Adversarial Training Effects on Clean Accuracy in Multimodal CodeT5 Variants

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does adversarial training impact the clean accuracy of multimodal CodeT5 variants on the MBXP Python benchmark compared to unimodal baselines. 14 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: How does adversarial training impact the clean accuracy of multimodal CodeT5 variants on the MBXP Python benchmark compared to unimodal baselines?.

2 Methodology

Systematic literature search across multiple databases yielded 8 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

8 papers retrieved. 14 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
MAT consistently achieves significantly greater robustness against multimodal attacks than the unimodal AT methods, FARE	×	0.11
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.00
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.00
We evaluate defense methods against the multimodal adversarial attack SGA [19], with perturbation constraints of ϵI = 2/	×	0.09
We fine-tune the pre-trained CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B models using MAT.	×	0.05
Adversarial images are generated via 2-step-PGD (perturbation size of 2/255 in l∞-norm), and adversarial texts using BER	×	0.06
We consider two types of augmentations: intra-modal and cross-modal.	×	0.01
Intra-modal augmentation enhances data points without considering image-text interactions (text $\rightarrow$text, image $\rightarrow$image), wh	×	0.09
Text augmentations include EDA [35] for basic word-level edits and LLM-based rewriting [8].	×	0.03
Cross-modal augmentation includes Image-to-text (I2T) generation using different models.	×	0.06
MAT largely improves multimodal robustness, highlighting the importance of considering multimodal perturbations in VL da	×	0.09
We leverage one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.12
Multimodal attacks, which perturb both image and text modalities, are significantly more effective [11, 19, 33, 37].	×	0.13
Existing defense strategies for VL models mainly focus on vision robustness, in which adversarial attacks perturb only t	✓	0.25

References

- http://arxiv.org/abs/2103.01400v3
- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2205.14230v2