Quantized InternVL Models: Inference Latency and Memory Trade-offs on Edge Devices

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the efficiency tradeoff between 7B and 13B InternVL models in terms of inference latency and memory usage when deployed on edge devices with quantized weights. Quantized neural networks are well known for reducing the latency, power consumption, and model size without significant harm to the performance. This makes them highly appropriate for systems with limited resources and low power capacity. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: AMED: Automatic Mixed-Precision Quantization for Edge Devices. Research question: What is the efficiency tradeoff between 7B and 13B InternVL models in terms of inference latency and memory usage when deployed on edge devices with quantized weights?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

15 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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.

References

- http://arxiv.org/abs/2205.15437v2
- http://arxiv.org/abs/2506.10461v1
- http://arxiv.org/abs/2502.00425v2