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 inference throughput trade-off (tokens per second) between SMoES MoE-VLMs and dense models of equal total parameters on multimodal reasoning tasks at 7B and 34B scales?.

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. 10 claims extracted, 8 verified. Tribunal: 7.3/10 $\rightarrow$ 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
SMoES consists of dynamic soft modality scores that capture layer-dependent fusion patterns	✓	0.37
SMoES includes an expert binning mechanism aligned with expert-parallel deployment	✓	0.21
SMoES uses an inter-bin mutual information regularization that encourages coherent modality specialization	✓	0.29
The method leverages attention-based or Gaussian-statistics modality scores to optimize mutual information regularizatio	✓	0.33
Experiments were conducted across four MoE-based VLMs and 16 benchmarks	✓	0.17
SMoES shows 0.9% average gain on multimodal tasks	×	0.10
SMoES shows 4.2% average gain on language-only tasks	×	0.10
SMoES achieves 56.1% reduction in EP communication overhead	✓	0.16
SMoES achieves 12.3% throughput improvement under realistic deployment	✓	0.17
SMoES improves both effectiveness and efficiency in MoE-VLMs	✓	0.15

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2402.14800v2
- http://arxiv.org/abs/2604.23996v1