SOVEREIGN: Mixture-of-Experts Models in Vision: Routing, Optimization, and Generalization

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Mixture-of-Experts (MoE) architectures enable conditional computation by routing inputs to multiple expert subnetworks and are often motivated as a mechanism for scaling large language models. In this project, we instead study MoE behavior in an image classification setting, focusing on predictive performance, expert utilization, and generalization. We compare dense, SoftMoE, and SparseMoE classifier heads on the CIFAR10 dataset under comparable model capacity. Both MoE variants achieve slightly higher validation accuracy than the dense baseline while maintaining balanced expert utilization th

1 Introduction

Analysis of: Mixture-of-Experts Models in Vision: Routing, Optimization, and Generalization. Research goal: How does the inference latency and throughput of SMoES-based 7B VLMs compare against dense VLMs and hard-routing MoE baselines on MMBench and SEED-Bench at varying batch sizes and sequence lengths?.

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. 6 claims extracted, 6 verified. Tribunal: 7.8/10 $\rightarrow$ 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
Mixture-of-Experts (MoE) architectures enable conditional computation by routing inputs to multiple expert subnetworks.	✓	0.29
Both SoftMoE and SparseMoE achieve slightly higher validation accuracy than the dense baseline on the CIFAR10 dataset.	✓	0.27
Balanced expert utilization is maintained through regularization, avoiding expert collapse.	✓	0.19
SoftMoE exhibits higher sharpness by Hessian-based metrics (largest eigenvalue and trace) compared to Dense and SparseMo	✓	0.24
Dense and SparseMoE lie in a similar curvature regime despite all models achieving comparable generalization performance	✓	0.31
Naively implemented conditional routing does not yield inference speedups on modern hardware at this scale.	✓	0.27

References

- http://arxiv.org/abs/2410.21465v3
- http://arxiv.org/abs/2604.23996v1
- http://arxiv.org/abs/2601.15021v1