SOVEREIGN: What is the impact of ReKV's streaming window size on VideoQA performance when evaluated on the VideoQA benchm

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Long streaming video QA remains challenging due to growing visual tokens and limited reasoning length of large language models (LLMs). KV-caching stores the Key-Value (KV) of the historical tokens via LLM prefill and enables more efficient streaming QA. However, existing methods cache every one or two frames, causing redundant memory usage and losing fine-grained spatial details within frame or temporal contexts across frames. This paper proposes MuKV, a method that features a multi-grained KV cache compression module and a semi-hierarchical retrieval approach to improve both efficiency and ac

1 Introduction

Analysis of: MuKV: Multi-Grained KV Cache Compression for Long Streaming Video Question-Answering. Research goal: What is the impact of ReKV's streaming window size on VideoQA performance when evaluated on the VideoQA benchmark with varying temporal context lengths?.

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

9 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 3.3/10 $\rightarrow$ REJECT (revision_round=0). 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/10.1016/j.ijinfomgt.2020.102168
- https://doi.org/10.48550/arxiv.2510.17364
- https://openalex.org/W7162219301