Robustness of Language-Multimodal Radio Models to Adversarial Perturbations in Wireless Sensing

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: How robust are current LMRMs to adversarial perturbations in wireless signal-sensing alignment tasks, as quantified by accuracy degradation metrics under controlled adversarial conditions. 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; 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 robust are current LMRMs to adversarial perturbations in wireless signal-sensing alignment tasks, as quantified by accuracy degradation metrics under controlled adversarial conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 13 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	×	0.07
MAT consistently achieves significantly greater robustness against multimodal attacks than unimodal AT methods on ALBEF	×	0.08
The study evaluates defense methods against the multimodal adversarial attack SGA with perturbation constraints of ϵI =	×	0.10
FARE is an unsupervised unimodal adversarial fine-tuning scheme for CLIP that focuses on obtaining a robust CLIP vision	×	0.03
TeCoA-ITR fine-tunes all parameters using a cross-modal objective to generate adversarial images, whereas the original T	×	0.04
The models CLIP-ViT-B/16, ALBEF-14M, and BLIP w/ ViT-B were fine-tuned using MAT with adversarial images generated via 2	×	0.06
Intra-modal augmentation enhances data points without considering image-text interactions, while cross-modal augmentatio	×	0.06
EDA is used as an intra-modal text augmentation technique for basic word-level edits.	×	0.02
Unimodal attacks perturb a single modality to mislead models, whereas multimodal attacks perturb both image and text mod	×	0.13
Existing defense strategies for vision-language models mainly focus on vision robustness where adversarial attacks pertu	✓	0.23
On the Flickr30k dataset using CLIP, the Fine-tune baseline achieved a robust accuracy of 0.6% against multimodal attacks	×	0.04
On the COCO dataset using CLIP, the Fine-tune baseline achieved a robust accuracy of 0.1% against multimodal attacks, whi	×	0.04
The MAT method (specifically MAT T $\rightarrow$I with Cross, PGD-2 and Cross, BERT) achieved a robust accuracy of 37.5% on Flickr30	×	0.03

References

- http://arxiv.org/abs/2104.09369v1
- http://arxiv.org/abs/2503.14504v2
- http://arxiv.org/abs/2405.18770v6