Does the selection of optimizer algorithms affect the robustness of vision-language models against adversarial

Assignee Research

June 10, 2026

Abstract

Vision-Language Model (VLM) have gained widespread adoption in Open-Vocabulary (OV) object detection and segmentation tasks. Despite they have shown promise on OV-related tasks, their effectiveness in conventional vision tasks has thus far been unevaluated. In this work, we present the systematic review of VLM-based detection and segmentation, view VLM as the foundational model and conduct comprehensive evaluations across multiple downstream tasks for the first time: 1) The evaluation spans eight detection scenarios (closed-set detection, domain adaptation, crowded objects, etc.) and eight seg

1 Introduction

This paper examines: Vision-Language Model for Object Detection and Segmentation: A Review and Evaluation. Research question: Does the selection of optimizer algorithms affect the robustness of vision-language models against adversarial perturbations in instance segmentation benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.3/10.

3 Results

9 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
Object detection and segmentation are fundamental tasks in computer vision, serving as essential components for percepti	×	0.11
Object detection and segmentation technologies form the backbone of various practical applications across multiple domai	×	0.07
Current VLMs fundamentally operate by aligning visual and textual features to achieve their broad and robust capabilitie	×	0.03
In object detection tasks, VLM-based detection aligns visual features with text descriptions through contrastive learnin	×	0.13
GLIP and GroundingDINO achieve generalization across unseen categories through pretraining on large-scale datasets such	×	0.07
In segmentation tasks, recent works have focused on transferring global multi-modal alignment capabilities of VLMs to fi	×	0.05
These advancements leverage diverse supervision strategies to facilitate dense prediction in pixel-wise segmentation tas	×	0.06
Current VLMs predominantly demonstrate strong performance on open-vocabulary (OV) tasks.	×	0.10
DA-Pro builds upon RegionCLIP by dynamically generating domain-specific detection heads through domain-relevant and doma	×	0.04
COUNTGD improves instance counting by augmenting the text prompts in GroundingDINO with visual exemplars of correspondin	×	0.02
GroundingDino (swin-T) achieves an APnovel of 57.1, APbase of 47.5, and AP of 49.8 on OV-COCO.	×	0.02
GroundingDino (swin-B) achieves an APnovel of 62.7, APbase of 55.4, and AP of 57.3 on OV-COCO.	×	0.02
GLIP-T (A) achieves an APnovel of 65.9, APbase of 58.2, and AP of 60.3 on OV-COCO.	×	0.02
GLIP-T (B) achieves an APnovel of 69.8, APbase of 59.3, and AP of 62.0 on OV-COCO.	×	0.03
GLIP-T (C) achieves an APnovel of 69.3, APbase of 60.5, and AP of 62.8 on OV-COCO.	×	0.02
Region CLIP (Res50) achieves an 4APnovel of 25.2, APbase of 31.4, and AP of 26.8 on OV-COCO.	×	0.02

References

- http://arxiv.org/abs/2504.09480v1
- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2410.20971v2