Novelty Assessment Report

Paper: SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer
PDF URL: https://openreview.net/pdf?id=mzAchylAtf
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We introduce SANA-Video, a small diffusion model that can efficiently generate videos up to 720×1280 resolution and minute-length duration. SANA-Video synthesizes high-resolution, high-quality and long videos with strong text-video alignment at a remarkably fast speed, deployable on RTX 5090 GPU. Two core designs ensure our efficient, effective and long video generation: (1) Linear DiT: We leverage linear attention as the core operation, which is more efficient than vanilla attention given the large number of tokens processed in video generation. (2) Constant-Memory KV cache for Block Linear Attention: we design block-wise autoregressive approach for long video generation by employing a constant-memory state, derived from the cumulative properties of linear attention. This KV cache provides the Linear DiT with global context at a fixed memory cost, eliminating the need for a traditional KV cache and enabling efficient, minute-long video generation. In addition, we explore effective data filters and model training strategies, narrowing the training cost to 12 days on 64 H100 GPUs, which is only 1\% of the cost of MovieGen. Given its low cost, SANA-Video achieves competitive performance compared to modern state-of-the-art small diffusion models (e.g., Wan 2.1-1.3B and SkyReel-V2-1.3B) while being 16x faster in measured latency. Moreover, SANA-Video can be deployed on RTX 5090 GPUs with NVFP4 precision, accelerating the inference speed of generating a 5-second 720p video from 71s to 29s (2.4x} speedup). In summary, SANA-Video enables low-cost, high-quality video generation. Code and model will be publicly released.

Core Task Landscape

This paper addresses: **Efficient Video Generation with Linear Attention**

A total of **36 papers** were analyzed and organized into a taxonomy with **25 categories**.

Taxonomy Overview

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

- **Linear Attention Mechanisms for Diffusion Transformers**
- **Sparse and Hybrid Attention Strategies**
- **State Space Models for Video Generation**
- **Consistency and Temporal Coherence Methods**
- **Efficient Inference and Deployment Optimization**
- **Video Compression with Linear Complexity**
- **Video Editing and Restoration with Linear Attention**
- **Multimodal and Cross-Modal Efficient Attention**
- **Domain-Specific Linear Attention Applications**

Complete Taxonomy Tree

- Efficient Video Generation with Linear Attention Survey Taxonomy
- Linear Attention Mechanisms for Diffusion Transformers
  - Gated Linear Attention in Diffusion Models (2 papers)
  - [1] Dig: Scalable and efficient diffusion models with gated linear attention (Lianghui Zhu, 2025) View paper
  - [23] InfiniteVL: Synergizing Linear and Sparse Attention for Highly-Efficient, Unlimited-Input Vision-Language Models (Hongyuan Tao, 2025) View paper
  - Block Linear Attention for Long Video Synthesis ★ (3 papers)
  - [0] SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer (Anon et al., 2026) View paper
  - [4] LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity (Hongjie Wang, 2024) View paper
  - [20] LinGen-Uni: A Universal Linear-Complexity Framework for High-Resolution Minute-Length Text-to-Video Generation (Hongjie Wang, 2025) View paper
  - Post-Training Linear Attention Adaptation (2 papers)
  - [6] Linvideo: A post-training framework towards o (n) attention in efficient video generation (Ge, 2025) View paper
  - [9] Attention Surgery: An Efficient Recipe to Linearize Your Video Diffusion Transformer (Ghafoorian, 2025) View paper
- Sparse and Hybrid Attention Strategies
  - Radial and Energy-Decay Sparse Attention (2 papers)
  - [3] Radial Attention: Sparse Attention with Energy Decay for Long Video Generation (X Li, 2025) View paper
  - [35] Radial Attention: Sparse Attention for Long Video Generation (X Li, n.d.) View paper
  - Trainable Sparse-Linear Fusion (1 papers)
  - [7] SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention (Zhang Jintao, 2025) View paper
  - Attention Tile Pruning for Video Diffusion (1 papers)
  - [21] Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile (Li DaCheng, 2025) View paper

- State Space Models for Video Generation
  - Bidirectional State Space Diffusion Models (3 papers)
  - [22] Pushing the Boundaries of State Space Models for Image and Video Generation (Hong Yi-cong, 2025) View paper
  - [26] Ssm Meets Video Diffusion Models: Efficient Long-Term Video Generation with Selective State Spaces (Yuta Oshima, n.d.) View paper
  - [29] Scaling Diffusion Mamba with Bidirectional SSMs for Efficient Image and Video Generation (Mo, 2024) View paper
  - Medical and Specialized Video SSMs (2 papers)
  - [19] Vivim: A Video Vision Mamba for Ultrasound Video Segmentation (Yang Yijun, 2025) View paper
  - [25] Thermal Video Enhancement Mamba: A Novel Approach to Thermal Video Enhancement for Real-World Applications (Sargis Hovhannisyan, 2025) View paper
- Consistency and Temporal Coherence Methods
  - Self-Attention Consistency for Long-Range Generation (1 papers)
  - [2] Storydiffusion: Consistent self-attention for long-range image and video generation (Ming-Ming Cheng, 2024) View paper
  - Memory-Guided Diffusion for Talking Videos (2 papers)
  - [11] Memo: Memory-guided diffusion for expressive talking video generation (Zheng, 2024) View paper
  - [12] Efficient Long-duration Talking Video Synthesis with Linear Diffusion Transformer under Multimodal Guidance (Zhang Haojie, 2024) View paper
  - Image Animation with Linear Transformers (1 papers)
  - [14] Consistent and Controllable Image Animation with Motion Linear Diffusion Transformers (Ma Xin, 2025) View paper
- Efficient Inference and Deployment Optimization
  - On-Device and Mobile Video Generation (2 papers)
  - [10] On-device Sora: Enabling Training-Free Diffusion-based Text-to-Video Generation for Mobile Devices (Kim, 2025) View paper
  - [27] MobileI2V: Fast and High-Resolution Image-to-Video on Mobile Devices (Shuai Zhang, 2025) View paper
  - Full-Dimensional Efficient Attention for Medical Video (1 papers)
  - [5] FEAT: Full-Dimensional Efficient Attention Transformer for Medical Video Generation (Wang Huihan, 2025) View paper
  - Plug-and-Play Linear Attention for Restoration (1 papers)
  - [16] Plug-and-Play Linear Attention for Pre-trained Image and Video Restoration Models (Nair, 2025) View paper
- Video Compression with Linear Complexity
  - Generative Implicit Video Compression (1 papers)
  - [8] Givic: Generative implicit video compression (Gao Ge, 2025) View paper
  - Learned Compression with Motion Compensation (1 papers)
  - [33] LVC-LGMC: Joint Local and Global Motion Compensation for Learned Video Compression (Jiang Wei, 2024) View paper
- Video Editing and Restoration with Linear Attention
  - Linear Attention for Video Editing (1 papers)
  - [17] VRWKV-Editor: Reducing quadratic complexity in transformer-based video editing (Hmamouche, 2025) View paper
  - Video Super-Resolution with Linear Attention (2 papers)
  - [30] VMG: Rethinking U-Net Architecture for Video Super-Resolution (Jun Tang, 2024) View paper
  - [31] Space-Time Video Super-Resolution With Neural Operator (Yuantong Zhang, 2025) View paper
  - DemoirÃ©ing with Linear Attention and Flow Matching (1 papers)
  - [24] MoirÃ©XNet: Adaptive Multi-Scale DemoirÃ©ing with Linear Attention Test-Time Training and Truncated Flow Matching Prior (Li, 2025) View paper
- Multimodal and Cross-Modal Efficient Attention
  - Efficient Vision-Language Models with Linear Attention (1 papers)
  - [28] PERCEIVER-VL: Efficient Vision-and-Language Modeling with Iterative Latent Attention (Zineng Tang, 2023) View paper
  - Slow-Fast Video Multi-Modal Large Language Models (1 papers)
  - [32] Slow-Fast Architecture for Video Multi-Modal Large Language Models (Shi Min, 2025) View paper
  - Multi-Linear Attention Networks for Video Features (1 papers)
  - [18] Generalizable multi-linear attention network (Tao Jin, 2021) View paper
- Domain-Specific Linear Attention Applications
  - Industrial Video Analysis with Linear Attention (1 papers)
  - [15] A global linear attention incorporated video transformer for robust sintering condition recognition (Leyuan Wu, 2025) View paper
  - Polynomial Mixer for Image and Video Generation (1 papers)
  - [34] PoM: Efficient Image and Video Generation with the Polynomial Mixer (Picard, 2024) View paper
  - Survey and Theoretical Foundations (2 papers)
  - [13] Paying Attention to Video Generation (Rishika Bhagwatkar, 2021) View paper
  - [36] A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention (J Zhang, n.d.) View paper

Narrative

Core task: efficient video generation with linear attention. The field addresses the computational bottleneck of quadratic attention in video diffusion transformers by exploring diverse architectural strategies. The taxonomy reveals several major branches: Linear Attention Mechanisms for Diffusion Transformers develop direct replacements for standard attention that scale linearly with sequence length, often through kernel approximations or block-wise decompositions (e.g., LinGen[4], LinVideo[6]). Sparse and Hybrid Attention Strategies selectively compute attention over subsets of tokens or combine local and global patterns to reduce complexity while preserving quality. State Space Models for Video Generation leverage recurrent formulations like Mamba to achieve linear scaling, trading the flexibility of attention for efficiency (e.g., Diffusion Mamba[29]). Additional branches tackle Consistency and Temporal Coherence Methods to maintain frame-to-frame stability, Efficient Inference and Deployment Optimization for real-time or on-device scenarios (e.g., On-device Sora[10]), and domain-specific applications ranging from video editing to multimodal synthesis.

Within the Linear Attention Mechanisms branch, a particularly active line of work focuses on block linear attention for long video synthesis, where models partition sequences into manageable chunks to balance memory and quality. SANA-Video[0] exemplifies this approach by employing block-structured linear attention to generate extended sequences efficiently, positioning itself alongside LinGen[4] and LinGen-Uni[20], which similarly decompose attention across temporal blocks. These methods contrast with global linear attention schemes (e.g., Global Linear Attention[15]) that apply a single linear operator across all frames, trading fine-grained temporal modeling for simplicity. Meanwhile, hybrid strategies like Radial Attention[3] and plug-and-play modules (Plug-and-Play Linear[16]) explore middle grounds, injecting linear components into existing architectures without full redesigns. The central tension across these

directions lies in balancing computational savings, temporal coherence, and the ability to capture long-range dependencies—challenges that SANA-Video[0] addresses through its block-wise design, which shares conceptual ground with LinGen[4] but may differ in block size, attention kernel, or training objectives.

Related Works in Same Category

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

1. LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity

Authors: Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, et al. (16 authors total) | **Year/Venue**: 2024 • Computer Vision and Pattern Recognition | **URL**: View paper

Abstract

Text-to-video generation enhances content creation but is highly computationally intensive: The computational cost of Diffusion Transformers (DiTs) scales quadratically in the number of pixels. This makes minute-length video generation extremely expensive, limiting most existing models to generating videos of only 10-20 seconds length. We propose a Linear-complexity text-to-video Generation (Lin-Gen) framework whose cost scales linearly in the number of pixels. For the first time, LinGen enables...

Relationship Analysis

Both papers belong to the Block Linear Attention for Long Video Synthesis category, employing linear attention mechanisms to reduce computational complexity from $O(N^2)$ to $O(N)$ for efficient video generation. They overlap in using block-wise approaches with constant-memory KV caching for minute-length video generation, with SANA-Video using ReLU-based linear attention with 3D RoPE and block causal attention, while LinGen replaces self-attention entirely with MATE blocks combining bidirectional Mamba2 and temporal Swin attention. The key difference is that SANA-Video focuses on autoregressive block training with monotonically increasing SNR sampling and leverages pre-trained T2I models, whereas LinGen introduces a novel MATE architecture with Rotary Major Scan and review tokens specifically designed to address Mamba's adjacency preservation issues.

---

2. LinGen-Uni: A Universal Linear-Complexity Framework for High-Resolution Minute-Length Text-to-Video Generation

Authors: Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, et al. (13 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Text-to-video generation enhances content creation but is highly computationally intensive: The computational cost of Diffusion Transformers (DiTs) scales quadratically in the number of pixels. This makes minute-length video generation extremely expensive, limiting most existing models to generating videos of only 10-20 seconds length. We propose a Linear-complexity text-to-video Generation (LinGen) framework whose cost scales linearly in the number of pixels. For the fir...

Relationship Analysis

Both papers belong to the same taxonomy category of block linear attention for long video generation, employing linear attention mechanisms to enable efficient minute-length video synthesis with constant-memory KV caching. They overlap in their core approach of replacing quadratic-complexity attention with linear alternatives and using block-wise processing for extended video generation. However, SANA-Video focuses on block-wise autoregressive training with monotonically increasing SNR sampling and self-forcing for long videos, while LinGen-Uni emphasizes a MATE block architecture combining bidirectional Mamba2 with Rotary Major Scan and TEmporal Swin Attention, plus a distillation framework to convert pre-trained DiTs to linear complexity.

Contributions Analysis

Overall novelty summary. SANA-Video introduces a linear diffusion transformer for efficient high-resolution, minute-length video generation, combining linear attention with a constant-memory KV cache for block-wise autoregressive synthesis. The paper resides in the 'Block Linear Attention for Long Video Synthesis' leaf, which contains only three papers including SANA-Video itself. This is a relatively sparse research direction within the broader taxonomy of 36 papers across the field, suggesting the specific combination of block linear attention and constant-memory caching for extended video generation remains an emerging area with limited prior exploration.

The taxonomy reveals that SANA-Video's parent branch, 'Linear Attention Mechanisms for Diffusion Transformers', encompasses neighboring approaches like gated linear attention and post-training adaptation methods. Adjacent branches explore sparse-linear fusion strategies and state space models, which offer alternative pathways to linear complexity. The taxonomy's scope notes clarify that block linear attention specifically targets minute-length synthesis through structured decomposition, distinguishing it from full-sequence linear methods and hybrid sparse approaches. SANA-Video's positioning suggests it bridges architectural efficiency (linear attention) with practical deployment constraints (constant-memory caching), connecting to but diverging from pure architectural innovations in neighboring leaves.

Among the three contributions analyzed, the linear DiT architecture examined 10 candidates with 2 appearing to provide overlapping prior work, while the constant-memory KV cache examined 5 candidates with none clearly refuting novelty. The training strategy contribution also examined 10 candidates with 2 potential overlaps. Given the limited search scope of 25 total candidates from semantic search, these statistics suggest the core architectural innovation (linear DiT) operates in a more crowded space, whereas the constant-memory caching mechanism for block attention appears less directly addressed in the examined literature. The analysis does not claim exhaustive coverage but indicates differential novelty across contributions within the sampled candidate set.

Based on the limited literature search, SANA-Video appears to occupy a sparsely populated niche combining block linear attention with constant-memory mechanisms for long video synthesis. The taxonomy structure and sibling paper count suggest this specific integration is relatively underexplored, though individual components (linear attention, block decomposition) have precedents in the examined candidates. The analysis reflects top-25 semantic matches and does not capture the full landscape of video generation research, particularly work outside the linear attention paradigm or published after the search cutoff.

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

Contribution 1: Linear DiT for efficient video generation

Description: The authors extend SANA's linear DiT design to video by replacing all attention modules with efficient linear attention, reducing complexity from $O(N^2)$ to $O(N)$. They integrate Rotary Position Embeddings (RoPE) and introduce a 1D temporal convolution to the Mix-FFN for improved spatio-temporal modeling.

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. On-device Sora: Enabling Training-Free Diffusion-based Text-to-Video Generation for Mobile Devices

URL: View paper

Brief Assessment

On-device Sora[10] focuses on inference optimization techniques (linear proportional leap, token merging, dynamic loading) for pre-trained models on mobile devices, not on architectural design of linear attention mechanisms for video generation training.

---

2. Photorealistic video generation with diffusion models

URL: View paper

Brief Assessment

Photorealistic Diffusion[52] uses window attention architecture for video generation, not linear attention mechanisms. The candidate's approach differs fundamentally from the original paper's linear attention design with O(N) complexity.

---

3. Memo: Memory-guided diffusion for expressive talking video generation

URL: View paper

Brief Assessment

Memo[11] focuses on audio-driven talking video generation using memory-guided temporal modules with linear attention for identity consistency, not on general-purpose video generation efficiency or extending SANA's linear DiT architecture to video domains.

---

4. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers

URL: View paper

Prior Art Analysis

SANA[53] demonstrates that linear attention mechanisms were already applied to diffusion transformers for efficient image generation before the original paper's video work. The candidate paper explicitly describes replacing all vanilla attention with linear attention (reducing complexity from $O(N^2)$ to $O(N)$) in their Linear DiT architecture, and integrating RoPE for improved modeling. While SANA[53] focuses on image generation rather than video, it establishes prior art for the core architectural innovation of using linear attention in diffusion transformers with RoPE integration, which the original paper extends to video.

Evidence

Evidence 1 - **Rationale**: Both papers describe replacing all attention modules with linear attention to reduce computational complexity from $O(N^2)$ to $O(N)$. The candidate paper explicitly states they 'first proposed linear dit', establishing prior work on this architectural design. - **Original**: we extend SANA (xie et al., 2025a) linear dit design to the video domain, addressing the significant computational bottleneck of traditional self-attention (o(n2)), as shown in fig. 1(d). By replacing all attention modules with our efficient linear attention, we reduce complexity too(n) - **Candidate**: we first proposed linear dit, which completely replaces the original self-attention with linear attention, achieving higher computational efficiency in high-resolution generation without compromising performance

Evidence 2 - **Rationale**: While the original paper adds temporal convolution specifically for video, the candidate paper already established the Mix-FFN design with convolution for aggregating local information in the linear attention framework. - **Original**: we introduce a 1d temporal convolution to the mix-ffn via a shortcut connection. this design allows us to effectively leverage pre-trained image models and efficiently adapt them for video generation by aggregating temporal features - **Candidate**: we employ mixffn to replace the original mlp-ffn, incorporating 3x3 depth-wise convolution to better aggregate token information

---

5. Efficient Long-duration Talking Video Synthesis with Linear Diffusion Transformer under Multimodal Guidance

URL: View paper

Brief Assessment

Efficient Talking Video[12] focuses on talking video synthesis with multimodal guidance (audio and portrait features), not general video generation frameworks with linear attention mechanisms as explored in the original paper.

---

6. Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile

URL: View paper

Brief Assessment

Efficient-vDiT[21] focuses on pruning 3D full attention based on tile-style repetitive patterns and uses sparse attention, not linear attention mechanisms. The candidate does not employ linear attention with O(N) complexity as described in the original paper.

---

7. Efficient diffusion transformer with step-wise dynamic attention mediators

URL: View paper

Brief Assessment

Dynamic Attention Mediators[55] focuses on step-wise dynamic attention mechanisms rather than linear attention architectures for video generation. The candidate's context fragments mention linear attention but do not provide sufficient detail to establish prior work on linear DiT designs with RoPE and temporal convolutions for video generation.

---

8. ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features

URL: View paper

Brief Assessment

ConceptAttention[54] focuses on interpretability of diffusion transformer attention layers for segmentation tasks, not on linear attention mechanisms for efficient video generation.

---

9. Radial Attention: Sparse Attention with Energy Decay for Long Video Generation

URL: View paper

Brief Assessment

Radial Attention[3] focuses on sparse attention with energy decay mechanisms for long video generation, not on extending linear DiT architectures from image to video domains with RoPE and temporal convolutions as described in the original paper.

---

10. LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity

URL: View paper

Prior Art Analysis

LinGen[4] demonstrates prior work on linear attention mechanisms for video generation with diffusion transformers. Both papers replace quadratic self-attention with linear-complexity alternatives to address computational bottlenecks in video generation. LinGen[4] explicitly states it 'replaces the computationally-dominant and quadratic-complexity block, self-attention, with a linear-complexity block' and achieves 'linear-complexity text-to-video generation' where 'cost scales linearly in the number of pixels.' This directly parallels the

original paper's claim of extending 'linear DiT design to video by replacing all attention modules with efficient linear attention, reducing complexity from O(N²) to O(N).' Both works target the same fundamental problem (quadratic attention complexity in video generation) with the same core solution approach (linear attention mechanisms), though they differ in specific architectural implementations (SANA uses ReLU linear attention with RoPE, while LinGen uses MA-TE blocks with Mamba2).

Evidence

Evidence 1 - **Rationale**: Both papers identify the same computational bottleneck (quadratic complexity of self-attention in DiTs) and propose the same fundamental solution (linear complexity mechanisms). This demonstrates that the concept of applying linear attention to video DiTs was not novel to the original paper. - **Original**: we extend sana (xie et al., 2025a) linear dit design to the video domain, addressing the significant computational bottleneck of traditional self-attention (o(n2)), as shown in fig. 1(d). by replacing all attention modules with our efficient linear attention, we reduce complexity too(n), which is cr... - **Candidate**: text-to-video generation enhances content creation but is highly computationally intensive: the computational cost of diffusion transformers (dits) scales quadratically in the number of pixels. this makes minute-length video generation extremely expensive, limiting most existing models to generating...

Evidence 2 - **Rationale**: This pair shows both papers explicitly replace self-attention with linear-complexity alternatives in video generation DiTs, demonstrating prior art for this architectural approach. - **Original**: linear dit.we extend sana (xie et al., 2025a) linear dit design to the video domain, addressing the significant computational bottleneck of traditional self-attention (o(n2)), as shown in fig. 1(d). by replacing all attention modules with our efficient linear attention, we reduce complexity too(n) - **Candidate**: it replaces the computationally-dominant and quadratic-complexity block, self-attention, with a linear-complexity block called mate, which consists of an ma-branch and a te-branch.

Evidence 3 - **Rationale**: Both papers articulate the identical motivation (quadratic complexity limiting long video generation) and solution direction (linear complexity frameworks), showing the original paper's contribution was not the first to address this problem with linear attention in video DiTs. - **Original**: scaling video generation to higher resolutions and longer sequences dramatically increases the number of tokens, making the o(n2) complexity of self-attention a major bottleneck in computation, speed, and memory. this underscores the need for efficient linear attention in video generation. - **Candidate**: the computational cost of diffusion transformers (dits) scales quadratically in the number of pixels. this makes minute-length video generation extremely expensive, limiting most existing models to generating videos of only 10-20 seconds length. we propose a linear-complexity text-to-video generatio...

Contribution 2: Constant-memory KV cache for block linear attention

Description: The authors reformulate causal linear attention to maintain a fixed-memory KV cache that provides global context at constant memory cost. This enables efficient minute-long video generation without the memory overhead of traditional KV caches used in full attention models.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. InfVSR: Breaking Length Limits of Generic Video Super-Resolution

URL: View paper

Brief Assessment

InfVSR[41] focuses on video super-resolution using a rolling KV-cache for causal self-attention in autoregressive inference, not on reformulating causal linear attention for constant-memory global context in video generation as the original paper does.

---

2. LongLive: Real-time Interactive Long Video Generation

URL: View paper

Brief Assessment

LongLive[37] uses a KV-recache mechanism with causal attention and frame-level attention sink for long video generation, which is architecturally different from the original paper's constant-memory KV cache derived from cumulative properties of block linear attention. The candidate focuses on refreshing cached states for interactive prompt switching rather than maintaining a fixed-memory global context through linear attention reformulation.

---

3. GoldFinch: High Performance RWKV/Transformer Hybrid with Linear Pre-Fill and Extreme KV-Cache Compression

URL: View paper

Brief Assessment

GoldFinch[38] focuses on a hybrid architecture combining linear attention with transformers for cache compression during autoregressive generation, not on block-wise video generation with constant-memory global context as in the original paper.

---

4. SneakPeek: Future-Guided Instructional Streaming Video Generation

URL: View paper

Brief Assessment

SneakPeek[39] focuses on instructional video generation with future-guided streaming and dual-region KV caching for causal prediction, not on reformulating causal linear attention to maintain constant-memory global context for long video generation.

---

5. LoRATv2: Enabling Low-Cost Temporal Modeling in One-Stream Trackers

URL: View paper

Brief Assessment

LoRATv2[40] uses KV caching for frame-wise causal attention in visual object tracking, not for video generation. The technical context (tracking vs. generation) and architectural integration differ fundamentally from the original paper's video generation framework.

Contribution 3: Efficient training strategy and data filtering

Description: The authors develop a multi-stage training approach that leverages pre-trained text-to-image models, applies resolution-specific data filtering criteria, and uses a coarse-to-fine training paradigm. This reduces training costs to approximately 1% of MovieGen while achieving competitive performance.

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. ID-Crafter: VLM-Grounded Online RL for Compositional Multi-Subject Video Generation

URL: View paper

Brief Assessment

ID-Crafter[51] focuses on multi-subject identity preservation in video generation using hierarchical attention and VLM guidance, not on efficient training strategies or data filtering for text-to-video diffusion models.

2. Mora: Enabling Generalist Video Generation via A Multi-Agent Framework

URL: View paper

Brief Assessment

Mora[46] focuses on multi-agent collaboration frameworks for video generation, not on efficient training strategies for diffusion models. The original paper's contribution centers on coarse-to-fine training paradigms and resolution-specific data filtering for diffusion transformers, while Mora[46] addresses agent coordination and data synthesis through large models rather than training efficiency optimization.

3. DomainDiff: Unified Two-Stage Optimization for Text-Video Retrieval

URL: View paper

Brief Assessment

DomainDiff[50] focuses on text-video retrieval tasks, not video generation. The candidate addresses cross-modal retrieval optimization, which is fundamentally different from the original paper's video generation training strategies and data filtering for diffusion models.

4. Vivid-VR: Distilling Concepts from Text-to-Video Diffusion Transformer for Photorealistic Video Restoration

URL: View paper

Brief Assessment

Vivid-VR[45] focuses on video restoration using a concept distillation training strategy for controllable generation pipelines, not on efficient training strategies for general text-to-video diffusion models or data filtering criteria for video generation at scale.

5. Waver: Wave your way to lifelike video generation

URL: View paper

Brief Assessment

Waver[44] focuses on data curation pipelines and MLLM-based quality filtering for video generation, but does not provide sufficient technical detail about multi-stage training approaches or resolution-specific filtering criteria that would directly challenge the original paper's novelty claims about their specific training methodology.

6. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

URL: View paper

Prior Art Analysis

Stable Video Diffusion[43] demonstrates that multi-stage training approaches leveraging pre-trained text-to-image models, combined with systematic data curation and filtering strategies, were already established for video generation. The candidate paper explicitly describes a three-stage training methodology (text-to-image pretraining, video pretraining, and high-quality video finetuning) with detailed data curation processes including captioning and filtering strategies. This directly challenges the novelty claim that the original paper was first to develop such approaches, as Stable Video Diffusion[43] presents these same core concepts in a systematic framework.

Evidence

Evidence 1 - **Rationale**: Both papers describe the approach of adapting pre-trained 2D image models for video generation. Stable Video Diffusion[43] establishes this as a known methodology, indicating that leveraging text-to-image pretraining for video models was already an established practice. - **Original**: stage1: v ae adaptation on text-to-image (t2i). training video dit models from scratch is resource-intensive due to the mismatch between image and video v aes. we address this by first efficiently adapting existing t2i models to new video v aes. - **Candidate**: recently, latent diffusion models trained for 2d image synthesis have been turned into generative video models by inserting temporal layers and finetuning them on small, high-quality video datasets.

Evidence 2 - **Rationale**: Both papers describe a staged approach where models are initialized from pre-trained image models and then fine-tuned on video data. Stable Video Diffusion[43] demonstrates this multi-stage finetuning strategy was already established in the field. - **Original**: stage2: continue pre-training from t2i model. initializing video linear dit from a pre-trained t2i model (xie et al., 2025a) is an efficient and effective way to leverage the well-learned visual and textual semantic knowledge. therefore, we initialize our sana-video with a model adapted from the fir... - **Candidate**: we then explore the impact of finetuning our base model on high-quality data and train a text-to-video model that is competitive with closed-source video generation.

7. Kandinsky 5.0: A Family of Foundation Models for Image and Video Generation

URL: View paper

Prior Art Analysis

Kandinsky[49] demonstrates that similar efficient training strategies and data filtering approaches for text-to-video models existed prior to the original paper. Both papers describe multi-stage training pipelines that leverage pre-trained models, apply resolution-specific data filtering, and use coarse-to-fine training paradigms. Kandinsky[49] explicitly mentions 'multi-stage training pipeline that involves extensive pre-training' and 'data curation lifecycle - including collection, processing, filtering and clustering', which directly parallels the original paper's claimed contributions of multi-stage training with data filtering criteria and coarse-to-fine approaches.

Evidence

Evidence 1 - **Rationale**: Both papers describe comprehensive data filtering and multi-stage training approaches. Kandinsky[49] explicitly details data curation (collection, processing, filtering, clustering) combined with multi-stage training, demonstrating that such efficient training strategies with data filtering were already established in prior work. - **Original**: we explore effective data filters and model training strategies, narrowing the training cost to12 days on 64 h100 gpus, which is only 1% of the cost of moviegen. - **Candidate**: we provide a comprehensive review of the data curation lifecycle - including collection, processing, filtering and clustering - for the multi-stage training pipeline that involves extensive pre-training and incorporates quality-enhancement techniques such as self-supervised fine-tuning (sft) and rei...

Evidence 2 - **Rationale**: The original paper attributes low training cost to pre-trained models and efficient data filtering. Kandinsky[49] describes the same approach: leveraging pre-training combined with comprehensive data filtering (collection, processing, filtering, clustering), showing this strategy existed in prior work. - **Original**: the low training cost is mainly attribute to three aspects: the powerful pre-trained text-to-image (t2i) model, efficient data filtering, and the efficient training strategy. - **Candidate**: we provide a comprehensive review of the data curation lifecycle - including collection, processing, filtering and clustering - for the multi-stage training pipeline that involves extensive pre-training

Evidence 3 - **Rationale**: The original paper's coarse-to-fine training approach (learning from abundant data then refining with higher-quality data) is reflected in Kandinsky[49]'s multi-stage training pipeline with data filtering and clustering, indicating this training paradigm was already established. - **Original**: This coarse-to-fine approach efficiently encourages sana-video to fast learn dynamic information with abundant data and then refine details using less, but higher-quality, data. - **Candidate**: we provide a comprehensive review of the data curation lifecycle - including collection, processing, filtering and clustering - for the multi-stage training pipeline

8. AMD-Hummingbird: Towards an Efficient Text-to-Video Model

URL: View paper

Brief Assessment

AMD-Hummingbird[48] focuses on model pruning (reducing U-Net from 1.4B to 0.7B parameters) and visual feedback learning for efficiency, rather than the multi-stage coarse-to-fine training paradigm with resolution-specific filtering that the original paper employs. The candidate's data processing uses LLMs and VQA models for prompt/video enhancement, which differs from the original's resolution-progressive training strategy.

---

9. VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models

URL: View paper

Brief Assessment

VideoCrafter2[42] focuses on leveraging low-quality videos and synthesized high-quality images through spatial-temporal module coupling analysis, rather than the multi-stage coarse-to-fine training paradigm with resolution-specific filtering criteria described in the original paper.

---

10. EffiVED: Efficient Video Editing via Text-instruction Diffusion Models

URL: View paper

Brief Assessment

EffiVED[47] focuses on video editing tasks using image editing datasets and open-world videos, not on general text-to-video generation training strategies. The candidate addresses data scarcity for video editing specifically, while the original contribution concerns efficient training for text-to-video diffusion models with resolution-specific filtering and coarse-to-fine paradigms.

Appendix: Text Similarity Detection

Textual similarity detection checked 26 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. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers

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] SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer View paper
- [1] Dig: Scalable and efficient diffusion models with gated linear attention View paper
- [2] Storydiffusion: Consistent self-attention for long-range image and video generation View paper
- [3] Radial Attention: Sparse Attention with Energy Decay for Long Video Generation View paper
- [4] LinGen: Towards High-Resolution Minute-Length Text-to-Video Generation with Linear Computational Complexity View paper
- [5] FEAT: Full-Dimensional Efficient Attention Transformer for Medical Video Generation View paper
- [6] Linvideo: A post-training framework towards o (n) attention in efficient video generation View paper
- [7] SLA: Beyond Sparsity in Diffusion Transformers via Fine-Tunable Sparse-Linear Attention View paper
- [8] Givic: Generative implicit video compression View paper
- [9] Attention Surgery: An Efficient Recipe to Linearize Your Video Diffusion Transformer View paper
- [10] On-device Sora: Enabling Training-Free Diffusion-based Text-to-Video Generation for Mobile Devices View paper
- [11] Memo: Memory-guided diffusion for expressive talking video generation View paper
- [12] Efficient Long-duration Talking Video Synthesis with Linear Diffusion Transformer under Multimodal Guidance View paper
- [13] Paying Attention to Video Generation View paper
- [14] Consistent and Controllable Image Animation with Motion Linear Diffusion Transformers View paper
- [15] A global linear attention incorporated video transformer for robust sintering condition recognition View paper
- [16] Plug-and-Play Linear Attention for Pre-trained Image and Video Restoration Models View paper
- [17] VRWKV-Editor: Reducing quadratic complexity in transformer-based video editing View paper
- [18] Generalizable multi-linear attention network View paper
- [19] Vivim: A Video Vision Mamba for Ultrasound Video Segmentation View paper
- [20] LinGen-Uni: A Universal Linear-Complexity Framework for High-Resolution Minute-Length Text-to-Video Generation View paper
- [21] Efficient-vDiT: Efficient Video Diffusion Transformers With Attention Tile View paper
- [22] Pushing the Boundaries of State Space Models for Image and Video Generation View paper
- [23] InfiniteVL: Synergizing Linear and Sparse Attention for Highly-Efficient, Unlimited-Input Vision-Language Models View paper
- [24] MoirÃ©XNet: Adaptive Multi-Scale DemoirÃ©ing with Linear Attention Test-Time Training and Truncated Flow Matching Prior View paper
- [25] Thermal Video Enhancement Mamba: A Novel Approach to Thermal Video Enhancement for Real-World Applications View paper
- [26] Ssm Meets Video Diffusion Models: Efficient Long-Term Video Generation with Selective State Spaces View paper
- [27] MobileI2V: Fast and High-Resolution Image-to-Video on Mobile Devices View paper
- [28] PERCEIVER-VL: Efficient Vision-and-Language Modeling with Iterative Latent Attention View paper
- [29] Scaling Diffusion Mamba with Bidirectional SSMs for Efficient Image and Video Generation View paper
- [30] VMG: Rethinking U-Net Architecture for Video Super-Resolution View paper
- [31] Space-Time Video Super-Resolution With Neural Operator View paper
- [32] Slow-Fast Architecture for Video Multi-Modal Large Language Models View paper
- [33] LVC-LGMC: Joint Local and Global Motion Compensation for Learned Video Compression View paper
- [34] PoM: Efficient Image and Video Generation with the Polynomial Mixer View paper
- [35] Radial Attention: Sparse Attention for Long Video Generation View paper
- [36] A Survey of Efficient Attention Methods: Hardware-efficient, Sparse, Compact, and Linear Attention View paper
- [37] LongLive: Real-time Interactive Long Video Generation View paper
- [38] GoldFinch: High Performance RWKV/Transformer Hybrid with Linear Pre-Fill and Extreme KV-Cache Compression View paper

- [39] SneakPeek: Future-Guided Instructional Streaming Video Generation View paper
- [40] LoRATv2: Enabling Low-Cost Temporal Modeling in One-Stream Trackers View paper
- [41] InfVSR: Breaking Length Limits of Generic Video Super-Resolution View paper
- [42] VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models View paper
- [43] Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets View paper
- [44] Waver: Wave your way to lifelike video generation View paper
- [45] Vivid-VR: Distilling Concepts from Text-to-Video Diffusion Transformer for Photorealistic Video Restoration View paper
- [46] Mora: Enabling Generalist Video Generation via A Multi-Agent Framework View paper
- [47] EffiVED: Efficient Video Editing via Text-instruction Diffusion Models View paper
- [48] AMD-Hummingbird: Towards an Efficient Text-to-Video Model View paper
- [49] Kandinsky 5.0: A Family of Foundation Models for Image and Video Generation View paper
- [50] DomainDiff: Unified Two-Stage Optimization for Text-Video Retrieval View paper
- [51] ID-Crafter: VLM-Grounded Online RL for Compositional Multi-Subject Video Generation View paper
- [52] Photorealistic video generation with diffusion models View paper
- [53] SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers View paper
- [54] ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features View paper
- [55] Efficient diffusion transformer with step-wise dynamic attention mediators View paper