Reinforcement Learning for Optimizations in Biomechanics

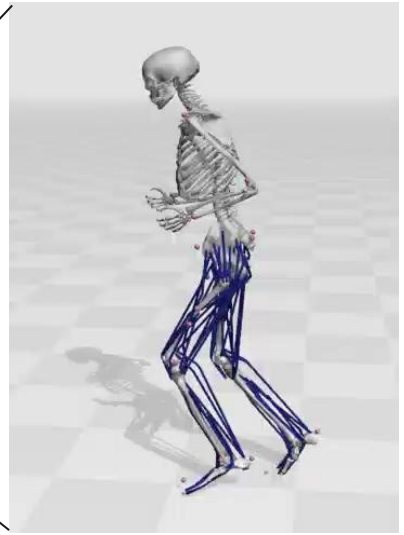
Humanoid Modeling Using Gait Data

For Original Presentation: Link

Oguzhan Esen
Okan Arif Güvenkaya

Project Overview

Data Collection → Data Processing → Inverse Kinematics



Target Movements

Inverse Kinematics → Reinforcement Learning

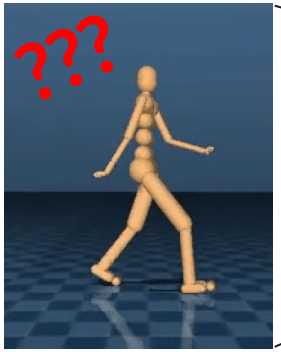
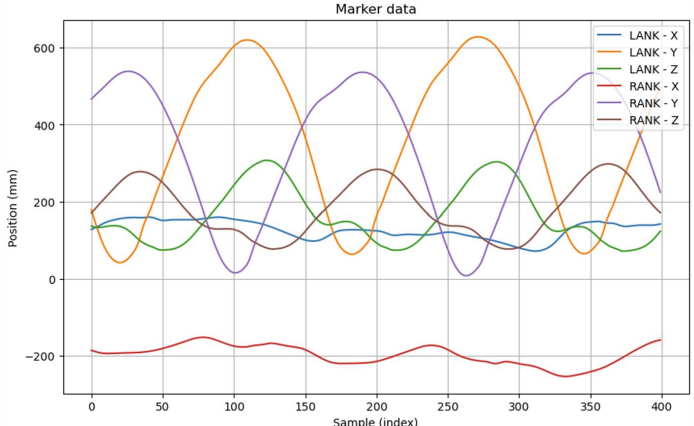


Image Link: Click

Data Collection

- Collected by MIRMI with Vicon Motion Capture System.
- Positions (X, Y, Z) of reflective markers are captured.
- These are the target movements for the training!



Research of Gheorghe Lisca and Kim Peper, TUM MIRMI

Data Processing for Static Trial

The static trial scales the OpenSim model to the subject's anatomy.

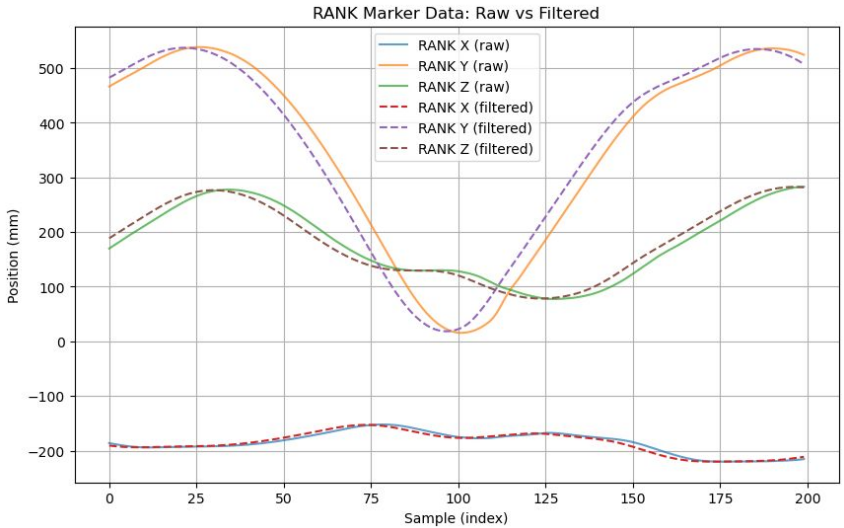
- Unnecessary markers were removed.
- Marker coordinates were rotated to match OpenSim's frame.
- Cleaned data was saved as a .trc file.
- OpenSim Scale Tool was used to generate a personalized model.

Research of Gheorghe Lisca and Kim Peper, TUM MIRMI

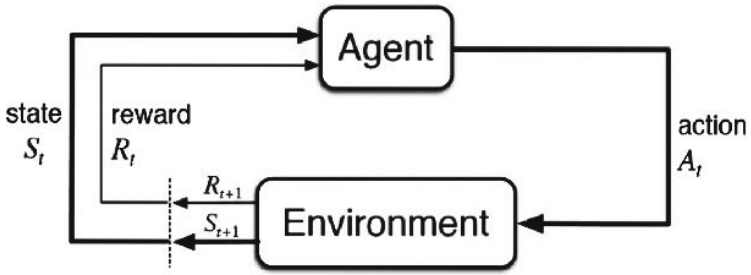
Data Processing for Dynamic Trial

The dynamic trial captures full-body motion during gait for kinematic analysis.

- Removed joints not used in humanoid model
- Applied moving mean filter
- Extracted gait cycles
- Rotated marker data
- Ran **OpenSim Inverse Kinematics (IK)**



How Reinforcement Learning Works?



TUM Reinforcement Learning for Optimization in Biomechanics Lecture 5 Slides, Gheorghe Lisca

- The agent makes a decision (**action**)
- The **environment** responds with a new **state** and a **reward**
- The **agent** uses this feedback to learn better actions
- This loop continues at every time step

The Goal: Maximize total reward over time

How It Works in Our Case?

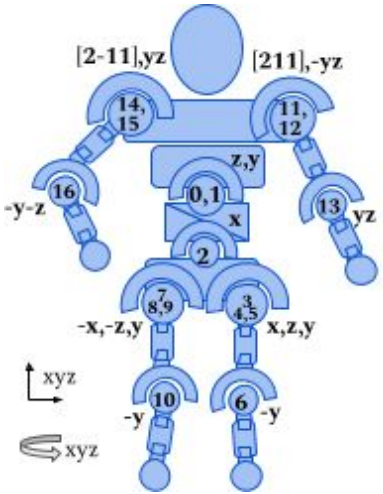


Image Link: Click

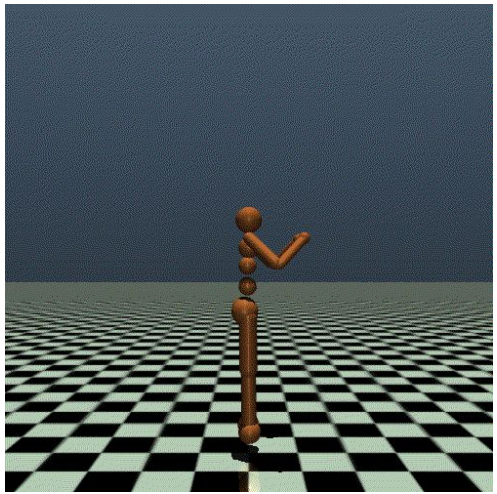


Image Link: Click

- **Agent:** The humanoid robot (from DeepMind's dm_control)
- **State:** Current joint angles of the robot
- **Action:** Joint movements chosen by the agent
- **Environment:** The physics simulation (MuJoCo engine via dm_control)

Reward Explanation

Actions that Increase Reward

- Joint angles closely match the real human motion
- The torso remains upright and balanced
- The agent uses minimal control effort (smooth, energy-efficient movements)

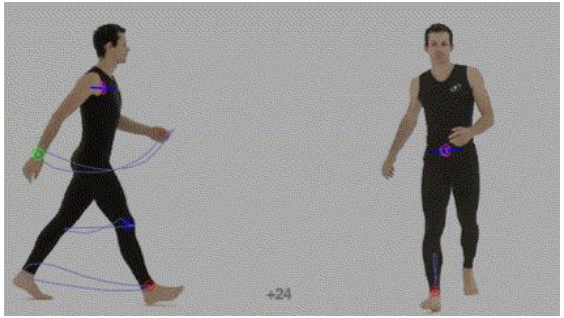


Image Link: Click

Image Link: Click

Reward Components

- R_track – Joint Tracking Reward
  - Based on deviation from real human joint angles
- R_upright – Posture Reward
  - Rewards keeping the torso upright
  - Uses a tolerance function on torso's Z-axis alignment
- R_control – Smooth Control Reward
  - Rewards lower actuator efforts
  - Less reward for jerky or overly strong movements

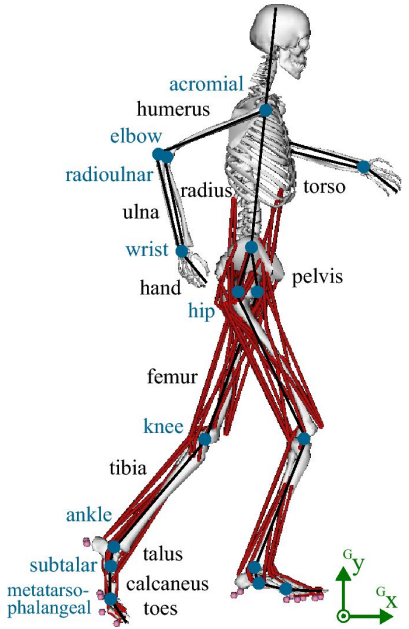


Image Link: Click

Total Reward: components are normalized and combined with weights to compute the final reward.

Episode Initialization

- Each episode begins from a real gait-cycle frame extracted from the .sto data.
- The humanoid's joint positions are set to match the human pose.
- Velocities are zeroed for a static start.
- This ensures a realistic and consistent initial state.

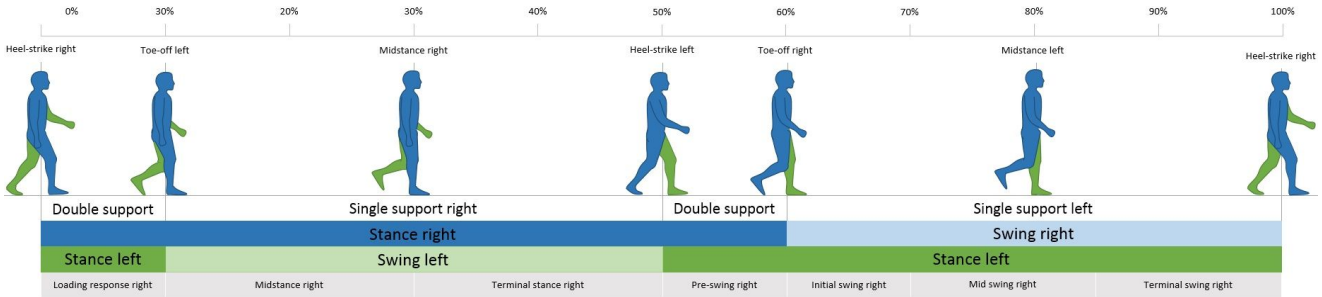
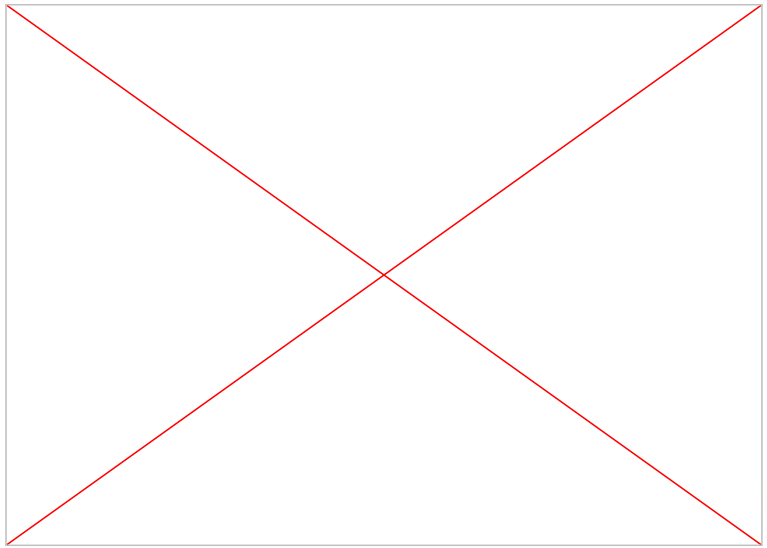
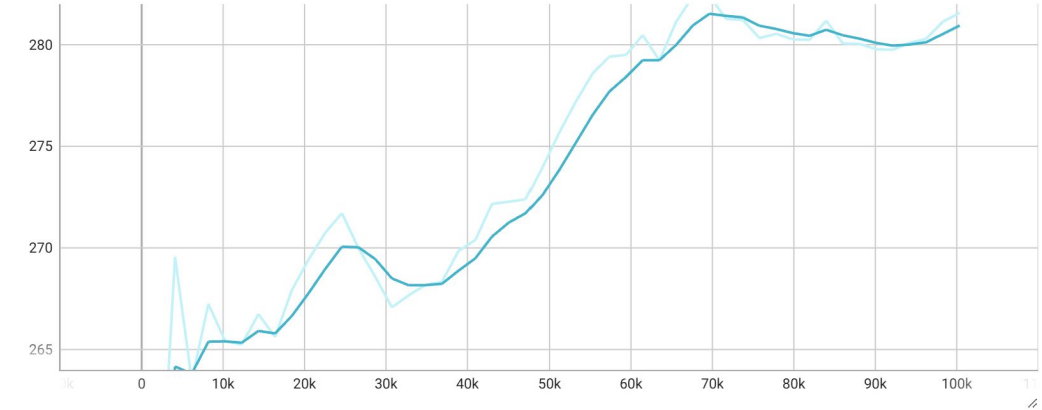


Image Link: Click

Results

algorithm:	PPO
n_time_steps:	100_000
gravity:	-9.8 m/s²

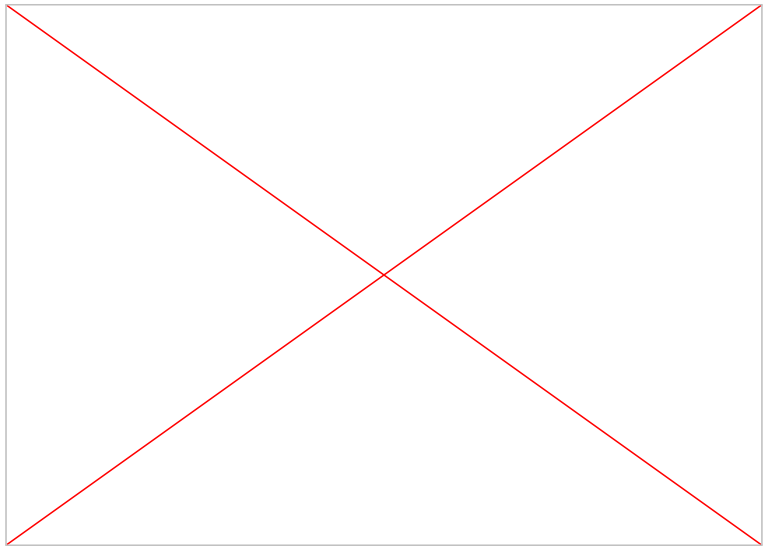


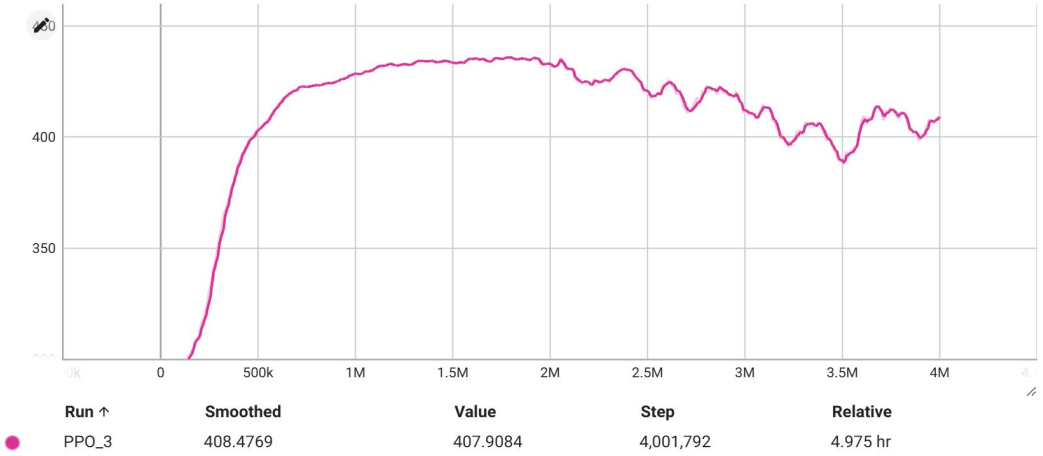


Run ↑	Smoothed	Value	Step	Relative
PPO_2	280.9432	281.5649	100,352	7.681 min

Results

algorithm:	PPO
n_time_steps:	4_000_000
gravity:	-9.8 m/s²

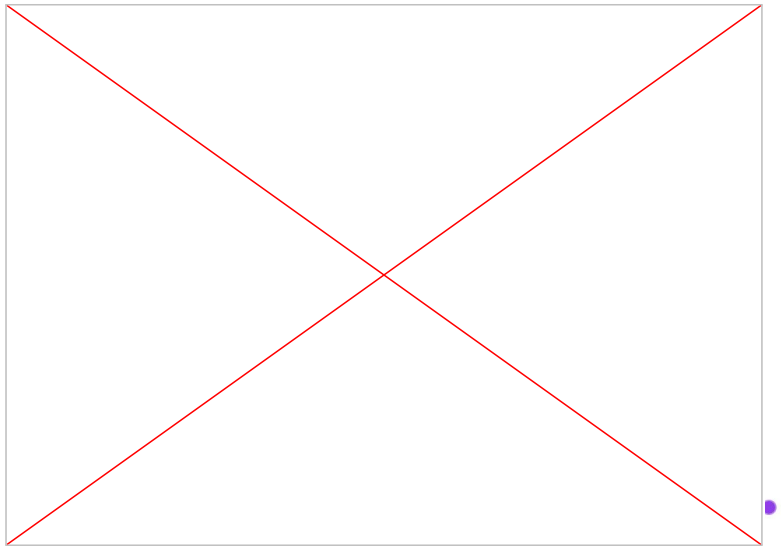


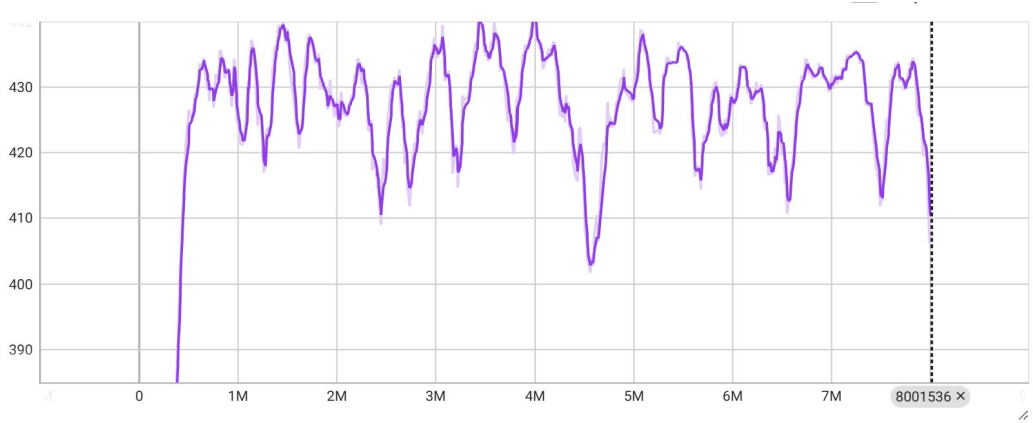


Run ↑	Smoothed	Value	Step	Relative
PPO_3	408.4769	407.9084	4,001,792	4.975 hr

Results

algorithm:	PPO
n_time_steps:	8_000_000
gravity:	-9.8 m/s²

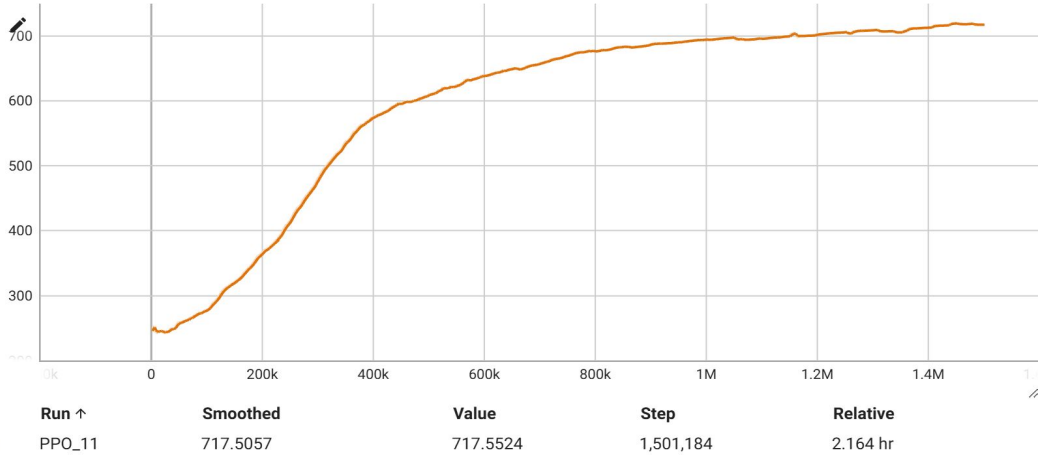




Run ↑	Smoothed	Value	Step	Relative
PPO_4	410.3334	407.4216	8,001,536	11.37 hr

Results

algorithm:	PPO
n_time_steps:	1_500_000
gravity:	-9.8 m/s²

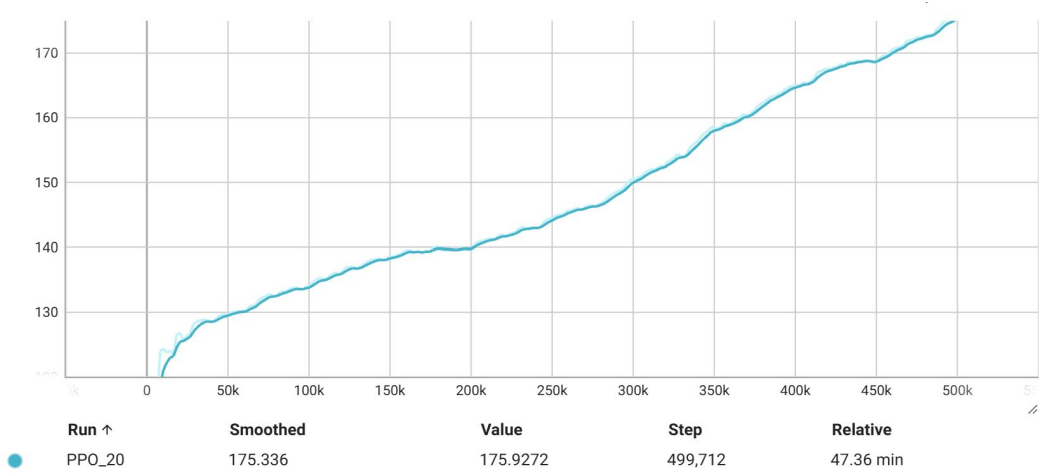


Run ↑	Smoothed	Value	Step	Relative
PPO_11	717.5057	717.5524	1,501,184	2.164 hr

Focus: standing up !

Results

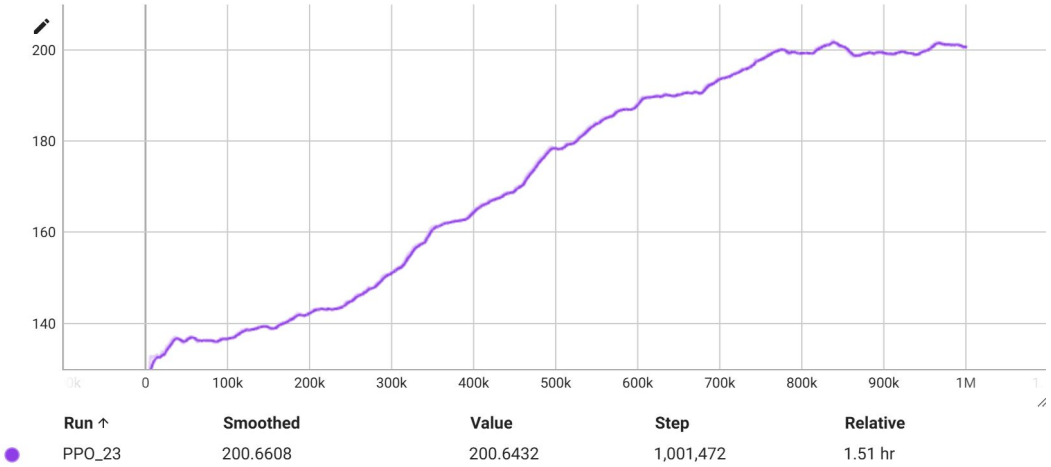
algorithm:	PPO
n_time_steps:	500_000
gravity:	-1.0 m/s²



Run ↑	Smoothed	Value	Step	Relative
PPO_20	175.336	175.9272	499,712	47.36 min

Results

algorithm:	PPO
n_time_steps:	1_000_000
gravity:	-1.0 m/s²



Run ↑	Smoothed	Value	Step	Relative
PPO_23	200.6608	200.6432	1,001,472	1.51 hr

Interpretation

- Despite extensive training, the agent could not fully replicate the complex human gait.
- The humanoid learned some partial walking behaviors, but achieving both stable standing and walking simultaneously was not successful.
- Possible reasons include model complexity, limitations in the reward function, and simulation constraints.

Discussion

- Challenges faced
  - High-dimensional joint space of humanoid robot
  - Reward function balancing between tracking and stability
- Potential overfitting to specific gait cycles

Future Experiments

- Incorporate dynamic data (forces, torques) into reward function
- Experiment with more advanced or hierarchical RL algorithms
- Use curriculum learning starting from simpler tasks
- Include sensor noise and external disturbances for robustness

Q&A