Novelty Assessment Report

Paper: Mixture of Contexts for Long Video Generation
PDF URL: https://openreview.net/pdf?id=y6XJZlEC2x
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-29

Abstract

Long video generation is fundamentally a long context memory problem: models must retain and retrieve salient events across a long range without collapsing or drifting. However, scaling diffusion transformers to generate long-context videos is fundamentally limited by the quadratic cost of self-attention, which makes memory and computation intractable and difficult to optimize for long sequences. We recast long-context video generation as an internal information retrieval task and propose a simple, learnable sparse attention routing module, Mixture of Contexts (MoC), as an effective long-term memory retrieval engine. In MoC, each query dynamically selects a few informative chunks plus mandatory anchors (caption, local windows) to attend to, with causal routing that prevents loop closures. As we scale the data and gradually sparsify the routing, the model allocates compute to salient history, preserving identities, actions, and scenes over minutes of content. Efficiency follows as a byproduct of retrieval (near-linear scaling), which enables practical training and synthesis, and the emergence of memory and consistency at the scale of minutes.

Core Task Landscape

This paper addresses: **long-context video generation with sparse attention routing**

A total of **45 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

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

- **Sparse Attention Mechanisms for Video Diffusion Transformers**
- **Attention Optimization and Distillation for Video Generation**
- **Long-Context and Temporal Coherence in Video Generation**
- **Controllable and Conditional Video Generation**
- **Multi-Shot and Narrative Video Generation**
- **Specialized Architectures and Attention Variants**
- **General Sparse Attention and Theoretical Foundations**
- **Few-Shot and Training-Free Video Synthesis**
- **Text-to-Video Generation with Sparse Mechanisms**
- **Human-Object Interaction and Egocentric Video**

Complete Taxonomy Tree

- long-context video generation with sparse attention routing Survey Taxonomy
- Sparse Attention Mechanisms for Video Diffusion Transformers
  - Dynamic and Adaptive Sparse Attention ★ (4 papers)
  - [0] Mixture of Contexts for Long Video Generation (Anon et al., 2026) View paper
  - [11] Training-free and adaptive sparse attention for efficient long video generation (Xia Yifei, 2025) View paper
  - [22] VORTA: Efficient Video Diffusion via Routing Sparse Attention (Sun Wen-hao, 2025) View paper
  - [32] MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation (Jia Weinan, 2025) View paper
  - Static Pattern Sparse Attention (5 papers)
  - [2] Radial Attention: Sparse Attention with Energy Decay for Long Video Generation (X Li, 2025) View paper
  - [23] Sparse-vDiT: Unleashing the Power of Sparse Attention to Accelerate Video Diffusion Transformers (Chen Pengtao, 2025) View paper
  - [29] Compact Attention: Exploiting Structured Spatio-Temporal Sparsity for Fast Video Generation (Li, 2025) View paper
  - [44] Radial Attention: Sparse Attention for Long Video Generation (X Li, n.d.) View paper
  - Training-Free Sparse Attention Acceleration (4 papers)
  - [12] Grouping First, Attending Smartly: Training-Free Acceleration for Diffusion Transformers (Ren, 2025) View paper
  - [19] DraftAttention: Fast Video Diffusion via Low-Resolution Attention Guidance (Shen Xuan, 2025) View paper
  - [21] SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference (Zhang Jintao, 2025) View paper
  - Hybrid and Multi-Scale Sparse Attention (3 papers)
  - [1] PAROAttention: Pattern-Aware ReOrdering for Efficient Sparse and Quantized Attention in Visual Generation Models (Zhao, 2025) View paper
  - [3] Bidirectional sparse attention for faster video diffusion training (Zhan, 2025) View paper
  - [30] PSA: Pyramid Sparse Attention for Efficient Video Understanding and Generation (Xiaolong Li, 2025) View paper
- Attention Optimization and Distillation for Video Generation
  - Step Distillation with Sparse Attention (2 papers)
  - [13] BLADE: Block-Sparse Attention Meets Step Distillation for Efficient Video Generation (Li, 2025) View paper

- [37] USV: Unified Sparsification for Accelerating Video Diffusion Models (Xinjian Wu, 2025) View paper
- Attention Routing and Guidance (2 papers)
- [25] XAttention: Block Sparse Attention with Antidiagonal Scoring (Xu Ruyi, 2025) View paper
- [31] Input-Aware Sparse Attention for Real-Time Co-Speech Video Generation (Lu, 2025) View paper

- Long-Context and Temporal Coherence in Video Generation
  - Long-Term Memory and Context Management (4 papers)
  - [16] EgoLCD: Egocentric Video Generation with Long Context Diffusion (Liuzhou Zhang, 2025) View paper
  - [24] Long Video Diffusion Generation with Segmented Cross-Attention and Content-Rich Video Data Curation (Xin Yan, 2025) View paper
  - [27] Lag-Relative Sparse Attention In Long Context Training (Huang Wanyi, 2025) View paper
  - [28] LongCat-Video Technical Report (Meituan LongCat Team, 2025) View paper
  - Autoregressive Generation and Drift Mitigation (2 papers)
  - [14] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation (Chen Ya-bo, 2025) View paper
  - [35] End-to-End Training for Autoregressive Video Diffusion via Self-Resampling (Yuwei Guo, 2025) View paper
  - Planning and Hierarchical Generation (1 papers)
  - [38] AutoScape: Geometry-Consistent Long-Horizon Scene Generation (Chen Jiacheng, 2025) View paper
- Controllable and Conditional Video Generation
  - Camera Pose and Trajectory Control (2 papers)
  - [17] CPA: Camera-pose-awareness Diffusion Transformer for Video Generation (Wang Yue-lei, 2024) View paper
  - [45] ZeroTrail: Zero-Shot Trajectory Control Framework for Video Diffusion Models (Y Lu, n.d.) View paper
  - Motion Transfer and Cloning (2 papers)
  - [9] MotionClone: Training-Free Motion Cloning for Controllable Video Generation (Ling, 2024) View paper
  - [36] VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models (Hyeonho Jeong, 2023) View paper
  - Conditional Synthesis with Spatial Guidance (2 papers)
  - [6] Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control (Mariam Hassan, 2025) View paper
  - [18] ConditionVideo: Training-Free Condition-Guided Text-to-Video Generation (Peng Bo, 2023) View paper
- Multi-Shot and Narrative Video Generation (1 papers)
  - [15] HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives (Ouyang Hao, 2025) View paper
- Specialized Architectures and Attention Variants
  - Fused and Hybrid Attention Architectures (2 papers)
  - [7] 4Real-Video-V2: Fused View-Time Attention and Feedforward Reconstruction for 4D Scene Generation (Wang Chaoyang, 2025) View paper
  - [8] Hunyuanvideo 1.5 technical report (Bing Wu, 2025) View paper
  - Multimodal and Vision-Language Sparse Attention (2 papers)
  - [5] InfiniteVL: Synergizing Linear and Sparse Attention for Highly-Efficient, Unlimited-Input Vision-Language Models (Hongyuan Tao, 2025) View paper
  - [34] VideoNSA: Native Sparse Attention Scales Video Understanding (Chai, 2025) View paper
  - Domain-Specific Sparse Attention Applications (3 papers)
  - [20] MaGGIe: Masked Guided Gradual Human Instance Matting (Chuong Huynh, 2024) View paper
  - [39] Temporally Consistent Object Editing in Videos using Extended Attention (Zamani, 2024) View paper
  - [40] Flow-Guided Sparse Transformer for Video Deblurring (Lin Jing, 2022) View paper
- General Sparse Attention and Theoretical Foundations (1 papers)
  - [10] Fractal reservoir structuring for large language model generative pathways: An empirical investigation with large language model (O Strickland, 2025) View paper
- Few-Shot and Training-Free Video Synthesis (2 papers)
  - [41] Motion Selective Prediction for Video Frame Synthesis (Prinet Veronique, 2022) View paper
  - [43] Adaptive Compact Attention For Few-shot Video-to-video Translation (Huang RiSheng, 2020) View paper
- Text-to-Video Generation with Sparse Mechanisms (1 papers)
  - [42] GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions (Wu, 2021) View paper
- Human-Object Interaction and Egocentric Video (1 papers)
  - [4] Controllable human-object interaction synthesis (Li, 2024) View paper

Narrative

Core task: long-context video generation with sparse attention routing. The field addresses the computational challenge of generating extended video sequences by reducing the quadratic cost of full attention in diffusion transformers. The taxonomy reveals several major branches: some focus on designing specialized sparse attention patterns for video diffusion transformers (including dynamic and adaptive routing strategies), while others emphasize attention optimization through distillation or architectural variants. Additional branches tackle long-context temporal coherence, controllable generation conditioned on various inputs, multi-shot narrative synthesis, and training-free or few-shot methods. Theoretical foundations and general sparse attention mechanisms form another strand, alongside specialized topics such as human-object interaction and egocentric video modeling. Representative works like Hunyuanvideo[8] and InfiniteVL[5] illustrate how large-scale systems integrate these sparse routing ideas, whereas methods such as PAROAttention[1] and Radial Attention[2] propose specific geometric or hierarchical sparsity patterns.

Within the dynamic and adaptive sparse attention cluster, several lines of work explore how to route attention based on content or learned policies rather than fixed patterns. Mixture of Contexts[0] exemplifies this adaptive approach by dynamically selecting relevant context subsets during generation, contrasting with methods like Bidirectional Sparse Attention[3] that impose structured bidirectional connectivity or VORTA[22] that leverages token-level routing. A key trade-off across these branches is between the flexibility of learned routing (which can better capture complex temporal dependencies) and the simplicity of predefined sparse masks (which offer predictable memory savings). Open questions include how to balance sparsity with quality for very long sequences and whether adaptive routing can generalize across diverse video domains. Mixture of Contexts[0] sits naturally among these adaptive strategies, sharing the goal of content-aware sparsity with neighbors like MoGA[32], yet differing in how context selection is orchestrated across layers and frames.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Training-free and adaptive sparse attention for efficient long video generation

Authors: Xia Yifei, Yifei Xia, Fu, Fangcheng, Suhan Ling, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Generating high-fidelity long videos with Diffusion Transformers (DiTs) is often hindered by significant latency, primarily due to the computational demands of attention mechanisms. For instance, generating an 8-second 720p video (110K tokens) with HunyuanVideo takes about 600 PFLOPs, with around 500 PFLOPs consumed by attention computations. To address this issue, we propose AdaSpa, the first Dynamic Pattern and Online Precise Search sparse attention method. Firstly, to realize the Dynamic Patt...

Relationship Analysis

Both papers belong to the Dynamic and Adaptive Sparse Attention category, focusing on learning sparse attention patterns for long-context video generation with diffusion transformers. They overlap in addressing the quadratic cost of self-attention through learned routing mechanisms that adapt to input content. The key difference is that the original paper (MoC) learns to route queries to content-aligned chunks (frames, shots, captions) with mandatory anchors and causal constraints, while the candidate paper (AdaSpa) proposes a training-free blockified sparse attention with LSE-cached online search that exploits invariance across denoising steps, requiring no fine-tuning or dataset profiling.

---

2. VORTA: Efficient Video Diffusion via Routing Sparse Attention

Authors: Sun Wen-hao, Tu, Rong-Cheng, Wenhao Sun, Ding Yifu, et al. (17 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Video diffusion transformers have achieved remarkable progress in high-quality video generation, but remain computationally expensive due to the quadratic complexity of attention over high-dimensional video sequences. Recent acceleration methods enhance the efficiency by exploiting the local sparsity of attention scores; yet they often struggle with accelerating the long-range computation. To address this problem, we propose VORTA, an acceleration framework with two novel components: 1) a sparse...

Relationship Analysis

Both papers belong to the Dynamic and Adaptive Sparse Attention category, focusing on learned routing mechanisms that adapt attention patterns based on input content for long-context video generation. They overlap in their core approach of dynamically selecting relevant context chunks through trainable routing modules to reduce quadratic attention costs while maintaining long-range dependencies. The key difference is that the original paper (MoC) uses content-aligned chunking with causal routing and mandatory anchors (text tokens, local windows) for minute-scale video generation, while the candidate paper (VORTA) employs a signal-aware router that switches between multiple sparse attention variants (sliding window, core-set attention) based on diffusion timestep embeddings and SNR levels, targeting general video diffusion acceleration.

---

3. MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation

Authors: Jia Weinan, Lu, Yuning, Weinan Jia, Huang, et al. (23 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Long video generation with Diffusion Transformers (DiTs) is bottlenecked by the quadratic scaling of full attention with sequence length. Since attention is highly redundant, outputs are dominated by a small subset of query-key pairs. Existing sparse methods rely on blockwise coarse estimation, whose accuracy-efficiency trade-offs are constrained by block size. This paper introduces Mixture-of-Groups Attention (MoGA), an efficient sparse attention that uses a lightweight, learnable token router ...

Relationship Analysis

Both papers belong to the Dynamic and Adaptive Sparse Attention category, learning sparse attention patterns for long-context video generation with diffusion transformers. They overlap in using learnable routing mechanisms to dynamically select relevant context chunks, with MoC employing top-k selection on mean-pooled chunk descriptors and MoGA using a lightweight linear router to assign tokens to groups. The key difference is that MoC focuses on content-aligned chunking (frames, shots, captions) with mandatory anchors and causal masking to prevent loop closures, while MoGA emphasizes mixture-of-groups architecture with group balancing loss and seamless integration with FlashAttention and sequence parallelism.

Contributions Analysis

Overall novelty summary. The paper proposes Mixture of Contexts (MoC), a learnable sparse attention routing module that dynamically selects informative chunks and mandatory anchors for long-context video generation. It resides in the 'Dynamic and Adaptive Sparse Attention' leaf, which contains four papers total, including the original work. This leaf sits within the broader 'Sparse Attention Mechanisms for Video Diffusion Transformers' branch, indicating a moderately populated research direction focused on reducing quadratic attention costs through content-aware sparsity rather than fixed patterns.

The taxonomy reveals neighboring leaves exploring static sparse patterns (five papers on radial, diagonal, or block structures), training-free acceleration (four papers on heuristic pruning), and hybrid multi-scale designs (three papers on hierarchical attention). The original paper diverges from these by emphasizing learned routing over predefined geometry or post-hoc pruning. Adjacent branches address attention optimization via step distillation and long-term memory management, suggesting the field balances efficiency gains with temporal coherence challenges. The scope notes clarify that dynamic methods like MoC differ from static patterns by adapting attention based on input content during inference.

Among eighteen candidates examined across three contributions, the core MoC framework (Contribution A) shows one refutable candidate out of ten examined, while the content-aligned chunking strategy (Contribution B) found no refutations among eight candidates. Contribution C (causal routing) was not evaluated against prior work. The limited search scope—eighteen papers rather than an exhaustive survey—means these statistics reflect top-K semantic matches and citation expansion, not comprehensive coverage. The single refutation for the main contribution suggests some overlap with existing adaptive routing ideas, though most examined candidates appear non-overlapping or unclear.

Given the moderate density of the dynamic sparse attention leaf and the limited literature search, the work appears to occupy a recognizable niche within an active but not overcrowded subfield. The analysis covers top semantic neighbors and immediate taxonomy siblings but does not claim exhaustive prior art review. The chunking and causal routing components show less direct overlap in the examined set, though the core routing framework encounters at least one closely related prior approach among the candidates surveyed.

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

Contribution 1: Mixture of Contexts (MoC) framework with learnable sparse attention routing

Description: The authors introduce MoC, a learnable sparse attention routing mechanism that reformulates long-context video generation as an internal information retrieval process. Each query dynamically selects a few informative chunks plus mandatory anchors

(caption, local windows) to attend to, with causal routing that prevents loop closures, enabling minute-scale video generation at near short-video computational cost.

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. Pack and Force Your Memory: Long-form and Consistent Video Generation

URL: View paper

Brief Assessment

Pack and Force[61] focuses on a different technical approach: it uses memorypack with cross-attention to text/image guidance for long-term memory, rather than MoC's learnable top-k routing among video chunks. The candidate does not demonstrate that similar sparse attention routing mechanisms existed prior to the original work.

---

2. MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation

URL: View paper

Prior Art Analysis

MoGA[32] demonstrates that a similar learnable sparse attention routing mechanism for long-context video generation was proposed prior to the original paper. Both papers reformulate long-context video generation as an information retrieval task where each query dynamically selects relevant chunks through learnable routing. MoGA[32] uses a lightweight token router (a single linear layer) to assign tokens to specialized groups, enabling groupwise attention for efficient long-context modeling. This approach achieves similar goals of reducing quadratic attention costs while maintaining long-range dependencies, with MoGA[32] achieving 71.25% sparsity and generating minute-level videos at 580k context length.

Evidence

Evidence 1 - **Rationale**: Both papers propose learnable sparse attention routing mechanisms that reformulate long-context video generation as a retrieval task, where queries dynamically select relevant context through learned routing rather than fixed patterns. - **Original**: we recast long-context video generation as an internal information retrieval task and propose a simple, learnable sparse attention routing module, mixture of contexts (moc), as an effective long-term memory retrieval engine. in moc, each query dynamically selects a few informative chunks plus mandat... - **Candidate**: this paper introduces mixtureof-groups attention (moga), an efficient sparse attention that uses a lightweight, learnable token router to precisely match tokens without blockwise estimation. through semantic-aware routing, moga enables effective long-range interactions.

Evidence 2 - **Rationale**: Both papers describe learnable routing mechanisms where queries are dynamically routed to relevant context segments through trainable components, demonstrating prior work on this approach. - **Original**: in this work, we reformulate long-context video generation as an internal information retrieval process, where each token dynamically accesses only the most relevant context through learnable sparse attention routing. to realize this, we propose an adaptive mixture of contexts (moc) framework that l... - **Candidate**: mogaaddresses the above challenge via efficient token routing, where a lightweight, trainable router assigns correlated tokens to groups and performs self-attention within each group. specifically, the router is a linear projection followed by softmax gating

Evidence 3 - **Rationale**: Both papers demonstrate that learned sparse routing enables minute-level video generation with maintained consistency across shots, showing that MoGA[32] achieved similar capabilities prior to the original work. - **Original**: we show that replacing dense self-attention with our adaptive mixture of contexts (moc) reframes long-video generation as internal in-context retrieval. a learned sparse context routing policy allocates compute to salient history and sustains cross-shot identities, actions, and layouts over minutes-... - **Candidate**: building on moga, we introduce a video generation model capable of producing minute-level, multi-shot, 480p videos at 24 fps with a context length of about 580k tokens. extensive evaluations show consistent improvements over state-of-the-art (sota) sparse attention baselines

---

3. Fractal reservoir structuring for large language model generative pathways: An empirical investigation with large language model

URL: View paper

Brief Assessment

Fractal Reservoir[10] focuses on fractal reservoir structuring for LLM generative pathways, not on sparse attention routing mechanisms for video generation or retrieval-based context selection in diffusion transformers.

---

4. InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing

URL: View paper

Brief Assessment

InfiniteTalk[62] addresses audio-driven video generation for dubbing tasks with temporal context frames for inter-chunk transitions, not learnable sparse attention routing for long-context video generation as an internal information retrieval process.

---

5. EgoLCD: Egocentric Video Generation with Long Context Diffusion

URL: View paper

Brief Assessment

EgoLCD[16] focuses on egocentric video generation with a sparse KV cache for memory management, not a learnable routing mechanism that dynamically selects chunks via top-k operations as in MoC.

---

6. Learning World Models for Interactive Video Generation

URL: View paper

Brief Assessment

Learning World Models[60] focuses on interactive video generation with action conditioning and memory retrieval for world modeling consistency, not on learnable sparse attention routing mechanisms for long-context video generation. The candidate addresses compounding errors and spatiotemporal coherence through explicit global state conditioning and retrieval-augmented generation, which is architecturally distinct from MoC's internal information retrieval via dynamic chunk selection with top-k routing.

---

7. Animate-a-story: Storytelling with retrieval-augmented video generation

URL: View paper

Brief Assessment

Animate a Story[59] focuses on retrieval-augmented video generation for storytelling with structure guidance from retrieved videos, not on learnable sparse attention routing mechanisms for long-context video generation. The candidate addresses a different technical problem (storytelling with external video retrieval) compared to the original's internal information retrieval through sparse attention.

---

8. Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis

URL: View paper

Brief Assessment

Salova[57] focuses on segment-level video retrieval for long-form video question answering, not on sparse attention routing for video generation. The candidate addresses video understanding/QA tasks, while the original paper targets video generation with diffusion transformers.

---

9. Generative World-Model Planning for Long-Horizon User Preference Evolution and Responsible Personalization

URL: View paper

Brief Assessment

Generative World Model[58] focuses on user preference evolution and personalization in recommendation systems, not on sparse attention routing mechanisms for video generation. The candidate's context fragments mention recommendation and retrieval systems, which are fundamentally different from the original paper's video generation architecture.

---

10. MotionRAG: Motion Retrieval-Augmented Image-to-Video Generation

URL: View paper

Brief Assessment

MotionRAG[56] focuses on motion retrieval and transfer for image-to-video generation using a context-aware motion adaptation module, not on sparse attention routing mechanisms for long-context video generation. The technical approaches are fundamentally different.

Contribution 2: Content-aligned chunking strategy for video sequences

Description: The authors propose a content-aligned chunking approach that partitions heterogeneous multi-modal video token streams along natural boundaries (frames, shots, text segments) rather than using uniform windows. This design preserves semantic coherence and enables more discriminative top-k retrieval while maintaining compatibility with existing video generation architectures.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Split Federated Learning for Real-Time Aerial Video Event Recognition in UAV-Based Geospatial Monitoring

URL: View paper

Brief Assessment

Split Federated UAV[50] focuses on UAV-based federated learning for real-time event recognition with dynamic video chunking for distributed processing, not on multi-modal token stream partitioning for video generation architectures.

---

2. Text-video retrieval via multi-modal hypergraph networks

URL: View paper

Brief Assessment

Multimodal Hypergraph[47] focuses on text-video retrieval tasks using chunk-level matching for query-video alignment, not on video generation architectures or token stream partitioning for generative models.

---

3. Semantic-Assisted Object Clustering for Multi-Modal Referring Video Segmentation.

URL: View paper

Brief Assessment

Semantic Object Clustering[53] focuses on multi-modal referring video segmentation with semantic-assisted object clustering, not on content-aligned chunking strategies for video token sequences in generation tasks.

---

4. Adaptive Chunking for VideoRAG Pipelines with a Newly Gathered Bilingual Educational Dataset

URL: View paper

Brief Assessment

Adaptive Chunking VideoRAG[51] focuses on educational video retrieval (VideoQA) using CLIP embeddings and SSIM scores for slide-based lectures, while the original paper addresses long-context video generation with diffusion transformers using frame/shot/caption boundaries for generative modeling.

---

5. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding

URL: View paper

Brief Assessment

Hippomm[49] focuses on temporal segmentation based on perceptual boundaries (SSIM, audio energy) for memory formation in audiovisual understanding, not on partitioning multi-modal token streams for video generation architectures. The original paper's content-aligned chunking serves video generation with diffusion transformers, while Hippomm[49] addresses event-based memory encoding for retrieval tasks.

---

6. Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities

URL: View paper

Brief Assessment

Mirasol3b[46] partitions video/audio into time-synchronized chunks for autoregressive processing, but does not address heterogeneous multi-modal token streams with natural boundaries (frames, shots, text segments) as in the original paper. The candidate focuses on time-aligned media modalities rather than content-aware chunking for discriminative retrieval in diffusion transformers.

---

7. Adaptive Token Boundaries: Integrating Human Chunking Mechanisms into Multimodal LLMs

URL: View paper

Brief Assessment

Adaptive Token Boundaries[52] focuses on tokenization for multimodal LLMs processing static image-text pairs in cognitive tasks, not video generation. The candidate's chunking operates on linguistic and visual tokens for VQA/captioning tasks, whereas the original paper addresses temporal video sequences with frames, shots, and captions for long-context video generation.

---

8. Towards training-free long video understanding: methods, benchmarks, and open challenges

URL: View paper

Brief Assessment

Training Free Long[55] focuses on training-free methods for long video understanding (e.g., keyframe selection, memory structures, agent-based reasoning) rather than proposing novel chunking strategies for video generation architectures. The candidate does not address content-aligned chunking for multi-modal token streams in generative models.

Contribution 3: Causal routing mechanism to prevent pathological loop closures

Description: The authors introduce a causal masking constraint at the routing stage that restricts each chunk to attend only to earlier positions in the sequence, transforming the routing graph into a directed acyclic graph. This design eliminates isolated feedback loops and ensures information flows strictly forward in time, resulting in smoother temporal dynamics and more stable training.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 22 papers and found 1 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. Pack and Force Your Memory: Long-form and Consistent Video Generation

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] Mixture of Contexts for Long Video Generation View paper
- [1] PAROAttention: Pattern-Aware ReOrdering for Efficient Sparse and Quantized Attention in Visual Generation Models View paper
- [2] Radial Attention: Sparse Attention with Energy Decay for Long Video Generation View paper
- [3] Bidirectional sparse attention for faster video diffusion training View paper
- [4] Controllable human-object interaction synthesis View paper
- [5] InfiniteVL: Synergizing Linear and Sparse Attention for Highly-Efficient, Unlimited-Input Vision-Language Models View paper
- [6] Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control View paper
- [7] 4Real-Video-V2: Fused View-Time Attention and Feedforward Reconstruction for 4D Scene Generation View paper
- [8] Hunyuanvideo 1.5 technical report View paper
- [9] MotionClone: Training-Free Motion Cloning for Controllable Video Generation View paper
- [10] Fractal reservoir structuring for large language model generative pathways: An empirical investigation with large language model View paper
- [11] Training-free and adaptive sparse attention for efficient long video generation View paper
- [12] Grouping First, Attending Smartly: Training-Free Acceleration for Diffusion Transformers View paper
- [13] BLADE: Block-Sparse Attention Meets Step Distillation for Efficient Video Generation View paper
- [14] Macro-from-micro planning for high-quality and parallelized autoregressive long video generation View paper
- [15] HoloCine: Holistic Generation of Cinematic Multi-Shot Long Video Narratives View paper
- [16] EgoLCD: Egocentric Video Generation with Long Context Diffusion View paper
- [17] CPA: Camera-pose-awareness Diffusion Transformer for Video Generation View paper
- [18] ConditionVideo: Training-Free Condition-Guided Text-to-Video Generation View paper
- [19] DraftAttention: Fast Video Diffusion via Low-Resolution Attention Guidance View paper
- [20] MaGGIe: Masked Guided Gradual Human Instance Matting View paper
- [21] SpargeAttention: Accurate and Training-free Sparse Attention Accelerating Any Model Inference View paper
- [22] VORTA: Efficient Video Diffusion via Routing Sparse Attention View paper
- [23] Sparse-vDiT: Unleashing the Power of Sparse Attention to Accelerate Video Diffusion Transformers View paper
- [24] Long Video Diffusion Generation with Segmented Cross-Attention and Content-Rich Video Data Curation View paper
- [25] XAttention: Block Sparse Attention with Antidiagonal Scoring View paper
- [26] Radial Attention: $O(n\log n)$ Sparse Attention with Energy Decay for Long Video Generation View paper
- [27] Lag-Relative Sparse Attention In Long Context Training View paper
- [28] LongCat-Video Technical Report View paper
- [29] Compact Attention: Exploiting Structured Spatio-Temporal Sparsity for Fast Video Generation View paper
- [30] PSA: Pyramid Sparse Attention for Efficient Video Understanding and Generation View paper
- [31] Input-Aware Sparse Attention for Real-Time Co-Speech Video Generation View paper
- [32] MoGA: Mixture-of-Groups Attention for End-to-End Long Video Generation View paper
- [33] SpargeAttn: Accurate Sparse Attention Accelerating Any Model Inference View paper
- [34] VideoNSA: Native Sparse Attention Scales Video Understanding View paper
- [35] End-to-End Training for Autoregressive Video Diffusion via Self-Resampling View paper
- [36] VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models View paper
- [37] USV: Unified Sparsification for Accelerating Video Diffusion Models View paper
- [38] AutoScape: Geometry-Consistent Long-Horizon Scene Generation View paper
- [39] Temporally Consistent Object Editing in Videos using Extended Attention View paper
- [40] Flow-Guided Sparse Transformer for Video Deblurring View paper
- [41] Motion Selective Prediction for Video Frame Synthesis View paper
- [42] GODIVA: Generating Open-DomaIn Videos from nAtural Descriptions View paper
- [43] Adaptive Compact Attention For Few-shot Video-to-video Translation View paper
- [44] Radial Attention: Sparse Attention for Long Video Generation View paper

- [45] ZeroTrail: Zero-Shot Trajectory Control Framework for Video Diffusion Models View paper
- [46] Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities View paper
- [47] Text-video retrieval via multi-modal hypergraph networks View paper
- [48] Text-video retrieval via variational multi-modal hypergraph networks View paper
- [49] Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding View paper
- [50] Split Federated Learning for Real-Time Aerial Video Event Recognition in UAV-Based Geospatial Monitoring View paper
- [51] Adaptive Chunking for VideoRAG Pipelines with a Newly Gathered Bilingual Educational Dataset View paper
- [52] Adaptive Token Boundaries: Integrating Human Chunking Mechanisms into Multimodal LLMs View paper
- [53] Semantic-Assisted Object Clustering for Multi-Modal Referring Video Segmentation. View paper
- [54] Towards Effective and Efficient Long Video Understanding of Multimodal Large Language Models via One-shot Clip Retrieval View paper
- [55] Towards training-free long video understanding: methods, benchmarks, and open challenges View paper
- [56] MotionRAG: Motion Retrieval-Augmented Image-to-Video Generation View paper
- [57] Salova: Segment-augmented long video assistant for targeted retrieval and routing in long-form video analysis View paper
- [58] Generative World-Model Planning for Long-Horizon User Preference Evolution and Responsible Personalization View paper
- [59] Animate-a-story: Storytelling with retrieval-augmented video generation View paper
- [60] Learning World Models for Interactive Video Generation View paper
- [61] Pack and Force Your Memory: Long-form and Consistent Video Generation View paper
- [62] InfiniteTalk: Audio-driven Video Generation for Sparse-Frame Video Dubbing View paper