SOVEREIGN: How does SMoES soft modality-guided routing compare to dense VLMs and standard MoE routing on multi-step reaso

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 SMoES soft modality-guided routing compare to dense VLMs and standard MoE routing on multi-step reasoning tasks (e.g., MathVista, MMMU) in terms of accuracy and inference latency when controlling for total parameter count?.

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. 13 claims extracted, 0 verified. Tribunal: 1.5/10 → 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 achieves a 10.3% reduction in TTFT and 10.5% reduction in TPOT on MMMU at batch size 1 compared to baseline.	×	0.02
SMoES achieves a 22.6% reduction in TTFT and 9.6% reduction in TPOT on MMMU at batch size 32 compared to baseline.	×	0.02
SMoES achieves a 9.2% reduction in TTFT and 9.7% reduction in TPOT on SQA-IMG at batch size 1 compared to baseline.	×	0.02
SMoES achieves a 20.0% reduction in TTFT and 10.6% reduction in TPOT on SQA-IMG at batch size 32 compared to baseline.	×	0.02
Using a Gaussian Mixture Model (GMM) estimator with k=2 components yields a +2.8% overall improvement over baseline on O	×	0.03
Using a GMM estimator with k=4 components yields a +2.1% overall improvement over baseline on OLMoE.	×	0.02
Using a GMM estimator with k=1 component yields a +2.0% overall improvement over baseline on OLMoE.	×	0.03
SMoES with gaussian-soft achieves a score of 0.800 on the 'No Specialization' metric in Table (p6).	×	0.06
SMoES with attention-soft achieves a score of 0.752 on the 'No Specialization' metric in Table (p6).	×	0.06
SMoES with gaussian-soft achieves a score of 0.754 on the 'No Specialization' metric in Table (p6) second instance.	×	0.05
SMoES achieves a 15.0% improvement on one metric and 99.3% on another in Table (p8) under PV:PT:DT=32:8:1.	×	0.02
SMoES achieves a 97.5% improvement on one metric and 99.7% on another in Table (p8) under PV:PT:DT=14:7:1.	×	0.02
SMoES reduces cross-GPU EP transfer ratio for vision and text tokens separately during prefill and decode phases.	×	0.05

References

- http://arxiv.org/abs/2603.26742v1
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2604.23996v1