SOVEREIGN: How robust are SMoES-based 7B VLMs to distribution shifts in visual inputs on SEED-Bench compared to dense and

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Vision-Language Models (VLMs) are increasingly used as perceptual modules for visual content reasoning, including through captioning and DeepFake detection. In this work, we expose a critical vulnerability of VLMs when exposed to subtle, structured perturbations in the frequency domain. Specifically, we highlight how these feature transformations undermine authenticity/DeepFake detection and automated image captioning tasks. We design targeted image transformations, operating in the frequency domain to systematically adjust VLM outputs when exposed to frequency-perturbed real and synthetic ima

1 Introduction

Analysis of: On the Reliability of Vision-Language Models Under Adversarial Frequency-Domain Perturbations. Research goal: How robust are SMoES-based 7B VLMs to distribution shifts in visual inputs on SEED-Bench compared to dense and hard-routing MoE baselines, measured via accuracy degradation under adversarial perturbations and domain variations?.

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, 0 verified. Tribunal: 3.0/10 → 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
The method enables manipulation of VLM captioning without introducing perceptible artifacts.	×	0.06
High-spatial frequency perturbations are applied to evaluate changes in caption verbosity using Δlength(YVLM) and ΔNtoke	×	0.07
The Qwen2-VL-7B-Instruct model shows a verbosity drift (Δlength(YVLM)) of 6.1467±93.6376 for SD3.5-Fantasy generated ima	×	0.02
For the COCO-2017 real dataset, the Qwen2-VL-7B-Instruct model exhibits a ΔNtokens(YVLM) value of -6.3925±12.6866.	×	0.02
The YVLM drift for CIFAKE generated images is 0.0502±0.1053 when evaluated with the Qwen2-VL-7B-Instruct model.	×	0.02
In the high-realism setting, the base image task achieves a Pr(I $>$ τ2) score of 0.6745 for real images.	×	0.03

References

- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2506.05429v1
- http://arxiv.org/abs/2103.15670v3