

Multimodal vs. Text-Only Performance of Large Multimodal Models on HumanEval-V

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How does the performance of LMMs vary when evaluated on HumanEval-V tasks with diagrams versus those with purely textual descriptions, as measured by accuracy and reasoning error rates. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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 LMMs vary when evaluated on HumanEval-V tasks with diagrams versus those with purely textual descriptions, as measured by accuracy and reasoning error rates?.

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 4.2/10.

3 Results

13 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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 encoding the problem context, a function signature defining the task’s input | × | 0.08 |
| The top-performing model, Claude 3.5 Sonnet, achieves 36.8% pass@1 on HumanEval-V. | × | 0.11 |
| The best open-weight model, Pixtral 124B, reaches 21.3% pass@1 on HumanEval-V. | × | 0.03 |
| Claude 3.5 Sonnet achieves a 74.3% pass rate with 100 samples on HumanEval-V. | × | 0.05 |
| Claude 3.5 Sonnet can reach 55.3% pass@1 with four self-refining iterations based on test case execution feedback on Hum | × | 0.05 |
| HumanEval-V offers a more diverse and complex set of diagrams spanning six task types. | × | 0.11 |
| The visual context must be essential for solving the task, with all relevant information contained in a single image. | × | 0.03 |
| Tasks should be designed around the visual context with minimal textual description. | × | 0.06 |
| The two-stage evaluation pipeline supports LMMs with limited coding abilities by first prompting them to generate a stru | × | 0.06 |
| Through extensive experiments with 22 LMMs, we observe the following key findings: (1) Our diagram D and function signat | × | 0.07 |

References

- <http://arxiv.org/abs/2309.10783v1>

- <http://arxiv.org/abs/2410.12381v3>
- <http://arxiv.org/abs/2505.06280v1>