SOVEREIGN: How does MMICL's zero-shot image-text retrieval accuracy on MSCOCO and Flickr30K compare to Flamingo, PaLI, an

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Since the resurgence of deep learning, vision-language models (VLMs) enhanced by large language models (LLMs) have grown exponentially in popularity. However, while LLMs can utilize extensive background knowledge and task information with in-context learning, most VLMs still struggle with understanding complex multi-modal prompts with multiple images, making VLMs less effective in downstream vision-language tasks. In this paper, we address the limitation above by 1) introducing vision-language Model with Multi-Modal In-Context Learning(MMICL), a new approach to allow the VLM to deal with multi

1 Introduction

Analysis of: MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. Research goal: How does MMICL's zero-shot image-text retrieval accuracy on MSCOCO and Flickr30K compare to Flamingo, PaLI, and BLIVA when varying the number of in-context examples?.

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

2 papers retrieved. 9 claims extracted, 9 verified. Tribunal: 8.5/10 $\rightarrow$ 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
Vision-language models (VLMs) enhanced by large language models (LLMs) have grown exponentially in popularity since the	✓	0.33
Most VLMs still struggle with understanding complex multi-modal prompts with multiple images, making VLMs less effective	✓	0.40
MMICL introduces a new approach to allow the VLM to deal with multi-modal inputs efficiently.	✓	0.25
MMICL proposes a novel context scheme to augment the in-context learning ability of the VLM.	✓	0.25
The Multi-modal In-Context Learning (MIC) dataset is constructed to enhance the VLM's ability to understand complex mult	✓	0.37
MMICL achieves new state-of-the-art zero-shot performance on a wide range of general vision-language tasks, especially f	✓	0.38
MMICL effectively tackles the challenge of complex multi-modal prompt understanding and emerges the impressive ICL abili	✓	0.33
MMICL successfully alleviates language bias in VLMs, a common issue for VLMs that often leads to hallucination when face	✓	0.33
The code, dataset, dataset tool, and model for MMICL are available at https://github.com/PKUnlp-icler/MIC.	✓	0.27

References

- https://doi.org/10.48550/arxiv.2309.07915
- https://doi.org/10.48550/arxiv.2311.07594