Activation-Aware Zero-Shot Quantization for Robust Multimodal LLM Visual Grounding

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: To what extent does activation-aware zero-shot quantization improve robustness to domain shift in multimodal LLM visual grounding tasks compared to standard post-training quantization on RefCOCO+. Multimodal Large Language Models (MLLM) are increasingly deployed in domains where both reliability and efficiency are critical. However, current models remain overconfident, producing highly certain but incorrect answers. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.3/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Evaluating the Impact of Post-Training Quantization on Reliable VQA with Multimodal LLMs. Research question: To what extent does activation-aware zero-shot quantization improve robustness to domain shift in multimodal LLM visual grounding tasks compared to standard post-training quantization on RefCOCO+?.

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 3.3/10.

3 Results

11 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.3/10.

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
Quantization consistently reduces both task accuracy and reliability.	×	0.09
As bit width decreases, accuracy drops and ECE increases, showing that model calibration deteriorates correspondingly.	×	0.06
Data-aware MBQ maintains higher accuracy and lower calibration error than HQQ, especially at 4 bits and below.	×	0.09
At int4, performance remains within about 2 percentage points of bf16, while int3 introduces severe confidence instabili	×	0.03
Reliability degradation mirrors accuracy loss: quantization noise directly perturbs the confidence distribution.	×	0.05
Selector significantly improves reliability for both quantized and bf16 models.	×	0.07
Compared to MaxProb, the Selector consistently lowers ECE and improves C@1 % across all bit widths.	×	0.05
The Selector restores reliability for both models to values comparable to the bf16 baseline, except for int3 quantizatio	×	0.05
The Selector acts as an efficient reliability-restoration mechanism without modifying base model weights.	×	0.03
Shifting from VQAv2 to AdVQA and VizWiz introduces progressively stronger multimodal distribution shifts that stress bot	×	0.05
Performance and reliability deteriorate roughly in proportion to their in-distribution degradation, indicating that quan	×	0.04
Data-aware MBQ quantization consistently yields smoother declines and greater reliability retention than HQQ, especially	×	0.09
The Selector enhances coverage and effective reliability throughout this progression, delaying but not preventing the co	×	0.03
Selector behavior remains tightly correlated with intrinsic confidences, suggesting that its benefit stems from stabiliz	×	0.04
For autoregressive decoders, intrinsic confidence can be defined as the joint softmax probability of all generated token	×	0.02
Higher joint probabilities usually indicate more reliable answers, but this measure is often overconfident and poorly ca	×	0.04

References

- http://arxiv.org/abs/2602.13289v1
- http://arxiv.org/abs/2509.16989v3
- http://arxiv.org/abs/2406.08155v2