One-to-Many Image-Text Training Enhances CLIP Robustness Against Joint Spectral Adversarial Attacks

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 18 peer-reviewed papers addressing the following research question: How does training with one-to-many image-text pairs affect the robustness accuracy of CLIP-based models under joint spectral adversarial perturbations compared to standard adversarial training. 12 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: A Multimodal Adversarial Attack Method via Frequency Domain Enhancement and Fine-Grained Cross-Modal Guidance. Research question: How does training with one-to-many image-text pairs affect the robustness accuracy of CLIP-based models under joint spectral adversarial perturbations compared to standard adversarial training?.

2 Methodology

Systematic literature search across multiple databases yielded 18 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.3/10.

3 Results

18 papers retrieved. 12 claims extracted; 10 independently verified. Quality review score: 7.3/10.

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
Vision-language pretraining (VLP) models are highly susceptible to transferable adversarial attacks.	✓	0.26
Ensemble-based guided attacks primarily rely on spatial-domain data augmentation.	✓	0.25
Reliance on spatial-domain data augmentation in ensemble-based attacks can lead to model overfitting to image details.	✓	0.23
Reliance on spatial-domain data augmentation limits the generalization capability of adversarial attacks.	✓	0.17
The proposed method modifies specific frequency components of input images.	✓	0.24
Modifying specific frequency components reduces detail interference.	×	0.10
Modifying specific frequency components enhances the stability of adversarial examples.	✓	0.15
The proposed method introduces a fine-grained feature extraction technique to optimize image-text alignment.	✓	0.26
Optimizing image-text alignment via fine-grained feature extraction improves the transferability of cross-modal attacks.	✓	0.28
The proposed method achieves superior attack transferability compared to baseline methods across fusion and alignment VL	✓	0.21
The proposed method achieves superior generalization performance across fusion and alignment VLP architectures.	✓	0.22
Experimental evaluation was conducted on the Flickr30K and MSCOCO datasets.	×	0.06

References

- https://arxiv.org/abs/2403.01944
- https://www.semanticscholar.org/paper/97d27200feda2e61466bfc69020e16b3166c1e51
- https://www.semanticscholar.org/paper/01dd34924be8e46e0dfed9ddf35a64b18dc7ec3e