One-to-Many Alignment Strategies and Robustness in Multimodal Large Language Models under Adversarial Perturbations

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How do one-to-many alignment strategies impact the robustness of multimodal large language models on the HumanEval-V benchmark under text-only adversarial perturbations. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. 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 do one-to-many alignment strategies impact the robustness of multimodal large language models on the HumanEval-V benchmark under text-only adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

12 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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	×	0.14
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.01
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.02
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	×	0.09
MAT T $\rightarrow$I (Cross, PGD-2) (Cross, BERT) All 83.7 67.5 77.4 61.4 72.2 51.1 37.5 24.8 8.79 -	×	0.02
TeCoA-ITR I (Cross, PGD-10) - All 83.1 68.2 77.7 61.9 64.7 42.7 27.5 17.6 10.29 $\times$ 1.17	×	0.03
Finetune 92.1 77.2 0.6 0.6 66.6 50.1 0.1 0.1	×	0.00
FARE 75.9 61.0 27.1 21.0 45.2 32.3 9.1 6.9	×	0.00
TeCoA-ITR 83.1 68.2 27.5 17.6 58.0 41.6 9.6 6.2	×	0.00
Finetune 89.5 77.7 2.5 1.3 69.9 53.6 1.0 0.7	×	0.00
TeCoA-ITR 85.4 69.3 35.5 21.9 64.8 48.6 14.2 9.5	×	0.01
Finetune 72.9 57.5 1.2 1.1	×	0.00
TeCoA-ITR 64.6 51.8 20.2 13.9	×	0.00

References

- http://arxiv.org/abs/2502.07987v3
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2408.09798v1