VLA-Adapter Model Size and Inference Latency Trade-offs on RoboBench

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: What is the trade-off between model size and inference latency when comparing VLA-Adapter with other lightweight multimodal action models on the RoboBench suite. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: VLA-Adapter: An Effective Paradigm for Tiny-Scale Vision-Language-Action Model. Research question: What is the trade-off between model size and inference latency when comparing VLA-Adapter with other lightweight multimodal action models on the RoboBench suite?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.8/10.

3 Results

12 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 2.8/10.

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
All experiments were run on 4 NVIDIA H100 GPUs.	×	0.01
On the LIBERO-Long benchmark, the B1 backbone (Prismatic VLM on Qwen2.5-0.5B) with OpenVLA-OFT achieved a success rate o	×	0.02
On the LIBERO-Long benchmark, the B1 backbone with VLA-Adapter achieved a success rate of 95.0%.	×	0.05
On the LIBERO-Long benchmark, the B2 backbone (Prismatic VLM on LLaMA2-7B) with OpenVLA-OFT achieved a success rate of 8	×	0.02
On the LIBERO-Long benchmark, the B2 backbone with VLA-Adapter achieved a success rate of 95.2%.	×	0.05
On the LIBERO-Long benchmark, the B3 backbone (OpenVLA-7B) with OpenVLA-OFT achieved a success rate of 94.5%.	×	0.01
On the LIBERO-Long benchmark, the B3 backbone with VLA-Adapter achieved a success rate of 95.4%.	×	0.05
VLA-Adapter using a 0.5B backbone requires 24.7GB of Training VRAM with a batch size of 8.	×	0.08
OpenVLA-OFT using a 7B backbone requires 62GB of Training VRAM with a batch size of 8.	×	0.02
VLA-Adapter using a 0.5B backbone achieves a throughput of 219.2Hz for an 8-dim chunk.	×	0.08
OpenVLA-OFT using a 7B backbone achieves a throughput of 71.4Hz for an 8-dim chunk.	×	0.02
VLA-Adapter achieved a performance score of 97.3% on the LIBERO benchmark.	×	0.07
OpenVLA-OFT achieved a performance score of 97.1% on the LIBERO benchmark.	×	0.02
The VLA-Adapter paradigm allows the backbone to remain frozen while only the Action-Query and Policy are trained from scr	×	0.07
OpenVLA-OFT is described as the existing state-of-the-art method on major benchmarks including LIBERO-Long prior to this	×	0.04

References

- http://arxiv.org/abs/2509.09372v2
- http://arxiv.org/abs/2502.00425v2
- http://arxiv.org/abs/2601.19634v1