Multimodal Model Performance on Visual Mathematical Reasoning Benchmarks

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do multimodal models like FLIP, GIT, and BLIP compare in terms of accuracy and robustness on visual mathematical reasoning benchmarks such as GSM8K-V and MATH-V. 7 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MATH-PT: A Math Reasoning Benchmark for European and Brazilian Portuguese. Research question: How do multimodal models like FLIP, GIT, and BLIP compare in terms of accuracy and robustness on visual mathematical reasoning benchmarks such as GSM8K-V and MATH-V?.

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 3.8/10.

3 Results

4 papers retrieved. 7 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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
The evaluation protocol uses a standardized prompting strategy adapted to each linguistic variant (European and Brazilia	×	0.07
For multiple-choice questions, the prompt instructs the model to solve the question, output only the letter of the corre	×	0.05
For questions containing figures (European Portuguese only), an additional block enumerating the referenced visual conte	×	0.03
For the Brazilian Portuguese subset (pt-BR), which consists exclusively of questions without figures, an analogous promp	×	0.13
For open-ended questions, the model is asked to place the final answer inside a \boxed{} command.	×	0.07
All models are evaluated in a zero-shot setting, without chain-of-thought supervision or few-shot examples.	×	0.02
Each question is submitted independently to the model using the appropriate prompt.	×	0.01

References

- http://arxiv.org/abs/2109.05633v1
- http://arxiv.org/abs/2604.25926v1
- http://arxiv.org/abs/2603.18589v1