SOVEREIGN: How does the SMoES soft modality-guided routing mechanism compare to dense baselines and hard routing variants

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 SMoES soft modality-guided routing mechanism compare to dense baselines and hard routing variants in terms of out-of-distribution robustness on the MMMU benchmark when scaling from 7B to 34B total parameters?.

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, 0 verified. Tribunal: 2.7/10 → REJECT (revision_round=0). 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 improves performance on diverse VLM and language-only tasks	×	0.06
SMoES achieves maximum throughput improvement of 2.8% for overall performance	×	0.06
SMoES reduces TTFT by 22.0% and TPOT by 9.0% for MMMU benchmark at batch size 32	×	0.04
SMoES reduces TTFT by 20.0% and TPOT by 10.6% for SQA-IMG benchmark at batch size 32	×	0.02
SMoES shows an average performance improvement of 2.8% for language tasks	×	0.08
SMoES achieves 31.1% improvement in inter-bin specialization with attention-soft method	×	0.13

References

- http://arxiv.org/abs/2510.13759v3
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2603.11114v1