Modality-Specific Static Quantization Effects on Multimodal LLM Reasoning in LaVIS Benchmarks

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the impact of full static quantization on the reasoning capabilities of multimodal large language models, as measured by accuracy on the LaVIS benchmark suite. Multimodal large language models (MLLMs) have garnered widespread attention due to their ability to understand multimodal input. However, their large parameter sizes and substantial computational demands severely hinder their practical deployment and application.While. 5 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MQuant: Unleashing the Inference Potential of Multimodal Large Language Models via Full Static Quantization. Research question: What is the impact of full static quantization on the reasoning capabilities of multimodal large language models, as measured by accuracy on the LaVIS benchmark suite?.

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 4.3/10.

3 Results

12 papers retrieved. 5 claims extracted; 1 independently verified. Quality review score: 4.3/10.

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
MQuant is a general PTQ framework designed for MLLMs that demonstrates near-lossless performance and significant speedup	×	0.10
Modality-specific Static Quantization (MSQ) and Attention-Invariant Flexible Switching (AIFS) are designed to accelerate	✓	0.20
Rotation Magnitude Suppression enhances quantization performance by addressing weight outliers caused by online rotation	×	0.12
The paper provides the first comprehensive analysis of quantization issues in MLLMs, identifying root causes of performa	×	0.05
Multimodal Large Language Models (MLLMs) consist of three main modules: Visual encoder, Vision-language projector, and L	×	0.12

References

- http://arxiv.org/abs/2502.00425v2
- http://arxiv.org/abs/2407.17856v4
- http://arxiv.org/abs/2511.02794v1