SOVEREIGN: AnyExperts: On-Demand Expert Allocation for Multimodal Language Models with Mixt

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Multimodal Mixture-of-Experts (MoE) models offer a promising path toward scalable and efficient large vision-language systems. However, existing approaches rely on rigid routing strategies (typically activating a fixed number of experts per token) ignoring the inherent heterogeneity in semantic importance across modalities. This leads to suboptimal compute allocation, where redundant tokens consume as many resources as critical ones. To address this, we propose AnyExperts, a novel on-demand, budget-aware dynamic routing framework that allocates a variable total number of expert slots per token

1 Introduction

Analysis of: AnyExperts: On-Demand Expert Allocation for Multimodal Language Models with Mixture of Expert. Research goal: Does dynamic expert specialization in MoE-VLMs improve compositional generalization on multi-step reasoning tasks compared to fixed routing, as measured by accuracy on the GQA and NLVR2 benchmarks?.

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

6 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.8/10 → APPROVE (revision_round=0). Policy: AUTO_APPROVE.

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
Existing multimodal Mixture-of-Experts models rely on rigid routing strategies that typically activate a fixed number of	✓	0.23
AnyExperts allocates a variable total number of expert slots per token based on its semantic importance.	✓	0.29
In AnyExperts, the total slots per token are constrained within a fixed range.	✓	0.21
Each slot in AnyExperts is filled by either a real expert or a virtual expert, with the virtual share capped at a small	✓	0.28
AnyExperts improves performance under the same compute budget across diverse tasks in visual understanding, audio unders	✓	0.29
On general image/video tasks, AnyExperts achieves comparable accuracy with 40% fewer real expert activations.	✓	0.26
On text-dense tasks (OCR and NLP), AnyExperts maintains performance while reducing real expert usage by 10%.	✓	0.27

References

- https://www.semanticscholar.org/paper/65331807c808ed9e8e7b4c11c095a2f9ccbec6b1
- https://www.semanticscholar.org/paper/e1d78957a52cf82aafa648f11794e8acf4184853
- https://www.semanticscholar.org/paper/01085ddff95b81bf6b438e5e03e9313483679a49