Multimodal Model Robustness Under Domain-Specific and General Adversarial Training

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the difference in robustness scores between multimodal models trained on domain-specific adversarial images versus general adversarial training when evaluated on MME benchmark. 13 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: What is the difference in robustness scores between multimodal models trained on domain-specific adversarial images versus general adversarial training when evaluated on MME benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

9 papers retrieved. 13 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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	×	0.08
MAT consistently achieves significantly greater robustness against multimodal attacks than unimodal AT methods on ALBEF	×	0.09
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	×	0.04
The models CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B were fine-tuned using MAT with adversarial images generated via 2	×	0.04
Intra-modal augmentation enhances data points without considering image-text interactions, while cross-modal augmentatio	×	0.07
EDA is used as an intra-modal text augmentation technique for basic word-level edits.	×	0.03
Unimodal attacks perturb a single modality to mislead models, whereas multimodal attacks perturb both image and text mod	×	0.15
Existing defense strategies for VL models mainly focus on vision robustness where adversarial attacks perturb only the i	✓	0.25
Solving the inner-maximization in Eq. 6 requires updating both modalities and involves a high computational cost.	×	0.02
Table (p6) reports a Finetune score of 92.1 and a FARE score of 75.9 for the first metric listed.	×	0.01
Table (p7) reports that MAT T $\rightarrow$I (Cross, PGD-2) achieved a score of 72.2 compared to TeCoA-ITR's 64.7 in one of the repo	×	0.01

References

- http://arxiv.org/abs/1905.11736v5
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2103.01400v3