SOVEREIGN: MoE-LLaVA: Mixture of Experts for Large Vision-Language Models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Recent advances demonstrate that scaling Large Vision-Language Models (LVLMs) effectively improves downstream task performances. However, existing scaling methods enable all model parameters to be active for each token in the calculation, which brings massive training and inferring costs. In this work, we propose a simple yet effective training strategy MoE-Tuning for LVLMs. This strategy innovatively addresses the common issue of performance degradation in multi-modal sparsity learning, consequently constructing a sparse model with an outrageous number of parameters but a constant computation

1 Introduction

Analysis of: MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. Research goal: How does SMoES's routing strategy compare to dense models and modality-agnostic MoE baselines in terms of Top-1 accuracy on the MMMU benchmark under visual vs. textual adversarial perturbations across model scales (e.g., 7B, 13B)?.

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

12 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
MoE-LLaVA, with only approximately 3B sparsely activated parameters, demonstrates performance comparable to LLaVA-1.5-7B	✓	0.34
MoE-LLaVA surpasses LLaVA-1.5-13B in object hallucination benchmark.	✓	0.27
MoE-Tuning addresses the common issue of performance degradation in multi-modal sparsity learning.	✓	0.27
MoE-LLaVA activates only the top-k experts through routers during deployment, keeping the remaining experts inactive.	✓	0.26
Scaling methods that enable all model parameters to be active for each token bring massive training and inferring costs.	✓	0.23

References

- https://www.semanticscholar.org/paper/cd1d7f5c4ce2d31ce9ee72db165a8272624da7d3
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2601.15021v1