Language Tokenization of Robot Actions in Vision-Language-Action Models and Its Throughput Trade-offs

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: To what extent does representing low-level robot actions as natural language tokens degrade the inference throughput of Vision-Language-Action models compared to continuous action head architectures. 11 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. 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: To what extent does representing low-level robot actions as natural language tokens degrade the inference throughput of Vision-Language-Action models compared to continuous action head architectures?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

15 papers retrieved. 11 claims extracted; 1 independently verified. Quality review score: 4.2/10.

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
VLA-Adapter achieves a success rate of 95.0% on LIBERO-Long with a B1 backbone, which is a 9.2% improvement over OpenVLA	×	0.07
VLA-Adapter achieves a success rate of 95.2% on LIBERO-Long with a B2 backbone, which is a 7.7% improvement over OpenVLA	×	0.07
VLA-Adapter achieves a success rate of 95.4% on LIBERO-Long with a B3 backbone, which is a 0.9% improvement over OpenVLA	×	0.07
VLA-Adapter uses 24.7GB of VRAM for training with an 8 batch size, which is 1/14$\times$ the VRAM usage of OpenVLA-OFT.	×	0.04
VLA-Adapter achieves a throughput of 219.2Hz with an 8-dim chunk, which is 3$\times$ the throughput of OpenVLA-OFT.	×	0.04
VLA-Adapter achieves a performance of 97.3% on LIBERO, which is comparable to the 97.1% performance of OpenVLA-OFT.	×	0.08
VLA-Adapter remains effective when the backbone is frozen, with only the ActionQuery and Policy trained from scratch.	×	0.05
Experiments are run on 4 NVIDIA H100 GPUs.	×	0.02
VLA-Adapter is compared with OpenVLA-OFT and SmolVLA in Table 3.	×	0.04
VLA-Adapter is evaluated on LIBERO-Long, LIBERO, CALVIN, and real-world data.	×	0.07
VLA-Adapter is designed to bridge the gap between vision-language representations and actions more effectively.	✓	0.16

References

- http://arxiv.org/abs/2603.13782v1
- http://arxiv.org/abs/2606.00229v1
- http://arxiv.org/abs/2509.09372v2