One-to-Many Relationship-Aware Defenses Enhance Vision-Language Model Alignment

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: Can one-to-many relationship-aware defenses in vision-language models improve alignment between visual and textual representations, as evaluated by retrieval performance on the Flickr30K benchmark. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. 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: Can one-to-many relationship-aware defenses in vision-language models improve alignment between visual and textual representations, as evaluated by retrieval performance on the Flickr30K benchmark under adversarial conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 16 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

16 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.7/10.

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	×	0.09
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.06
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.04
MAT largely improves multimodal robustness.	×	0.06
Adversarial images are generated via 2-step-PGD (perturbation size of 2/255 in l∞-norm).	×	0.04
Adversarial texts are generated using BERT-attack (1-token perturbation).	×	0.04
Intra-modal augmentation enhances data points without considering image-text interactions (text $\rightarrow$text, image $\rightarrow$image).	×	0.06
Cross-modal augmentation enhances data points by leveraging the other modality (image $\leftrightarrow$text).	×	0.11
EDA [35] is used for basic word-level edits in text augmentation.	×	0.02
LLM-based rewriting [8] is used for text augmentation.	×	0.03
Image-to-text (I2T) generation is used for cross-modal augmentation.	×	0.12
MAT is designed to be both effective and efficient through extensive analysis.	×	0.04
MAT leverages one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.11
Unimodal attacks, such as gradient-based image attacks [20] and BERT-Attack for text [17], perturb a single modality to	×	0.06
Multimodal attacks, which perturb both image and text modalities, are significantly more effective [11, 19, 33, 37].	×	0.11
Existing defense strategies for VL models mainly focus on vision robustness, in which adversarial attacks perturb only t	✓	0.24

References

- http://arxiv.org/abs/2601.03594v1
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2403.10883v2