Robustness of GRACE-LLaVA-1.5-7B-INT4 and Qwen-VL-Chat-INT4 under Adversarial Visual Perturbations

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the robustness of GRACE-LLaVA-1.5-7B-INT4 compare to that of other quantized multimodal models like Qwen-VL-Chat-INT4 on adversarial visual perturbations across language understanding. 12 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. Research question: How does the robustness of GRACE-LLaVA-1.5-7B-INT4 compare to that of other quantized multimodal models like Qwen-VL-Chat-INT4 on adversarial visual perturbations across language understanding benchmarks such as VQA and COCO-Caption?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

8 papers retrieved. 12 claims extracted; 3 independently verified. Quality review score: 4.5/10.

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
LLaVA-Mini uses only 1 vision token per image fed into the LLM backbone, compared to 576 tokens in LLaVA-v1.5.	✓	0.19
LLaVA-Mini achieves a vision token compression rate of 0.17%.	×	0.11
LLaVA-Mini achieves a 77% reduction in FLOPs compared to LLaVA-v1.5.	×	0.05
LLaVA-Mini reduces GPU memory usage per image from 360 MB to 0.6 MB.	×	0.03
LLaVA-Mini decreases image understanding inference latency from 100 ms to 40 ms.	×	0.03
LLaVA-Mini enables processing of long videos exceeding 10,000 frames (over 3 hours) on an NVIDIA RTX 3090 with 24GB of m	×	0.03
LLaVA-Mini was evaluated on 11 image-based and 7 video-based understanding benchmarks.	✓	0.16
LLaVA-Mini achieves performance comparable to LLaVA-v1.5 across the evaluated benchmarks.	×	0.06
In LLaVA architecture, attention devoted to vision tokens decreases sharply as layers deepen, shifting towards input ins	×	0.07
Removing vision tokens entirely in some later layers of the LLM does not eliminate the model's visual understanding capa	×	0.11
LLaVA-Mini introduces a modality pre-fusion module before the LLM to fuse visual information into instruction text.	✓	0.24
LLaVA-v1.5 uses 576 vision tokens for image processing.	×	0.10

References

- http://arxiv.org/abs/2311.05437v1
- http://arxiv.org/abs/2103.15670v3
- http://arxiv.org/abs/2501.03895v2