SOVEREIGN: What is the throughput trade-off (inference latency vs. accuracy) when scaling expert count in vision-language

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

As a fundamental and challenging task in bridging language and vision domains, Image-Text Retrieval (ITR) aims at searching for the target instances that are semantically relevant to the given query from the other modality, and its key challenge is to measure the semantic similarity across different modalities. Although significant progress has been achieved, existing approaches typically suffer from two major limitations: (1) It hurts the accuracy of the representation by directly exploiting the bottom-up attention based region-level features where each region is equally treated. (2) It limit

1 Introduction

Analysis of: USER: Unified Semantic Enhancement With Momentum Contrast for Image-Text Retrieval. Research goal: What is the throughput trade-off (inference latency vs. accuracy) when scaling expert count in vision-language models evaluated on domain-shifted captioning datasets like Flickr30K to MSCOCO?.

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

1 papers retrieved. 10 claims extracted, 1 verified. Tribunal: 5.2/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
The proposed USER method belongs to the global-level matching method category for image-text retrieval.	×	0.12
In USER, the momentum coefficient is set to 0.999 (m = 0.999).	×	0.03
The sizes of dynamic queues for the MSCOCO and Flickr30K datasets are set to 4096 and 2048, respectively.	×	0.03
USER uses Faster-RCNN with ResNet-101 to extract 36 region proposals for image encoding.	×	0.03
In USER, BERT-base is used as the text encoder to extract 768-dimensional word-level features.	×	0.04
The dimensionality of the joint embedding space in USER is 1024 (dJ = 1024).	×	0.00
USER incorporates two Global representation based Semantic Enhancement (GSE) modules: Self-Guided Enhancement (SGE) and	✓	0.31
Compared with the existing best method NAAF, USER improves the R@1 metric by 5% and 2.4% on caption retrieval and image	×	0.04
USER achieves over 60 times faster inference speed compared to NAAF without using external knowledge or pre-trained mode	×	0.04
In USER's training, $\gamma = 90$ and $\epsilon = 0.5$ are set empirically in Eqs. (17)-(19).	×	0.03

References

- https://arxiv.org/abs/2301.06844