EfficientViT and Standard ViT Backbones in Multimodal Models for False Negative Reduction

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of using EfficientViT versus standard ViT backbones in multimodal models on false negative reduction rates when evaluated on the Flickr30k and ECCV Caption benchmarks. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. Research question: What is the impact of using EfficientViT versus standard ViT backbones in multimodal models on false negative reduction rates when evaluated on the Flickr30k and ECCV Caption benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.5/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 6.5/10.

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.

References

- http://arxiv.org/abs/2204.03359v5
- http://arxiv.org/abs/2305.07027v1
- http://arxiv.org/abs/2602.08071v1