Leveraging Sequentiality in Reinforcement Learning from a Single Demonstration

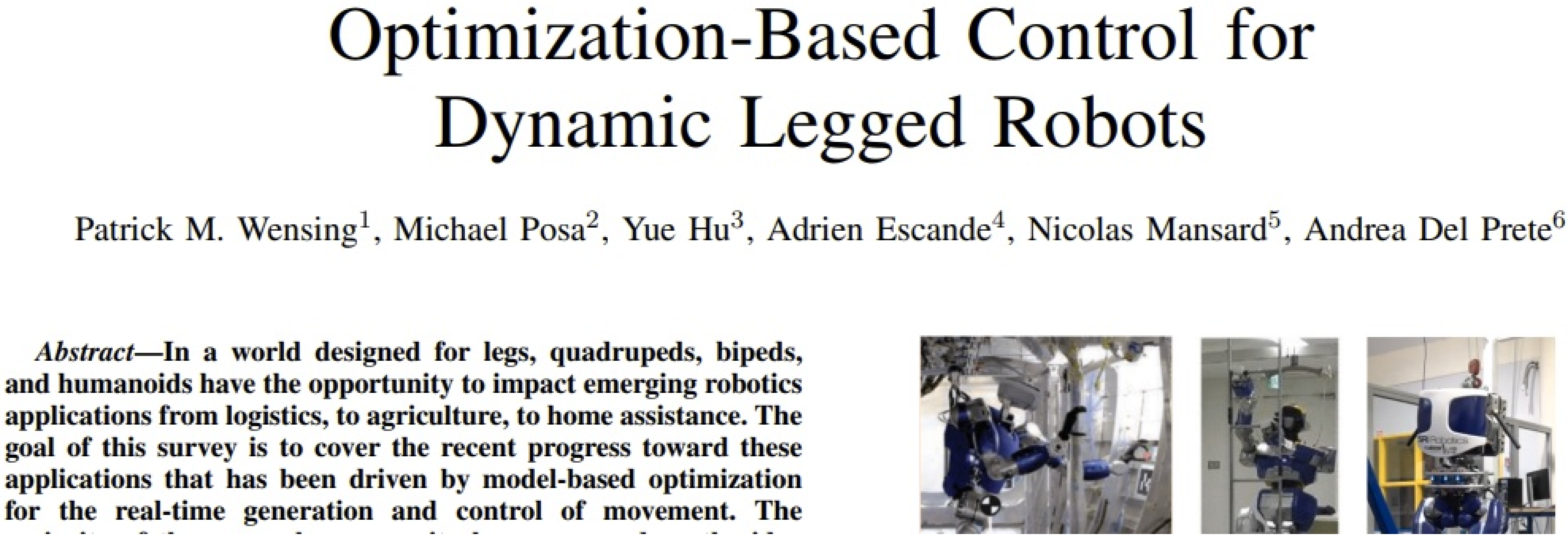
Alexandre Chenu (ISIR)

Olivier Serris (ISIR)

Olivier Sigaud (ISIR)

Nicolas Perrin-Gilbert (ISIR)

REINFORCEMENT LEARNING inside the CONTROL LOOP?

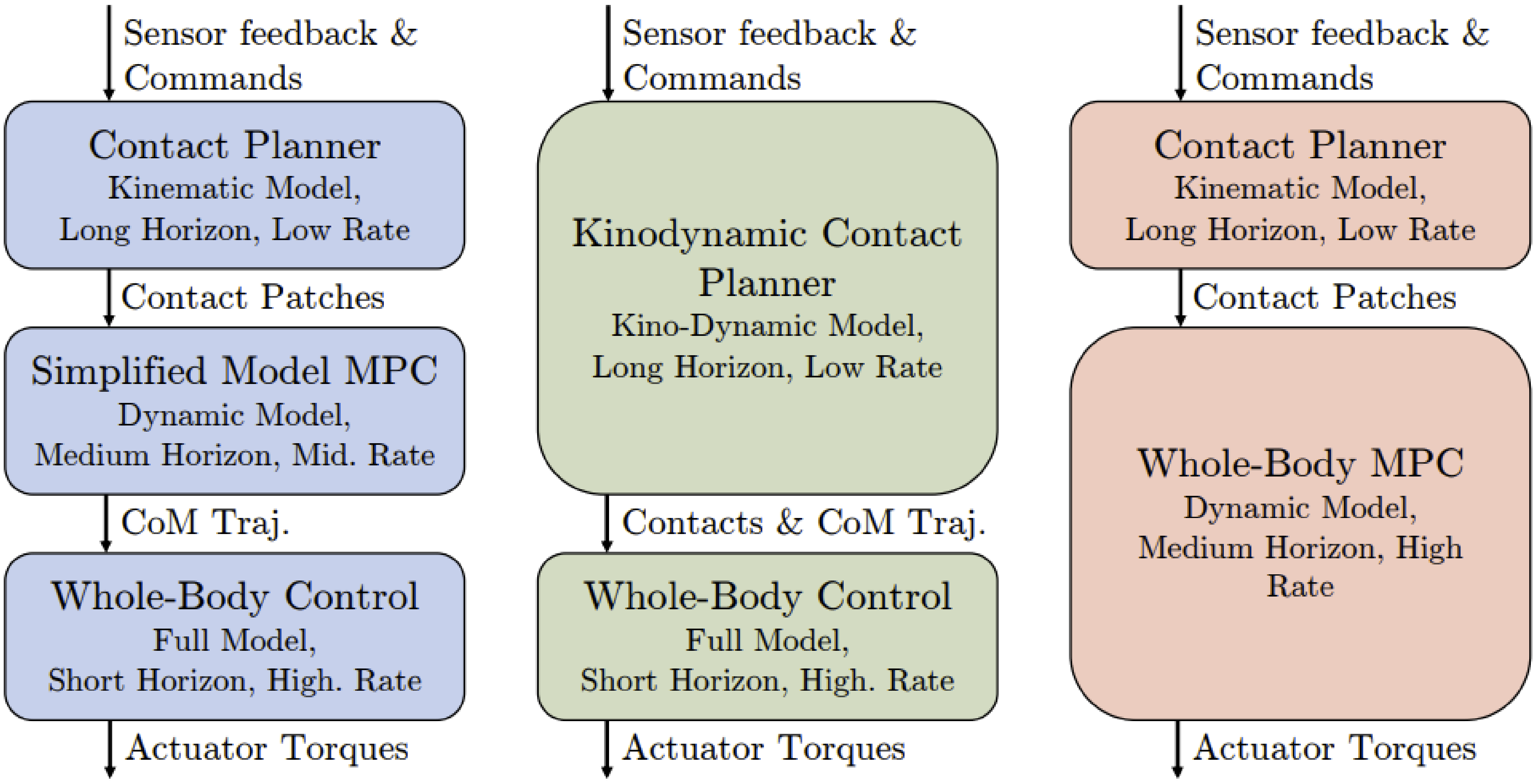


Optimization-Based Control for Dynamic Legged Robots

Patrick M. Wensing[1], Michael Posa[2], Yue Hu[3], Adrien Escande[4], Nicolas Mansard[5], Andrea Del Prete[6]

Abstract—In a world designed for legs, quadrupeds, bipeds, and humanoids have the opportunity to impact emerging robotics applications from logistics, to agriculture, to home assistance. The goal of this survey is to cover the recent progress toward these applications that has been driven by model-based optimization for the real-time generation and control of movement. The

(2022)



Sensor feedback & Commands

Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

Simplified Model MPC
Dynamic Model,
Medium Horizon, Mid. Rate

CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

Sensor feedback & Commands

Kinodynamic Contact Planner
Kino-Dynamic Model,
Long Horizon, Low Rate

Contacts & CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

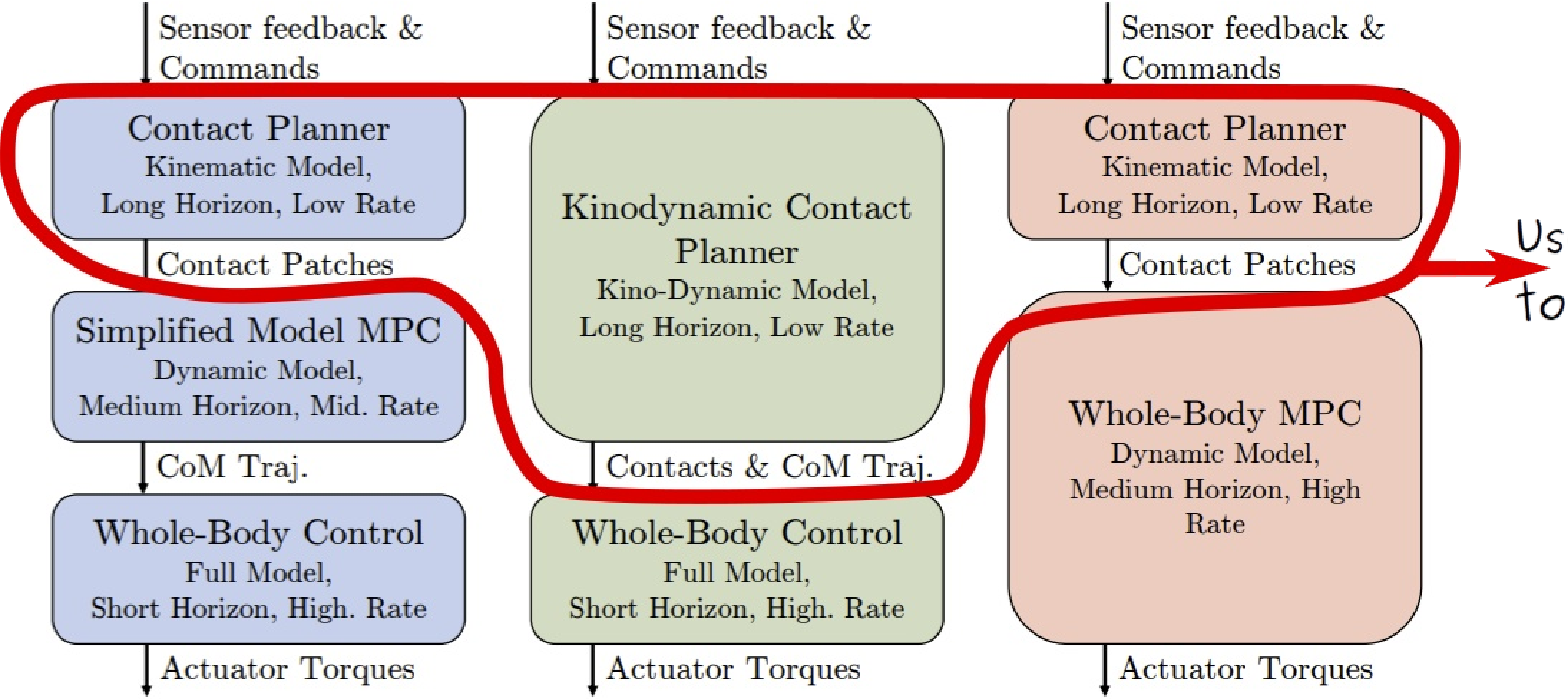
Sensor feedback & Commands

Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

Whole-Body MPC
Dynamic Model,
Medium Horizon, High Rate

Actuator Torques



Sensor feedback & Commands

Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

Simplified Model MPC
Dynamic Model,
Medium Horizon, Mid. Rate

CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

Sensor feedback & Commands

Kinodynamic Contact Planner
Kino-Dynamic Model,
Long Horizon, Low Rate

Contacts & CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

Sensor feedback & Commands

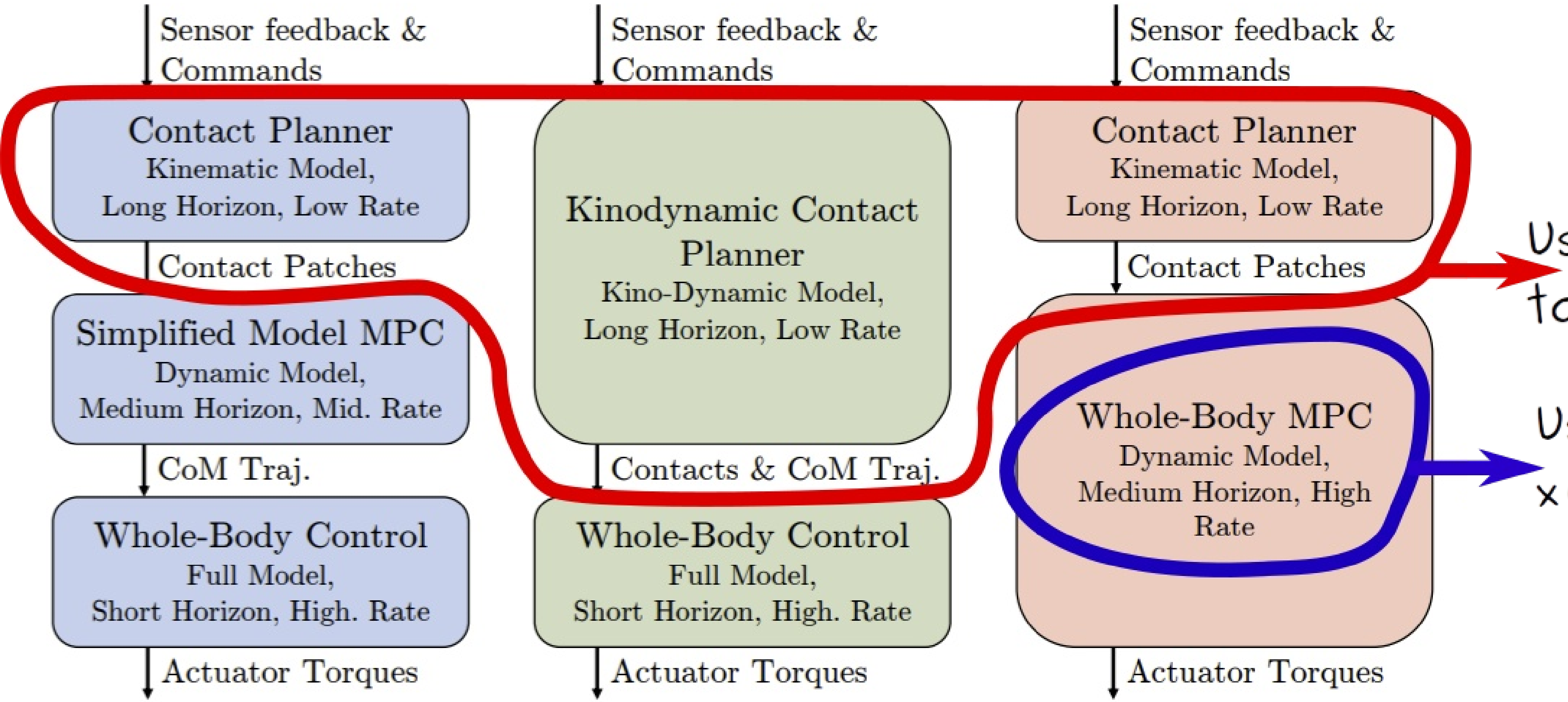
Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

Whole-Body MPC
Dynamic Model,
Medium Horizon, High Rate

Actuator Torques

Use motion planning
to get a rough plan



Sensor feedback & Commands

Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

Simplified Model MPC
Dynamic Model,
Medium Horizon, Mid. Rate

CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

Sensor feedback & Commands

Kinodynamic Contact Planner
Kino-Dynamic Model,
Long Horizon, Low Rate

Contacts & CoM Traj.

Whole-Body Control
Full Model,
Short Horizon, High. Rate

Actuator Torques

Sensor feedback & Commands

Contact Planner
Kinematic Model,
Long Horizon, Low Rate

Contact Patches

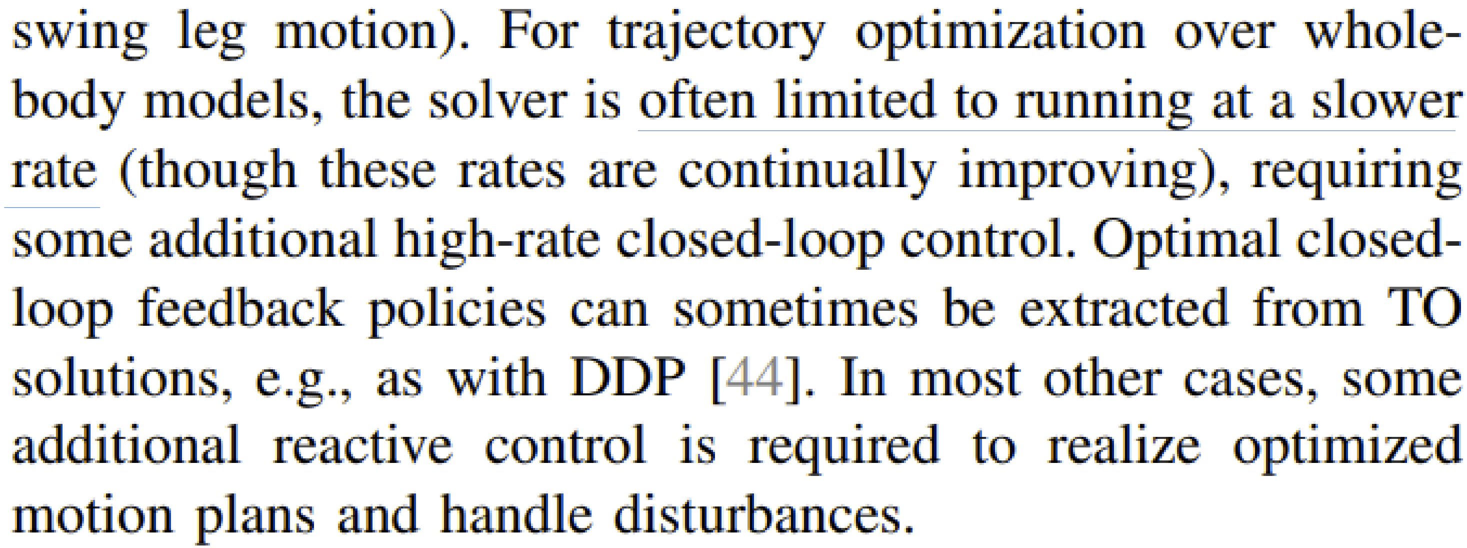
Whole-Body MPC
Dynamic Model,
Medium Horizon, High Rate

Actuator Torques

Use motion planning to get a rough plan

Use OC to generate x(t), u(t) ref.

x(t), u(t) is typically sent to a whole-body controller/stabilizer, or to inverse dynamics if the reference is only x(t).

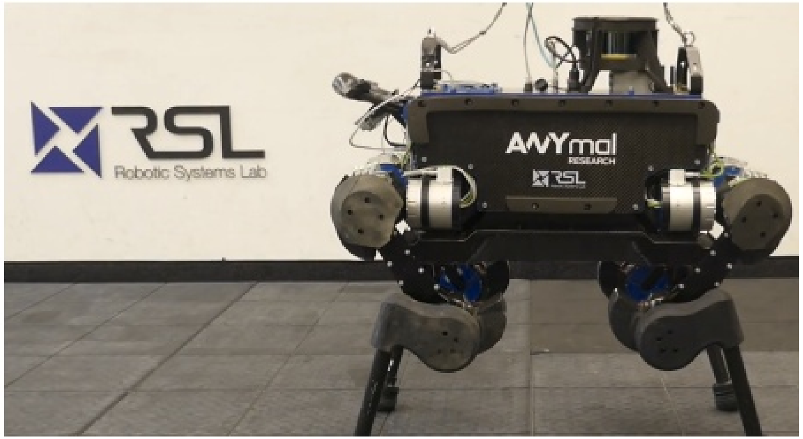


swing leg motion). For trajectory optimization over whole-body models, the solver is often limited to running at a slower rate (though these rates are continually improving), requiring some additional high-rate closed-loop control. Optimal closed-loop feedback policies can sometimes be extracted from TO solutions, e.g., as with DDP [44]. In most other cases, some additional reactive control is required to realize optimized motion plans and handle disturbances.

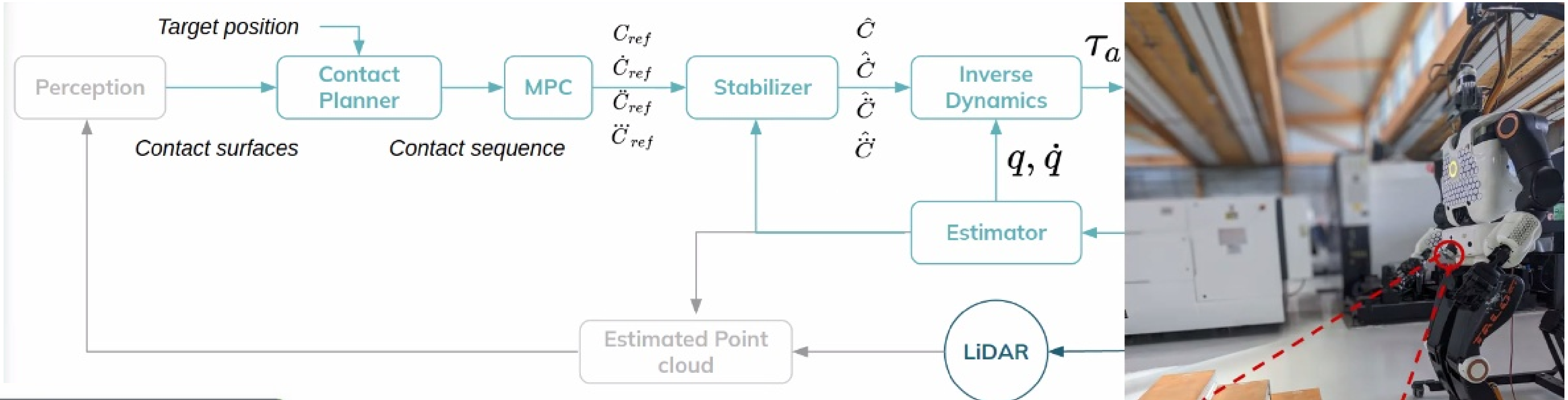
Feedback MPC for Torque-Controlled Legged Robots

Ruben Grandia[1], Farbod Farshidian[1], René Ranftl[2], Marco Hutter[1]

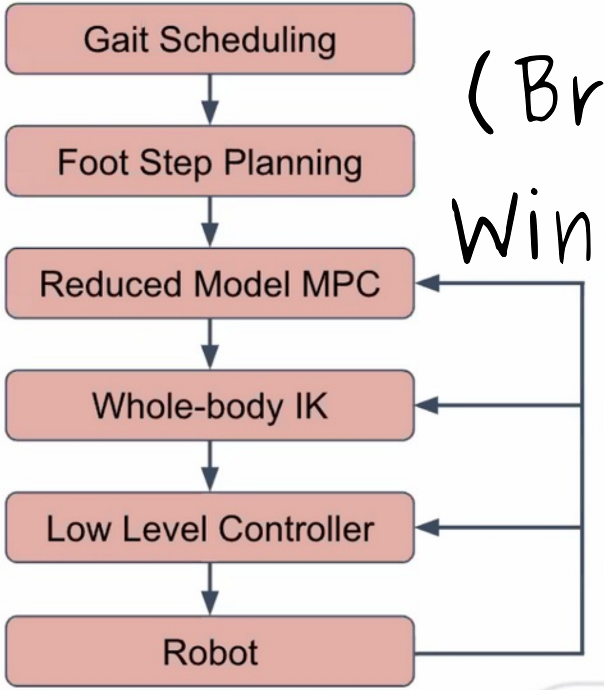
Abstract— **The computational power of mobile robots is currently insufficient to achieve torque level whole-body Model Predictive Control (MPC) at the update rates required for complex dynamic systems such as legged robots. This problem is commonly circumvented by using a fast tracking controller to compensate for model errors between updates. In this work, we show that the feedback policy from a Differential Dynamic Programming (DDP) based MPC algorithm is a viable alternative to bridge the gap between the low MPC update rate and the actuation command rate. We propose to**



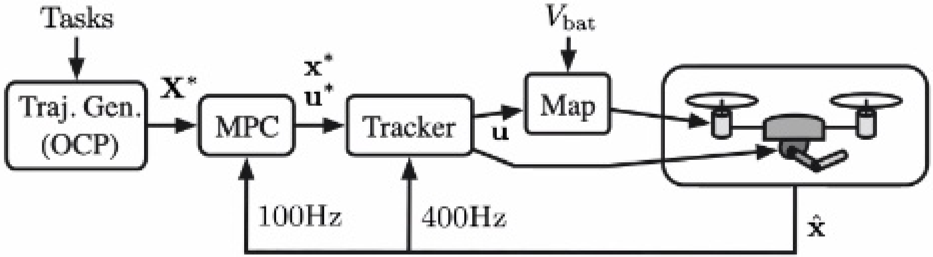
(2019)



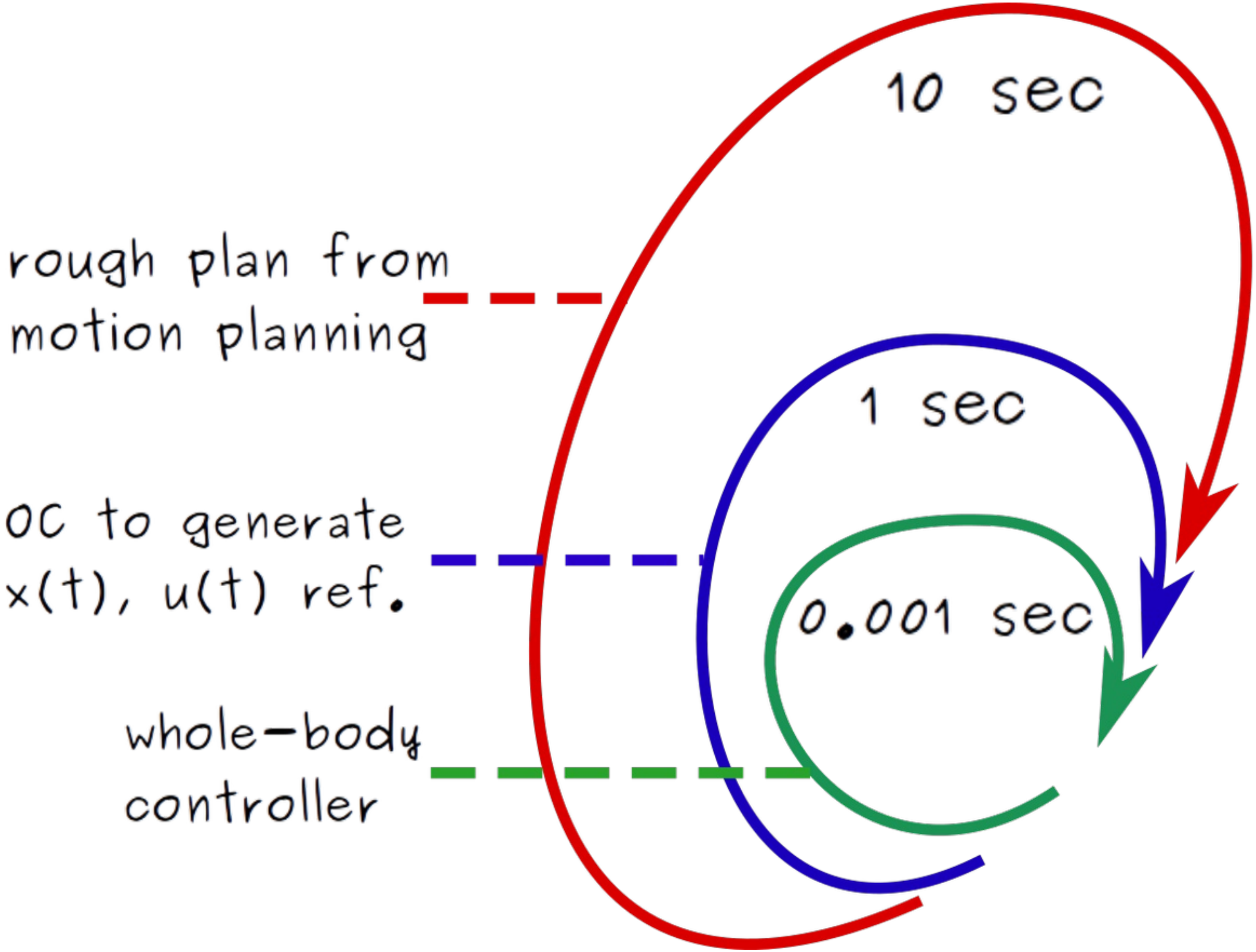
(Pierre Fernbach)



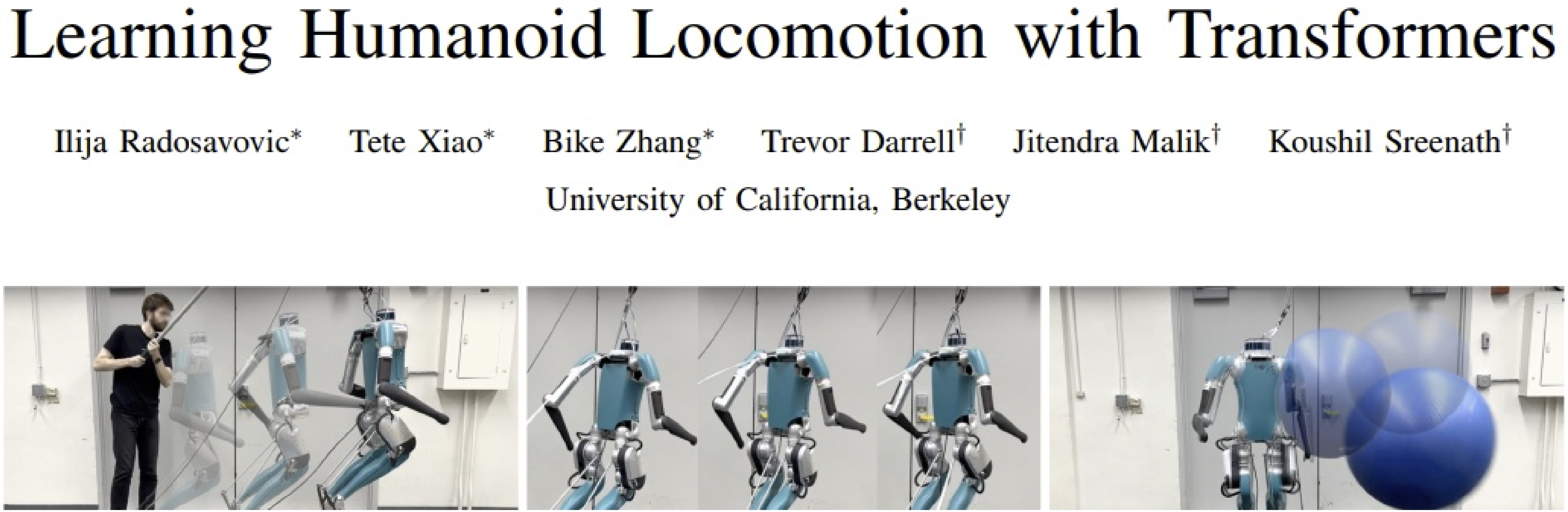
(Bruce Wingo)



(Joan Solà)


10 sec
1 sec
0.001 sec
rough plan from motion planning
OC to generate x(t), u(t) ref.
whole-body controller

Remark: the whole-body controller (WBC) can be a neural network policy



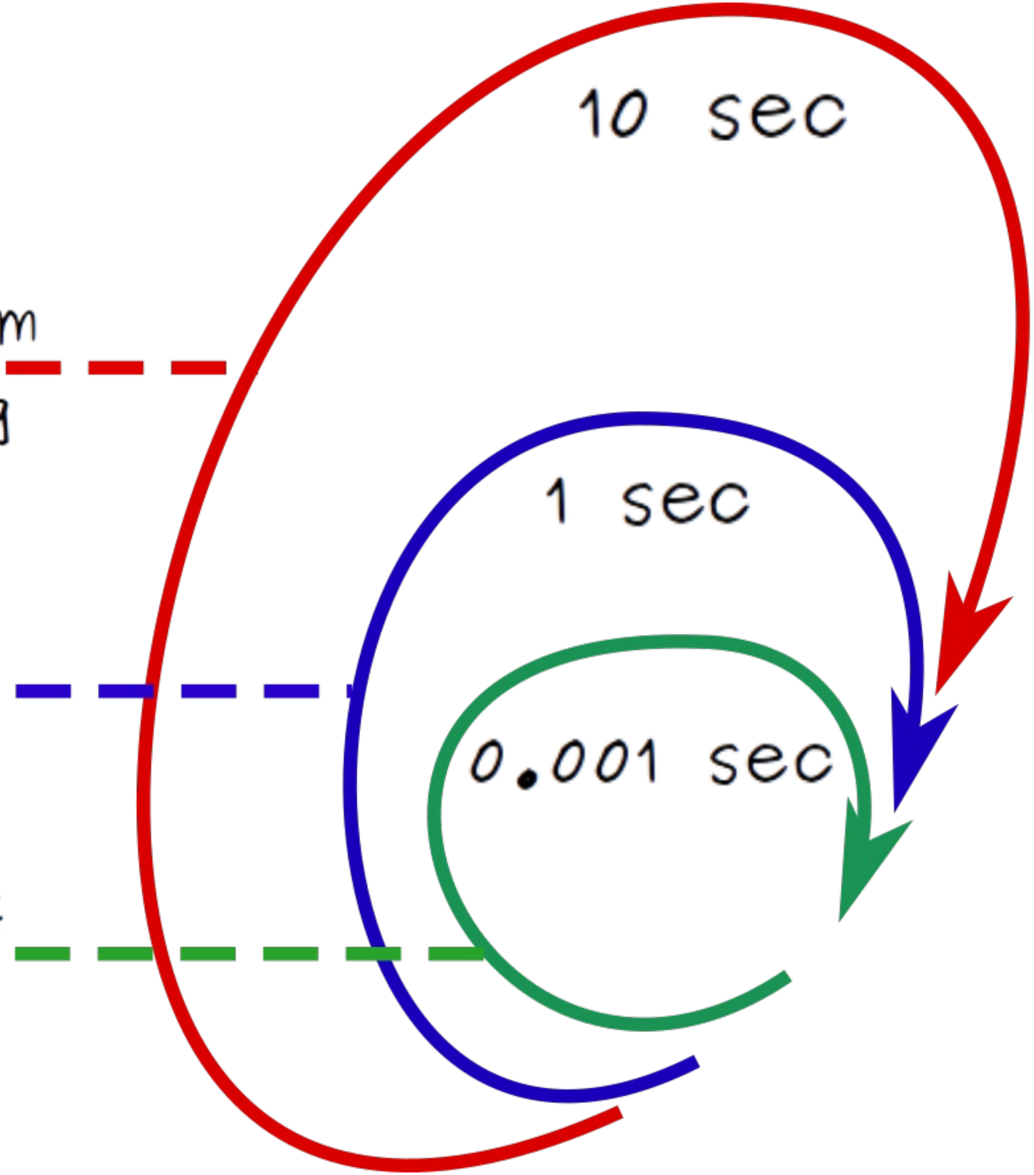
Learning Humanoid Locomotion with Transformers

Ilija Radosavovic* Tete Xiao* Bike Zhang* Trevor Darrell† Jitendra Malik† Koushil Sreenath†

University of California, Berkeley

(2023)

Proposition:



rough plan from motion planning

~~OC to generate x(t), u(t) ref.~~

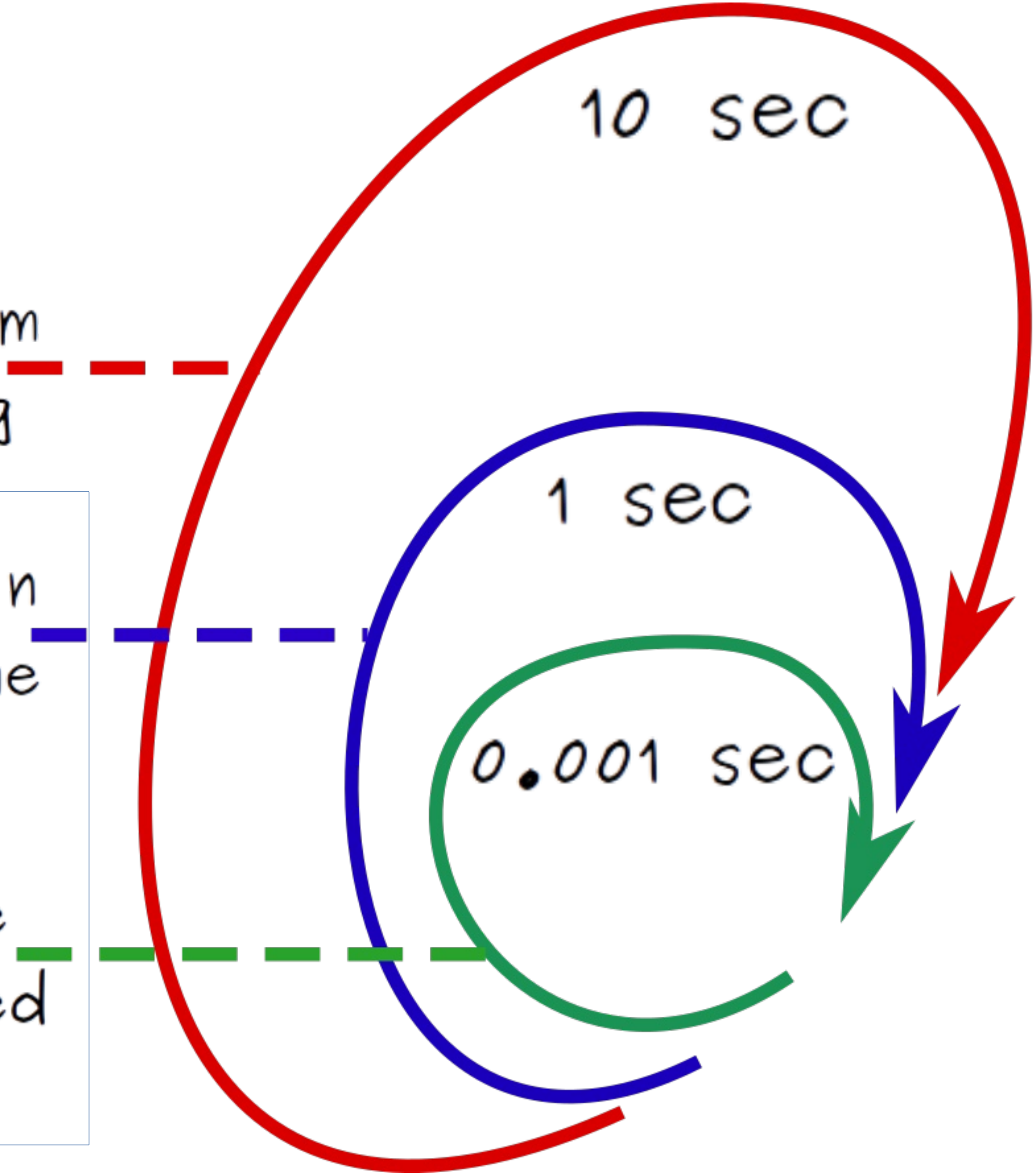
~~whole-body controller~~

Proposition:

rough plan from motion planning

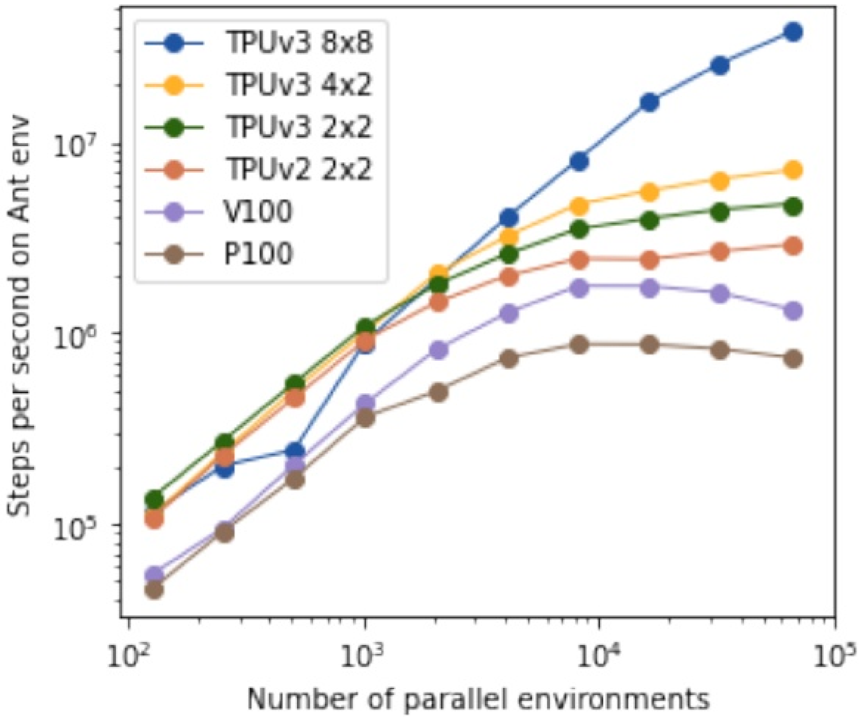
Use RL to train the WBC on the local plan

Apply the fine-tuned WBC

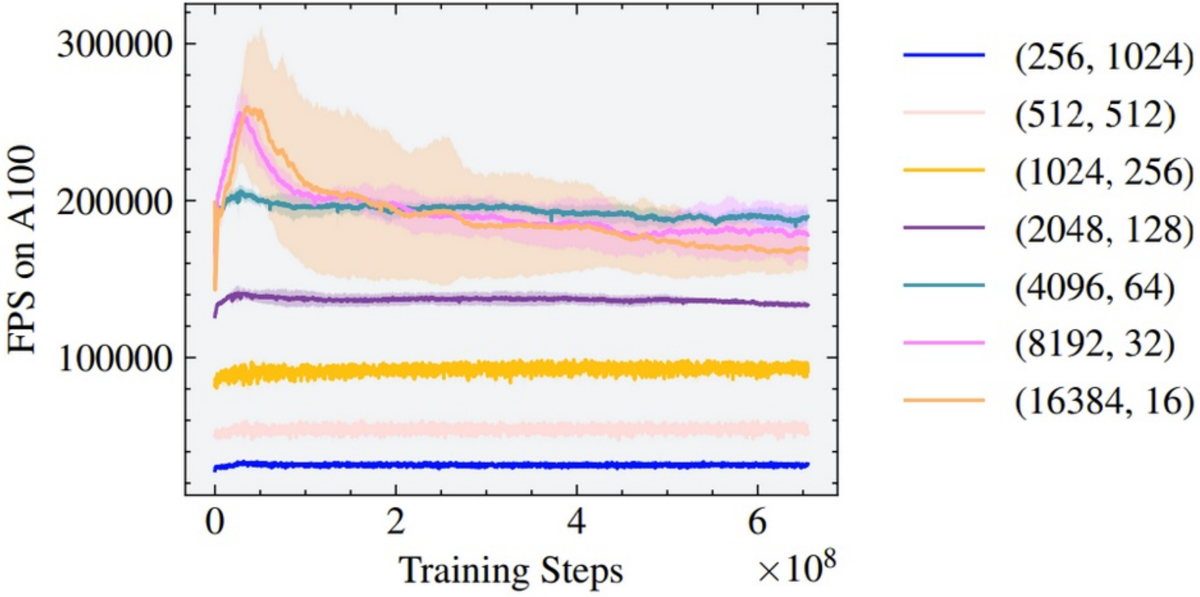


How many steps of training can we do in 1 second?

Brax (ant)



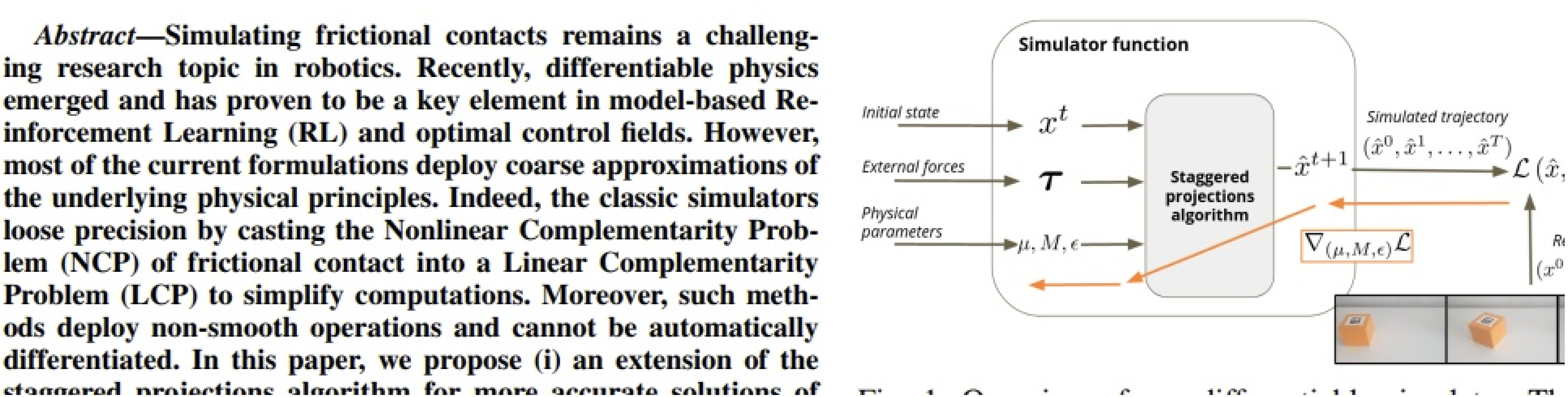
Isaac Gym (humanoid)



(b) Total number of environment steps per second

Differentiable simulation for physical system identification

Quentin Le Lidec[1], Igor Kalevatykh[1], Ivan Laptev[1], Cordelia Schmid[1] and Justin Carpentier[1]



Abstract—Simulating frictional contacts remains a challenging research topic in robotics. Recently, differentiable physics emerged and has proven to be a key element in model-based Reinforcement Learning (RL) and optimal control fields. However, most of the current formulations deploy coarse approximations of the underlying physical principles. Indeed, the classic simulators loose precision by casting the Nonlinear Complementarity Problem (NCP) of frictional contact into a Linear Complementarity Problem (LCP) to simplify computations. Moreover, such methods deploy non-smooth operations and cannot be automatically differentiated. In this paper, we propose (i) an extension of the staggered projections algorithm for more accurate solutions of

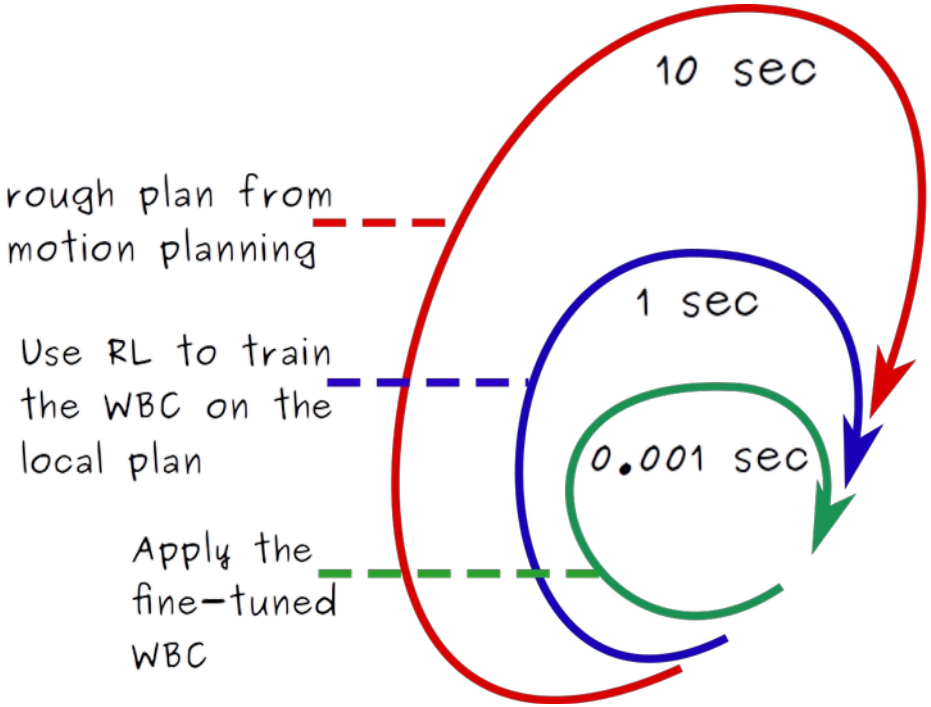
Fig. 1: Overview of our differentiable simulator. Th

(2021)

How many steps of training can we do in 1 second?

About 200k to 500k

RL FROM A SINGLE DEMONSTRATION


10 sec
1 sec
0.001 sec
rough plan from motion planning
Use RL to train the WBC on the local plan
Apply the fine-tuned WBC

To make this approach possible, we must **learn from a single rough demonstration in 500k steps.**

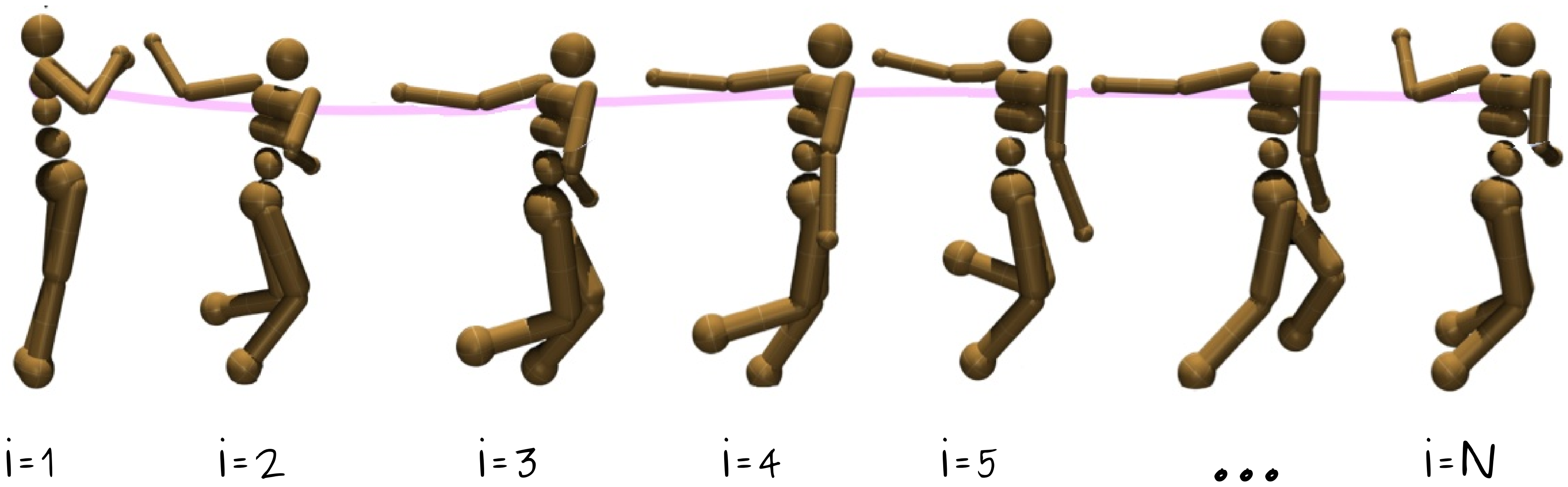
Leveraging Sequentiality in Reinforcement Learning from a Single Demonstration

Alexandre Chenu (ISIR)

Olivier Serris (ISIR)

Olivier Sigaud (ISIR)

Nicolas Perrin-Gilbert (ISIR)

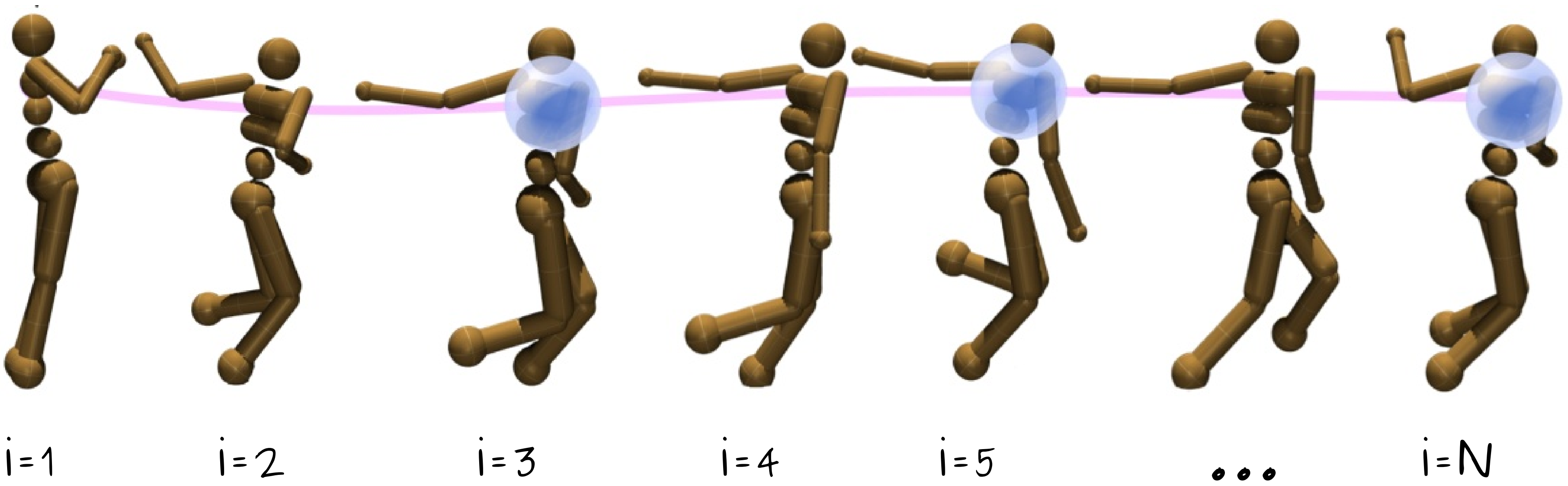

i=1
i=2
i=3
i=4
i=5
...
i=N

The plan, or demonstration, is a sequence of states.

Similarly to **multiple shooting**, we wish to train **almost independenly** a sequence of skills and then chain them.

The main reward is reaching targets, but these target **cannot be small high-dimensional spheres.**

We define low dimensional targets.



This has a big consequence on the difficulty of chaining skills: **their independence is lost.**

To densify the rewards, we use a mechanism called **Hindsight Experience Replay (HER)**, which requires the target to be part of the input:

$$\mu(s, g)$$

But, we must always **"prepare" for the next skill**, so the information of the next target must be available.

$$\mu(s, g, i)$$

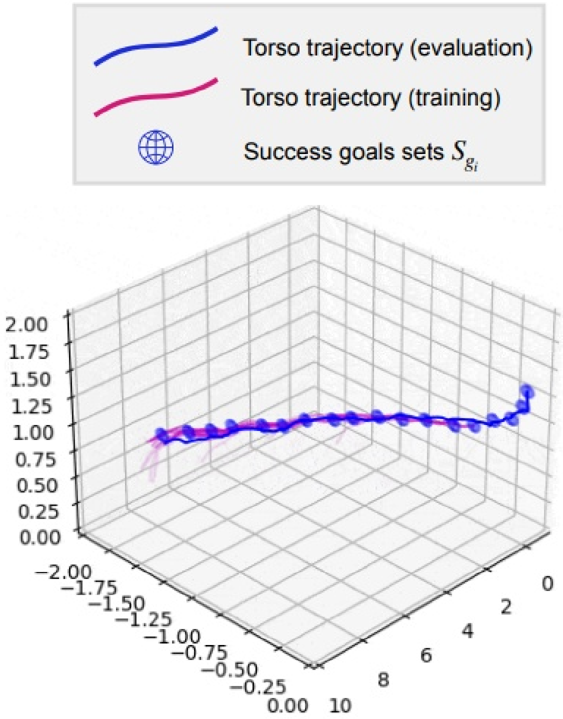
$\mu(s, g, i)$

This formulation is tricky: the current target is free (to use HER), but there is also a fixed sequence of targets.

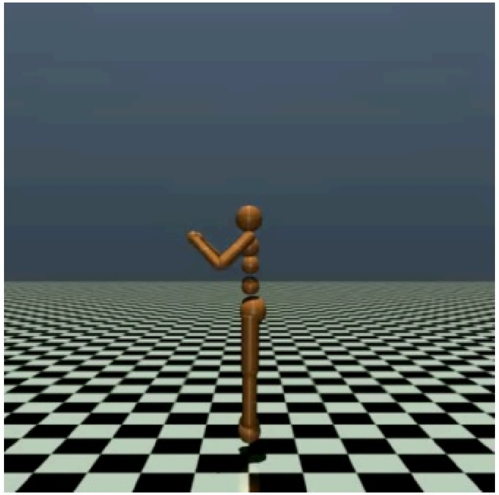
Our article shows how to do this properly, handling both the **target relabelling of HER** and the **backward value propagation** coming from the fixed target sequence.

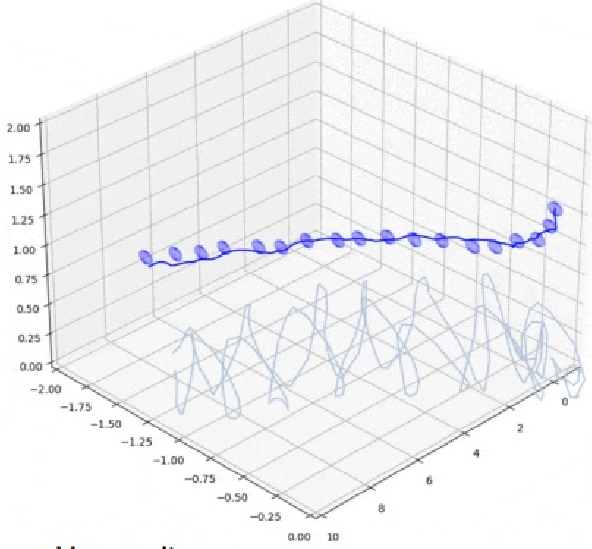
Experiments

Humanoid Locomotion



Training





Sequential goal-reaching result (after 1 257 000 training steps)

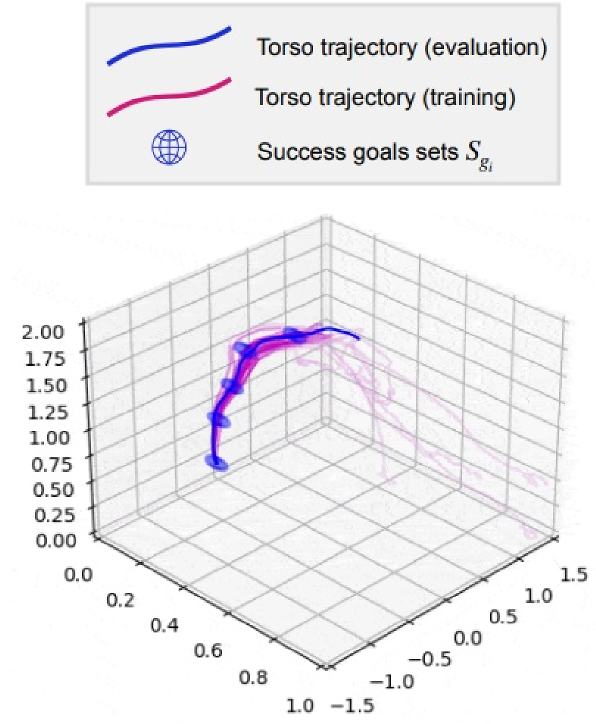
Torso trajectory

Foot trajectories

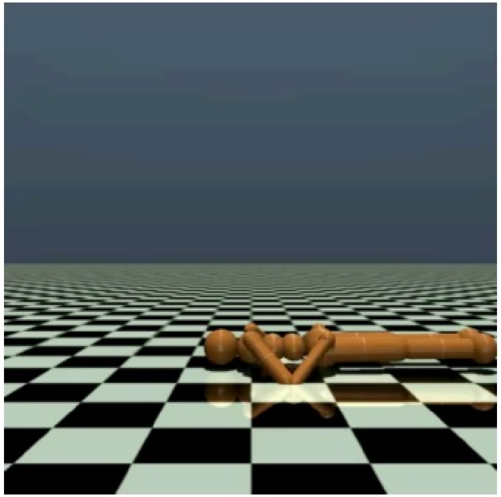
Success goals sets S_{g_i}

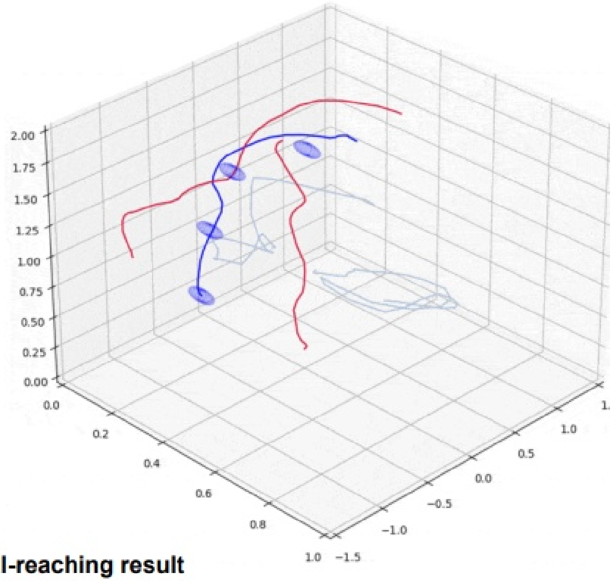
Experiments

Humanoid Stand-up



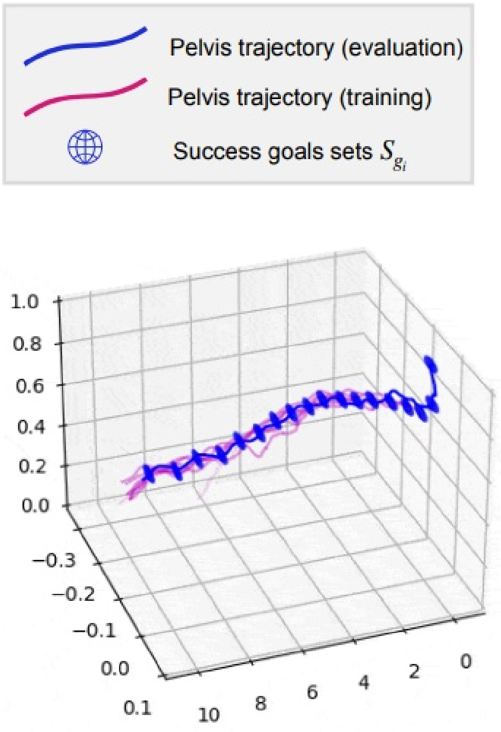
Training



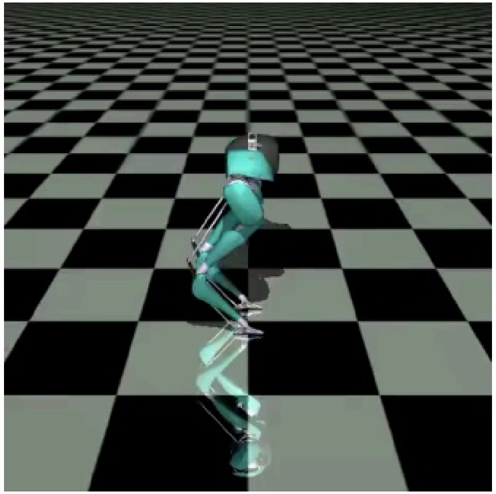


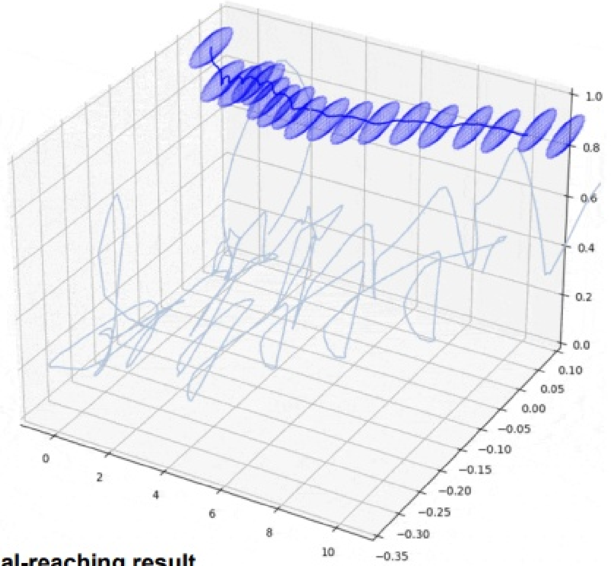
Sequential goal-reaching result (after 138 000 training steps)

Cassie Run



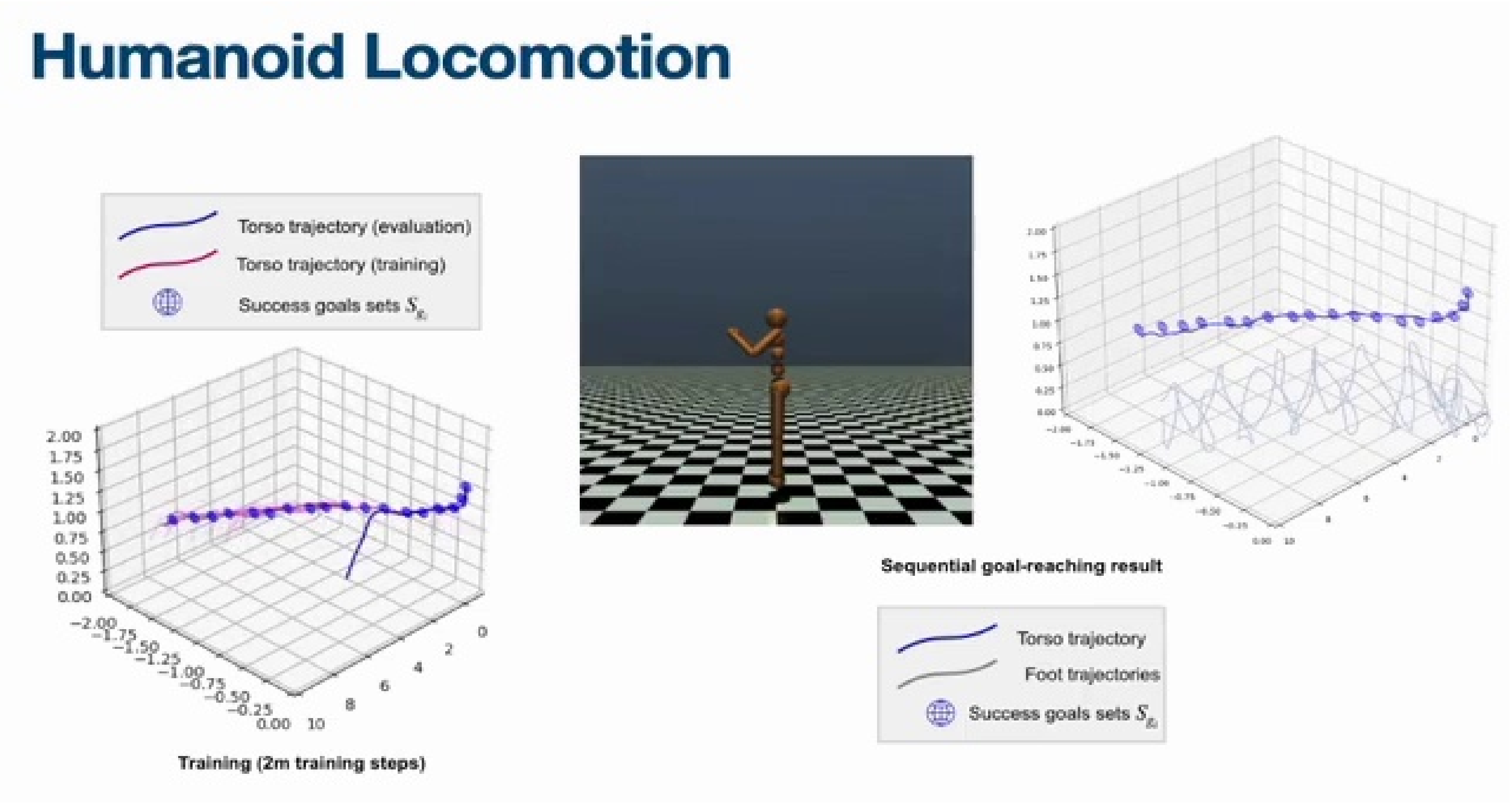
Training





Sequential goal-reaching result (after 718 000 training steps)

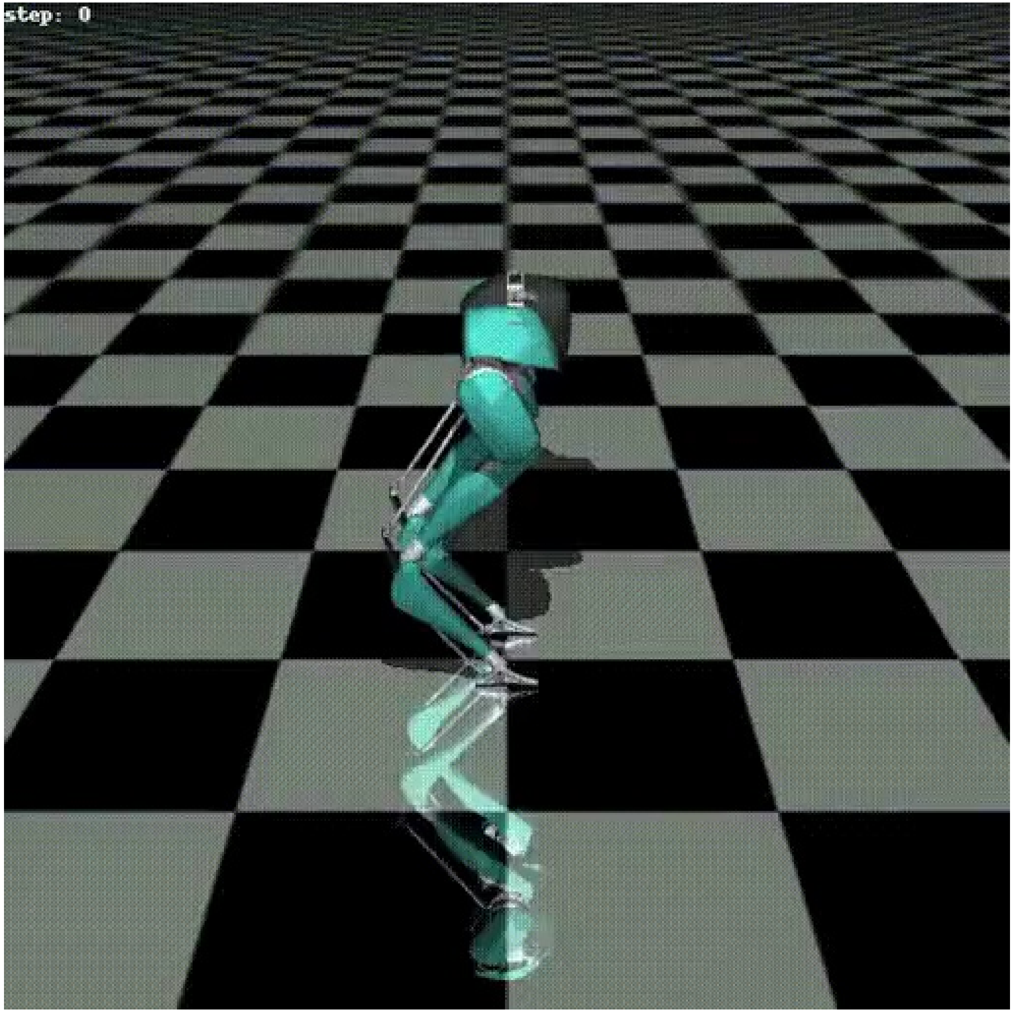
Experiments

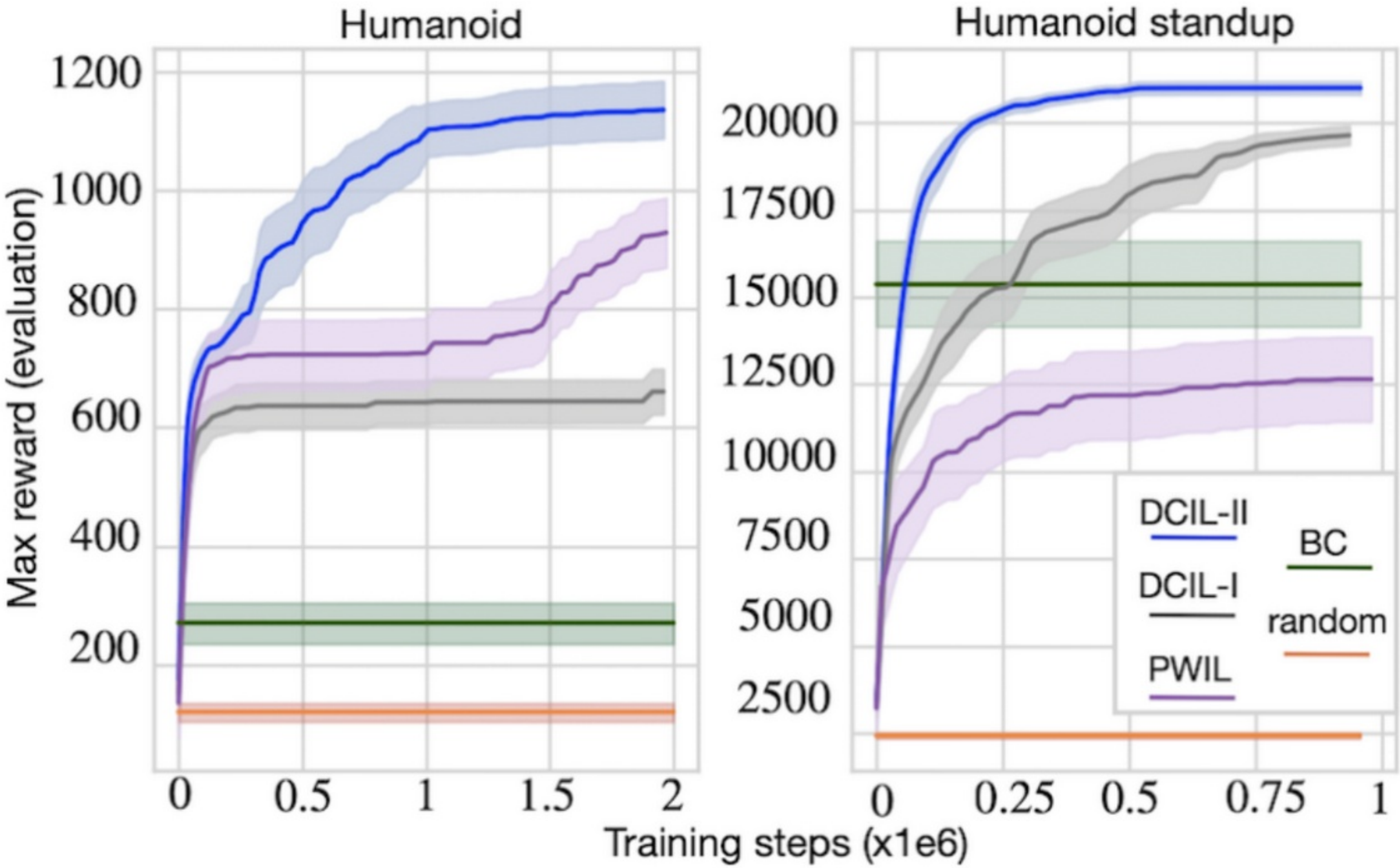


Humanoid Locomotion

Training (2m training steps)

Sequential goal-reaching result


step: 0

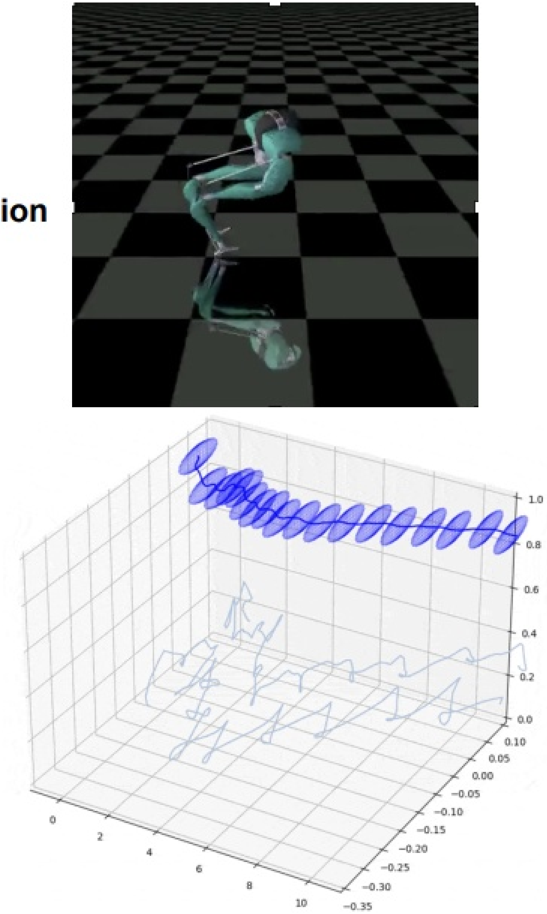


PWIL: Primal Wasserstein Imitation Learning, Dadashi et al. (2020)

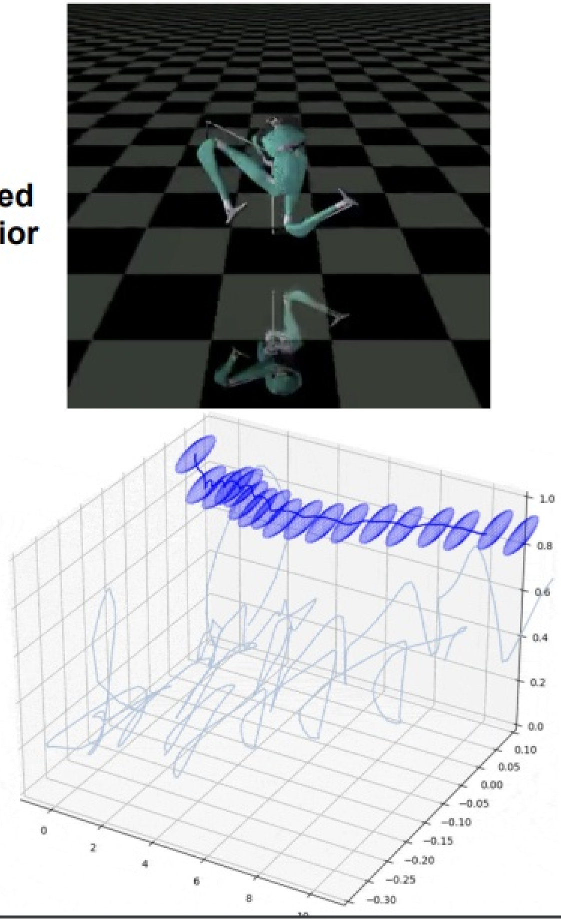
Experiments

Demonstrated VS learned behaviors

Demonstration



Learned behavior



Pelvis trajectory

Foot trajectories

Success goals sets S_{g_i}

Thanks :)