Can multimodal contrastive learning (e.g., using ALIGN or CLIP) enhance the robustness of synthetic data-gener

Assignee Research

June 10, 2026

Abstract

Pre-trained vision-language (VL) models are highly vulnerable to adversarial attacks. However, existing defense methods primarily focus on image classification, overlooking two key aspects of VL tasks: multimodal attacks, where both image and text can be perturbed, and the one-to-many relationship of images and texts, where a single image can correspond to multiple textual descriptions and vice versa (1:N and N:1). This work is the first to explore defense strategies against multimodal attacks in VL tasks, whereas prior VL defense methods focus on vision robustness. We propose multimodal adver

1 Introduction

This paper examines: Multimodal Adversarial Defense for Vision-Language Models by Leveraging One-To-Many Relationships. Research question: Can multimodal contrastive learning (e.g., using ALIGN or CLIP) enhance the robustness of synthetic data-generated models against adversarial attacks, as evaluated by adversarial accuracy on the RSICD dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

15 papers retrieved. 15 claims extracted; 1 independently verified. Quality review score: 4.8/10.

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	×	0.04
TeCoA-ITR fine-tunes all parameters using a cross-modal objective to generate adversarial images, whereas the original T	×	0.07
The models CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B were fine-tuned using MAT.	×	0.06
Adversarial images in the training process were generated via 2-step-PGD with a perturbation size of 2/255 in l∞-norm.	×	0.03
Adversarial texts in the training process were generated using BERT-attack with a 1-token perturbation.	×	0.06
Intra-modal augmentation enhances data points without considering image-text interactions.	×	0.08
Cross-modal augmentation enhances data points by leveraging the other modality (image $\leftrightarrow$text).	×	0.12
EDA is used for basic word-level edits as an intra-modal text augmentation technique.	×	0.04
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.	×	0.12
Existing defense strategies for VL models mainly focus on vision robustness where adversarial attacks perturb only the i	✓	0.24
The proposed MAT method leverages one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.11

References

- http://arxiv.org/abs/2603.09625v2
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2403.10883v2