Robotics & Intelligent Systems Expo (RISEx 2025), Nov. 13-14, 2025, Edmonton, Canada

INSTRUCT ROBOT: A MODEL-FREE FRAMEWORK FOR MAPPING NATURAL LANGUAGE INSTRUCTIONS INTO ROBOT MOTION

Iury Cleveston, Alana C. Santana, Paula D. P. Costa, Ricardo R. Gudwin, Alexandre S. Simões, Esther L. Colombini
University of Campinas/ Institute of Computing, Campinas, Brazil.
Email: iury.cleveston@ic.unicamp.br

INTRODUCTION

Translating natural language commands into robot actions remains a key challenge. Current methods are often impractical for complex robots, as they require vast datasets and known kinematics, limiting them to few degrees of freedom (DoF) [1]. We introduce InstructRobot, a framework that maps language to motion without these dependencies. It uses reinforcement learning (RL) to jointly learn language and motor control, simplifying training and removing the need for annotated data [2]. We validated our approach on a 26-DoF NAO robot, demonstrating its robustness in complex manipulation tasks.

MATERIALS AND METHODS

The InstructRobot framework consists of two main blocks: the Environment and the Agent, as shown in Fig. 1. The Environment is designed to generate task instructions and their corresponding rewards, removing the need to collect trajectory data. It comprises an Instructional Set with natural language commands and a Reward Generator that associates a reward function with each instruction.

The Agent learns a policy π (a|s) that maps a multimodal state s to a motor action a. The state includes the active language instruction, visual data from RGB cameras, and proprioceptive information (joint angles). The agent's architecture integrates multiple modules: a Transformer-based Language System to process instructions, a Perceptual System with convolutional neural networks (CNNs) to process images, and linear networks for proprioception. The representations from both modalities are aligned and fed into actor-critic modules, which are trained end-to-end using the Proximal Policy Optimization (PPO) algorithm to maximize the episodic accumulated return.

RESULTS AND DISCUSSION

To validate InstructRobot's ability to learn inverse kinematics without a prior model, we designed a series of experiments with a 26-DoF NAO robot. The task was to touch colored cubes based on natural language instructions, with performance rigorously measured by the mean accumulated reward per episode. The final experimental design involved a triangular object layout to improve reachability and, crucially, an enriched agent state that included proprioceptive data (joint angles). This configuration proved decisive, leading to robust quantitative results. In the multi-instruction validation scenario, the agent achieved a stable mean episodic reward of 3.1, corresponding to a task success rate of approximately 95%. This performance was built upon a foundational single-instruction test where the agent achieved a mean reward of 2.4. These numerical results provide strong evidence that the framework can indeed learn a complex kinematic model from scratch, successfully translating language commands into precise motor actions using only the reward signal as guidance.

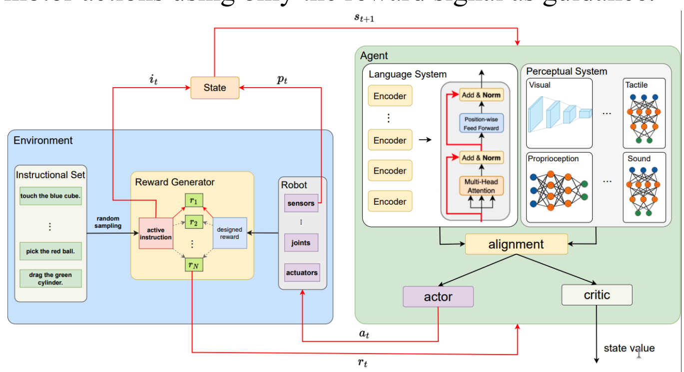

Fig 1. *The InstructRobot framework comprises two main blocks: environment and agent.*

CONCLUSIONS

We presented and experimentally validated the InstructRobot framework, confirming that a complex robot can acquire motor skills directly from natural language through reinforcement learning. Our primary contribution is demonstrating that a 26-DoF robot can learn its own inverse kinematics for manipulation tasks, effectively bypassing the conventional requirements for large datasets and predefined motion models. The validation showed that incorporating proprioceptive feedback was crucial, directly leading to a significant increase in performance and task success rates, which highlights the importance of rich state representations in reward-driven learning.

REFERENCES

[1] Jiang Y et al. VIMA: General Robot Manipulation with Multimodal Prompts. *arXiv preprint arXiv:2210.03094*, 2022.
[2] Ahn M et al. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv preprint arXiv:2204.01691*, 2022.