Quantization-Aware Training Preserves Multimodal Accuracy in Vision-Language Models

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does quantization-aware training influence multimodal benchmark performance on ScienceQA compared to post-training quantization. 15 claims were extracted from source literature; 9 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Gated Relational Alignment via Confidence-based Distillation for Efficient VLMs. Research question: How does quantization-aware training influence multimodal benchmark performance on ScienceQA compared to post-training quantization?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 6.8/10.

3 Results

12 papers retrieved. 15 claims extracted; 9 independently verified. Quality review score: 6.8/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Vision-Language Models (VLMs) achieve strong multimodal performance but are costly to deploy.	✓	0.24
Post-training quantization often causes significant accuracy loss in Vision-Language Models.	✓	0.21
Quantization-aware training for Vision-Language Models remains underexplored.	✓	0.15
GRACE is a framework that unifies knowledge distillation and Quantization-Aware Training (QAT) under the Information Bot	✓	0.22
GRACE introduces confidence-gated decoupled distillation to filter unreliable supervision.	✓	0.22
GRACE uses relational centered kernel alignment to transfer visual token structures.	✓	0.23
GRACE employs an adaptive controller via Lagrangian relaxation to balance fidelity against capacity constraints.	✓	0.21
On the SQA benchmark, the INT4 LLaVA-1.5-7B model achieved a score of 70.1.	×	0.09
On the SQA benchmark, the FP16 LLaVA-1.5-7B baseline achieved a score of 66.8.	×	0.09
On the MMBench benchmark, the INT4 Qwen2-VL-2B model achieved a score of 76.9.	×	0.10
On the MMBench benchmark, the FP16 Qwen2-VL-2B baseline achieved a score of 72.6.	×	0.09
Using real INT4 kernels, GRACE achieves 3x throughput compared to the baseline.	×	0.08
Using real INT4 kernels, GRACE achieves a 54% memory reduction compared to the baseline.	×	0.13
GRACE significantly outperforms existing quantization methods on extensive benchmarks.	✓	0.19
Code and data for GRACE are available at https://github.com/ForeverBlue816/GRACE.	✓	0.19

References

- https://doi.org/10.48550/arxiv.2307.13721
- https://doi.org/10.48550/arxiv.2408.01319
- https://openalex.org/W7127203421