How does the robustness of CLAM's latent action models to noisy or occluded inputs vary between multimodal and

Assignee Research

June 10, 2026

Abstract

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 one-to-many relationship of images and texts, where a single image can correspond to multiple textual descriptions and vice versa (1:N and N:1). This work is the first to explore defense strategies against multimodal attacks in VL tasks, whereas prior VL defense methods focus on vision robustness. We propose multimodal adver

1 Introduction

This paper examines: Multimodal Adversarial Defense for Vision-Language Models by Leveraging One-To-Many Relationships. Research question: How does the robustness of CLAM's latent action models to noisy or occluded inputs vary between multimodal and unimodal training, as evaluated by degradation in task completion rates on BridgeData V2 under adversarial conditions?.

2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

10 papers retrieved. 16 claims extracted; 4 independently verified. Quality review score: 5.2/10.

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.	×	0.00
The improvements are substantial and consistent for ALBEF on both datasets.	×	0.02
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.	×	0.15
Unimodal adversarial training assumes a deterministic image-to-label mapping.	×	0.12
Multimodal attacks are significantly more effective than unimodal attacks.	×	0.12
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.	✓	0.23
Adversarial attacks on VL models are categorized into unimodal and multimodal.	✓	0.15
Unimodal attacks perturb a single modality to mislead the models.	×	0.04
Multimodal attacks perturb both image and text modalities.	✓	0.19
Adversarial attacks on VL models include gradient-based image attacks and BERT-Attack for text.	×	0.12
Mao et al. and Wang et al. proposed robust fine-tuning methods for zero-shot image classification on CLIP.	×	0.04

References

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

- http://arxiv.org/abs/2505.04999v1
- http://arxiv.org/abs/2206.13405v2