SOVEREIGN: To what extent does AnyExperts' dynamic expert allocation mitigate representational collapse and maintain per-

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: To what extent does AnyExperts' dynamic expert allocation mitigate representational collapse and maintain per-expert specialization across varying numbers of experts (8 to 64) on multimodal reasoning tasks from the ARO dataset?.

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

12 papers retrieved. 10 claims extracted, 0 verified. Tribunal: 4.3/10 → 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
The AnyExperts framework uses 1.1 trillion multimodal tokens during pretraining.	×	0.04
The AnyExperts framework uses 30 million image–text, 10 million video–text, and 230 thousand hours of speech–text instru	×	0.05
Ling-mini-2.0 employs 256 experts and activates exactly 8 experts per token in a uniform manner.	×	0.05
In AnyExperts, the number of activated experts per token is dynamically adjusted between Kmin = 8 and Kmax = 12 during t	×	0.05
In AnyExperts, the average number of activated experts per token during inference is reduced to 7.2 without degrading do	×	0.07
The AnyExperts framework achieves an average of 79.73 on MMBench for image understanding.	×	0.05
The AnyExperts framework achieves an average of 85.10 on OCRBench.	×	0.03
The AnyExperts framework achieves an average of 90.17 on ARC-C.	×	0.05
The AnyExperts framework achieves an average of 66.40 on MVBench.	×	0.04
The AnyExperts framework achieves an average of 60.59 on VideoMME.	×	0.02

References

- http://arxiv.org/abs/2602.09258v1
- http://arxiv.org/abs/2511.18314v1

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