Mistral-Large-2 Inference Efficiency Scaling on MBPP Code Generation

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 17 peer-reviewed papers addressing the following research question: How does the inference efficiency of Mistral-Large-2 scale with model size when generating code on the MBPP benchmark, as measured by tokens per second and latency metrics. Large-scale video generative models, capable of creating realistic videos of diverse visual concepts, are strong candidates for general-purpose physical world simulators. However, their adherence to physical commonsense across real-world actions remains unclear (e.g., playing. 10 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. Research question: How does the inference efficiency of Mistral-Large-2 scale with model size when generating code on the MBPP benchmark, as measured by tokens per second and latency metrics?.

2 Methodology

Systematic literature search across multiple databases yielded 17 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

17 papers retrieved. 10 claims extracted; 1 independently verified. Quality review score: 4.4/10.

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
VideoPhy-2 evaluates physical commonsense and semantic adherence to text conditioning prompts for diverse real-world act	✓	0.22
The evaluation methodology uses human annotators to rate videos on a 5-point scale.	×	0.06
The VideoPhy-2 pipeline generates text prompts from seed actions using an LLM.	×	0.05
The VideoPhy-2 pipeline uses a VLM to caption generated videos and extract candidate physical rules.	×	0.11
Wan2.1-14B achieved a score of 32.6 on the 'All' metric and 21.9 on the 'Hard' metric.	×	0.02
CogVideoX-5B achieved a score of 31.5 on the 'All' metric and 36.2 on the 'Hard' metric.	×	0.02
Sora achieved a score of 23.3 on the 'All' metric and 5.3 on the 'Hard' metric.	×	0.02
Ray2 achieved a score of 21.0 on the 'All' metric and 18.5 on the 'Hard' metric.	×	0.02
The dataset includes examples where generated videos violate physical rules such as Conservation of Momentum and Gravity	×	0.12
Human annotators verify rule violations, suggest missing rules, and assess semantic adherence to the input prompt.	×	0.06

References

- https://arxiv.org/abs/2503.06800
- http://arxiv.org/abs/2601.06142v1
- http://arxiv.org/abs/2511.12188v1