SOVEREIGN: SMoES: Soft Modality-Guided Expert Specialization in MoE-VLMs

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 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: What is the scaling efficiency of soft modality-guided routing in SMoES when increasing total parameter count from 7B to 13B+ on multimodal benchmarks, measured by accuracy-per-parameter and FLOPs per inference step?.

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

10 papers retrieved. 6 claims extracted, 2 verified. Tribunal: 6.0/10 $\rightarrow$ 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 0.9% average gain on multimodal tasks.	×	0.10
SMoES achieves 4.2% average gain on language-only tasks.	×	0.10
SMoES reduces EP communication overhead by 56.1%.	×	0.11
SMoES improves throughput by 12.3% under realistic deployment.	×	0.10
SMoES uses attention-based or Gaussian-statistics modality scores to optimize mutual information regularization.	✓	0.28
SMoES consists of dynamic soft modality scores, an expert binning mechanism, and inter-bin mutual information regulariza	✓	0.34

References

- http://arxiv.org/abs/2603.11114v1
- https://www.semanticscholar.org/paper/e6c3e40973c9f51ebbd36d13dbb6b2470ae5c9b7
- http://arxiv.org/abs/2604.23996v1