Grok-4 and State-of-the-Art VLMs on GSM8K-V Under Adversarial Visual Distortions

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the accuracy difference between Grok-4 and other state-of-the-art VLMs on GSM8K-V when evaluated under adversarial visual distortions. 6 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: Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. Research question: What is the accuracy difference between Grok-4 and other state-of-the-art VLMs on GSM8K-V when evaluated under adversarial visual distortions?.

2 Methodology

Systematic literature search across multiple databases yielded 12 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

12 papers retrieved. 6 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
Single-stage training produces VLMs that maintain or outperform multi-stage models, saving considerable compute.	×	0.04
Single-stage training improves aggregate performance with a p-value of 0.00558.	×	0.03
Eliminating the first stage of multi-stage training saves 20-25% of training cost.	×	0.02
Full finetuning through visual backbones dramatically degrades performance across almost all benchmarks, especially on l	×	0.04
The authors' models reproduce the performance reported in Liu et al. (2023b).	×	0.01
The authors' family of VLMs at the 7-13B scale strictly outperform InstructBLIP and LLaVa v1.5, the state-of-the-art in	✓	0.29

References

- http://arxiv.org/abs/2402.07865v2
- http://arxiv.org/abs/2509.25160v1
- http://arxiv.org/abs/2604.12659v1