Video-LLaVA-8B and LLaVA-NeXT Robustness to Visual Noise in Diagram-Based Code Generation

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: How do Video-LLaVA-8B and LLaVA-NeXT differ in robustness against visual noise when performing diagram-based code generation on the HumanEval-V dataset. 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: HumanEval-V: Benchmarking High-Level Visual Reasoning with Complex Diagrams in Coding Tasks. Research question: How do Video-LLaVA-8B and LLaVA-NeXT differ in robustness against visual noise when performing diagram-based code generation on the HumanEval-V dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 9 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

9 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
HumanEval-V consists of 253 human-annotated coding tasks.	✓	0.16
Each task in HumanEval-V features a diagram encoding the problem context, a function signature defining the task's input	×	0.07
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.	×	0.03
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	×	0.04
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types, demanding versatile capabilities	×	0.14
The visual context in HumanEval-V must be essential for solving the task, with all relevant information contained in a s	×	0.07
Tasks in HumanEval-V should be designed around the visual context with minimal textual description.	×	0.05
The two-stage evaluation pipeline in HumanEval-V supports LMMs with limited coding abilities by first prompting them to	×	0.07
Extensive experiments with 22 LMMs were conducted on HumanEval-V.	×	0.14

References

- http://arxiv.org/abs/2405.18770v6
- http://arxiv.org/abs/2410.12381v3
- http://arxiv.org/abs/2103.15670v3