

INT4 Quantization Impact on LLaVA-UHD Performance Across SEED-Bench Visual Reasoning Tasks

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 3 peer-reviewed papers addressing the following research question: How does INT4 quantization of LLaVA-UHD affect its performance on SEED-Bench compared to FP16 precision across different visual reasoning subtasks. 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. 6 claims were extracted from source literature; 6 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Efficient multimodal large language models: a survey. Research question: How does INT4 quantization of LLaVA-UHD affect its performance on SEED-Bench compared to FP16 precision across different visual reasoning subtasks?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

3 papers retrieved. 6 claims extracted; 6 independently verified. Quality review score: 8.2/10.

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 answer	✓	0.46
The extensive model size and high training and inference costs have hindered the widespread application of MLLMs in acad	✓	0.39
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 in the survey.	✓	0.36

References

- <https://doi.org/10.1007/s44267-025-00099-6>
- <https://doi.org/10.48550/arxiv.2409.02889>
- <https://doi.org/10.48550/arxiv.2501.07783>