How does the corruption robustness of audio-visual deception detection models compare to vision-language model

Assignee Research

June 10, 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

This paper examines: On the Reliability of Vision-Language Models Under Adversarial Frequency-Domain Perturbations. Research question: How does the corruption robustness of audio-visual deception detection models compare to vision-language models on perturbed multimodal benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.0/10.

3 Results

10 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.0/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

References

- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2508.19294v2