Does applying the PowerInfer optimization strategy to LLaVA result in measurable degradation in VQA accuracy s

Assignee Research

May 29, 2026

Abstract

Abstract In the past years, multimodal large language models (MLLMs) have demonstrated remarkable performance in tasks such as visual question answering and visual understanding and reasoning. However, the extensive model size and high training and inference costs have hindered the widespread application of MLLMs in academia and industry. Thus, studying efficient and lightweight MLLMs has enormous potential, especially in edge computing scenarios. In this survey, we provide a comprehensive and systematic review of the current state of efficient MLLMs. Specifically, this survey summarizes the t

1 Introduction

This paper examines: Efficient multimodal large language models: a survey. Research question: Does applying the PowerInfer optimization strategy to LLaVA result in measurable degradation in VQA accuracy scores on the GQA benchmark compared to full-precision dense inference on single-GPU setups?.

2 Methodology

Systematic literature search across multiple databases yielded 7 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

7 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.5/10.

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
Multimodal large language models (MLLMs) have demonstrated remarkable performance in tasks such as visual question answe	✓	0.46
The extensive model size and high training and inference costs have hindered the widespread application of MLLMs in acad	✓	0.41
Studying efficient and lightweight MLLMs has enormous potential, especially in edge computing scenarios.	✓	0.36
This survey provides a comprehensive and systematic review of the current state of efficient MLLMs.	✓	0.31
The survey summarizes the timeline of representative efficient MLLMs, the current state of research in structures and st	✓	0.43
The limitations of current efficient MLLM research and promising future directions are discussed.	✓	0.35

References

- https://doi.org/10.1007/s44267-025-00099-6
- https://doi.org/10.22541/au.176348756.61222219/v1
- https://doi.org/10.48550/arxiv.2501.03265