Multimodal Dataset Distillation and Adversarial Robustness in Vision-Language Models

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of multimodal dataset distillation on adversarial robustness when evaluated using both textual and visual adversarial attacks on models like CLIP or BLIP. 13 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. 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 impact of multimodal dataset distillation on adversarial robustness when evaluated using both textual and visual adversarial attacks on models like CLIP or BLIP?.

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 2.8/10.

3 Results

13 papers retrieved. 13 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods, FARE	$\times$	0.12
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	$\times$	0.06
The improvements are substantial and consistent for ALBEF on both datasets.	$\times$	0.04
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	$\times$	0.10
MAT T $\rightarrow$I (Cross, PGD-2) (Cross, BERT) All achieves 83.7, 67.5, 77.4, 61.4, 72.2, 51.1, 37.5, 24.8, 8.79.	$\times$	0.02
TeCoA-ITR I (Cross, PGD-10) - All achieves 83.1, 68.2, 77.7, 61.9, 64.7, 42.7, 27.5, 17.6, 10.29 $\times$ 1.17.	$\times$	0.01
Finetune achieves 92.1, 77.2, 0.6, 0.6, 66.6, 50.1, 0.1, 0.1.	$\times$	0.00
FARE achieves 75.9, 61.0, 27.1, 21.0, 45.2, 32.3, 9.1, 6.9.	$\times$	0.01
TeCoA-ITR achieves 83.1, 68.2, 27.5, 17.6, 58.0, 41.6, 9.6, 6.2.	$\times$	0.00
Finetune achieves 89.5, 77.7, 2.5, 1.3, 69.9, 53.6, 1.0, 0.7.	$\times$	0.00
TeCoA-ITR achieves 85.4, 69.3, 35.5, 21.9, 64.8, 48.6, 14.2, 9.5.	$\times$	0.00
Finetune achieves 72.9, 57.5, 1.2, 1.1.	$\times$	0.00
TeCoA-ITR achieves 64.6, 51.8, 20.2, 13.9.	$\times$	0.00

References

- http://arxiv.org/abs/2405.18770v6

- http://arxiv.org/abs/2110.06166v4
- http://arxiv.org/abs/2403.10883v2