Token Pruning and Bit-Width Reduction Trade-offs in Vision-Language Model Grounding Performance

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the correlation between bit-width reduction in vision-language models and the degradation of grounding performance on RefCOCO+ versus the gain in inference throughput. Benchmark accuracy is often implicitly assumed to reflect grounded visual understanding in vision-language models (VLMs), yet it remains unclear to what extent such scores truly reflect reliance on visual evidence. Motivated by a surprising observation that removing a. 7 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Seeing without Looking: Do Vision-Language Benchmarks Really Test Vision?. Research question: What is the correlation between bit-width reduction in vision-language models and the degradation of grounding performance on RefCOCO+ versus the gain in inference throughput?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.2/10.

3 Results

13 papers retrieved. 7 claims extracted; 2 independently verified. Quality review score: 5.2/10.

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
Removing a substantial fraction of image tokens only degrades model performance very slightly on a widely used hallucina	✓	0.27
VLMs do incorporate visual input, but their predictions are less sensitive to the loss of fine-grained visual evidence t	✓	0.33
Randomly removing a substantial fraction of image tokens often leads to little degradation in overall accuracy.	×	0.12
For Qwen3-4B and LLaVA-1.5-7B, accuracy decreases approximately linearly as the drop ratio increases, but the magnitude	×	0.01
When drop ratio σ = 0.75, performance decreases by only about 3% compared to the baseline for Qwen3-4B and LLaVA-1.5-7B.	×	0.01
Qwen3-32B and Gemma3-12B do not exhibit a monotonic decline with increasing token removal.	×	0.03
When σ = 0.25, Qwen3-32B and Gemma3-12B slightly outperform their baseline accuracy.	×	0.03

References

- http://arxiv.org/abs/2404.07214v4
- http://arxiv.org/abs/2605.22903v1

- http://arxiv.org/abs/2510.06243v2