Multimodal Model Performance on HumanEval-V Across Diagram Modalities and Complexity Levels

Assignee Research

June 9, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the performance of multimodal models vary when evaluated on HumanEval-V with different diagram modalities (e.g., flowcharts, UML, graphs) while controlling for complexity. 13 claims were extracted from source literature; 1 was 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: HumanEval-V: Benchmarking High-Level Visual Reasoning with Complex Diagrams in Coding Tasks. Research question: How does the performance of multimodal models vary when evaluated on HumanEval-V with different diagram modalities (e.g., flowcharts, UML, graphs) while controlling for complexity?.

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 3.8/10.

3 Results

14 papers retrieved. 13 claims extracted; 1 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
HumanEval-V consists of 253 human-annotated coding tasks.	✓	0.17
Each task in HumanEval-V features a diagram, a function signature, and test cases.	×	0.12
HumanEval-V diagrams span six task types.	×	0.12
Claude 3.5 Sonnet achieves a 36.8% pass@1 score on HumanEval-V.	×	0.11
Pixtral 124B achieves a 21.3% pass@1 score on HumanEval-V.	×	0.03
Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples.	×	0.04
Claude 3.5 Sonnet reaches a 55.3% pass@1 score with four self-refining iterations based on test case execution feedback.	×	0.04
Experiments were conducted with 22 Large Multimodal Models (LMMs).	×	0.14
The evaluation pipeline includes a variant where the model generates a structured diagram description before coding.	×	0.05
The Intermediate Textual Representation method produces a problem specification consisting of Problem Restatement, Visua	×	0.04
GPT-4o achieves a 27.7% pass@1 score in the baseline setting according to Table (p5).	×	0.01
Gemini 1.5 Pro achieves a 22.9% pass@1 score in the baseline setting according to Table (p5).	×	0.02
Pixtral 124B achieves a 16.6% pass@1 score in the baseline setting according to Table (p5).	×	0.02

References

- http://arxiv.org/abs/2306.09265v1

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2506.02073v1