SOVEREIGN: Does iterative retrieval with visual frame reranking mitigate video content drift degradation more effectively

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Recent developments in modeling language and vision have been successfully applied to image question answering. It is both crucial and natural to extend this research direction to the video domain for video question answering (VideoQA). Compared to the image domain where large scale and fully annotated benchmark datasets exists, VideoQA datasets are limited to small scale and are automatically generated, etc. These limitations restrict their applicability in practice. Here we introduce ActivityNet-QA, a fully annotated and large scale VideoQA dataset. The dataset consists of 58,000 QA pairs on

1 Introduction

Analysis of: ActivityNet-QA: A Dataset for Understanding Complex Web Videos via Question Answering. Research goal: Does iterative retrieval with visual frame reranking mitigate video content drift degradation more effectively than single-pass dense retrieval across different vision-language model families on VideoQA 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

10 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 3.8/10 → 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/10.1007/s11263-020-01359-2
- https://doi.org/10.48550/arxiv.2402.19473
- https://doi.org/10.1609/aaai.v33i01.33019127