Project Description

Continuous Control from Open-Vocabulary Feedback

Yunus Emre Balci
yubal22@student.sdu.dk

September 2, 2025

Contents

Problem. This project aims to enable agents in MuJoCo environments to follow natural language instructions without relying on vision-based processing. In embodied AI, teaching agents to follow natural language instructions usually requires expensive visual simulators like IsaacGym or vision-based reward functions like CLIP. This creates computational bottlenecks and limits scalability. This project aims to develop a method that enables agents in MuJoCo environments to understand and execute open-vocabulary instructions without needing visual processing, by combining recent advances in motion-language models with hierarchical reinforcement learning.

Data. We will evaluate our approach using standard MuJoCo locomotion and manipulation environments such as:

- Humanoid (walking, waving, dancing)
- Half-Cheetah (running, jumping)
- Ant (navigation tasks)
- Object manipulation environments (ball kicking, door opening)

Methods. We will combine MotionGPT [1], which treats human motion as a foreign language using VQ-VAE tokenization and T5-based models, with the hierarchical structure from AnySkill [2]. Specifically, we will replace AnySkill's CLIP-based visual reward mechanism with MotionGPT's direct motion-language alignment. The system will have a low-level controller that learns atomic motions from mocap data, and a high-level policy that combines these motions based on language instructions.

Evaluation. The system will be evaluated on: (1) success rate in following diverse language instructions, (2) motion naturalness and physical plausibility scores, (3) computational efficiency compared to vision-based baselines, and (4) generalization to unseen instructions. Performance metrics will be normalized between 0 and 100, where higher scores indicate better instruction following and motion quality.

Distribution of work. The following is an estimate of the percentage-wise distribution of workload for this project:

- Literature Survey: 20%
- Implementation of proof of concept: 30%
- Integration and testing: 35%
- Report writing: 15%

References

[1] Jiang, B., et al. "MotionGPT: Human Motion as a Foreign Language". In: NeurIPS 2023. arXiv: 2306.14795. url: http://arxiv.org/abs/2306.14795.

[2] Cui, Y., et al. ”AnySkill: Learning Open-Vocabulary Physical Skill for Interactive Agents”. In: CVPR 2024. url: https://anyskill.github.io.

[3] ADIN Lab. ”ObjectRL: A Unified Framework for Object-Centric Reinforcement Learning”. GitHub: adinlab/objectrl. 2024.