SOVEREIGN: How does the Dynamic Clue Bottleneck architecture affect inference throughput (tokens/sec) and FLOPs efficienc

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 Dynamic Clue Bottleneck architecture affect inference throughput (tokens/sec) and FLOPs efficiency relative to Top-2 modality-agnostic MoE-VLMs on VCR adversarial splits under matched expert counts?.

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. 11 claims extracted, 0 verified. Tribunal: 5.0/10 → REVISE (revision_round=1). 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 dataset at batch size 1.	×	0.02
SMoES achieves a 22.0% reduction in TTFT for MMMU dataset at batch size 16.	×	0.03
SMoES achieves a 9.2% reduction in TTFT for SQA-IMG dataset at batch size 1.	×	0.03
SMoES achieves a 16.6% reduction in TTFT for SQA-IMG dataset at batch size 8.	×	0.03
SMoES achieves up to 22.0% reduction in TTFT across different batch sizes for multimodal tasks.	×	0.04
Language-only tasks show a +4.3% improvement with k=2 compared to k=1 baseline in computational efficiency.	×	0.09
SMoES shows significant improvements in expert specialization with reduced cross-GPU transfer ratios.	×	0.08
Decode stage maintains stable improvement ratios across batch sizes due to fewer activated experts.	×	0.05
SMoES achieves 10.3% latency reduction in Time to First Token (TTFT) for MMMU at batch size 1.	×	0.01
SMoES achieves 9.2% latency reduction in TTFT for SQA-IMG at batch size 1.	×	0.02
SMoES reduces Time Per Output Token (TPOT) by 9.7% for SQA-IMG at batch size 1.	×	0.01

References

- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2006.01205v2
- http://arxiv.org/abs/2603.11114v1