SOVEREIGN: What is the scaling efficiency of SMoES-based MoE-VLMs in terms of downstream task accuracy per additional exp

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: What is the scaling efficiency of SMoES-based MoE-VLMs in terms of downstream task accuracy per additional expert on VQA and image captioning benchmarks relative to dense VLMs of equivalent 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

9 papers retrieved. 11 claims extracted, 0 verified. Tribunal: 2.2/10 → 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.	×	0.05
SMoES achieves a 22.0% reduction in TTFT and 9.0% reduction in TPOT on MMMU at batch size 8 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 16.6% reduction in TTFT and 11.3% reduction in TPOT on SQA-IMG at batch size 8 compared to baseline.	×	0.03
SMoES with k=2 improves multimodal task performance by 1.2% and language task performance by 5.6% over baseline.	×	0.08
SMoES with k=4 improves multimodal task performance by 0.8% and language task performance by 4.3% over baseline.	×	0.08
SMoES with k=1 improves multimodal task performance by 0.6% and language task performance by 4.3% over baseline.	×	0.08
SMoES achieves a 15.0% improvement on one metric and 99.3% on another in the PV:PT:DT=32:8:1 configuration.	×	0.02
SMoES achieves a 97.5% improvement on one metric and 99.7% on another in the PV:PT:DT=14:7:1 configuration.	×	0.02
SMoES reduces TTFT by 22.6% on MMMU at batch size 16 compared to baseline.	×	0.02
SMoES reduces TTFT by 22.4% on SQA-IMG at batch size 16 compared to baseline.	×	0.02

References

- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2605.15484v1
- http://arxiv.org/abs/2603.11114v1