SOVEREIGN: What is the scaling efficiency in terms of memory usage and QA accuracy when applying value-based embedder tra

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

We propose ReKV, a novel training-free approach that enables efficient streaming video question-answering (StreamingVQA), by seamlessly integrating with existing Video Large Language Models (Video-LLMs). Traditional VideoQA systems struggle with long videos, as they must process entire videos before responding to queries, and repeat this process for each new question. In contrast, our approach analyzes long videos in a streaming manner, allowing for prompt responses as soon as user queries are received. Building on a common Video-LLM, we first incorporate a sliding-window attention mechanism,

1 Introduction

Analysis of: Streaming Video Question-Answering with In-context Video KV-Cache Retrieval. Research goal: What is the scaling efficiency in terms of memory usage and QA accuracy when applying value-based embedder training (as in Q-RAG) to a 70B model for multi-step retrieval, compared to fine-tuning a smaller retriever model on the same multi-hop QA task?.

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

13 papers retrieved. 18 claims extracted, 0 verified. Tribunal: 5.0/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
MLVUdev-mc is the multiple-choice subset of the MLVU-dev benchmark focused on evaluating long-form video understanding o	×	0.03
QAEGO4Dtest-mc is the multiple-choice subset of the QAEGO4D-test benchmark focusing on question-answering in long egocen	×	0.05
EgoSchema is a diagnostic benchmark for long VideoQA featuring over 5000 multiple-choice questions.	×	0.02
ActivityNet-QA encompasses human-annotated QA pairs on 5,800 videos derived from the ActivityNet dataset.	×	0.05
RVS-Ego and RVS-Movie are Streaming VideoQA benchmarks constructed using long videos from Ego4D and MovieNet datasets re	×	0.06
CGBenchmc is the multiple-choice subset of CG-Bench designed for clue-grounded question answering in long videos.	×	0.07
Uniform Sampling achieves the lowest recall and poorest VideoQA accuracy compared to other retrieval methods.	×	0.05
Oracle Retrieval delivers the highest VideoQA accuracy and has 100% recall.	×	0.02
External Retrieval and Internal Retrieval both surpass Uniform Sampling in VideoQA accuracy.	×	0.03
Experiments are conducted on NVIDIA A100 (80GB) GPUs with FP16 precision.	×	0.02
For video modeling, the video stream is processed at 0.5 FPS.	×	0.08
SigLIP-SO400M is used as the retriever for external video KV-Cache retrieval.	×	0.13
For internal KV-Cache retrieval, the block size is set to 1 and the number of retrieved frames is set to 64 by default.	×	0.07
LLaVA-OV-0.5B with Uniform Sampling achieves 42.6 VideoQA accuracy and 6.1 recall on QAEGO4Dtest-mc.	×	0.03
LLaVA-OV-0.5B with External Retrieval achieves 48.0 VideoQA accuracy and 58.1 recall on QAEGO4Dtest-mc.	×	0.03
LLaVA-OV-0.5B with Internal Retrieval achieves 50.0 VideoQA accuracy and 63.4 recall on QAEGO4Dtest-mc.	×	0.02
LLaVA-OV-0.5B with Oracle Retrieval achieves 52.0 VideoQA accuracy and 100 recall on QAEGO4Dtest-mc.	×	0.02
LLaVA-OV-7B with Uniform Sampling achieves 53.0 VideoQA accuracy and 6.1 recall on QAEGO4Dtest-mc.	×	0.03

References

- http://arxiv.org/abs/2404.14464v1
- https://arxiv.org/abs/2503.00540
- http://arxiv.org/abs/2511.07328v2