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 impact of varying expert count in SMoES-based MoE-VLMs on MMMU score variance and expert routing robustness compared to dense VLMs?.

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

9 papers retrieved. 6 claims extracted, 6 verified. Tribunal: 8.2/10 $\rightarrow$ 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
Existing routing strategies in MoE-VLMs are either hand-crafted or modality-agnostic, relying on idealized priors that i	✓	0.39
SMoES consists of dynamic soft modality scores that capture layer-dependent fusion patterns, an expert binning mechanism	✓	0.49
SMoES uses attention-based or Gaussian-statistics modality scores to optimize mutual information regularization.	✓	0.30
Experiments across four MoE-based VLMs and 16 benchmarks demonstrate 0.9% and 4.2% average gain on multimodal and langua	✓	0.28
SMoES achieves 56.1% reduction in EP communication overhead and 12.3% throughput improvement under realistic deployment.	✓	0.26
Aligning routing with modality-aware expert specialization unlocks MoE-VLM capacity and efficiency.	✓	0.34

References

- http://arxiv.org/abs/2506.14646v2
- http://arxiv.org/abs/2601.15021v1
- http://arxiv.org/abs/2604.23996v1