Multimodal Model Accuracy Under Varying Visual Encoder Resolutions in Code Generation

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of visual encoder resolution on the accuracy of multimodal models when interpreting complex diagrams in code generation tasks. 7 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: Investigating Redundancy in Multimodal Large Language Models with Multiple Vision Encoders. Research question: What is the impact of visual encoder resolution on the accuracy of multimodal models when interpreting complex diagrams in code generation tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 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

13 papers retrieved. 7 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
Eagle (Shi et al., 2024) represents MLLMs designed with a larger ensemble of encoders (typically 4 or 5), including CLIP	×	0.02
Cambrian-1 (Tong et al., 2024a) introduces a vision-centric approach with a novel fusion mechanism called Spatial Vision	×	0.02
Cambrian-1 uses cross-attention with learnable queries to integrate features from multiple encoders, including CLIP (Rad	×	0.02
The benchmark categorization proposed by Cambrian-1 (Tong et al., 2024a) groups common benchmarks into four distinct cat	×	0.06
All evaluations are performed using standardized protocols, primarily leveraging VLMEvalKit (Duan et al., 2025) for cons	×	0.02
Performance of multi-encoder MLLMs degrades gracefully rather than catastrophically when encoders are masked.	✓	0.18
Encoder redundancy is observed if removing one or more encoders does not harm or even improves performance.	×	0.11

References

- http://arxiv.org/abs/2507.03262v4
- http://arxiv.org/abs/2506.06276v1
- http://arxiv.org/abs/2408.07303v2