SOVEREIGN: How does SMoES's inference throughput (tokens/sec) on Winoground compare to modality-agnostic MoE-VLMs when sc

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's inference throughput (tokens/sec) on Winoground compare to modality-agnostic MoE-VLMs when scaling the number of experts from 16 to 128 with top-2 routing?.

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

9 papers retrieved. 12 claims extracted, 0 verified. Tribunal: 1.0/10 $\rightarrow$ 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 MMMU benchmark compared to the baseline.	×	0.03
SMoES achieves a 10.5% reduction in Time Per Output Token (TPOT) for MMMU benchmark compared to the baseline.	×	0.03
SMoES achieves a 22.0% reduction in TTFT for MMMU benchmark at batch size 16 compared to the baseline.	×	0.02
SMoES achieves a 9.0% reduction in TPOT for MMMU benchmark at batch size 16 compared to the baseline.	×	0.02
SMoES achieves a 9.2% reduction in TTFT for SQA-IMG benchmark compared to the baseline.	×	0.01
SMoES achieves a 9.7% reduction in TPOT for SQA-IMG benchmark compared to the baseline.	×	0.01
SMoES achieves a 16.6% reduction in TTFT for SQA-IMG benchmark at batch size 8 compared to the baseline.	×	0.02
SMoES achieves a 11.3% reduction in TPOT for SQA-IMG benchmark at batch size 8 compared to the baseline.	×	0.01
SMoES with attention-soft routing achieves 31.9% accuracy on MMMU benchmark.	×	0.05
SMoES with gaussian-soft routing achieves 32.9% accuracy on MMMU benchmark.	×	0.05
SMoES with attention-soft routing achieves 30.0% accuracy on SQA-IMG benchmark.	×	0.05
SMoES with gaussian-soft routing achieves 32.1% accuracy on SQA-IMG benchmark.	×	0.05

References

- http://arxiv.org/abs/2410.17954v2

- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2303.07226v1