Scaling Sub-Module Attention Mechanisms and Zero-Shot Detection Robustness in Multimodal Models

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the effect of scaling the sub-module attention mechanism on the robustness of zero-shot detection performance (using AP scores) in large-scale multimodal models like CLIP or BLIP. 8 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Object Detection with Multimodal Large Vision-Language Models: An In-depth Review. Research question: What is the effect of scaling the sub-module attention mechanism on the robustness of zero-shot detection performance (using AP scores) in large-scale multimodal models like CLIP or BLIP?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

15 papers retrieved. 8 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
Traditional object detection techniques like Background Subtraction and Color Segmentation are susceptible to changes in	×	0.06
Machine Learning and Deep Learning methods have transformed object detection tasks over the past 15 years.	×	0.10
Modern ML/DL object detection models significantly surpass the capabilities of traditional methods.	×	0.08
Single Shot MultiBox Detector (SSD) processes images in one shot to detect objects, delivering locations and class predi	×	0.04
YOLO divides images into grids where each grid predicts bounding boxes and probabilities, enabling rapid real-time detec	×	0.05
Fast R-CNN and Faster R-CNN enhance detection by using region proposal networks and shared convolutional features.	×	0.05
Mask R-CNN builds on Faster R-CNN by adding a segmentation overlay that provides precise pixel-level object outlines.	×	0.06
RetinaNet uses a focal loss to focus on hard-to-detect objects and balance the detection of various object sizes.	×	0.03

References

- http://arxiv.org/abs/2410.01534v2
- http://arxiv.org/abs/2504.09480v1
- http://arxiv.org/abs/2508.19294v2