SOVEREIGN: Does SMoES routing with soft modality guidance generalize robustness to unseen adversarial perturbations on mu

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Machine learning (ML) systems have introduced significant advances in various fields, due to the introduction of highly complex models. Despite their success, it has been shown multiple times that machine learning models are prone to imperceptible perturbations that can severely degrade their accuracy. So far, existing studies have primarily focused on models where supervision across all classes were available. In constrast, Zero-shot Learning (ZSL) and Generalized Zero-shot Learning (GZSL) tasks inherently lack supervision across all classes. In this paper, we present a study aimed on evaluat

1 Introduction

Analysis of: A Deep Dive into Adversarial Robustness in Zero-Shot Learning. Research goal: Does SMoES routing with soft modality guidance generalize robustness to unseen adversarial perturbations on multimodal reasoning benchmarks like MathVista and MMBench compared to hard modality-agnostic routing?.

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

11 papers retrieved. 8 claims extracted, 0 verified. Tribunal: 3.7/10 → 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.

5 Extracted Claims

Claim	Verified	Confidence
Evaluation is performed on three widely used ZSL/GZSL datasets: Caltech-UCSD-Birds 200-2011 (CUB), Animals with Attribut	×	0.06
CUB is a medium-sized fine-grained dataset with 312 attributes, 200 classes, and a total of 11788 images.	×	0.01
SUN is a medium-sized fine-grained dataset with 102 attributes, 717 classes, and a total of 14340 images.	×	0.01
AWA2 is a larger-scale dataset with 85 attributes, 50 classes, and a total of 37322 images.	×	0.01
The ResNet-101 feature extractor is used to produce AWA2 dataset embeddings.	×	0.01
PyTorch is used for the experiments.	×	0.03
The ALE model is formulated as F(x, y; W) = θ(x)W^T φ(y) where θ(x) is the visual and φ(y) is the class embeddings.	×	0.03
This defense is inherently more effective against l0-norm attacks.	×	0.03

References

- http://arxiv.org/abs/2008.07651v1
- http://arxiv.org/abs/1712.07019v1
- http://arxiv.org/abs/2009.13954v2