Visual Diagram Encoders in Multimodal Code Models: Performance on HumanEval-V

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the integration of visual diagram encoders in multimodal code models impact the accuracy of code generation tasks on HumanEval-V compared to text-only baselines like CodeLlama. Recent multimodal large language models (MLLMs) increasingly integrate multiple vision encoders to improve performance on various benchmarks, assuming that diverse pretraining objectives yield complementary visual signals. However, we show this assumption often fails in practice. 11 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: How does the integration of visual diagram encoders in multimodal code models impact the accuracy of code generation tasks on HumanEval-V compared to text-only baselines like CodeLlama?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 11 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
Eagle primarily uses channel concatenation for feature fusion.	×	0.02
Cambrian-1 (Tong et al., 2024a) introduces a vision-centric approach with a novel fusion mechanism called Spatial Vision	×	0.04
SVA uses cross-attention with learnable queries to integrate features from multiple encoders, including CLIP (Radford et	×	0.03
These model architectures provide an excellent testbed for investigating redundancy in systems with many specialized enc	×	0.09
To assess MLLM performance and analyze redundancy across diverse capabilities, we adopt the benchmark categorization pro	×	0.07
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.	✓	0.17
We consider MLLMs based on the prevalent 'ViT-adapter-LLM' architecture (Liu et al., 2024; Bai et al., 2025; Zhu et al.,	×	0.02
The output response Y of a multi-encoder MLLM with a set of n vision encoders En = {E1, . . . , En} is generated as: Y =	×	0.06
Encoder redundancy is observed if removing one or more encoders does not harm or even improves performance.	×	0.10

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2507.03262v4
- http://arxiv.org/abs/2407.04973v1