Novelty Assessment Report

Paper: VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling
PDF URL: https://openreview.net/pdf?id=MUjdNcfNPv
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

Long-context video modeling is critical for multimodal large language models (MLLMs), enabling them to process movies, online video streams, and so on. Despite its advances, handling long videos remains challenging due to the difficulty in efficiently understanding the extremely long video context. This paper aims to address this issue from aspects of the model architecture, training data, training strategy, and evaluation benchmark. First, we propose a novel Hierarchical video token Compression (HiCo) method, which leverages visual redundancy in long videos to compress long video context from Clip-level to Video-level, reducing the computation significantly while preserving essential details, achieving an extreme compression ratio of approximately 1/50 with almost no performance loss. Second, we introduce a multi-stage short-to-long learning scheme, a large-scale dataset of real-world long videos named LongVid, and a challenging "Multi-Hop Needle-In- A-Video-Haystack" benchmark. Finally, we build a powerful video MLLM named VideoChat-Flash, which shows a leading performance on both mainstream long and short video benchmarks at the 2B and 7B model scales. It first gets 99.1% accuracy over 10,000 frames in NIAH among open-source models.

Core Task Landscape

This paper addresses: **Long-Context Video Modeling in Multimodal Large Language Models**
A total of **50 papers** were analyzed and organized into a taxonomy with **28 categories**.

Taxonomy Overview

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

- **Token Compression and Efficiency Mechanisms**
- **Temporal Reasoning and Grounding**
- **Long-Context Extension and Scalability**
- **Multimodal Integration and Cross-Modal Reasoning**
- **Training Data and Learning Strategies**
- **Benchmarking and Evaluation**
- **Foundation Models and Pretraining**
- **Domain-Specific Applications**
- **Architectural Enhancements and Encoder Integration**
- **Surveys and Meta-Analysis**

Complete Taxonomy Tree

- Long-Context Video Modeling in Multimodal Large Language Models Survey Taxonomy
- Token Compression and Efficiency Mechanisms
  - Hierarchical and Adaptive Compression ★ (4 papers)
  - [0] VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling (Anon et al., 2026) View paper
  - [1] Longvlm: Efficient long video understanding via large language models (Weng, 2024) View paper
  - [23] Longvu: Spatiotemporal adaptive compression for long video-language understanding (Shen, 2024) View paper
  - [30] Video-xl: Extra-long vision language model for hour-scale video understanding (Yan Shu, 2025) View paper
  - Spatiotemporal Token Reduction (3 papers)
  - [2] Token-Efficient Long Video Understanding for Multimodal LLMs (Jiang Jindong, 2025) View paper
  - [7] VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs (Cheng, 2024) View paper
  - Streaming and Online Processing (2 papers)
  - [4] Streaming long video understanding with large language models (Qian Rui, 2024) View paper
  - [35] Ma-lmm: Memory-augmented large multimodal model for long-term video understanding (Bo HE, 2024) View paper
  - Slow-Fast and Multi-Resolution Architectures (1 papers)
  - [3] Slow-fast architecture for video multi-modal large language models (Shi Min, 2025) View paper
- Temporal Reasoning and Grounding
  - Temporal Localization and Grounding (4 papers)
  - [25] Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability (Chen Shi-min, 2024) View paper
  - [27] Momentor: Advancing video large language model with fine-grained temporal reasoning (Qian Long, 2024) View paper
  - [43] TimeLens: Rethinking Video Temporal Grounding with Multimodal LLMs (Jun Zhang, 2025) View paper
  - [44] Enrich and Detect: Video Temporal Grounding with Multimodal LLMs (Pramanick, 2025) View paper
  - Frame-Aware and Chain-of-Thought Reasoning (2 papers)
  - [12] VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding (Wang Shihao, 2025) View paper

    - [15] Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs via Frame-Aware Reasoning (Ghazanfari, 2025) View paper
- Long-Context Extension and Scalability
    - System-Level Parallelism and Training (1 papers)
    - [20] Longvila: Scaling long-context visual language models for long videos (Chen, 2024) View paper
    - Positional Encoding and Context Extension (1 papers)
    - [34] V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding (Ge, 2025) View paper
    - Infinite and Extreme-Length Video Understanding (1 papers)
    - [22] Infinite Video Understanding (Zhang, 2025) View paper
- Multimodal Integration and Cross-Modal Reasoning
    - Audio-Visual-Speech Integration (1 papers)
    - [21] Watch and Listen: Understanding Audio-Visual-Speech Moments with Multimodal LLM (Li, 2025) View paper
    - LLM-Based Reasoning and Knowledge Integration (3 papers)
    - [5] Understanding long videos with multimodal language models (Ranasinghe, 2024) View paper
    - [19] Understanding long videos in one multimodal language model pass (Kanchana Ranasinghe, 2024) View paper
    - [33] Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms (Liao, 2024) View paper
    - Agent-Based and Tool-Augmented Systems (1 papers)
    - [11] VideoAgent: Long-Form Video Understanding with Large Language Model as Agent (Wang Xiao-han, 2024) View paper
- Training Data and Learning Strategies
    - Multi-Stage and Progressive Training (2 papers)
    - [24] Internvideo2. 5: Empowering video mllms with long and rich context modeling (Wang Yi, 2025) View paper
    - [46] Kangaroo: A powerful video-language model supporting long-context video input (Liu Jiajun, 2024) View paper
    - Data Curation and Annotation (1 papers)
    - [40] MM-Ego: Towards Building Egocentric Multimodal LLMs (Ye, 2024) View paper
    - Transfer Learning from Image to Video (2 papers)
    - [37] From image to video, what do we need in multimodal llms? (Huang Su-yuan, 2024) View paper
    - [41] TC-LLaVA: Rethinking the Transfer from Image to Video Understanding with Temporal Considerations (Gao Mingze, 2024) View paper
- Benchmarking and Evaluation
    - Long-Context and Interleaved Video Benchmarks (2 papers)
    - [16] Longvideobench: A benchmark for long-context interleaved video-language understanding (Wu, 2024) View paper
    - [38] Hourvideo: 1-hour video-language understanding (Chandrasegaran, 2024) View paper
    - Fine-Grained Temporal Understanding Benchmarks (3 papers)
    - [14] Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models (Cai, 2024) View paper
    - [17] Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos (Chiara Plizzari, 2025) View paper
    - [18] Temporalbench: Towards fine-grained temporal understanding for multimodal video models (M Cai, 2024) View paper
    - Comprehensive Multimodal Video Benchmarks (2 papers)
    - [6] Mvbench: A comprehensive multi-modal video understanding benchmark (Kunchang Li, 2024) View paper
    - [8] Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis (Chaoyou Fu, 2025) View paper
    - Spatial-Temporal Object and Action Understanding (2 papers)
    - [13] Videorefer suite: Advancing spatial-temporal object understanding with video llm (Yuqian Yuan, 2025) View paper
    - [36] Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models (Himangi Mittal, 2024) View paper
- Foundation Models and Pretraining
    - Video Foundation Models (1 papers)
    - [9] Internvideo2: Scaling foundation models for multimodal video understanding (Yi Wang, 2024) View paper
    - Interleaved Visual-Textual Processing (1 papers)
    - [10] Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens (Shen, 2024) View paper
- Domain-Specific Applications
    - Egocentric Video Understanding (1 papers)
    - [29] Vinci: A real-time embodied smart assistant based on egocentric vision-language model (Huang Yi-fei, 2024) View paper
    - Video Summarization and Generation (2 papers)
    - [28] V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning (Hang Hua, 2025) View paper
    - [45] Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation (Zhanyu Wang, 2024) View paper
    - Embodied AI and Robotics (3 papers)
    - [31] A survey on vision-language-action models for embodied ai (Ma, 2024) View paper
    - [42] Online Reasoning Video Segmentation with Just-in-Time Digital Twins (Shen Yiqing, 2025) View paper
    - [48] Robovqa: Multimodal long-horizon reasoning for robotics (Pierre Sermanet, 2024) View paper
    - Specialized Domain Benchmarks (1 papers)
    - [32] SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models (XIA Haotian, 2024) View paper
- Architectural Enhancements and Encoder Integration
    - Dual-Encoder Architectures (1 papers)
    - [50] VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding (Maaz, 2024) View paper
- Surveys and Meta-Analysis
    - Token Compression Surveys (1 papers)
    - [47] When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios (Kele Shao, 2025) View paper

- General Video Understanding Surveys (2 papers)
- [26] Do language models understand time? (Xi Ding, 2025) View paper
- [49] A Survey on Visual Understanding Multimodal Large Language Models (Liu, 2025) View paper

Narrative

Core task: long-context video modeling in multimodal large language models. The field addresses how to enable large language models to process and reason over extended video sequences, often spanning minutes or hours, by integrating visual encoders with language backbones. The taxonomy reveals several interrelated branches: Token Compression and Efficiency Mechanisms focus on reducing the computational burden of dense frame representations through hierarchical or adaptive strategies (e.g., VideoChat-Flash[0], LongVU[23]); Temporal Reasoning and Grounding emphasize understanding event sequences and temporal relationships (e.g., TemporalBench[14], TimeMarker[25]); Long-Context Extension and Scalability explore architectural modifications to handle longer inputs (e.g., LongViLA[20], Infinite Video[22]); while Multimodal Integration and Cross-Modal Reasoning tackle the fusion of vision, language, and sometimes audio (e.g., Watch and Listen[21]). Additional branches cover training strategies, benchmarking efforts like Video-MME[8] and LongVideoBench[16], foundation model pretraining (e.g., Internvideo[9]), domain-specific applications such as robotics (RoboVQA[48]) and egocentric video (MM-Ego[40]), and architectural enhancements that refine encoder designs.

A particularly active line of work centers on hierarchical and adaptive compression, where methods dynamically allocate tokens based on content importance or temporal structure. VideoChat-Flash[0] exemplifies this approach by employing adaptive compression to balance efficiency and detail retention, closely aligning with LongVU[23] and Video-XL[30], which similarly pursue token-efficient representations for long videos. In contrast, works like Slow-Fast Architecture[3] and Token-Efficient Long Video[2] explore dual-rate processing or explicit token budgeting to manage computational costs. Another contrasting theme emerges in temporal grounding: some studies prioritize fine-grained event localization (TemporalBench Fine-Grained[18], VideoRefer Suite[13]), while others focus on holistic narrative understanding across extended timelines (Understanding Long Videos[5], HourVideo[38]). VideoChat-Flash[0] sits within the compression-focused cluster, sharing design principles with LongVU[23] and Video-XL[30], yet its hierarchical strategy distinguishes it from simpler uniform sampling or fixed-rate approaches, positioning it as a bridge between efficiency-driven and reasoning-oriented paradigms.

Related Works in Same Category

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

1. Longvlm: Efficient long video understanding via large language models

Authors: Weng, Yuetian, Han, Mingfei, Yuetian Weng, et al. (15 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Empowered by Large Language Models (LLMs), recent advancements in Video-based LLMs (VideoLLMs) have driven progress in various video understanding tasks. These models encode video representations through pooling or query aggregation over a vast number of visual tokens, making computational and memory costs affordable. Despite successfully providing an overall comprehension of video content, existing VideoLLMs still face challenges in achieving detailed understanding due to overlooking local info...

Relationship Analysis

Both papers belong to the Hierarchical and Adaptive Compression category, employing multi-level compression strategies to handle long-context videos efficiently. They overlap in their use of hierarchical token reduction mechanisms—VideoChat-Flash uses clip-level and video-level compression stages, while LongVLM decomposes videos into short-term segments with hierarchical token merging. The key difference is that VideoChat-Flash focuses on progressive visual dropout in LLM layers and achieves ~1/50 compression ratio, whereas LongVLM emphasizes integrating global semantics into local segment features and uses bipartite soft matching for token merging.

---

2. Longvu: Spatiotemporal adaptive compression for long video-language understanding

Authors: Shen, Xiaoqian, Xiong Yunyang, Xiaoqian Shen, Zhao Changsheng, et al. (42 authors total) | **Year/Venue**: 2024 | **URL**: View paper

Abstract

Multimodal Large Language Models (MLLMs) have shown promising progress in understanding and analyzing video content. However, processing long videos remains a significant challenge constrained by LLM's context size. To address this limitation, we propose LongVU, a spatiotemporal adaptive compression mechanism thats reduces the number of video tokens while preserving visual details of long videos. Our idea is based on leveraging cross-modal query and inter-frame dependencies to adaptively reduce ...

Relationship Analysis

Both papers belong to the Hierarchical and Adaptive Compression category, employing multi-level compression strategies to handle long videos in MLLMs. They share overlapping approaches in using hierarchical compression mechanisms—VideoChat-Flash uses clip-level and video-level compression stages, while LongVU employs temporal reduction, cross-modal query-based selection, and spatial token compression. The key differences are that VideoChat-Flash focuses on a two-stage hierarchical framework (HiCo) with progressive visual dropout in LLM layers and introduces a multi-hop NIAH benchmark, whereas LongVU emphasizes spatiotemporal adaptive compression leveraging DINOv2 features for temporal reduction and text-guided cross-modal queries for selective frame preservation.

---

3. Video-xl: Extra-long vision language model for hour-scale video understanding

Authors: Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Long video understanding poses a significant challenge for current Multi-modal Large Language Models (MLLMs). Notably, the MLLMs are constrained by their limited context lengths and the substantial costs while processing long videos. Although several existing methods attempt to reduce visual tokens, their strategies encounter severe bottleneck, restricting MLLMsâ□□ ability to perceive fine-grained visual details. In this work, we propose Video-XL, a novel approach that leverages MLLMsâ□□ inheren...

Relationship Analysis

Both papers belong to the Hierarchical and Adaptive Compression category, employing multi-level compression strategies to handle long-context video modeling in MLLMs. They overlap in their use of hierarchical compression mechanisms—VideoChat-Flash uses clip-level and video-level compression stages, while Video-XL employs interval-based compression with Visual Summarization Tokens (VSTs). The key difference is that VideoChat-Flash focuses on leveraging visual redundancy through spatio-temporal attention and token merging at the clip level followed by progressive dropout in the LLM, whereas Video-XL leverages the LLM's inherent KV sparsification capability with dynamic interval partitioning based on semantic consistency and depth scores.

Contributions Analysis

Overall novelty summary. The paper proposes a hierarchical video token compression method (HiCo) achieving approximately 1/50 compression ratio, alongside a multi-stage short-to-long training scheme, the LongVid dataset, and a Multi-Hop Needle-In-A-Video-Haystack benchmark. It resides in the 'Hierarchical and Adaptive Compression' leaf under 'Token Compression and Efficiency Mechanisms', sharing this space with three sibling papers (LongVU, Video-XL, and one other). This leaf represents a moderately populated research direction within a broader taxonomy of 50 papers across approximately 36 topics, indicating active but not overcrowded exploration of adaptive compression strategies for long-context video understanding.

The taxonomy reveals neighboring leaves focused on spatiotemporal token reduction via dedicated temporal encoders, streaming architectures with constant token budgets, and slow-fast dual-pathway designs. These adjacent directions emphasize different trade-offs: spatiotemporal methods prioritize inter-frame dependency modeling, while streaming approaches target online processing constraints. The paper's hierarchical compression strategy bridges efficiency-driven token reduction and reasoning-oriented temporal understanding, contrasting with fixed-rule compression or single-pathway methods. Its position suggests engagement with both compression efficiency and preservation of semantic detail across extended video timelines, distinguishing it from purely architectural or temporal grounding approaches in sibling branches.

Among 30 candidates examined, the HiCo compression method shows no clear refutation across 10 candidates, suggesting relative novelty in its specific hierarchical design. The Multi-Hop Needle benchmark similarly appears unrefuted across 10 candidates, indicating potential originality in evaluation methodology. However, the LongVid dataset and short-to-long learning strategy encountered one refutable candidate among 10 examined, pointing to existing work in progressive training or large-scale long-video data curation. The limited search scope means these findings reflect top-30 semantic matches rather than exhaustive coverage, leaving open the possibility of additional relevant prior work in less semantically similar papers.

Given the moderate density of the hierarchical compression leaf and the mixed contribution-level findings, the work appears to offer incremental advances in compression architecture and benchmarking while building on established paradigms in multi-stage training and dataset construction. The analysis is constrained by the top-30 candidate scope and does not capture potential overlaps in broader compression literature or domain-specific long-video datasets outside the semantic search radius.

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

Contribution 1: Hierarchical video token Compression (HiCo) method

Description: A two-stage compression approach that first reduces inter-frame redundancy at the clip level using spatio-temporal attention and token merging, then performs video-level compression by discarding task-irrelevant tokens during LLM processing. This achieves approximately 1/50 compression ratio with minimal performance loss.

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. HoliTom: Holistic Token Merging for Fast Video Large Language Models

URL: View paper

Brief Assessment

HoliTom[62] focuses on token merging strategies (outer-llm and inner-llm pruning) rather than the specific two-stage hierarchical compression paradigm (clip-level spatio-temporal attention + video-level task-irrelevant token discarding) proposed in the original paper. The candidate does not demonstrate prior work using this exact hierarchical compression approach with spatio-temporal attention at clip-level followed by LLM-based video-level compression.

---

2. Efficient Video Transformers via Spatial-temporal Token Merging for Action Recognition

URL: View paper

Brief Assessment

Spatial-Temporal Token Merging[67] focuses on merging tokens within video transformers for action recognition tasks, not on hierarchical compression for long-context video modeling in multimodal LLMs. The candidate addresses efficiency in video transformers, while the original targets long-video understanding in MLLMs with a two-stage compression paradigm (clip-level and video-level).

---

3. Framefusion: Combining similarity and importance for video token reduction on large vision language models

URL: View paper

Brief Assessment

FrameFusion[61] focuses on similarity-based token merging combined with importance-based pruning within LVLMs, rather than the two-stage hierarchical compression paradigm (clip-level spatio-temporal attention + video-level task-irrelevant token discarding) proposed in the original paper.

---

4. Progressive Growing of Video Tokenizers for Temporally Compact Latent Spaces

URL: View paper

Brief Assessment

Progressive Video Tokenizers[64] focuses on temporal compression in video tokenizers for latent diffusion models, using progressive growing techniques. The original paper's HiCo addresses video token compression for multimodal LLMs with clip-level and video-level stages, which is a different application domain and architectural approach.

---

5. B-vllm: A vision large language model with balanced spatio-temporal tokens

URL: View paper

Brief Assessment

B-VLLM[59] focuses on balancing spatio-temporal tokens through adaptive frame selection and spatial token sampling, rather than the two-stage hierarchical compression (clip-level and video-level) proposed in the original paper.

---

6. RESTHT: relation-enhanced spatialâ□□temporal hierarchical transformer for video captioning

URL: View paper

Brief Assessment

RESTHT[63] focuses on relation-enhanced spatial-temporal modeling for video captioning, not on hierarchical token compression for long-context video understanding. The candidate's approach addresses different technical challenges (relation modeling for captioning) compared to the original's compression-based efficiency framework.

---

7. Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs

URL: View paper

Brief Assessment

Multi-Granular Token Merging[66] focuses on training-free token reduction for video LLMs using spatial and temporal merging, while HiCo is a trainable two-stage compression method integrated into model architecture with specific clip-level and video-level designs.

---

8. STPM: Spatial-Temporal Token Pruning and Merging for Complex Activity Recognition

URL: View paper

Brief Assessment

STPM[68] focuses on token pruning/merging for complex activity recognition in videos where humans occupy small spatial regions, not on hierarchical compression for long-context video modeling in MLLMs.

---

9. The devil is in temporal token: High quality video reasoning segmentation

URL: View paper

Brief Assessment

Temporal Token Devil[60] focuses on video reasoning segmentation using hierarchical tokens for spatial-temporal feature encoding in segmentation tasks, not general video token compression for long-context video modeling in MLLMs.

---

10. Midframe-centric token merging for efficient video transformer

URL: View paper

Brief Assessment

Midframe Token Merging[65] focuses on token merging within video transformers for efficiency, not on hierarchical compression across clip-level and video-level stages as proposed in the original paper's HiCo method.

Contribution 2: LongVid dataset and short-to-long learning strategy

Description: A large-scale training corpus containing 114,228 long videos with 3,444,849 question-answering pairs covering five task types, combined with a multi-stage training strategy that progresses from image and short video data to joint short and long video instruction tuning.

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. Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection

URL: View paper

Brief Assessment

Unified Video Comprehension[54] focuses on moment retrieval and highlight detection tasks using multi-granularity feature integration, not on constructing large-scale long video datasets or multi-stage short-to-long training strategies for general video understanding.

---

2. Physformer: Facial video-based physiological measurement with temporal difference transformer

URL: View paper

Brief Assessment

PhysFormer[52] addresses facial video-based physiological measurement using temporal transformers, not long video understanding or multi-stage training strategies for video question-answering tasks.

---

3. Video-xl: Extra-long vision language model for hour-scale video understanding

URL: View paper

Brief Assessment

Video-XL[30] focuses on a different compression mechanism (VST tokens with KV sparsification) rather than multi-stage short-to-long learning strategies. While it uses curriculum learning for compression ratios and composite data curation, these are distinct from the original paper's progressive training from image/short video to joint short/long video instruction tuning with the LongVid dataset.

---

4. Adaptive curriculum learning for video captioning

URL: View paper

Brief Assessment

Adaptive Curriculum Learning[55] focuses on video captioning with curriculum learning based on caption difficulty, not on multi-stage short-to-long video understanding or long video QA datasets.

---

5. Spatialladder: Progressive training for spatial reasoning in vision-language models

URL: View paper

Brief Assessment

SpatialLadder[51] focuses on spatial reasoning across single-image, multi-view, and video modalities with a three-stage progressive training framework, not on general long video understanding with short-to-long learning strategies.

---

6. Kangaroo: A powerful video-language model supporting long-context video input

URL: View paper

Prior Art Analysis

Kangaroo[46] demonstrates prior work on multi-stage short-to-long learning strategies for long video understanding datasets. The candidate paper presents a curriculum training pipeline that progressively increases resolution and frame count, combined with a large-scale video dataset construction process. Both papers employ similar multi-stage training approaches: starting with image/short video data, then progressing to longer videos through curriculum learning. The candidate's data curation system and curriculum training strategy (stages I-IV) closely parallel the original paper's short-to-long learning scheme and LongVid dataset construction.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly describe multi-stage training strategies that progressively adapt models from short to long video understanding, demonstrating prior work on this approach. - **Original**: we introduce a multi-stageshort-to-long learningscheme, a large-scale dataset of real-world long videos namedlongvid, and a challenging"multi-hop needle-ina-video-haystack"benchmark. - **Candidate**: we design a curriculum training pipeline with gradually increasing resolution and number of input frames to accommodate long videos.

Evidence 2 - **Rationale**: Both papers describe nearly identical initial training stages that begin with image-text alignment followed by short video pre-training, establishing prior work on this multi-stage approach. - **Original**: stage-1: video-language alignment.in this stage, we freeze the visual encoder and the large language model while training the compressor and the mlp to align the language with the compressed visual features. we use 0.5 million image-text pairs and 0.5 million short video-text pairs, and sample 4 fra... - **Candidate**: stage i: visual-language pre-training. we start with image/video pre-training to connect fundamental language concepts and visual elements. first, we utilize the re-captioned image dataset for image pre-training. each image is represented as a single-frame video and is randomly assigned a timestamp ...

Evidence 3 - **Rationale**: Both papers describe constructing large-scale long video datasets by aggregating multiple existing video sources, showing prior work on this data curation approach. - **Original**: to construct the longvid dataset, we follow three core steps: (1)first, for data source selection, we leverage diverse existing long video datasets that include ego4d (grauman et al., 2022), howto100m (miech et al., 2019), hd-vila (xue et al., 2022), and miradata (ju et al., 2024); these datasets col... - **Candidate**: to build a high-quality dataset for video pre-training, we gather videos from webvid [43], panda-70m [44] with english captions and youku-mplug [45], chinaopen [46] with chinese captions. in addition, a number of internal videos are incorporated to enrich the diversity of video data.

---

7. Kwai keye-vl 1.5 technical report

URL: View paper

Brief Assessment

Kwai Keye-VL[56] focuses on a different multi-stage training approach with slowfast video encoding and progressive context extension (8k to 128k tokens), rather than the specific LongVid dataset construction and short-to-long learning strategy described in the original paper.

---

8. Clearvid: Curriculum learning for video description

URL: View paper

Brief Assessment

ClearVid[58] focuses on curriculum learning for video description generation using noise and dropout strategies, not on constructing large-scale long video datasets or multi-stage short-to-long training for long video understanding tasks.

---

9. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models

URL: View paper

Brief Assessment

Grounded-VideoLLM[53] focuses on fine-grained temporal grounding with a different multi-stage training approach (image→video→grounding tasks), not the specific short-to-long video progression described in the original paper's contribution.

---

10. Efficient VideoMAE via Temporal Progressive Training

URL: View paper

Brief Assessment

Efficient VideoMAE[57] focuses on temporal progressive training for masked autoencoder pre-training in video recognition, not on building long-video instruction-tuning datasets or multimodal large language models for long video understanding.

Contribution 3: Multi-Hop Needle-In-A-Video-Haystack benchmark

Description: A new evaluation benchmark that requires models to follow a reasoning path through multiple images inserted into long videos, with wrong paths as distractors. This tests both retrieval and complex reasoning abilities more robustly than previous single-hop approaches.

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. Hopper: Multi-hop transformer for spatiotemporal reasoning

URL: View paper

Brief Assessment

Hopper[74] focuses on object permanence reasoning in videos using multi-hop transformers for spatiotemporal reasoning, not on multi-hop needle-in-haystack benchmarks for video retrieval and reasoning as described in the original paper's contribution.

---

2. Ddog: optimizing multi-hop inference via dual-driven retrieval and reasoning path

URL: View paper

Brief Assessment

DDOG[75] focuses on multi-hop reasoning paths for retrieval and reasoning in text/knowledge graph contexts, not video-based needle-in-haystack evaluation benchmarks. The candidate addresses different modalities and problem domains.

---

3. Omchat: A recipe to train multimodal language models with strong long context and video understanding

URL: View paper

Brief Assessment

OMChat[78] proposes a different benchmark called 'Temporal Visual Needle in a Haystack' that focuses on identifying emoji sequences in videos, whereas the original paper's multi-hop benchmark requires following reasoning paths through multiple images with distractors. These are distinct evaluation approaches for different aspects of long video understanding.

---

4. VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos

URL: View paper

Brief Assessment

VRBench[76] focuses on multi-step reasoning in long narrative videos with stepwise annotations and process-level evaluation, not on needle-in-haystack retrieval tasks with distractor paths as described in the original paper's multi-hop NIAH benchmark.

---

5. Video-of-thought: Step-by-step video reasoning from perception to cognition

URL: View paper

Brief Assessment

Video-of-Thought[69] focuses on step-by-step video reasoning using spatial-temporal scene graphs for complex video QA tasks, not on multi-hop needle-in-haystack evaluation benchmarks for long-context video retrieval.

---

6. STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training

URL: View paper

Brief Assessment

STEP[77] focuses on compositional reasoning through spatio-temporal scene graphs and self-training, not on multi-hop needle-in-haystack evaluation benchmarks for long video retrieval.

---

7. SceneRAG: Scene-level Retrieval-Augmented Generation for Video Understanding

URL: View paper

Brief Assessment

SceneRAG[71] focuses on scene-level segmentation and retrieval-augmented generation for video understanding, not on multi-hop needle-in-haystack evaluation benchmarks. The candidate does not present a comparable benchmark for testing multi-hop reasoning with inserted needles and distractors.

---

8. Overview of the NLPCC 2025 shared task 4: multi-modal, multilingual, and multi-hop medical instructional video question answering challenge

URL: View paper

Brief Assessment

NLPCC Medical VideoQA[73] focuses on medical instructional video question answering with multilingual support and knowledge graphs, not on general video understanding with needle-in-haystack evaluation methodology.

---

9. Grounded multi-hop videoqa in long-form egocentric videos

URL: View paper

Brief Assessment

Grounded Multi-Hop VideoQA[70] focuses on multi-hop question answering in egocentric videos with temporal grounding of scattered evidence, not on needle-in-haystack evaluation benchmarks for testing retrieval across inserted images in long videos.

---

10. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding

URL: View paper

Brief Assessment

Logic-in-Frames[72] focuses on keyframe selection via visual semantic-logical search for video understanding, not on multi-hop reasoning benchmarks. The candidate does not present a comparable multi-hop needle-in-haystack evaluation methodology.

Appendix: Text Similarity Detection

Textual similarity detection checked 32 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. Longvlm: Efficient long video understanding via large language models

Detected in: Core Task (sibling)

⚠ **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] VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling View paper
- [1] Longvlm: Efficient long video understanding via large language models View paper
- [2] Token-Efficient Long Video Understanding for Multimodal LLMs View paper
- [3] Slow-fast architecture for video multi-modal large language models View paper
- [4] Streaming long video understanding with large language models View paper
- [5] Understanding long videos with multimodal language models View paper
- [6] Mvbench: A comprehensive multi-modal video understanding benchmark View paper
- [7] VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs View paper
- [8] Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis View paper
- [9] Internvideo2: Scaling foundation models for multimodal video understanding View paper
- [10] Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens View paper
- [11] VideoAgent: Long-Form Video Understanding with Large Language Model as Agent View paper
- [12] VideoITG: Multimodal Video Understanding with Instructed Temporal Grounding View paper
- [13] Videorefer suite: Advancing spatial-temporal object understanding with video llm View paper
- [14] Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models View paper
- [15] Chain-of-Frames: Advancing Video Understanding in Multimodal LLMs via Frame-Aware Reasoning View paper
- [16] Longvideobench: A benchmark for long-context interleaved video-language understanding View paper
- [17] Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos View paper
- [18] Temporalbench: Towards fine-grained temporal understanding for multimodal video models View paper
- [19] Understanding long videos in one multimodal language model pass View paper
- [20] Longvila: Scaling long-context visual language models for long videos View paper
- [21] Watch and Listen: Understanding Audio-Visual-Speech Moments with Multimodal LLM View paper
- [22] Infinite Video Understanding View paper
- [23] Longvu: Spatiotemporal adaptive compression for long video-language understanding View paper
- [24] Internvideo2. 5: Empowering video mllms with long and rich context modeling View paper
- [25] Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability View paper
- [26] Do language models understand time? View paper
- [27] Momentor: Advancing video large language model with fine-grained temporal reasoning View paper
- [28] V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning View paper

- [29] Vinci: A real-time embodied smart assistant based on egocentric vision-language model View paper
- [30] Video-xl: Extra-long vision language model for hour-scale video understanding View paper
- [31] A survey on vision-language-action models for embodied ai View paper
- [32] SPORTU: A Comprehensive Sports Understanding Benchmark for Multimodal Large Language Models View paper
- [33] Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms View paper
- [34] V2pe: Improving multimodal long-context capability of vision-language models with variable visual position encoding View paper
- [35] Ma-lmm: Memory-augmented large multimodal model for long-term video understanding View paper
- [36] Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models View paper
- [37] From image to video, what do we need in multimodal llms? View paper
- [38] Hourvideo: 1-hour video-language understanding View paper
- [39] STORM: Token-Efficient Long Video Understanding for Multimodal LLMs View paper
- [40] MM-Ego: Towards Building Egocentric Multimodal LLMs View paper
- [41] TC-LLaVA: Rethinking the Transfer from Image to Video Understanding with Temporal Considerations View paper
- [42] Online Reasoning Video Segmentation with Just-in-Time Digital Twins View paper
- [43] TimeLens: Rethinking Video Temporal Grounding with Multimodal LLMs View paper
- [44] Enrich and Detect: Video Temporal Grounding with Multimodal LLMs View paper
- [45] Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation View paper
- [46] Kangaroo: A powerful video-language model supporting long-context video input View paper
- [47] When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios View paper
- [48] Robovqa: Multimodal long-horizon reasoning for robotics View paper
- [49] A Survey on Visual Understanding Multimodal Large Language Models View paper
- [50] VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding View paper
- [51] Spatialladder: Progressive training for spatial reasoning in vision-language models View paper
- [52] Physformer: Facial video-based physiological measurement with temporal difference transformer View paper
- [53] Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models View paper
- [54] Bridging the gap: A unified video comprehension framework for moment retrieval and highlight detection View paper
- [55] Adaptive curriculum learning for video captioning View paper
- [56] Kwai keye-vl 1.5 technical report View paper
- [57] Efficient VideoMAE via Temporal Progressive Training View paper
- [58] Clearvid: Curriculum learning for video description View paper
- [59] B-vllm: A vision large language model with balanced spatio-temporal tokens View paper
- [60] The devil is in temporal token: High quality video reasoning segmentation View paper
- [61] Framefusion: Combining similarity and importance for video token reduction on large vision language models View paper
- [62] HoliTom: Holistic Token Merging for Fast Video Large Language Models View paper
- [63] RESTHT: relation-enhanced spatialâ□□temporal hierarchical transformer for video captioning View paper
- [64] Progressive Growing of Video Tokenizers for Temporally Compact Latent Spaces View paper
- [65] Midframe-centric token merging for efficient video transformer View paper
- [66] Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs View paper
- [67] Efficient Video Transformers via Spatial-temporal Token Merging for Action Recognition View paper
- [68] STPM: Spatial-Temporal Token Pruning and Merging for Complex Activity Recognition View paper
- [69] Video-of-thought: Step-by-step video reasoning from perception to cognition View paper
- [70] Grounded multi-hop videoqa in long-form egocentric videos View paper
- [71] SceneRAG: Scene-level Retrieval-Augmented Generation for Video Understanding View paper
- [72] Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding View paper
- [73] Overview of the NLPCC 2025 shared task 4: multi-modal, multilingual, and multi-hop medical instructional video question answering challenge View paper
- [74] Hopper: Multi-hop transformer for spatiotemporal reasoning View paper
- [75] Ddog: optimizing multi-hop inference via dual-driven retrieval and reasoning path View paper
- [76] VRBench: A Benchmark for Multi-Step Reasoning in Long Narrative Videos View paper
- [77] STEP: Enhancing Video-LLMs' Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training View paper
- [78] Omchat: A recipe to train multimodal language models with strong long context and video understanding View paper