SOVEREIGN: What is the inference throughput and memory cost trade-off for MoE-LLaVA under adversarial textual perturbatio

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Following the recent popularity of Large Language Models (LLMs), several attempts have been made to extend them to the visual domain. From having a visual assistant that could guide us through unfamiliar environments to generative models that produce images using only a high-level text description, the vision-language model (VLM) applications will significantly impact our relationship with technology. However, there are many challenges that need to be addressed to improve the reliability of those models. While language is discrete, vision evolves in a much higher dimensional space in which con

1 Introduction

Analysis of: An Introduction to Vision-Language Modeling. Research goal: What is the inference throughput and memory cost trade-off for MoE-LLaVA under adversarial textual perturbations on MMMU, relative to modality-agnostic MoE baselines, at 7B versus 13B model scales?.

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

9 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.5/10 $\rightarrow$ 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
Several attempts have been made to extend Large Language Models (LLMs) to the visual domain following their recent popul	✓	0.24
Vision-language model (VLM) applications will significantly impact our relationship with technology.	✓	0.30
There are many challenges that need to be addressed to improve the reliability of vision-language models.	✓	0.25
Language is discrete, while vision evolves in a much higher dimensional space in which concepts cannot always be easily	✓	0.32
This work primarily focuses on mapping images to language, but also discusses extending VLMs to videos.	✓	0.30

References

- https://doi.org/10.48550/arxiv.2401.08092
- https://doi.org/10.48550/arxiv.2405.02246
- https://doi.org/10.48550/arxiv.2405.17247