Multimodal Adversarial Robustness and Generation Quality Trade-offs in Text-to-Image Code Synthesis

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the trade-off between robustness and generation quality for multimodal code models when evaluated on text-to-image code synthesis tasks under adversarial perturbations. 14 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.3/10. 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: What is the trade-off between robustness and generation quality for multimodal code models when evaluated on text-to-image code synthesis tasks under adversarial perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.3/10.

3 Results

14 papers retrieved. 14 claims extracted; 3 independently verified. Quality review score: 5.3/10.

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	×	0.08
The improvements are substantial and consistent for CLIP on Flickr30k and COCO.	×	0.02
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.03
MAT largely improves multimodal robustness.	×	0.06
MAT is both effective and efficient.	×	0.04
MAT highlights the importance of considering multimodal perturbations in VL data.	×	0.08
MAT leverages one-to-many (1:N) image-text relationships via augmentations to enhance robustness.	×	0.14
Unimodal adversarial training assumes a deterministic image-to-label mapping.	×	0.10
Multimodal attacks, which perturb both image and text modalities, are significantly more effective.	✓	0.17
Developing defense strategies against multimodal attacks for VL tasks remains largely unexplored.	✓	0.20
Existing defense strategies for VL models mainly focus on vision robustness.	✓	0.24
Adversarial attacks on VL models are categorized into unimodal and multimodal.	×	0.14
Unimodal attacks, such as gradient-based image attacks and BERT-Attack for text, perturb a single modality to mislead th	×	0.07
Mao et al. and Wang et al. approached zero-shot image classification on CLIP by proposing robust fine-tuning methods.	×	0.06

References

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2604.16532v1
- http://arxiv.org/abs/2511.18488v2