SOVEREIGN: What is the inference latency overhead of VideoRAG's retrieval-augmented approach compared to vanilla video un

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large Language Models (LLMs) suffer from hallucinations and outdated knowledge due to their reliance on static training data. Retrieval-Augmented Generation (RAG) mitigates these issues by integrating external dynamic information for improved factual grounding. With advances in multimodal learning, Multimodal RAG extends this approach by incorporating multiple modalities such as text, images, audio, and video to enhance the generated outputs. However, cross-modal alignment and reasoning introduce unique challenges beyond those in unimodal RAG. This survey offers a structured and comprehensive

1 Introduction

Analysis of: Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation. Research goal: What is the inference latency overhead of VideoRAG's retrieval-augmented approach compared to vanilla video understanding models when processing extreme-length video contexts (>100K frames) on Adept-1B and similar multimodal benchmarks?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

4 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 6.0/10 → REVISE (revision_round=1). Policy: ESCALATE_TO_OWNER.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

References

- https://doi.org/10.48550/arxiv.2502.08826
- https://doi.org/10.18653/v1/2025.findings-acl.861
- https://doi.org/10.36227/techrxiv.176341513.38473003/v1