Benchmark for Physically-Aware Vision-Language-Action Manipulation

Nuriev Kamil
Innopolis University
Kazan, Russia
k.nuriev@innopolis.university

Novikov Egor
Innopolis University
Kazan, Russia
email@domain.com

***Abstract*—We present a benchmark for evaluating physically-aware manipulation in Vision-Language-Action (VLA) models. The benchmark focuses on assessing how well robotic policies account for physical constraints such as stability, inertia, friction, and fluid behavior during manipulation tasks. We introduce structured evaluation scenarios and quantitative metrics designed to measure physical optimality beyond binary task success.**

***Index Terms*—robotic manipulation, vision-language-action models, benchmark, physical reasoning, robotic control**

I. Introduction

II. Related Work

III. Benchmark Design

A. Task Scenarios

B. Evaluation Metrics

IV. Experimental Setup

A. Robotic Platform

B. Compared Models

V. Results

VI. Discussion

VII. Conclusion

Acknowledgment

References

References

[1] M. J. Kim, C. Finn, and P. Liang, “Fine-tuning vision-language-action models: Optimizing speed and success,” *arXiv preprint arXiv:2502.19645*, 2025.

[2] A. Azzolini, J. Bai, H. Brandon, J. Cao, P. Chattopadhyay, H. Chen *et al.*, “Cosmos-reason1: From physical common sense to embodied reasoning,” *arXiv preprint arXiv:2503.15558*, 2025.

[3] G. C. Kang, J. Kim, K. Shim, J. K. Lee, and B. T. Zhang, “Clip-rt: Learning language-conditioned robotic policies from natural language supervision,” *arXiv preprint arXiv:2411.00508*, 2024.