# SOVEREIGN: How does the performance of VideoRAG compare to temporal video question answering models on long-form video un

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

**Abstract**

We present HERO, a novel framework for large-scale video+language omnirepresentation learning. HERO encodes multimodal inputs in a hierarchical structure, where local context of a video frame is captured by a Cross-modal Transformer via multimodal fusion, and global video context is captured by a Temporal Transformer. In addition to standard Masked Language Modeling (MLM) and Masked Frame Modeling (MFM) objectives, we design two new pre-training tasks: (i) Video-Subtitle Matching (VSM), where the model predicts both global and local temporal alignment; and (ii) Frame Order Modeling (FOM), wher

## 1 Introduction

Analysis of: HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. Research goal: How does the performance of VideoRAG compare to temporal video question answering models on long-form video understanding tasks when evaluated with METEOR scores across 10x context scaling from 32K to 128K tokens?.

## 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

8 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 8.8/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

## 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 |
|---|---|---|
| HERO encodes multimodal inputs in a hierarchical structure, where local context of a video frame is captured by a Cross- | ✓ | 0.42 |
| HERO uses standard Masked Language Modeling (MLM) and Masked Frame Modeling (MFM) objectives for pre-training | ✓ | 0.28 |
| HERO introduces two new pre-training tasks: Video-Subtitle Matching (VSM) and Frame Order Modeling (FOM) | ✓ | 0.30 |
| HERO is jointly trained on HowTo100M and large-scale TV datasets | ✓ | 0.25 |
| HERO achieves new state of the art on multiple benchmarks over Text-based Video/Video-moment Retrieval, Video Question A | ✓ | 0.41 |
| Two new challenging benchmarks How2QA and How2R for Video QA and Retrieval were introduced | ✓ | 0.23 |

## References


- https://doi.org/10.18653/v1/2020.emnlp-main.161
- https://doi.org/10.3389/frai.2024.1430984
- https://doi.org/10.1145/3355390