SOVEREIGN: Does soft MoE routing with token-level gating improve ANLS on InfographicsVQA and ChartQA compared to hard top

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: Does soft MoE routing with token-level gating improve ANLS on InfographicsVQA and ChartQA compared to hard top-2 routing in MoE-VLMs under distribution shift (e.g., style transfer or domain adaptation)?.

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. 8 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
The model 'SMoES' improves computational efficiency under expert-parallel deployment	×	0.13
Larger batch sizes in Prefill increase communication proportion, yielding greater gains	×	0.04
The 'sequential order' is used in the baseline, which is equivalent to a random order since there are no order constrain	×	0.01
Experts exhibit significant specialization	×	0.06
Data goes through DeDup (deduplication) to avoid transmitting duplicate tokens routed to experts on the same device in t	×	0.02
SMoES achieves a latency decrease compared to the soft routing baseline at different batch sizes on edge-side deployment	×	0.05
SMoES achieves better performance on multiple benchmark datasets	×	0.03
Text token transfer ratio differs between prefill and decode phases for vision and text tokens	×	0.03

References

- http://arxiv.org/abs/1803.07724v1
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2407.04255v1