SOVEREIGN: How does the inference throughput and FLOPs efficiency of SMoE with dynamic routing signatures compare against

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 the inference throughput and FLOPs efficiency of SMoE with dynamic routing signatures compare against modality-agnostic fixed-ratio MoE baselines on SNLI-VE when evaluated under batch-size and latency constraints?.

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. 10 claims extracted, 0 verified. Tribunal: 3.3/10 → 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 for MMMU benchmark compared to the baseline.	×	0.02
SMoES achieves a 9.2% reduction in TTFT and 9.7% reduction in TPOT for SQA-IMG benchmark compared to the baseline.	×	0.02
SMoES shows consistent latency improvements across different batch sizes from 1 to 32.	×	0.02
The TTFT and TPOT for both MMMU and SQA-IMG show reductions at all batch sizes tested.	×	0.01
The expert-parallel deployment of SMoES maintains stable improvement ratios for Decode across batch sizes.	×	0.08
SMoES reduces communication overhead in Prefill with larger batch sizes resulting in greater gains.	×	0.04
GMM estimator with different n-components was evaluated on OLMoE showing performance variance.	×	0.02
Ablation studies on inter-bin specialization objectives show the impact of KL divergence and mutual information.	×	0.11
SMoES achieves cross-GPU expert transfer ratios that vary between prefill and decode phases for vision and text tokens.	×	0.04
SMoES with attention-soft and gaussian-soft configurations shows performance on various metrics including accuracy, FLOP	×	0.07

References

- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2603.11114v1