Adversarial Robustness of Large Multimodal Models on HumanEval-V Tasks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How robust are LMMs on HumanEval-V tasks when evaluated under adversarial conditions (e.g., perturbed diagrams or ambiguous instructions), and how do different alignment techniques mitigate. 9 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Multimodal Adversarial Defense for Vision-Language Models by Leveraging One-To-Many Relationships. Research question: How robust are LMMs on HumanEval-V tasks when evaluated under adversarial conditions (e.g., perturbed diagrams or ambiguous instructions), and how do different alignment techniques mitigate performance degradation?.

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 4.5/10.

3 Results

14 papers retrieved. 9 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods, FARE	×	0.10
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.05
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.04
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	×	0.08
MAT is designed to be both effective and efficient.	×	0.03
MAT leverages one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.13
Multimodal attacks, which perturb both image and text modalities, are significantly more effective than unimodal attacks	✓	0.17
Existing defense strategies for VL models mainly focus on vision robustness, in which adversarial attacks perturb only t	✓	0.26
Mao et al. [21] and Wang et al. [34] approached zero-shot image classification on CLIP by proposing robust fine-tuning m	×	0.04

References

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2103.15670v3