Quantization-Aware Training Effects on LLaVA-UHD Edge Deployment Efficiency

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 7 peer-reviewed papers addressing the following research question: What is the impact of quantization-aware training on the inference latency and memory requirements of LLaVA-UHD when deployed on edge devices. Large foundation models, including large language models (LLMs), vision transformers (ViTs), diffusion, and LLM-based multimodal models, are revolutionizing the entire machine learning lifecycle, from training to deployment. However, the substantial advancements in versatility. 7 claims were extracted from source literature; 7 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: A Survey of Resource-efficient LLM and Multimodal Foundation Models. Research question: What is the impact of quantization-aware training on the inference latency and memory requirements of LLaVA-UHD when deployed on edge devices?.

2 Methodology

Systematic literature search across multiple databases yielded 7 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

7 papers retrieved. 7 claims extracted; 7 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
Large foundation models include large language models (LLMs), vision transformers (ViTs), diffusion models, and LLM-base	✓	0.29
Large foundation models are revolutionizing the machine learning lifecycle from training to deployment.	✓	0.27
Advancements in versatility and performance of large foundation models come at a significant cost in terms of hardware r	✓	0.30
There has been a considerable focus on developing resource-efficient strategies to support the growth of large models in	✓	0.34
This survey examines both algorithmic and systemic aspects of resource-efficient strategies.	✓	0.19
The survey encompasses topics ranging from model architectures and training/serving algorithms to practical system desig	✓	0.23
The goal of the survey is to provide an understanding of how current approaches tackle resource challenges posed by larg	✓	0.28

References

- https://doi.org/10.48550/arxiv.2401.08092
- https://doi.org/10.1038/s41467-025-61040-5
- https://doi.org/10.48550/arxiv.2405.10739