Quantization-Aware Training Performance Across LLaVA Model Versions on VQA and GQA

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: How do different LLaVA model versions compare in terms of quantization-aware training effectiveness on standard multimodal reasoning benchmarks like VQA and GQA. Recent advances in multimodal vision-language models (VLMs) have enabled joint reasoning over visual and textual information, yet their application to planetary science remains largely unexplored. A key hindrance is the absence of large-scale datasets that pair real planetary. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-LE: Large Language-and-Vision Assistant for Lunar Exploration. Research question: How do different LLaVA model versions compare in terms of quantization-aware training effectiveness on standard multimodal reasoning benchmarks like VQA and GQA?.

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 3.3/10.

3 Results

13 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
LLaVA-LE was evaluated on a benchmark of 50 lunar patches from the LUCID stage 1 dataset.	×	0.11
The evaluation benchmark included 190 questions generated by a language-only GPT-5.1 model.	×	0.07
LLaVA-LE Stage 2 achieved an average overall score of 0.921 compared to judge scores.	×	0.07
Base LLaVA achieved an average overall score of 0.278 compared to judge scores.	×	0.06
LLaVA-LE Stage 1 achieved an average overall score of 0.443 compared to judge scores.	×	0.07
LLaVA-LE Stage 2 showed a 3.3$\times$ improvement over Base LLaVA.	×	0.09
LLaVA-LE Stage 2 showed a 2.1$\times$ improvement over LLaVA-LE Stage 1.	×	0.08
LLaVA-LE Stage 2 scored 1.070 on Reasoning questions, exceeding the judge's reference score.	×	0.15
LUCID dataset contains 96K samples derived from co-registered lunar remote sensing observations.	×	0.08
LUCID Stage 1 contains 76K samples used for concept alignment.	×	0.06
LUCID Stage 2 contains approximately 20K samples used for instruction tuning.	×	0.10

References

- http://arxiv.org/abs/2306.00890v1
- http://arxiv.org/abs/2603.24696v1
- http://arxiv.org/abs/2509.23661v3