SOVEREIGN: How does SMoES's cross-modal reasoning accuracy on Winoground compare to modality-agnostic MoE-VLMs when evalu

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Current cross-modal retrieval systems are evaluated using R@K measure which does not leverage semantic relationships rather strictly follows the manually marked image text query pairs. Therefore, current systems do not generalize well for the unseen data in the wild. To handle this, we propose a new measure, SemanticMap, to evaluate the performance of cross-modal systems. Our proposed measure evaluates the semantic similarity between the image and text representations in the latent embedding space. We also propose a novel cross-modal retrieval system using a single stream network for bidirecti

1 Introduction

Analysis of: Do Cross Modal Systems Leverage Semantic Relationships?. Research goal: How does SMoES's cross-modal reasoning accuracy on Winoground compare to modality-agnostic MoE-VLMs when evaluated on the hardest subset of examples (those with low unimodal bias)?.

2 Methodology

Multi-query arXiv search (1 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

9 papers retrieved. 8 claims extracted, 4 verified. Tribunal: 5.8/10 $\rightarrow$ 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
The proposed system is the first to employ a single stream network for cross modal retrieval systems.	✓	0.34
The proposed system is evaluated on MSCOCO and Flickr30K datasets.	✓	0.15
The proposed system has shown comparable results to current state-of-the-art methods.	✓	0.22
The system uses extended center loss during training.	×	0.09
Text descriptions are encoded as images which enables the use of a single stream network for both text and images.	✓	0.30
Most existing techniques require separate networks for each modality which increases computational complexity.	×	0.03
The system employs a single stream network to extract representation from multiple modalities without pairwise or triple	×	0.08
The system uses a pairwise loss function Lc(xi, x±i) = d(xi, x±i) where d is a distance metric.	×	0.02

References

- http://arxiv.org/abs/1909.01976v1
- http://arxiv.org/abs/2507.08804v1
- http://arxiv.org/abs/2504.16021v1