Graph-Augmented Attention in Pyramidal Multimodal Memory for Long-Horizon Video Understanding

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the graph-augmented attention mechanism in pyramidal multimodal memory models compare to standard Vision-Language Models in terms of inference throughput (tokens per second) on long-horizon. While multimodal large language models have demonstrated impressive short-term reasoning, they struggle with long-horizon video understanding due to limited context windows and static memory mechanisms that fail to mirror human cognitive efficiency. Existing paradigms typically. 18 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: From Verbatim to Gist: Distilling Pyramidal Multimodal Memory via Semantic Information Bottleneck for Long-Horizon Video Agents. Research question: How does the graph-augmented attention mechanism in pyramidal multimodal memory models compare to standard Vision-Language Models in terms of inference throughput (tokens per second) on long-horizon video understanding benchmarks like Ego4D or ActivityNet?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

11 papers retrieved. 18 claims extracted; 1 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
MM-Mem uses Qwen3-VL-8B as the base model.	×	0.03
For text retrieval, bge-large-en-v1.5 and bge-reranker-v2-m3 are used.	×	0.02
For visual retrieval, clip-level retrieval by jointly scoring keyframes per clip is used.	×	0.02
Models are served with vLLM, and fine-tuning is performed under SWIFT with SIB-GRPO.	×	0.05
Hyperparameters include β1 = 0.1, β2 = 0.3, and temperature to 0.0.	×	0.00
MM-Mem consistently outperforms prior agent systems on Video-MME, MLVU, and HD-EPIC++ benchmarks.	×	0.06
MM-Mem yields a 5.1% relative gain on Video-MME and a 7.1% gain on MLVU in M-Avg compared to Vgent.	×	0.03
MM-Mem surpasses all compared open-source MLLMs and is competitive with proprietary models such as Gemini 1.5 Pro.	×	0.09
MM-Mem improves over Flash-VStream by 5.9% and 5.2% in terms of Accuracy and Score on VStream-QA-Ego.	×	0.02
MM-Mem achieves 30.28% accuracy on HD-EPIC++, outperforming all baselines.	×	0.04
MM-Mem exceeds Qwen3-VL-8B by +4.40 points on HD-EPIC++.	×	0.02
MM-Mem surpasses LLaVA-Video-7B on HD-EPIC++.	×	0.04
MM-Mem is a hierarchical pyramidal multimodal memory architecture inspired by Fuzzy-Trace Theory (FTT).	✓	0.20
MM-Mem emphasizes low-level visual details while preserving high-level semantic attributes.	×	0.09
MM-Mem converts raw videos into structured textual memories for efficiency.	×	0.03
MM-Mem is highly dynamic, unlike most existing static systems.	×	0.06
MM-Mem is inspired by the complementarity between visual and textual modalities and the distinction between verbatim and	×	0.09
MM-Mem is realized through cross-modal fusion rather than a rigid one-to-one layer mapping.	×	0.03

References

- http://arxiv.org/abs/2604.00086v1
- http://arxiv.org/abs/2603.01455v3
- http://arxiv.org/abs/2605.17065v1