Dynamic Graph Pooling Effects on Inference Latency in Large-Scale Code Generation

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the impact of dynamic graph pooling on the inference latency of large-scale code generation models when evaluated on the HumanEval benchmark with varying input sequence lengths. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.0/10. 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: What is the impact of dynamic graph pooling on the inference latency of large-scale code generation models when evaluated on the HumanEval benchmark with varying input sequence lengths?.

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 5.0/10.

3 Results

14 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 5.0/10.

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.	✓	0.18
HumanEval-V offers a more diverse and complex set of diagrams spanning six task types.	×	0.10
Claude 3.5 Sonnet achieves 36.8% pass@1 on HumanEval-V.	×	0.12
Pixtral 124B reaches 21.3% pass@1 on HumanEval-V.	×	0.03
Current LMMs exhibit stronger vision-to-language alignment than vision-to-code.	×	0.06
LMMs' performance can be enhanced through sampling or iterative self-refinement.	×	0.03
HumanEval-V uses code generation tasks for evaluation instead of multiple-choice or short-answer questions.	×	0.08
HumanEval-V requires comprehensive logical thinking and visual understanding with minimal chance of correct guesses.	×	0.05
The best performance of LMMs occurs when they serve as diagram describers, with GPT-4o acting as the coder model.	×	0.04

References

- http://arxiv.org/abs/2004.13542v4
- http://arxiv.org/abs/2602.02159v1
- http://arxiv.org/abs/2410.12381v3