Parameter Count and Token Latency in VLA-Adapter vs. OpenVLA Fine-Tuning on RoboBench

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the parameter count of VLA-Adapter correlate with token generation latency on the RoboBench suite compared to full fine-tuning of OpenVLA. 9 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: OpenVLA: An Open-Source Vision-Language-Action Model. Research question: How does the parameter count of VLA-Adapter correlate with token generation latency on the RoboBench suite compared to full fine-tuning of OpenVLA?.

2 Methodology

Systematic literature search across multiple databases yielded 14 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

14 papers retrieved. 9 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
OpenVLA was evaluated on 170 rollouts (17 tasks with 10 trials each) for BridgeData V2 experiments.	×	0.02
OpenVLA was evaluated on 60 rollouts (12 tasks with 5 trials each) for Google robot experiments.	×	0.04
OpenVLA was compared to RT-1-X, RT-2-X, and Octo in terms of performance.	×	0.04
RT-1-X has 35M parameters.	×	0.03
Octo has 93M parameters.	×	0.02
RT-2-X has 55B parameters.	×	0.08
4-bit quantization matches the performance of bfloat16 inference.	×	0.03
4-bit quantization reduces the GPU memory footprint by more than half.	×	0.02
OpenVLA achieves an 8x reduction in compute compared to full fine-tuning.	×	0.05

References

- https://arxiv.org/abs/2605.06175
- https://arxiv.org/abs/2406.09246
- https://arxiv.org/abs/2603.07404