Activation-Aware Weight Quantization in LLaVA-1.5 on the GQA Benchmark

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does activation-aware weight quantization affect LLaVA-1.5 performance on the GQA benchmark compared to standard post-training quantization methods. We present LLaVA-OneVision-1.5, a novel family of Large Multimodal Models (LMMs) that achieve state-of-the-art performance with significantly reduced computational and financial costs. Different from the existing works, LLaVA-OneVision-1.5 provides an open, efficient, and. 19 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-OneVision-1.5: Fully Open Framework for Democratized Multimodal Training. Research question: How does activation-aware weight quantization affect LLaVA-1.5 performance on the GQA benchmark compared to standard post-training quantization methods.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.4/10.

3 Results

8 papers retrieved. 19 claims extracted; 0 independently verified. Quality review score: 2.4/10.

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-OV-1.5 achieves 67.7 accuracy on MMStar benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 68.2 accuracy on MMStar benchmark	×	0.02
Qwen2.5-VL achieves 68.3 accuracy on MMStar benchmark	×	0.07
LLaVA-OV-1.5 achieves 84.1 accuracy on MMBenchen benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 85.7 accuracy on MMBenchen benchmark	×	0.02
Qwen2.5-VL achieves 85.7 accuracy on MMBenchen benchmark	×	0.07
LLaVA-OV-1.5 achieves 81.0 accuracy on MMBenchcn benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 84.2 accuracy on MMBenchcn benchmark	×	0.02
The offline parallel data packing method achieves up to an 11$\times$ compression ratio on 85 million pretraining samples	×	0.08
LLaVA-OV-1.5 achieves 61.7 accuracy on MME-RealWorlden benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 63.4 accuracy on MME-RealWorlden benchmark	×	0.04
LLaVA-OV-1.5 achieves 56.1 accuracy on MME-RealWorldcn benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 56.1 accuracy on MME-RealWorldcn benchmark	×	0.02
LLaVA-OV-1.5 achieves 77.3 accuracy on SeedBenchimage benchmark	×	0.01
LLaVA-OV-1.5 achieves 80.7 accuracy on CV-Bench benchmark	×	0.01
LLaVA-OV-1.5 RL achieves 72.3 accuracy on MathVistamini benchmark	×	0.02
LLaVA-OV-1.5 achieves 86.5 accuracy on ChartQA benchmark	×	0.01
LLaVA-OV-1.5 achieves 95.0 accuracy on DocVQA benchmark	×	0.01
LLaVA-OV-1.5 achieves 62.2 accuracy on PixmoCount benchmark	×	0.01

References

- http://arxiv.org/abs/2509.16989v3
- http://arxiv.org/abs/2509.23661v3
- http://arxiv.org/abs/2406.16299v1