# ShadowKV Method for Multimodal Retrieval Accuracy with Restricted KV Cache

Assignee Research

June 11, 2026

**Abstract**

With the widespread deployment of long-context large language models (LLMs), there has been a growing demand for efficient support of high-throughput inference. However, as the key-value (KV) cache expands with the sequence length, the increasing memory footprint and the need to access it for each token generation both result in low throughput when serving long-context LLMs. While various dynamic sparse attention methods have been proposed to speed up inference while maintaining generation quality, they either fail to sufficiently reduce GPU memory consumption or introduce significant decoding

## 1 Introduction

This paper examines: ShadowKV: KV Cache in Shadows for High-Throughput Long-Context LLM Inference. Research question: Does the ShadowKV method preserve multimodal retrieval accuracy on long-document benchmarks when the KV cache size is restricted to less than 20% of the full sequence length?.

## 2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.1/10.

## 3 Results

9 papers retrieved. 6 claims extracted; 5 independently verified. Quality review score: 8.1/10.

## 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 |
|---|---|---|
| ShadowKV is a high-throughput long-context LLM inference system that stores the low-rank key cache and offloads the valu | ✓ | 0.44 |
| ShadowKV employs an accurate KV selection strategy that reconstructs minimal sparse KV pairs on-the-fly to minimize deco | ✓ | 0.30 |
| ShadowKV was evaluated on benchmarks including RULER, LongBench, and Needle In A Haystack, and models like Llama-3.1-8B, | ✓ | 0.39 |
| ShadowKV can support up to 6$\times$ larger batch sizes and boost throughput by up to 3.04$\times$ on an A100 GPU without sacrificing | ✓ | 0.26 |
| ShadowKV surpasses the performance achievable with infinite batch size under the assumption of infinite GPU memory. | ✓ | 0.23 |
| The code for ShadowKV is available. | × | 0.05 |

## References


- https://openalex.org/W7162045077
- https://doi.org/10.48550/arxiv.2410.21465
- https://doi.org/10.18653/v1/2025.emnlp-main.1079