GRACE-LLaVA Quantization and Model Scaling on Adversarial Visual Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the performance of GRACE-LLaVA-1.5-7B-INT4 scale with model size (e.g., 7B vs. 13B) on adversarial visual perturbation tasks compared to unquantized models, as measured by accuracy on. 8 claims were extracted from source literature; 8 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: Quantizing Diffusion Models for Scalable and Efficient Generative Inference Across Diverse Hardware Platforms. Research question: How does the performance of GRACE-LLaVA-1.5-7B-INT4 scale with model size (e.g., 7B vs. 13B) on adversarial visual perturbation tasks compared to unquantized models, as measured by accuracy on benchmarks like HatefulMemes and TextVQA?.

2 Methodology

Systematic literature search across multiple databases yielded 4 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

4 papers retrieved. 8 claims extracted; 8 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
Diffusion models have recently emerged as a dominant framework in generative modeling, achieving unprecedented performan	✓	0.38
Diffusion models are computationally intensive and memory-heavy, which significantly hinders their deployment in real-wo	✓	0.35
Quantization—the process of reducing the numerical precision of model weights, activations, or gradients—offers a promis	✓	0.40
Quantizing diffusion models poses unique challenges that differ markedly from those encountered in traditional classific	✓	0.33
These challenges arise from the multi-step nature of the generative process, the sensitivity of score-based sampling to	✓	0.39
This review provides a comprehensive and detailed exploration of the current landscape of diffusion model quantization.	✓	0.28
We systematically examine the theoretical underpinnings of diffusion processes and how they interact with various quanti	✓	0.40
We analyze the trade-offs between model accuracy,	✓	0.18

References

- https://doi.org/10.36227/techrxiv.175372917.71320154/v1
- https://doi.org/10.20944/preprints202512.0118.v1
- https://openalex.org/W7125352730