

Sparse-Only-1.7B Performance Degradation on HumanEval-V Under Adversarial Noise

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the performance degradation of Sparse-Only-1.7B on HumanEval-V tasks when adversarial noise is introduced to the visual context. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. 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 degradation of Sparse-Only-1.7B on HumanEval-V tasks when adversarial noise is introduced to the visual context?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.5/10.

3 Results

14 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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	×	0.08
For the Qwen2-VL-7B-Instruct model on the SD3.5-Fantasy dataset, the mean change in caption string length (Δ length) is 6	×	0.01
For the Qwen2-VL-7B-Instruct model on the SD1.5-ImageNet dataset, the mean change in caption string length (Δ length) is	×	0.02
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 (Δ length) is -34.6	×	0.01
Applying high-spatial frequency perturbation to SD3.5-Fantasy generated images increases the mean probability score $\Pr(I)$	×	0.06
Applying high-spatial frequency perturbation to COCO-2017 real images decreases the mean probability score $\Pr(I)$ from 9.	×	0.06
Under the 'High realism' configuration ($\alpha_1=0.85$, $\alpha_2=1.00$), the proportion of generated images classified as real ($\Pr(I)$)	×	0.03
Under the 'High realism' configuration, the YVLM drift for generated images is 0.0912 ± 0.1230 .	×	0.02
The study utilizes the Qwen2-VL-7B-Instruct and Qwen2-VL-2B-Instruct models for evaluation.	×	0.03

References

- <http://arxiv.org/abs/2207.02159v4>
- <http://arxiv.org/abs/2507.22398v3>
- <http://arxiv.org/abs/2306.02080v3>