

SOVEREIGN: Does increasing the number of experts in SMOES improve robustness to cloaking-style adversarial perturbations

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 increasing the number of experts in SMOES improve robustness to cloaking-style adversarial perturbations on MMMU accuracy compared to fixed top-1 routing in modality-agnostic MoE?.

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

11 papers retrieved. 6 claims extracted, 0 verified. Tribunal: 3.4/10 → 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 Time to First Token (TTFT) for MMR task at batch size 1	×	0.02
SMoES achieves a 9.2% reduction in TTFT for SQA-IMG task at batch size 1	×	0.03
The baseline Time to First Token (TTFT) for MMR at batch size 1 is 2.810 seconds	×	0.02
The baseline Time to First Token (TTFT) for SQA-IMG task is 1.493 seconds	×	0.02
The baseline Time Per Output Token (TPOT) for SQA-IMG task is 0.766 seconds	×	0.02
SMoES achieves a 22.0% reduction in prefill stage latency (TTFT) at batch size 32	×	0.02

References

- <http://arxiv.org/abs/2602.09258v1>
- <http://arxiv.org/abs/2604.23996v1>
- <http://arxiv.org/abs/2603.11114v1>