Impact of Finetuning Dataset Size on Pass@1 Scores in HLCE Benchmark ICPC Problems

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What is the impact of finetuning dataset size on the pass@1 scores for ICPC World Finals problems in the HLCE benchmark when controlling for model parameter scale. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.1/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: What-If World: A Causal Benchmark for General World Models in Embodied Scenarios. Research question: What is the impact of finetuning dataset size on the pass@1 scores for ICPC World Finals problems in the HLCE benchmark when controlling for model parameter scale?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.1/10.

3 Results

4 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 4.1/10.

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
Existing video-generation benchmarks satisfy neither the state fixation requirement nor the single-variable change requi	×	0.07
Existing video-generation benchmarks only test each video model against a single prompt in isolation.	×	0.13
The What-If World benchmark construction pipeline filters clips from the nuScenes and DROID datasets.	×	0.07
The benchmark extracts the causal branching point (x0) at the action onset to lock the initial state.	×	0.04
The benchmark constructs contrastive prompt pairs (p+, p-) that share scene context and differ only in one physical vari	×	0.09
The outcome of the action is withheld from the prompts in the benchmark construction.	×	0.06
The evaluation uses a VLM judge based on Gemini 3.1 Pro.	×	0.01
The VLM judge receives video pairs (V+, V-) and prompt pairs (p+, p-) to answer primitive-conditioned binary questions.	×	0.06
The evaluation uses binary scoring instead of Likert scoring to avoid position, verbosity, and central-tendency biases.	×	0.02
The VLM judge was validated against 421 human-annotated samples.	×	0.02
The VLM judge agrees with human labels on 82.30% of decisions averaged across dimensions.	×	0.00
Grok Imagine achieved the highest rank (1) in the APEO Framework evaluation.	×	0.01
Grok Imagine achieved an Adherence score of 69.2 and a Physics score of 49.2.	×	0.05
Veo 3.1 achieved the second highest rank (2) with an average physics score (pAvg) of 50.9.	×	0.03
Seedance 2.0 achieved the third highest rank (3) with an Adherence score of 63.0.	×	0.03
CogVideoX1.5-5B achieved the lowest rank (9) with an average physics score (pAvg) of 22.7.	×	0.03

References

- http://arxiv.org/abs/2605.27589v1
- http://arxiv.org/abs/2405.19595v1
- http://arxiv.org/abs/hep-ex/0509008v3