

One-to-Many Defense Strategy Effects on Zero-Shot Image-Text Retrieval Under Multimodal Adversarial Attacks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the one-to-many relationship defense strategy impact zero-shot image-text retrieval accuracy on MS-COCO under simultaneous multimodal adversarial perturbations compared to standard. 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. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. 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 does the one-to-many relationship defense strategy impact zero-shot image-text retrieval accuracy on MS-COCO under simultaneous multimodal adversarial perturbations compared to standard adversarial training?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

13 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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	×	0.13
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.04
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.03
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	×	0.10
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 -	×	0.03
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	×	0.02
TeCoA-ITR 83.1 68.2 27.5 17.6 58.0 41.6 9.6 6.2	×	0.02
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	×	0.00
Finetune 72.9 57.5 1.2 1.1	×	0.00
TeCoA-ITR 64.6 51.8 20.2 13.9	×	0.02

References

- <http://arxiv.org/abs/2110.06166v4>

- <http://arxiv.org/abs/2402.18400v2>
- <http://arxiv.org/abs/2405.18770v6>