

# SOVEREIGN: SMOES: Soft Modality-Guided Expert Specialization in MoE-VLMs

SOVEREIGN Research Kernel

Autonomous draft — Owner review required before publication

May 27, 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 inference throughput of SMOES-based MoE-VLMs with 4, 8, and 16 experts compare to dense VLMs on MMMU accuracy under fixed FLOPs budgets?.

## 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. 5 claims extracted, 3 verified. Tribunal: 6.8/10 → 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 0.9% average gain on multimodal tasks and a 4.2% average gain on language-only tasks across four MoE-based	✓	0.20
SMoES reduces expert-parallel (EP) communication overhead by 56.1%.	×	0.13
SMoES improves throughput by 12.3% under realistic deployment.	×	0.10
Existing routing strategies in MoE-VLMs are either hand-crafted or modality-agnostic, relying on idealized priors that i	✓	0.39
SMoES uses attention-based or Gaussian-statistics modality scores to optimize mutual information regularization.	✓	0.31

## References

- <http://arxiv.org/abs/2605.15484v1>
- <http://arxiv.org/abs/2602.18846v2>
- <http://arxiv.org/abs/2604.23996v1>