Adaptive Weighting Schemes and Contrastive Losses in Adversarial Robustness of Vision-Language Models

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: What is the comparative impact of adaptive weighting schemes versus standard contrastive losses on the alignment robustness of vision-language models when evaluated against adversarial attacks on. 15 claims were extracted from source literature; 1 was 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: What is the comparative impact of adaptive weighting schemes versus standard contrastive losses on the alignment robustness of vision-language models when evaluated against adversarial attacks on image-text retrieval tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 10 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

10 papers retrieved. 15 claims extracted; 1 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
The study evaluates defense methods against the multimodal adversarial attack SGA with perturbation constraints of ϵI =	×	0.11
FARE is an unsupervised unimodal adversarial fine-tuning scheme for CLIP that focuses on obtaining a robust CLIP vision	×	0.03
TeCoA-ITR fine-tunes all parameters using a cross-modal objective to generate adversarial images, whereas the original T	×	0.06
The models CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B were fine-tuned using the proposed MAT method.	×	0.06
Adversarial images in the training process were generated via 2-step-PGD with a perturbation size of 2/255 in l∞-norm.	×	0.04
Adversarial texts in the training process were generated using BERT-attack with a 1-token perturbation.	×	0.05
Intra-modal augmentation enhances data points without considering image-text interactions (e.g., text $\rightarrow$ text, image $\rightarrow$ im	×	0.07
Cross-modal augmentation enhances data points by leveraging the other modality (image $\leftrightarrow$ text).	×	0.08
EDA is used as an intra-modal text augmentation technique for basic word-level edits.	×	0.03
MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods FARE	×	0.08
MAT consistently achieves significantly greater robustness against multimodal attacks than unimodal AT methods on ALBEF	×	0.09
Unimodal attacks perturb a single modality to mislead models, while multimodal attacks perturb both image and text modal	×	0.12
Multimodal attacks are significantly more effective than unimodal attacks on vision-language models.	×	0.15
Existing defense strategies for vision-language models mainly focus on vision robustness where adversarial attacks pertu	✓	0.23
The proposed MAT method leverages one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.10

References

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2504.09480v1
- http://arxiv.org/abs/2403.10883v2