-Video Generation: Recent Advances and Future Directions

URL: View paper

Brief Assessment

Text-to-Video Survey[58] is a survey paper that reviews existing methods in the text-to-video generation field. The provided context fragments mention inference-time approaches and multimodal planning but do not contain sufficient technical detail about specific methods that would refute the novelty of SteinsGate's combination of MLLM-based temporal action binding with Video Path Integral for autoregressive video continuation.

5. Acdc: Autoregressive coherent multimodal generation using diffusion correction

URL: View paper

Brief Assessment

Acdc[55] focuses on combining autoregressive models with diffusion models for multimodal generation (story and video), using SDEdit for local correction. SteinsGate addresses video continuation with temporal causality modeling via Video Path Integral, a fundamentally different technical approach for different goals.

6. CoS: Chain-of-Shot Prompting for Long Video Understanding

URL: View paper

Brief Assessment

CoS[60] focuses on shot selection for long video understanding using binary coding and test-time visual prompt optimization, not on autoregressive video continuation or converting diffusion models for video generation tasks.

7. Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation

URL: View paper

Brief Assessment

Midas[57] focuses on real-time autoregressive video generation for interactive digital humans using multimodal control (audio, pose, text), while SteinsGate addresses multi-action long video generation with temporal causality modeling. The technical approaches differ fundamentally: Midas[57] uses an autoregressive LLM backbone with diffusion rendering for streaming generation, whereas SteinsGate introduces Video Path Integral as a temporal guidance technique to convert pre-trained text-and-image-to-video models into autoregressive continuation models without training.

8. Mirasol3B: A Multimodal Autoregressive Model for Time-Aligned and Contextual Modalities

URL: View paper

Brief Assessment

Mirasol3B[56] focuses on multimodal learning combining video, audio, and text through separate autoregressive components for time-synchronized and context modalities. It does not address inference-time video continuation using diffusion models or temporal action binding for video generation.

9. Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization

URL: View paper

Brief Assessment

Training-free Guidance[52] focuses on spatial layout control and object trajectory planning using multimodal models and noise inversion for text-to-video generation, not on autoregressive video continuation with temporal causality modeling as in the original paper.

10. VideoPoet: A Large Language Model for Zero-Shot Video Generation

URL: View paper

Brief Assessment

VideoPoet[54] is a large language model trained for video generation using discrete tokens, not an inference-time method that converts pre-trained diffusion models. It requires extensive pretraining on multimodal data rather than being a plug-and-play solution.

Contribution 3: Video Path Integral temporal guidance technique

Description: A temporal guidance method that integrates multiple image-to-video paths from historical frames during sampling to explicitly propagate spatio-temporal information from history into future video generation, thereby enforcing temporal causality in pre-trained diffusion models.

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. Diffueraser: A diffusion model for video inpainting

URL: View paper

Brief Assessment

Diffueraser[64] focuses on video inpainting using diffusion models with prior incorporation and temporal smoothing for frame consistency, not on temporal guidance for general video generation via historical frame propagation as in the original paper.

2. Video diffusion models are strong video inpainter

URL: View paper

Brief Assessment

Video Inpainter[62] focuses on video inpainting using first-frame filling with noise latent propagation, not on general temporal guidance for video continuation or multi-action generation. The candidate's approach propagates noise latents only to the first frame for inpainting tasks, whereas the original paper's Video Path Integral integrates multiple i2v paths across all historical frames for autoregressive video continuation.

3. Sparsectrl: Adding sparse controls to text-to-video diffusion models

URL: View paper

Brief Assessment

Sparsectrl[65] focuses on adding sparse spatial controls (sketch/depth/RGB) to video generation via an add-on encoder, not on temporal guidance for propagating historical frame information through path integration. The candidate addresses spatial controllability rather than temporal causality enforcement through historical trajectory integration.

4. Progressive autoregressive video diffusion models

URL: View paper

Brief Assessment

Progressive Autoregressive[69] uses progressive noise levels across frames for autoregressive generation, not path integral guidance that integrates multiple i2v paths from historical frames. The mechanisms for temporal causality enforcement differ fundamentally.

5. Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution

URL: View paper

Brief Assessment

Upscale-A-Video[70] focuses on video super-resolution using flow-guided recurrent latent propagation for temporal consistency, not on temporal guidance for video generation via historical frame path integration as in the original paper.

6. Lavie: High-quality video generation with cascaded latent diffusion models

URL: View paper

Brief Assessment

Lavie[61] focuses on cascaded latent diffusion models with temporal self-attention and joint image-video training for text-to-video generation. It does not propose a path integral-based temporal guidance method that integrates multiple image-to-video paths from historical frames during sampling to enforce temporal causality.

7. Frame context packing and drift prevention in next-frame-prediction video diffusion models

URL: View paper

Brief Assessment

Frame Context Packing[68] focuses on frame compression and drift prevention through packing schedules and history discretization, not on temporal guidance via historical frame propagation. The candidate does not demonstrate prior work on integrating multiple image-to-video paths for temporal causality enforcement.

8. Video Diffusion Models

URL: View paper

Brief Assessment

Video Diffusion Models[63] focuses on extending standard image diffusion architectures for video generation and introduces conditional sampling for spatial/temporal extension, but does not describe a path integral-based temporal guidance method that integrates multiple image-to-video paths from historical frames during sampling.

9. Diffusion forcing: Next-token prediction meets full-sequence diffusion

URL: View paper

Brief Assessment

Diffusion Forcing[66] does not describe a video path integral technique that integrates multiple image-to-video paths from historical frames. Instead, it presents diffusion forcing as a method where tokens are denoised with independent noise levels in a causal architecture.

10. Synchronized Multiâ□□Frame Diffusion for Temporally Consistent Video Stylization

URL: View paper

Brief Assessment

Synchronized Multi-Frame[67] focuses on video stylization using synchronized multi-frame diffusion with optical flow-based information sharing during denoising. The original paper's Video Path Integral integrates multiple image-to-video paths from historical frames for temporal causality in video generation, which is a different technical approach and application domain.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] SteinsGate: Adding Causality to Diffusions for Long Video Generation via Path Integral View paper
- [1] Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising View paper
- [2] Long-horizon visual imitation learning via plan and code reflection View paper
- [3] Video-of-thought: Step-by-step video reasoning from perception to cognition View paper
- [4] Mind the Time: Temporally-Controlled Multi-Event Video Generation View paper
- [5] Slowfast-vgen: Slow-fast learning for action-driven long video generation View paper
- [6] Mavin: Multi-action video generation with diffusion models via transition video infilling View paper
- [7] Video prediction by modeling videos as continuous multi-dimensional processes View paper
- [8] Mixture of Contexts for Long Video Generation View paper
- [9] Vid2World: Crafting Video Diffusion Models to Interactive World Models View paper
- [10] RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation View paper
- [11] LAVA: Long-horizon Visual Action based Food Acquisition View paper
- [12] Generative World-Model Planning for Long-Horizon User Preference Evolution and Responsible Personalization View paper
- [13] Robovqa: Multimodal long-horizon reasoning for robotics View paper
- [14] Conditional Video Generation Guided by Multimodal Inputs: A Comprehensive Survey View paper
- [15] Vchain: Chain-of-visual-thought for reasoning in video generation View paper
- [16] Infinity-RoPE: Action-Controllable Infinite Video Generation Emerges From Autoregressive Self-Rollout View paper

- [17] End-to-end video scene graph generation with temporal propagation transformer View paper
- [18] HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models View paper
- [19] STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training View paper
- [20] Vavim and vavam: Autonomous driving through video generative modeling View paper
- [21] Causal Motion Tokenizer for Streaming Motion Generation View paper
- [22] VQ-VLA: Improving Vision-Language-Action Models via Scaling Vector-Quantized Action Tokenizers View paper
- [23] LoHoVLA: A Unified Vision-Language-Action Model for Long-Horizon Embodied Tasks View paper
- [24] Object-centric Video Representation for Long-term Action Anticipation View paper
- [25] Improving Tropical Cyclone Forecasting With Video Diffusion Models View paper
- [26] Adamsformer for spatial action localization in the future View paper
- [27] Generative skill chaining: Long-horizon skill planning with diffusion models View paper
- [28] Predictive autonomy for UAV remote sensing: A survey of video prediction View paper
- [29] Plan-X: Instruct Video Generation via Semantic Planning View paper
- [30] Rethinking learning approaches for long-term action anticipation View paper
- [31] Action anticipation using pairwise human-object interactions and transformers View paper
- [32] Frame-wise Action Representations for Long Videos via Sequence Contrastive Learning View paper
- [33] Video + CLIP Baseline for Ego4D Long-term Action Anticipation View paper
- [34] : pseudo-image sequence evolution-based 3D pose prediction View paper
- [35] VideoGen-of-Thought: Step-by-step generating multi-shot video with minimal manual intervention View paper
- [36] Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation View paper
- [37] Learning a Generative Model for Multiâ□□Step Humanâ□□Object Interactions from Videos View paper
- [38] Long Context Tuning for Video Generation View paper
- [39] Towards Efficient Video Understanding and Generation: Free Training Signals to Faster Inference View paper
- [40] SeqVLA: Sequential Task Execution for Long-Horizon Manipulation with Completion-Aware Vision-Language-Action Model View paper
- [41] SneakPeek: Future-Guided Instructional Streaming Video Generation View paper
- [42] FantasyHSI: Video-Generation-Centric 4D Human Synthesis In Any Scene through A Graph-based Multi-Agent Framework View paper
- [43] Glocalnet: Class-aware long-term human motion synthesis View paper
- [44] MIND-V: Hierarchical Video Generation for Long-Horizon Robotic Manipulation with RL-based Physical Alignment View paper
- [45] Knot Forcing: Taming Autoregressive Video Diffusion Models for Real-time Infinite Interactive Portrait Animation View paper
- [46] AlcheMinT: Fine-grained Temporal Control for Multi-Reference Consistent Video Generation View paper
- [47] Keystate-Driven Long-Term Generation of Bimanual Object Manipulation Sequences View paper
- [48] Inference-based GAN Video Generation View paper
- [49] BEYOND SINGLE-STEP: MULTI-FRAME ACTION-CONDITIONED VIDEO GENERATION FOR REINFORCE-MENT LEARNING ENVIRONMENTS View paper
- [50] Video Language Planning View paper
- [51] Emu3: Next-token prediction is all you need View paper
- [52] Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization View paper
- [53] Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception View paper
- [54] VideoPoet: A Large Language Model for Zero-Shot Video Generation View paper
- [55] Acdc: Autoregressive coherent multimodal generation using diffusion correction View paper
- [56] Mirasol3B: A Multimodal Autoregressive Model for Time-Aligned and Contextual Modalities View paper
- [57] Midas: Multimodal interactive digital-human synthesis via real-time autoregressive video generation View paper
- [58] Generative AI for Text-to-Video Generation: Recent Advances and Future Directions View paper
- [59] Test-Time Temporal Sampling for Efficient MLLM Video Understanding View paper
- [60] CoS: Chain-of-Shot Prompting for Long Video Understanding View paper
- [61] Lavie: High-quality video generation with cascaded latent diffusion models View paper
- [62] Video diffusion models are strong video inpainter View paper
- [63] Video Diffusion Models View paper
- [64] Diffueraser: A diffusion model for video inpainting View paper
- [65] Sparsectrl: Adding sparse controls to text-to-video diffusion models View paper
- [66] Diffusion forcing: Next-token prediction meets full-sequence diffusion View paper
- [67] Synchronized Multiâ□□Frame Diffusion for Temporally Consistent Video Stylization View paper
- [68] Frame context packing and drift prevention in next-frame-prediction video diffusion models View paper
- [69] Progressive autoregressive video diffusion models View paper
- [70] Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution View paper
- [71] Ar-diffusion: Asynchronous video generation with auto-regressive diffusion View paper
- [72] Seine: Short-to-long video diffusion model for generative transition and prediction View paper
- [73] FilmWeaver: Weaving Consistent Multi-Shot Videos with Cache-Guided Autoregressive Diffusion View paper
- [74] Long-context autoregressive video modeling with next-frame prediction View paper
- [75] Learning Real-World Action-Video Dynamics with Heterogeneous Masked Autoregression View paper
- [76] Autoregressive Video Generation without Vector Quantization View paper
- [77] MAGI-1: Autoregressive Video Generation at Scale View paper
- [78] Videotetris: Towards compositional text-to-video generation View paper