Sparse-Only-1.7B Robustness Decline Under Frequency-Domain DeepFake Attacks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the performance drop of Sparse-Only-1.7B on DeepFake detection tasks when visual inputs are altered with structured frequency-domain attacks. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: On the Reliability of Vision-Language Models Under Adversarial Frequency-Domain Perturbations. Research question: What is the performance drop of Sparse-Only-1.7B on DeepFake detection tasks when visual inputs are altered with structured frequency-domain attacks?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
The proposed method enables the manipulation of Vision-Language Model (VLM) captioning without introducing perceptible a	×	0.09
For the Qwen2-VL-7B-Instruct model on the SD3.5-Fantasy dataset, the mean change in caption string length (Δlength) is 6	×	0.01
For the Qwen2-VL-7B-Instruct model on the SD1.5-ImageNet dataset, the mean change in caption string length (Δlength) is	×	0.01
For the Qwen2-VL-7B-Instruct model on the CIFAKE dataset, the mean YVLM drift is 0.0502.	×	0.01
For the Qwen2-VL-7B-Instruct model on the COCO-2017 dataset, the mean change in caption string length (Δlength) is -34.6	×	0.01
Applying high-spatial frequency perturbation to SD3.5-Fantasy generated images increased the mean perceived realness sco	×	0.06
Applying high-spatial frequency perturbation to COCO-2017 real images decreased the mean perceived realness score (Pr) f	×	0.06
Under the 'High realism' perturbation setting (α1=0.85, α2=1.00), the proportion of generated images classified as real	×	0.03
Under the 'High realism' perturbation setting, the YVLM drift for generated images is 0.0912 $\pm$ 0.1230.	×	0.02
The study evaluates changes in caption verbosity using metrics Δlength(YVLM) and ΔNtokens(YVLM).	×	0.02

References

- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2604.03558v1
- http://arxiv.org/abs/2305.06564v4