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: How does SMoES routing compare to dense baselines and hard-routed MoE-VLMs on inference throughput (tokens/sec) versus ANLS accuracy when scaling from 7B to 13B+ parameters on DocVQA?.

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. 5 claims extracted, 5 verified. Tribunal: 7.8/10 $\rightarrow$ 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 achieves a 0.9% average gain on multimodal tasks and a 4.2% average gain on language-only tasks across four MoE-ba	✓	0.22
SMoES reduces expert-parallel (EP) communication overhead by 56.1%.	✓	0.15
SMoES achieves a 12.3% throughput improvement under realistic deployment.	✓	0.17
Existing routing strategies in MoE-VLMs are either hand-crafted or modality-agnostic, relying on idealized priors that i	✓	0.40
SMoES consists of dynamic soft modality scores, an expert binning mechanism aligned with expert-parallel deployment, and	✓	0.42

References

- http://arxiv.org/abs/2502.03692v1
- http://arxiv.org/abs/1309.1755v1
- http://arxiv.org/abs/2604.23996v1