# SOVEREIGN: How does HERO's cross-modal fusion efficiency compare to other multimodal architectures when processing high-r

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

**Abstract**

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3x3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16-19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respective

## 1 Introduction

Analysis of: Very Deep Convolutional Networks for Large-Scale Image Recognition. Research goal: How does HERO's cross-modal fusion efficiency compare to other multimodal architectures when processing high-resolution video frames with corresponding subtitles?.

## 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: 2.0/10 $\rightarrow$ 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/10.1109/euvip53989.2022.9922810
- https://doi.org/10.1186/s40537-021-00444-8
- https://doi.org/10.48550/arxiv.1409.1556