SOVEREIGN: How does the expert specialization in SMoES-based MoE-VLMs affect cross-modal reasoning robustness on multimod

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Mixture-of-Experts (MoE) has become a prevalent backbone for large vision-language models (VLMs), yet how modality-specific signals should guide expert routing remains under-explored. Existing routing strategies are either hand-crafted or modality-agnostic, relying on idealized priors that ignore the layer-dependent modality fusion patterns in MoE-VLMs and provide little guidance for expert specialization. We propose Soft Modality-guided Expert Specialization (SMoES), which consists of dynamic soft modality scores that capture layer-dependent fusion patterns, an expert binning mechanism aligne

1 Introduction

Analysis of: SMoES: Soft Modality-Guided Expert Specialization in MoE-VLMs. Research goal: How does the expert specialization in SMoES-based MoE-VLMs affect cross-modal reasoning robustness on multimodal benchmarks like MathVista or MMBench when compared to modality-agnostic routing?.

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, 1 verified. Tribunal: 4.7/10 → REVISE (revision_round=1). 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
SMoES achieves a 10.3% reduction in Time to First Token (TTFT) and a 10.5% reduction in Time Per Output Token (TPOT) for	×	0.03
OLMoE with the Gaussian Mixture Model (GMM) estimator with different n-components shows improved performance over the ba	×	0.05
The cross-GPU expert transfer ratio for vision and text tokens differs between prefill and decode phases.	×	0.05
Attention-soft and gaussian-soft methods in SMoES show significant specialization in modality-specific tasks.	✓	0.15
SMoES reduces latency by 22.0% for SQA-IMG and 10.3% for MMMM in comparison to the baseline.	×	0.02

References

- http://arxiv.org/abs/2508.19294v2
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2604.00086v1