SOVEREIGN: Does MoE-LLaVA's routing strategy improve cross-modal robustness to textual adversarial perturbations (e.g., s

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Following the success in advancing natural language processing and understanding, transformers are expected to bring revolutionary changes to computer vision. This work provides a comprehensive study on the robustness of vision transformers (ViTs) against adversarial perturbations. Tested on various white-box and transfer attack settings, we find that ViTs possess better adversarial robustness when compared with MLP-Mixer and convolutional neural networks (CNNs) including ConvNeXt, and this observation also holds for certified robustness. Through frequency analysis and feature visualization, w

1 Introduction

Analysis of: On the Adversarial Robustness of Vision Transformers. Research goal: Does MoE-LLaVA's routing strategy improve cross-modal robustness to textual adversarial perturbations (e.g., synonym substitution, typographical attacks) on the MMMU benchmark compared to dense and modality-agnostic MoE baselines, and how does this vary with model scale (7B vs. 13B)?.

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. 0 claims extracted, 0 verified. Tribunal: 2.2/10 → 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

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2605.15484v1
- http://arxiv.org/abs/2103.15670v3