SOVEREIGN: What is the inference throughput and memory efficiency trade-off of SMoES MoE-VLMs relative to dense VLMs of e

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

The field of natural language processing (NLP) has made significant strides in recent years, particularly in the development of large-scale vision-language models (VLMs). These models aim to bridge the gap between text and visual information, enabling a more comprehensive understanding of multimedia data. However, as these models become larger and more complex, they also become more challenging to train and deploy. One approach to addressing this challenge is the use of sparsely-gated mixture-of-experts (MoE) techniques, which divide the model into smaller, specialized sub-models that can join

1 Introduction

Analysis of: Scaling Vision-Language Models with Sparse Mixture of Experts. Research goal: What is the inference throughput and memory efficiency trade-off of SMoES MoE-VLMs relative to dense VLMs of equivalent parameter count (7B and 34B) on multimodal reasoning benchmarks like MMMU and MathVista?.

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

11 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 7.8/10 → 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
The field of natural language processing (NLP) has made significant strides in recent years, particularly in the develop	✓	0.40
sparsely-gated mixture-of-experts (MoE) techniques divide the model into smaller, specialized sub-models that can jointl	✓	0.38
MoE can achieve state-of-the-art performance on a range of benchmarks over dense models of equivalent computational cost	✓	0.28
The research offers insights into stabilizing the training of MoE models	✓	0.22
The research offers insights into understanding the impact of MoE on model interpretability	✓	0.21
The research offers insights into balancing the trade-offs between compute performance when scaling VLMs	✓	0.26

References

- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.48550/arxiv.2409.12191
- https://doi.org/10.18653/v1/2023.findings-emnlp.758