Large Multimodal Models Trained on HumanEval-V Generalize to Visual Reasoning Benchmarks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do LMMs trained on HumanEval-V generalize to other visual reasoning benchmarks (e.g., DVQA, DiagramsVQA) in terms of cross-domain robustness and accuracy consistency. 12 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: HumanEval-V: Benchmarking High-Level Visual Reasoning with Complex Diagrams in Coding Tasks. Research question: How do LMMs trained on HumanEval-V generalize to other visual reasoning benchmarks (e.g., DVQA, DiagramsVQA) in terms of cross-domain robustness and accuracy consistency?.

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. 12 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
HumanEval-V consists of 253 human-annotated coding tasks.	×	0.15
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.	×	0.10
The best open-weight model, Pixtral 124B, reaches 21.3% pass@1 on HumanEval-V.	×	0.02
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	×	0.04
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types, demanding versatile capabilities	×	0.13
HumanEval-V uses code generation tasks for evaluation instead of the multiple-choice or short-answer questions commonly	×	0.12
The visual context in HumanEval-V must be essential for solving the task, with all relevant information contained in a s	×	0.05
Tasks in HumanEval-V should be designed around the visual context with minimal textual description.	×	0.04
HumanEval-V utilizes a two-stage evaluation pipeline that supports LMMs with limited coding abilities by first prompting	×	0.07
Extensive experiments with 22 LMMs were conducted on HumanEval-V.	✓	0.16

References

- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2508.17298v2
- http://arxiv.org/abs/2505.04921v2