Weight-Only Quantization Trade-Offs in LLaVA for Cross-Modality Reasoning on TextVQA

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the trade-off between inference latency and cross-modality reasoning performance when applying weight-only quantization to LLaVA on the TextVQA dataset. Vision systems to see and reason about the compositional nature of visual scenes are fundamental to understanding our world. The complex relations between objects and their locations, ambiguities, and variations in the real-world environment can be better described in human. 11 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.6/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Foundational Models Defining a New Era in Vision: A Survey and Outlook. Research question: What is the trade-off between inference latency and cross-modality reasoning performance when applying weight-only quantization to LLaVA on the TextVQA dataset?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

11 papers retrieved. 11 claims extracted; 9 independently verified. Quality review score: 7.6/10.

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
Foundational models are learned to bridge the gap between modalities such as vision, language, audio, and depth coupled	✓	0.28
Foundational models facilitate contextual reasoning, generalization, and prompt capabilities at test time.	✓	0.23
The output of foundational models can be modified through human-provided prompts without retraining.	✓	0.22
Foundational models can segment a particular object by providing a bounding box as a prompt.	✓	0.17
Foundational models can engage in interactive dialogues by answering questions about an image or video scene.	✓	0.16
Foundational models can manipulate robot behavior through language instructions.	×	0.15
Foundational model architectures combine different modalities including vision, text, and audio.	✓	0.18
Common training objectives for foundational models include contrastive and generative approaches.	×	0.13
Common prompting patterns for foundational models include textual, visual, and heterogeneous types.	✓	0.17
Open challenges for foundational models in computer vision include difficulties in evaluations and benchmarking.	✓	0.19
Open challenges for foundational models include gaps in their real-world understanding.	✓	0.20

References

- https://doi.org/10.1007/s44267-024-00070-x
- https://doi.org/10.48550/arxiv.2307.13721
- https://doi.org/10.48550/arxiv.2405.10739