Static Quantization Robustness Trade-offs in Multimodal Models on AdvBench

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does full static quantization disproportionately reduce robustness against adversarial visual perturbations in multimodal models as measured by accuracy drops on the AdvBench 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. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.8/10. 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: Does full static quantization disproportionately reduce robustness against adversarial visual perturbations in multimodal models as measured by accuracy drops on the AdvBench suite?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.8/10.

3 Results

13 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.8/10.

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
MQuant achieves near-lossless performance with significant speedup in multimodal large language models (MLLMs).	✓	0.19
The first comprehensive analysis of quantization issues in MLLMs reveals the root causes of performance collapse and ide	×	0.07
Modality-specific Static Quantization (MSQ) and Attention-Invariant Flexible Switching (AIFS) are designed to accelerate	✓	0.21
Rotation Magnitude Suppression (RMS) is proposed to enhance quantization performance by addressing weight outliers cause	×	0.14
The MQuant framework demonstrates near-lossless performance and significant speedup in MLLMs.	×	0.06
The visual encoder in MLLMs processes visual inputs and compresses them into more compact patch features using a Vision	×	0.03
The vision-language projector in MLLMs maps visual patch features into the textual feature space.	×	0.05
The large language model in MLLMs handles multi-modal tokens and performs reasoning, with capabilities such as zero-shot	×	0.08
Commonly used open-source LLMs include the Llama series, Qwen, InternLM, MiniCPM, and ChatGLM.	×	0.02
Flamingo pioneered connecting pre-trained language models with visual encoders.	×	0.06
The performance of MQuant on InternVL2-8B with W8A8 quantization is 77.49 and 90.27, with a speedup of 785 and 2156 resp	×	0.03
The performance of MQuant on Qwen-VL-Chat-9.6B with W4A8 quantization is 61.16 and 59.31, with a speedup of 483 and 1691	×	0.06

References

- `http://arxiv.org/abs/2502.00425v2`

- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2405.18770v6