Graph-Augmented Attention Trade-Offs in Multimodal Video Agents with Memory Distillation

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the trade-off between inference latency and reasoning accuracy when applying graph-augmented attention with different memory distillation ratios in multimodal video agents. 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. 15 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: What is the trade-off between inference latency and reasoning accuracy when applying graph-augmented attention with different memory distillation ratios in multimodal video agents?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

16 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 3.8/10.

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.	×	0.04
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.	×	0.03
Models are served with vLLM, and fine-tuning is performed under SWIFT with SIB-GRPO.	×	0.05
Hyperparameters are provided in Appendix D.	×	0.00
MM-Mem consistently outperforms prior agent systems.	×	0.12
MM-Mem yields a 5.1% relative gain on Video-MME (both w/o- and w/-subtitle) and a 7.1% gain on MLVU in M-Avg.	×	0.03
MM-Mem surpasses all compared open-source MLLMs (e.g., Qwen2-VL-72B) and is competitive with strong proprietary models s	×	0.03
MM-Mem improves over the previous best method Flash-VStream by 5.9% and 5.2% in terms of Accuracy and Score, respectivel	×	0.06
MM-Mem achieves 30.28% accuracy on HD-EPIC++.	×	0.06
MM-Mem exceeds the strongest competitor (Qwen3-VL-8B) by +4.40 points (30.28 vs. 25.88) on HD-EPIC++.	×	0.02
MM-Mem surpasses LLaVA-Video-7B on HD-EPIC++.	×	0.04
MM-Mem is a novel hierarchical pyramidal multimodal memory architecture inspired by the principles of FTT.	✓	0.16
MM-Mem is inspired by the complementarity between visual and textual modalities and the distinction between verbatim and	×	0.10
This connection is realized through cross-modal fusion rather than a rigid one-to-one layer mapping.	×	0.05

References

- http://arxiv.org/abs/2510.04514v2

- http://arxiv.org/abs/2603.01455v3
- http://arxiv.org/abs/2503.11495v1