SOVEREIGN: How does layer-wise modality specialization in MoE VLMs influence inference throughput and MMBench accuracy tr

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Real world deployments often expose modern object recognition models to domain shifts that precipitate a severe drop in accuracy. Such shifts encompass (i) variations in low level image statistics, (ii) changes in object pose and viewpoint, (iii) partial occlusion, and (iv) visual confusion across adjacent classes. To mitigate this degradation, we introduce the Re-Thinking Vision Language Model (RT-VLM) framework. The foundation of this framework is a unique synthetic dataset generation pipeline that produces images annotated with"4-Clues": precise bounding boxes, class names, detailed object-

1 Introduction

Analysis of: RT-VLM: Re-Thinking Vision Language Model with 4-Clues for Real-World Object Recognition Robustness. Research goal: How does layer-wise modality specialization in MoE VLMs influence inference throughput and MMBench accuracy trade-offs under distribution shift from natural to synthetic image inputs?.

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

3 papers retrieved. 0 claims extracted, 0 verified. Tribunal: 1.7/10 → REJECT (revision_round=0). 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.

References

- https://www.semanticscholar.org/paper/b022abf574e5ffe319607abbc58f8c2fe61a7e0e
- https://arxiv.org/abs/2509.05333
- https://arxiv.org/abs/2503.06003