Advancing Robot Intelligence with Reinforcement Learning

Ashish Kumar

What is reinforcement learning?

Imitation Learning

- **Used in:** VLMs, video generation, image segmentation, etc
- **Characteristics:**
  - Relies heavily on off-policy data
  - Learns from good examples
  - Gives generalists

Reinforcement learning

- **Used in:** Game playing, reasoning, robotics
- **Characteristics:**
  - On policy data
  - Uses good as well as bad data
  - Gives Specialists

What has reinforcement learning achieved?

- **AlphaGo** [1]: Learned to exceed what humans can do
- **Reasoning in LLMs** [2]: Coherent over long horizon
- **Dexterity in robotics** [rest of the talk]: Reliability

What enabled these successes ?

- Well specified rewards
- Scalable evaluation of policies

[1] Silver, David, et al. "Mastering the game of Go with deep neural networks and tree search." *nature* (2016)

[2] Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv:2501.12948* (2025)

Zooming in on robotics

What enabled these success ?

- Well specified rewards —> Programmatically calculate them
- Scalable evaluation of policies —> Simulation

Simulation to real?

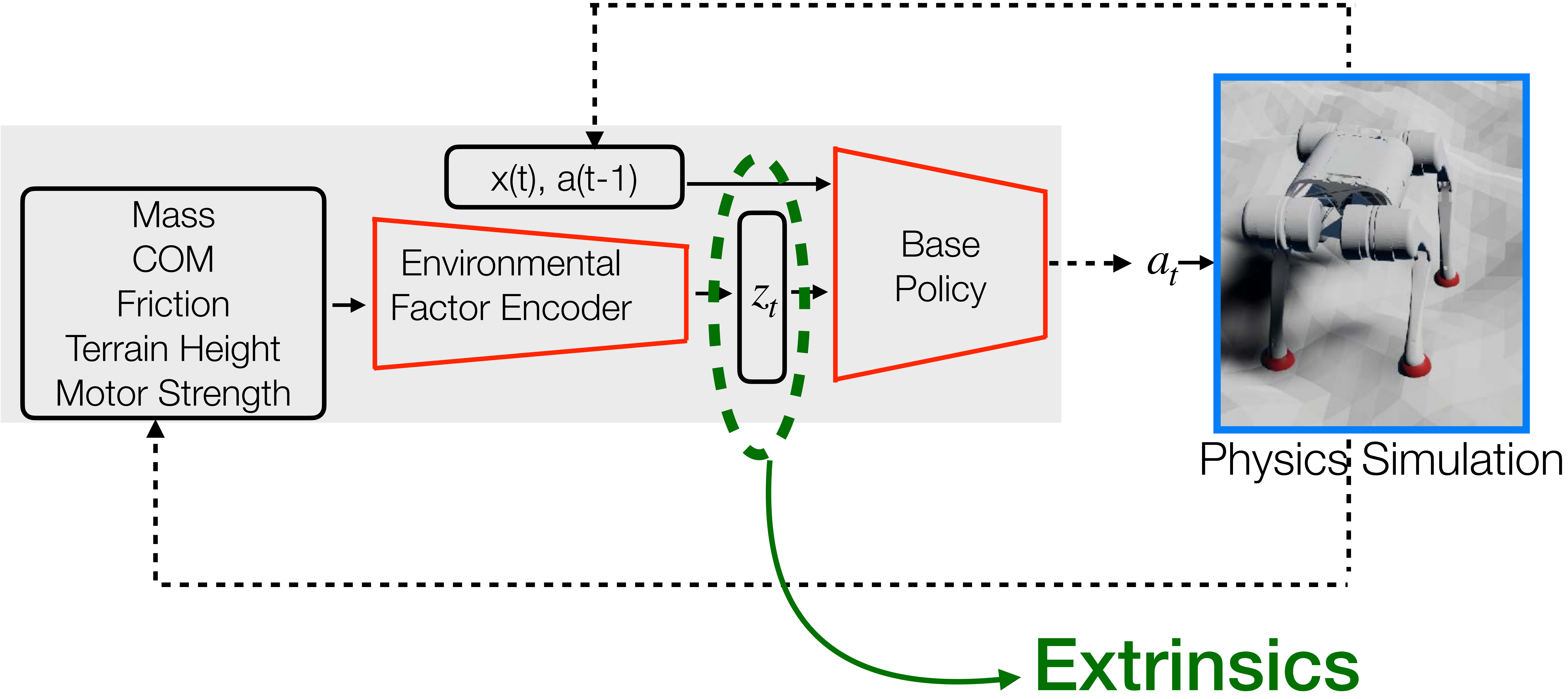

A

One Policy to Walk them All

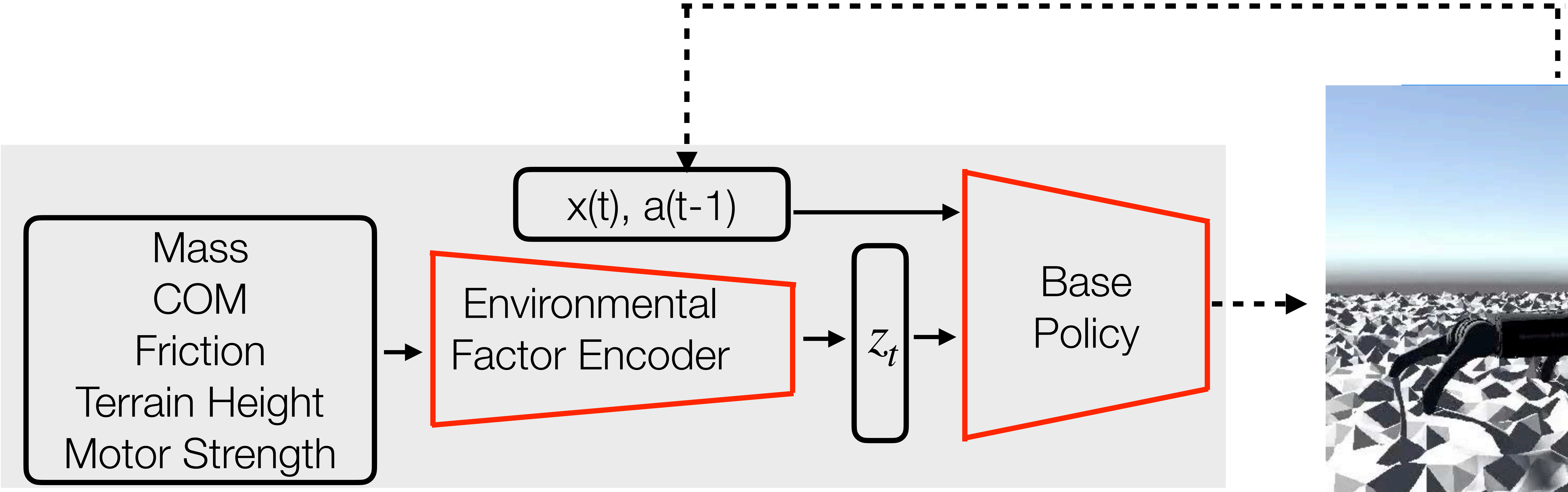
Rapid Motor Adaptation for Legged Robots

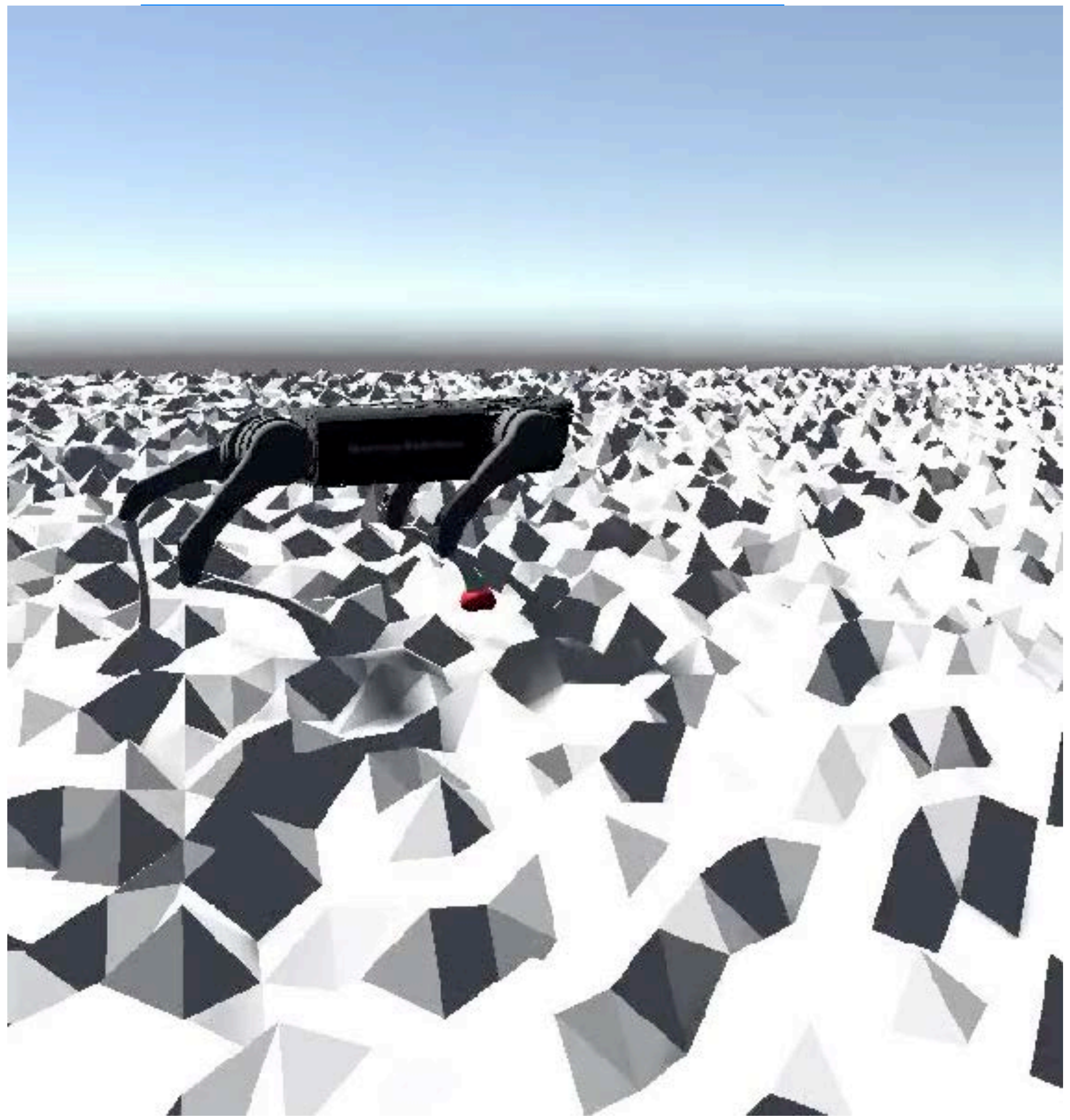
Ashish Kumar	Zipeng Fu	Deepak Pathak	Jitendra Malik
UC Berkeley	CMU	CMU	UC Berkeley/FAIR

Learn to Walk in Simulation



Learn to Walk in Simulation





- Important: Reward Function minimizes work and ground impact (Biomechanics and Energetics)

Reward Function

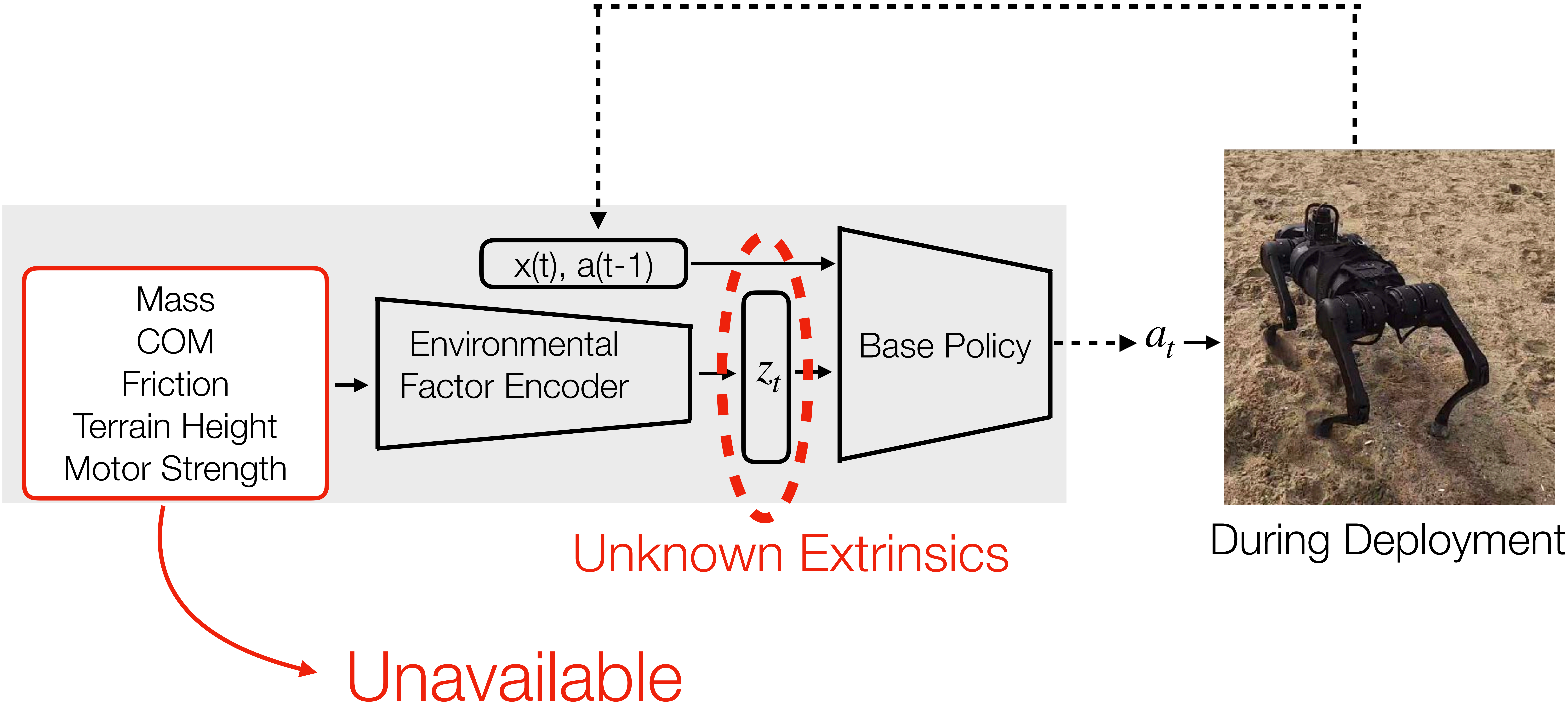
1) Forward: $\min(v_x^t, 0.35)$
2) Lateral Movement and Rotation: $-\|v_y^t\|^2 - \|\omega_{\mathtt{yaw}}^t\|^2$
3) Work: $-|\boldsymbol{\tau}^T \cdot (\mathbf{q}^t - \mathbf{q}^{t-1})|$
4) Ground Impact: $-\|\mathbf{f}^t - \mathbf{f}^{t-1}\|^2$
5) Smoothness: $-\|\boldsymbol{\tau}^t - \boldsymbol{\tau}^{t-1}\|^2$
6) Action Magnitude: $-\|\mathbf{a}^t\|^2$
7) Joint Speed: $-\|\dot{\mathbf{q}}^t\|^2$
8) Orientation: $-\|\boldsymbol{\theta}_{\mathtt{roll,\ pitch}}^t\|^2$
9) Z Acceleration: $-\|v_z^t\|^2$
10) Foot Slip: $-\|\mathrm{diag}(\mathbf{g}^t) \cdot \mathbf{v_f}^t\|^2$

Forward Walking

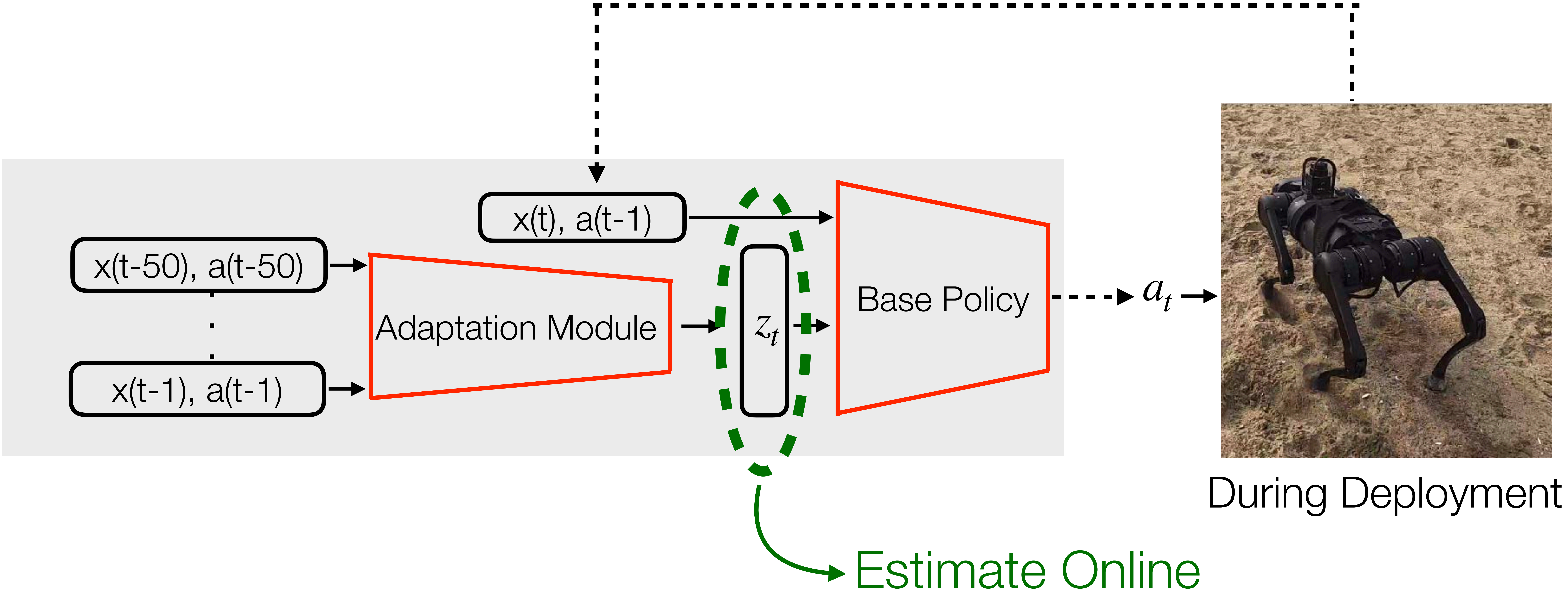
Energetics

Stability + Minimize hardware damage

How can we deploy it?

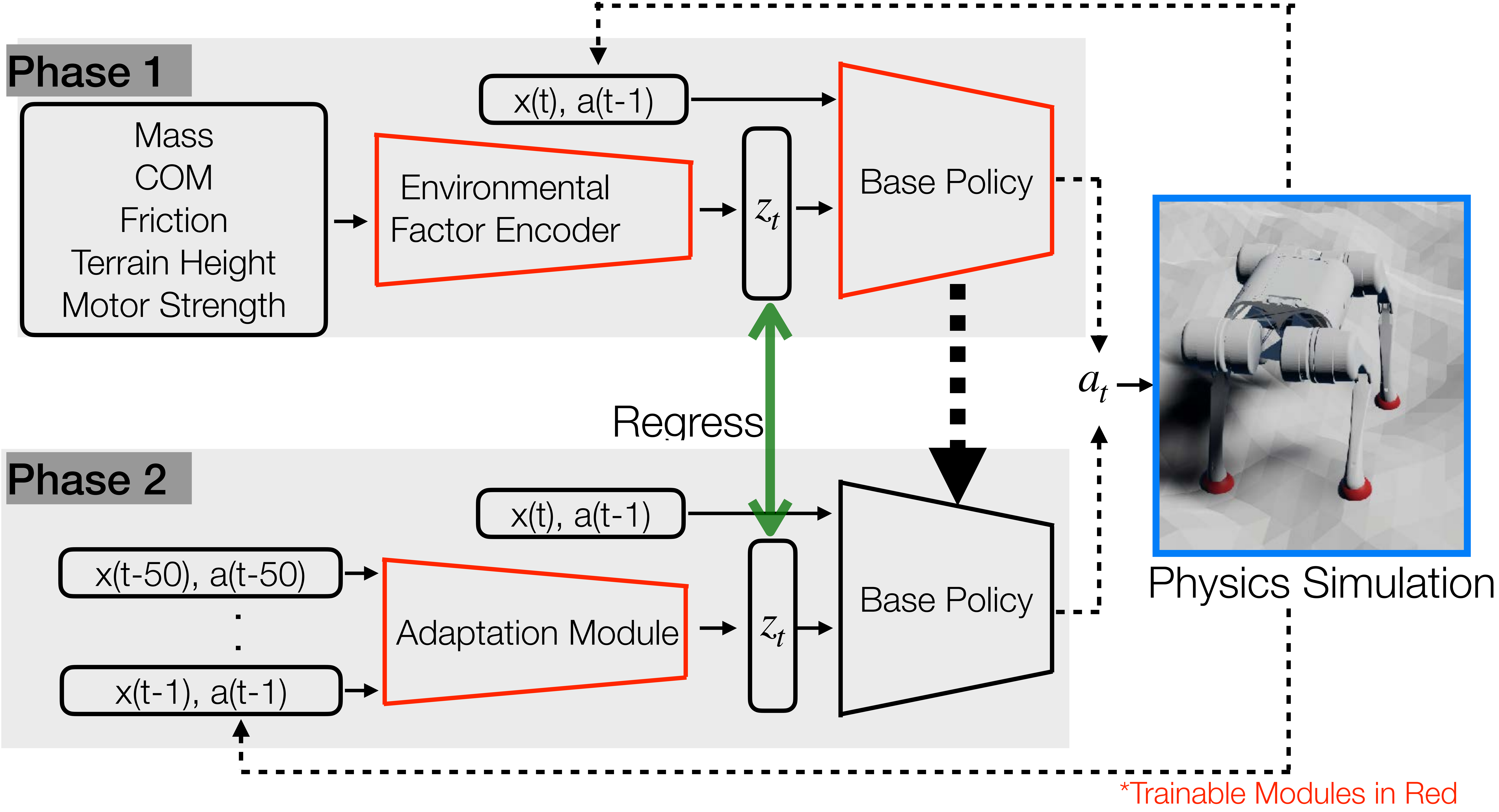


Key Insight — Extrinsics from Observation History

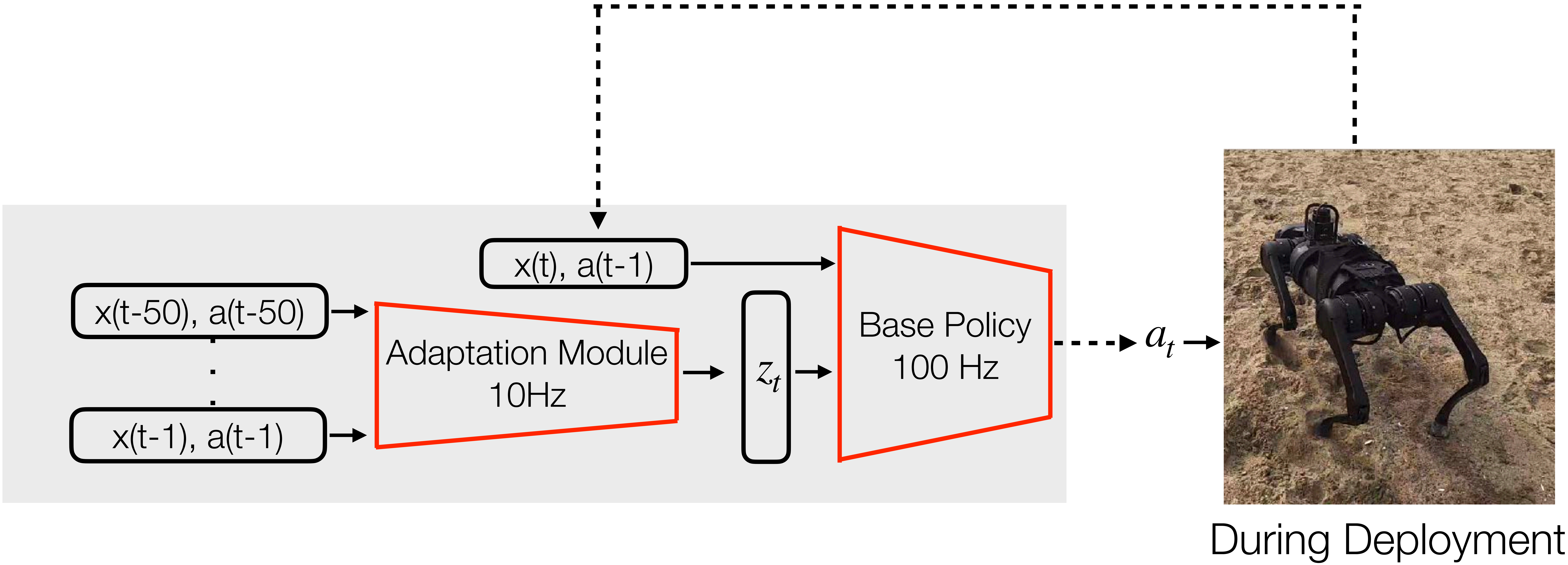


- Discrepancy b/w expected movement and actual measured movement
- Continuously estimate these extrinsics online

Training Summary



Test Time



x(t), a(t-1)
x(t-50), a(t-50)
x(t-1), a(t-1)
Adaptation Module
10Hz
z_t
Base Policy
100 Hz
a_t
During Deployment

Indoors Evaluation


MOVING SERVICES

Oily surface and plastic wrapped feet


SLOW
Oily surface

5kg payload throw


SLOW
5kg payload throw

Planks - uneven footholds and moving surfaces

RMA - Indoors

Comparison to RMA w/o Adaptation


RMA
8kg payload
RMA w/o Adaptation

RMA

RMA w/o Adaptation

Analysis of Adaptation Module

Oily surface and plastic wrapped feet


RL
RR
FL
FR
Knee Tau
10
0
z_1, z_5 value
0.5
0


5kg payload throw


RL
RR
FL
FR
Knee Tau
10
0
z_2, z_7 value
1
0.4

Quantitative Comparison

Existing Attempts at Learning General Policies

	Success (%)	TTF	Reward	Distance (m)	Samples	Torque	Jerk	Ground Impact
Robust [49, 38]	62.4	0.80	4.62	1.13	0	527.59	122.50	4.20
SysID [54]	56.5	0.74	4.82	1.17	0	565.85	149.75	4.03
AWR [39]	41.7	0.65	4.17	0.95	40k	599.71	162.60	4.02
RMA w/o Adapt	52.1	0.75	4.72	1.15	0	524.18	106.25	4.55
RMA	73.5	0.85	5.22	1.34	0	500.00	92.85	4.27
Expert	76.2	0.86	5.23	1.35	0	485.07	85.56	3.90

- Domain Randomization (Robust)
- System Identification
- Fine tuning at test time in the real world

Why do we need vision?

Perception enables *precise* and *agile* locomotion

Legged Locomotion in Challenging Terrains using Egocentric Vision

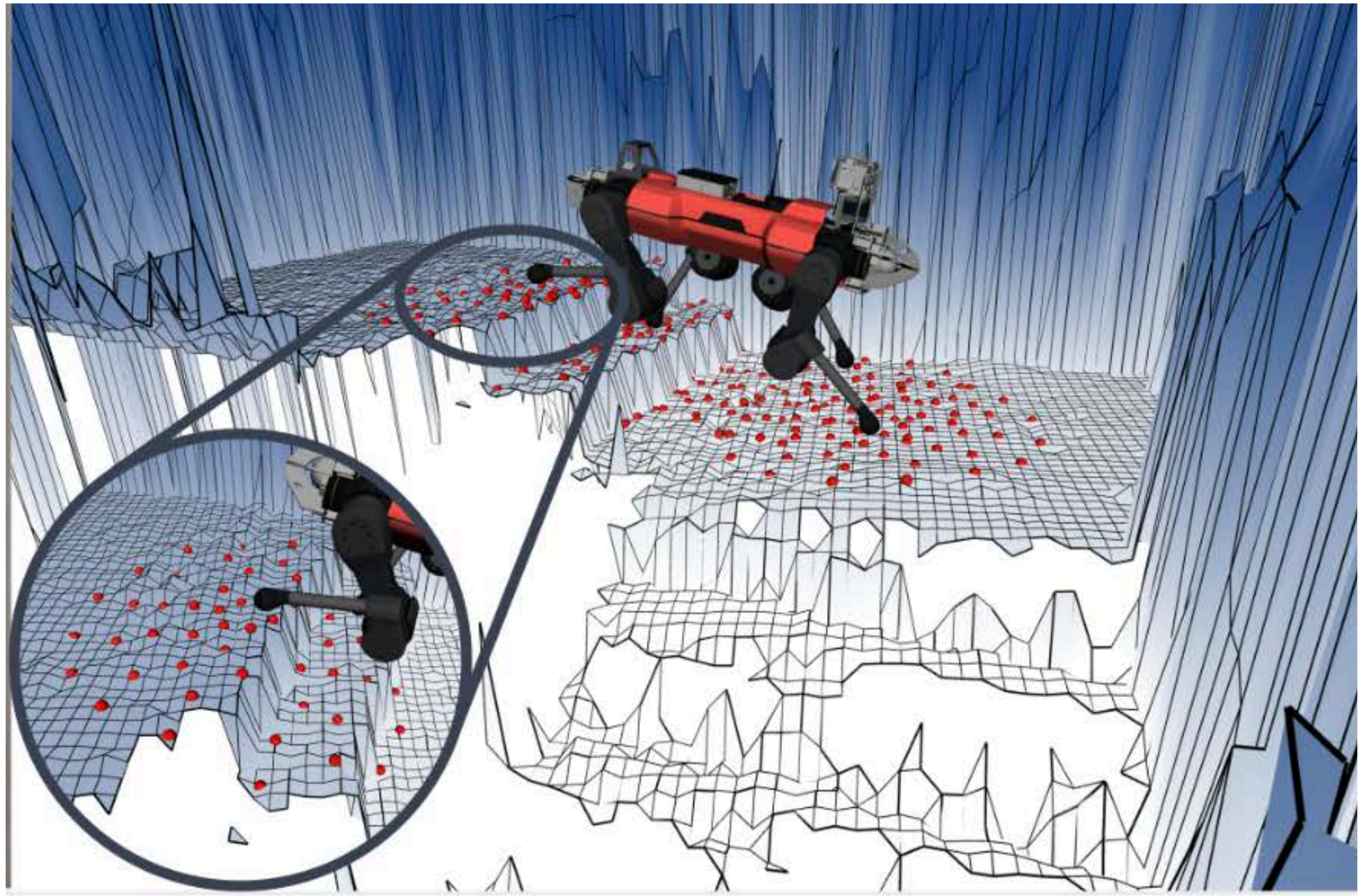
Ananye Agarwal*
CMU

Ashish Kumar*
UC Berkeley

Jitendra Malik†
UC Berkeley

Deepak Pathak†
CMU

Typical approach: build terrain maps from vision

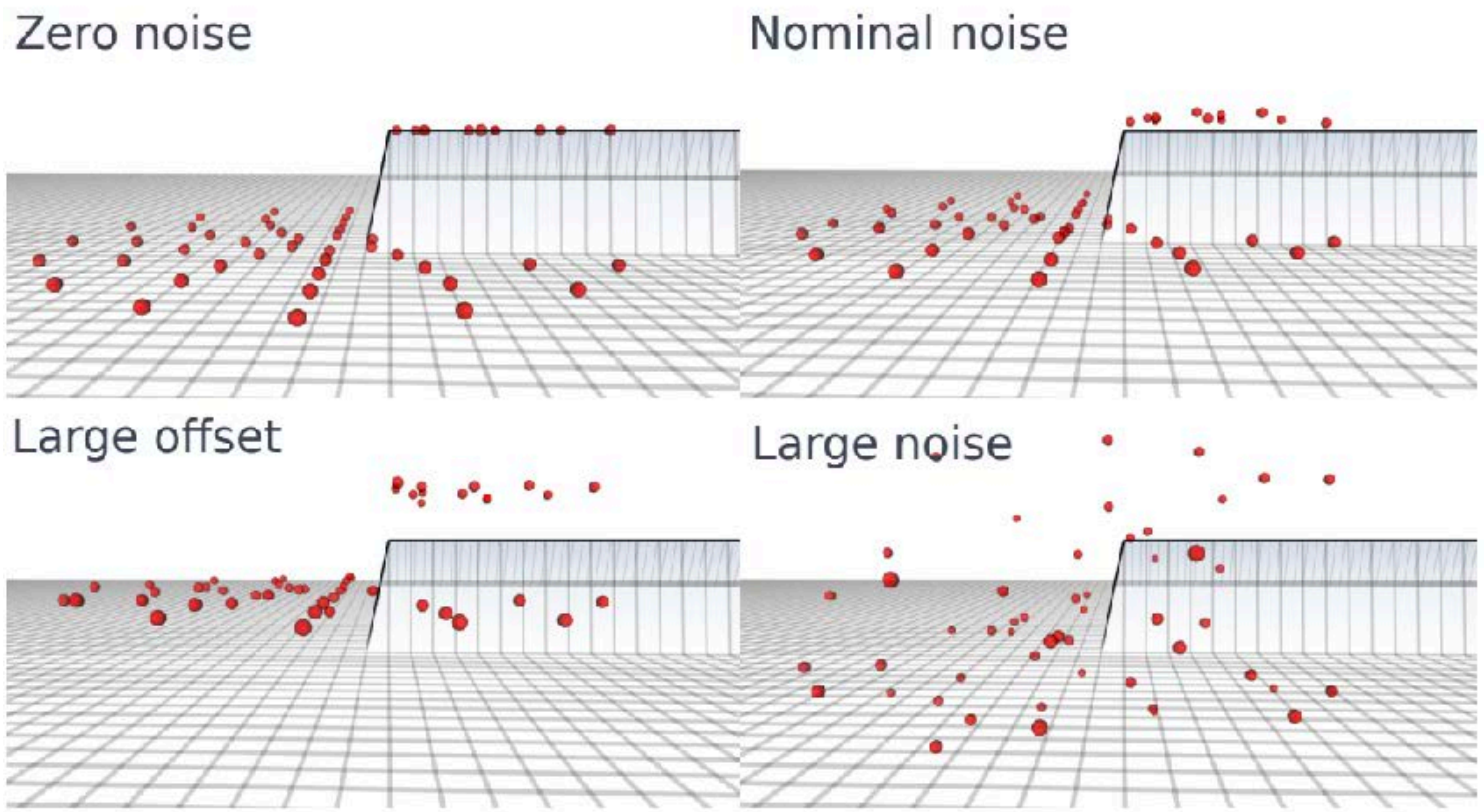


Miki, Takahiro, et al. "Learning robust perceptive locomotion for quadrupedal robots in the wild." *Science Robotics* 7.62 (2022)

Kim, Donghyun, et al. "Vision aided dynamic exploration of unstructured terrain with a small-scale quadruped robot." ICRA 2020.

But terrain maps are very noisy!

Noisy maps wipe out signal => degraded performance



Miki, Takahiro, et al. "Learning robust perceptive locomotion for quadrupedal robots in the wild." Science Robotics 7.62 (2022)

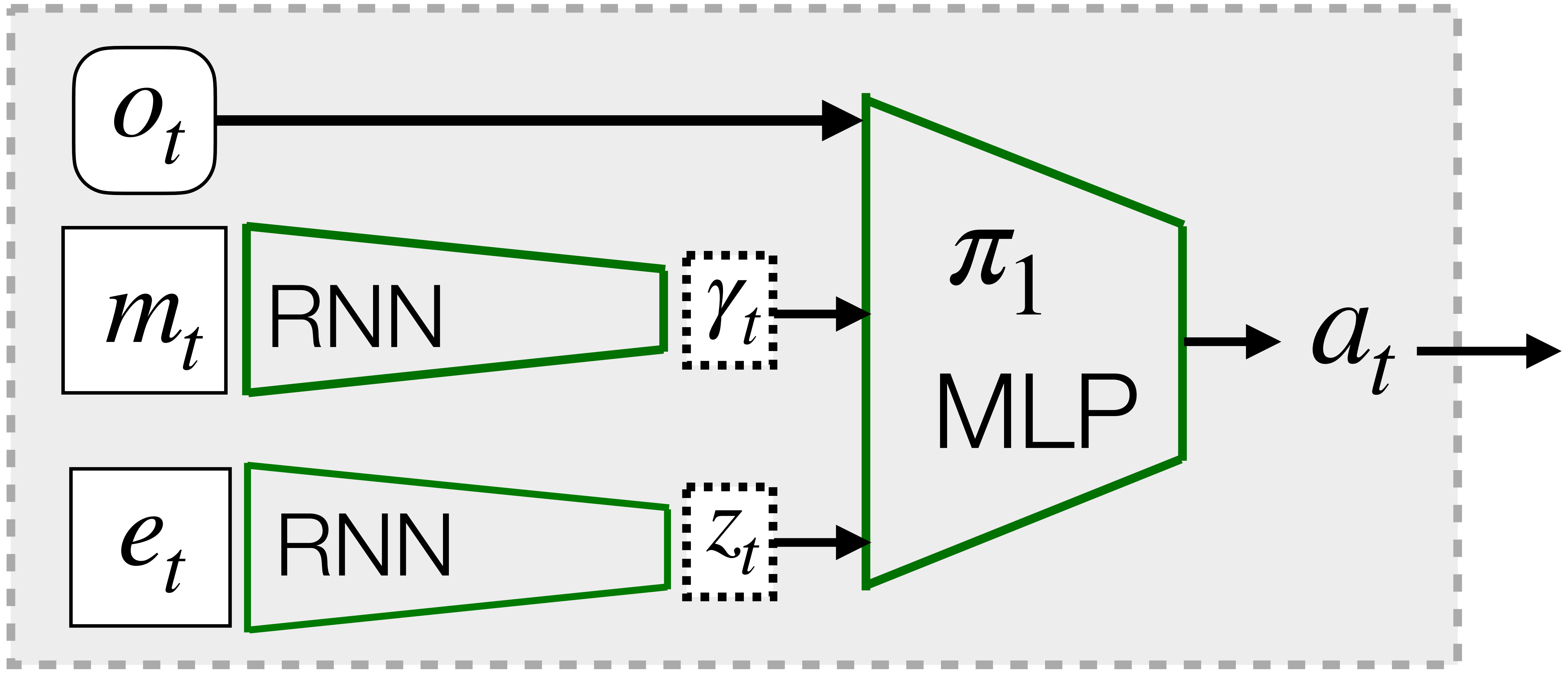
?

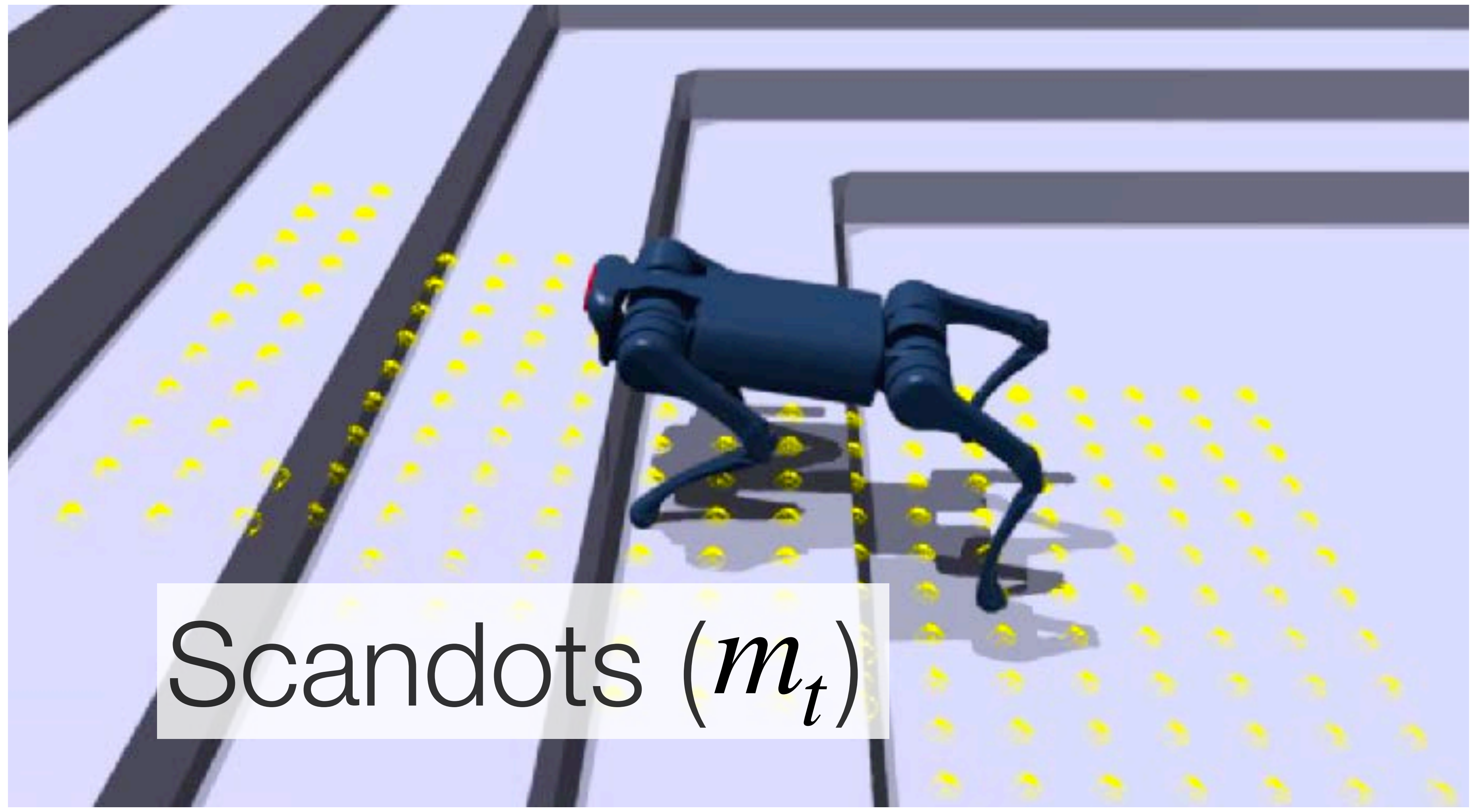
Do we really need terrain maps?

We tightly couple vision and control


Stepping Stones (~15 cm apart)

Phase 1: Learning to Walk with Privileged Terrain Information



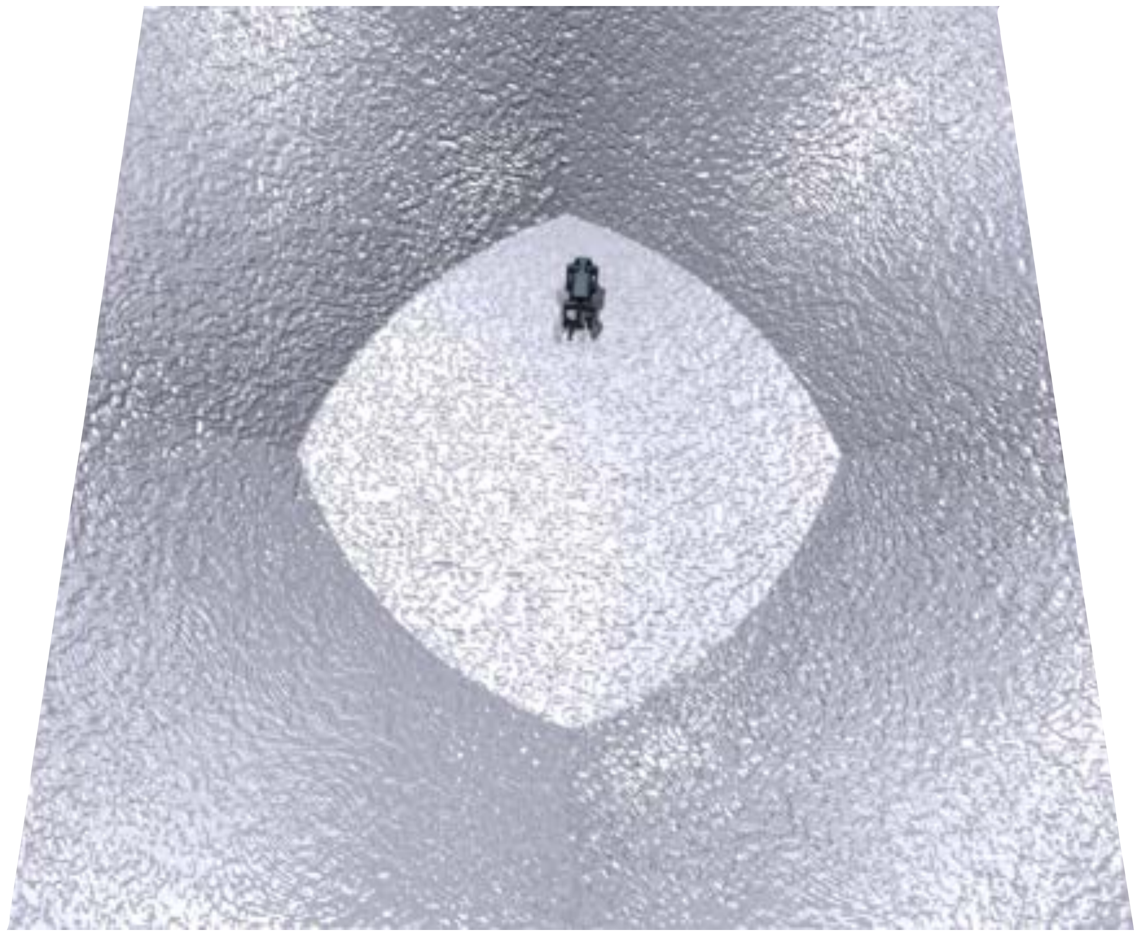


PPO in IsaacGym

Reward = (Track Velocity) + (Minimize Energy)

Training Terrains

Stairs

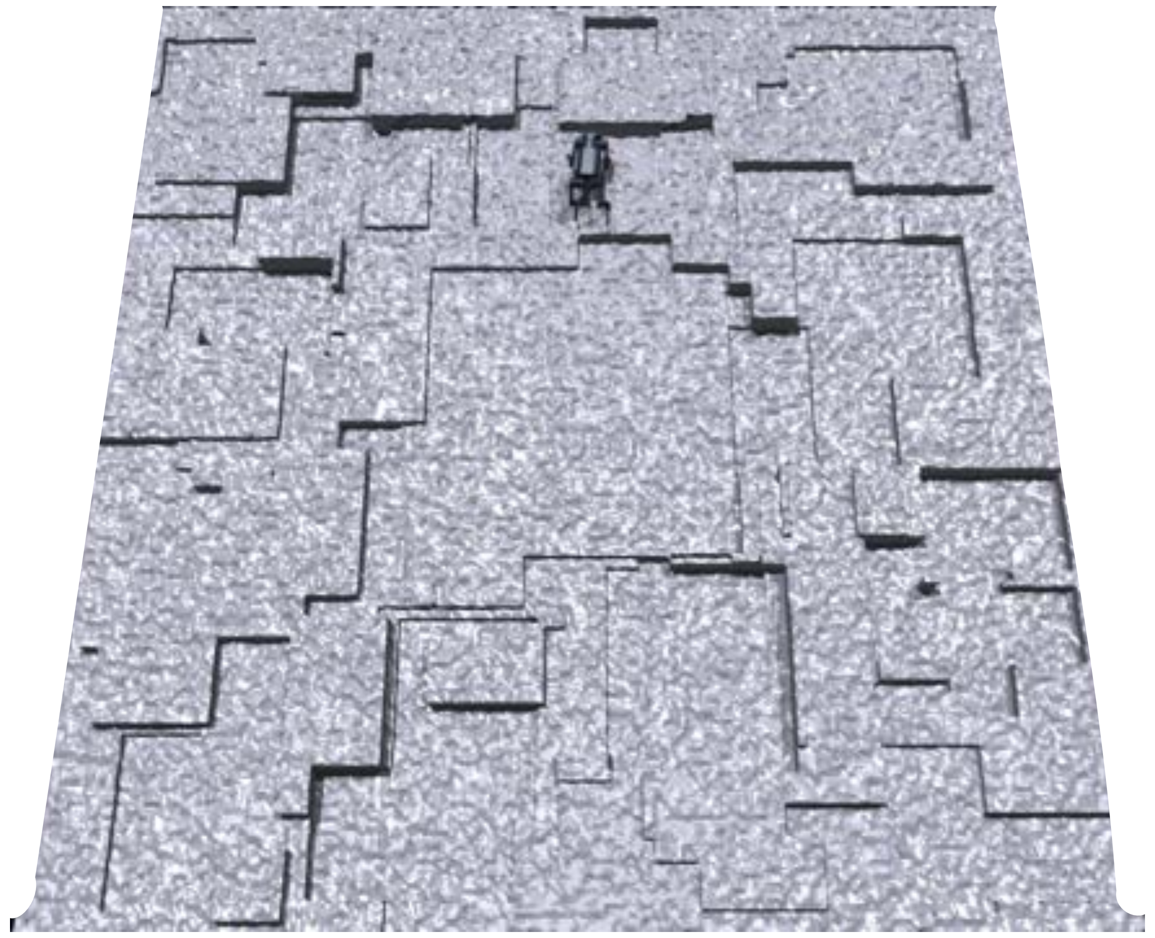


Slopes

Stepping Stones

Rough Flat

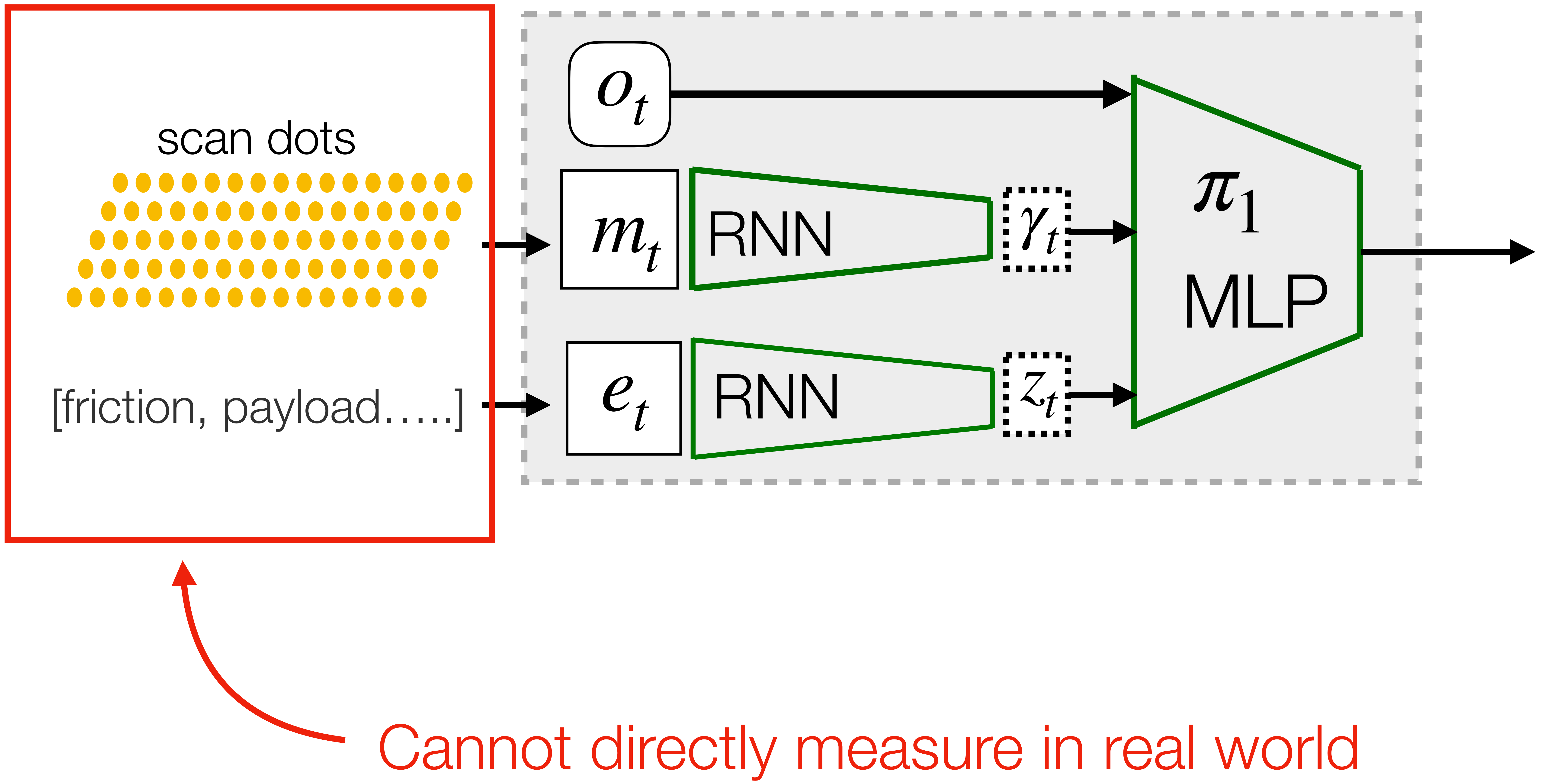
Gaps



Discrete Obstacles

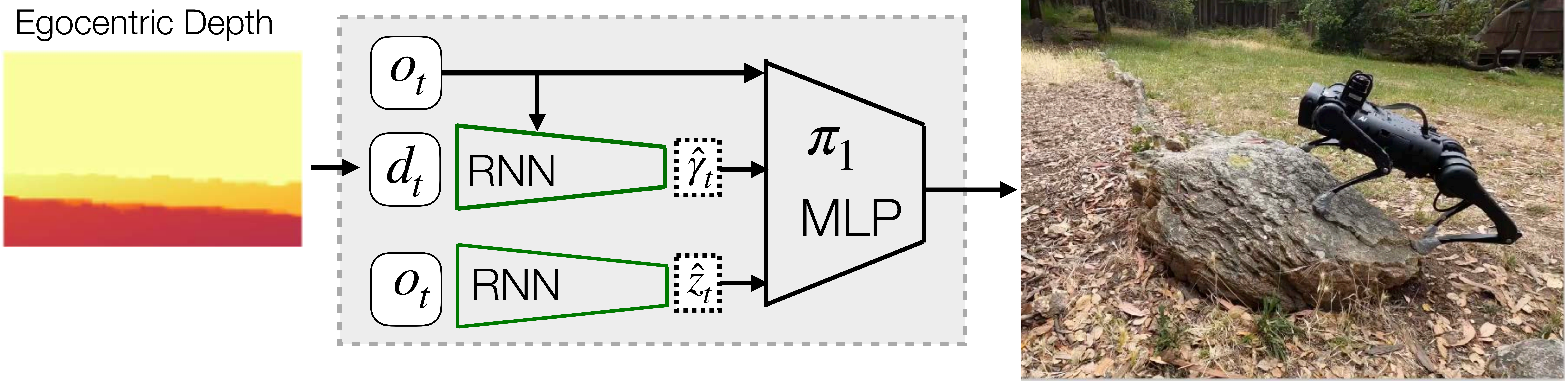
Phase 1 Policy

How do we deploy it?



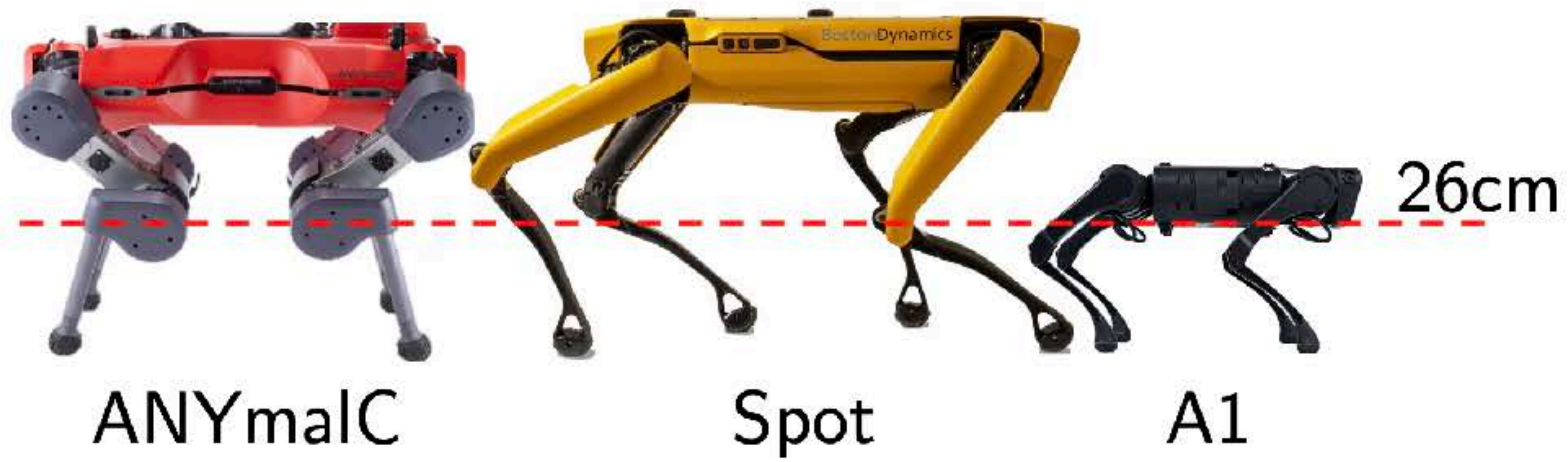
Phase 2: Learning to Walk with Egocentric Depth

Deployment Policy

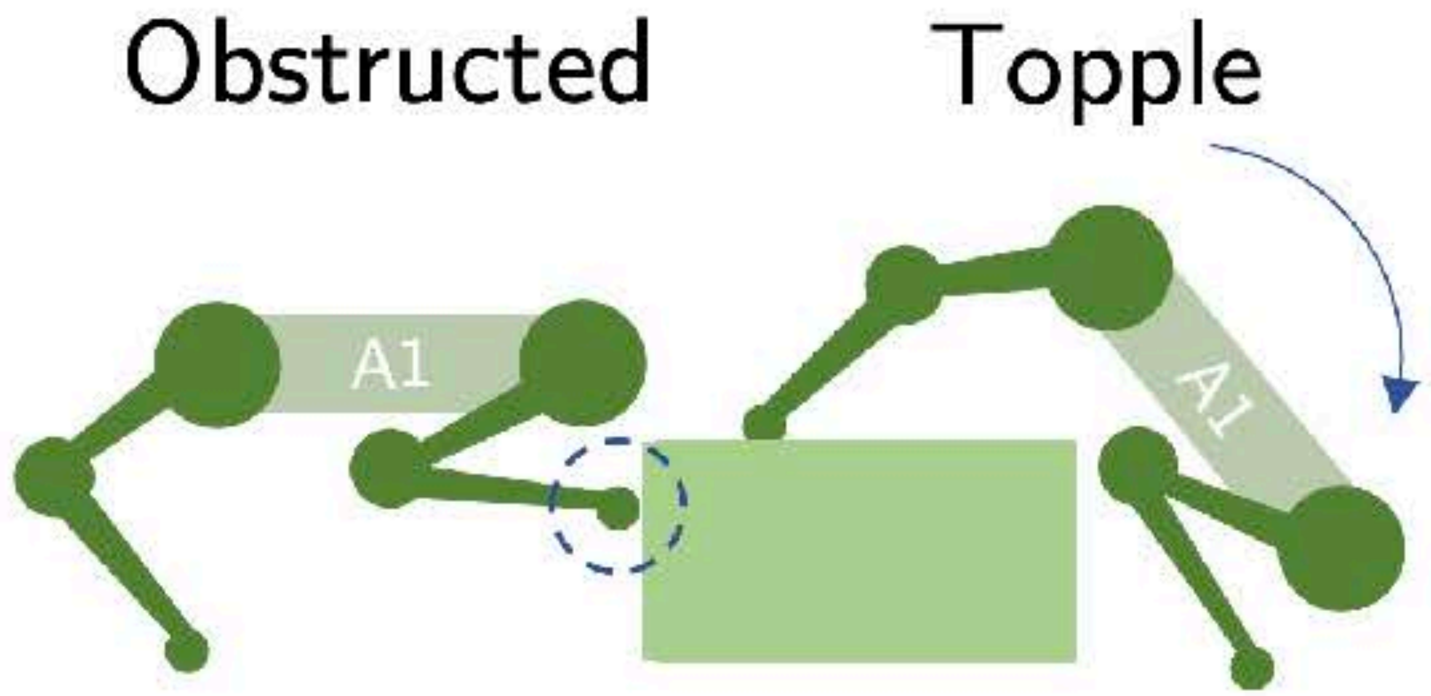


Stairs are designed for humans

Stairs are challenging for small robots



(a) Robot size comparison



(b) Challenges due to size

No predefined gait => emergent hip abduction


Stepping Stones (~20 cm apart)
Emergent footstep planning

Map Free, Gait Free


Live demo at New Zealand (CoRL 2022)
Live demo at New Orleans (CVPR 2022)

Comparisons to baselines

Performance is better without maps

Average Distance

	Blind	Map Based	Ours
Slopes	34.72	36.14	43.98
Stepping Stones	1.02	1.09	18.83
Stairs	16.64	6.74	31.24
Discrete Obstacles	32.41	29.08	40.13

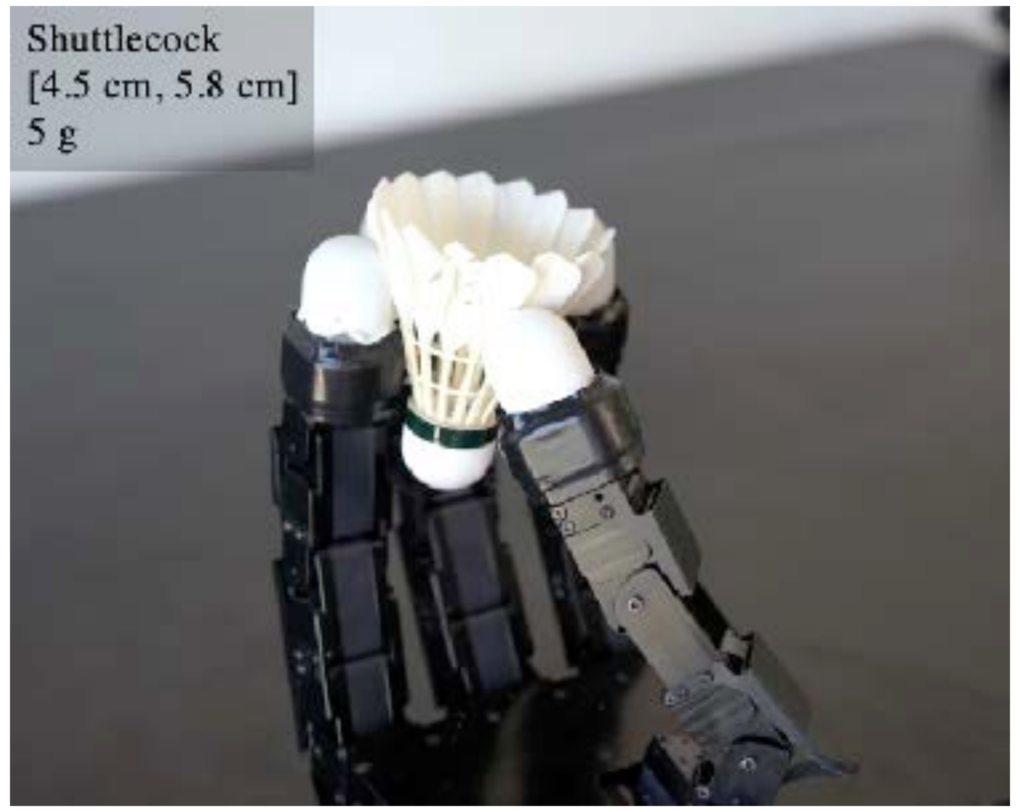
Performance gap is larger on challenging terrains

Application to dexterous manipulation

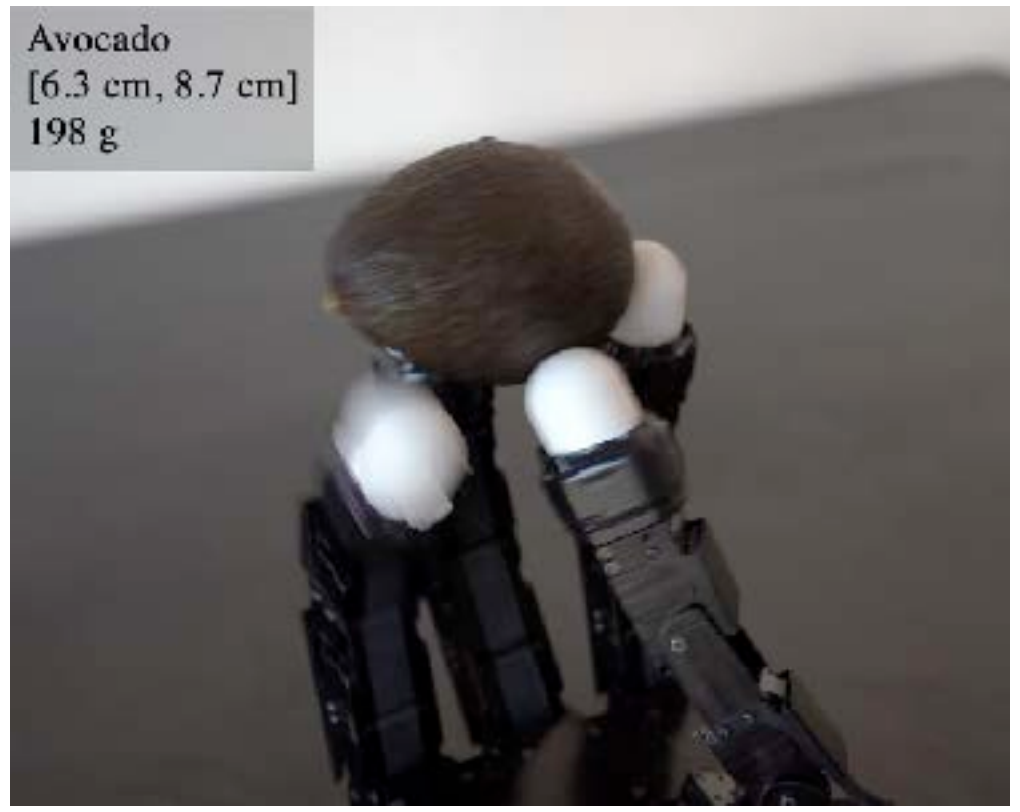
Qi* et. al. In-Hand Object Rotation via Rapid Motor Adaptation. CoRL 2022.

Generalization across Objects with Different Physical Properties

Weight

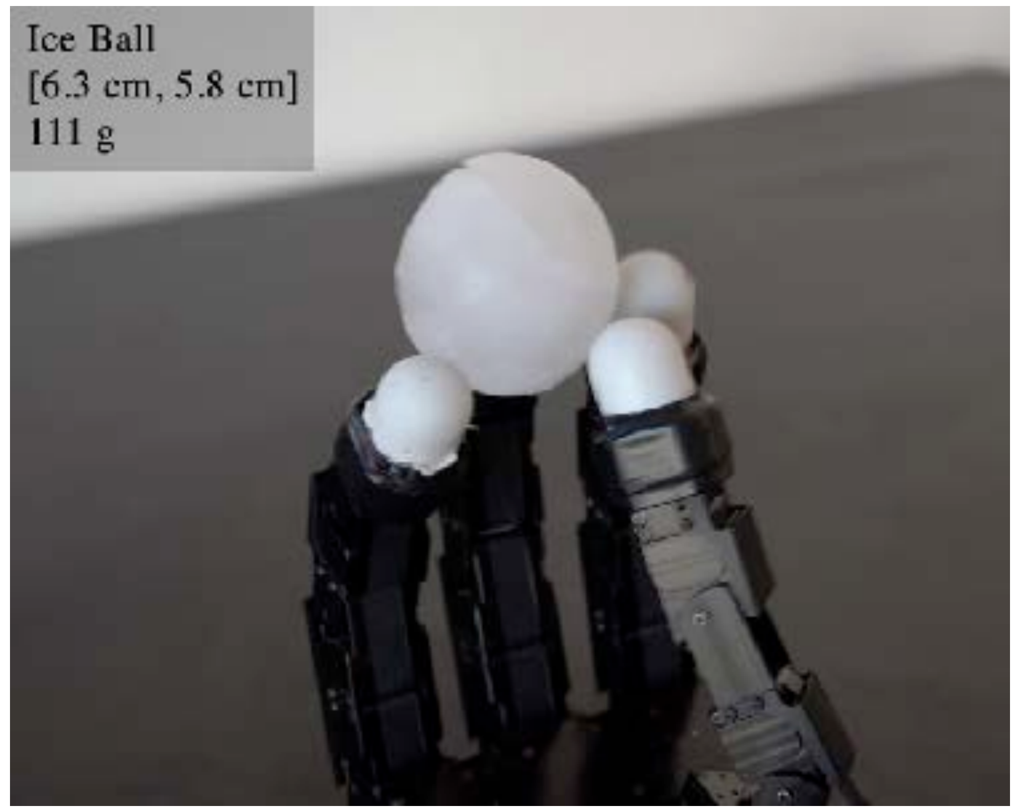


5 g

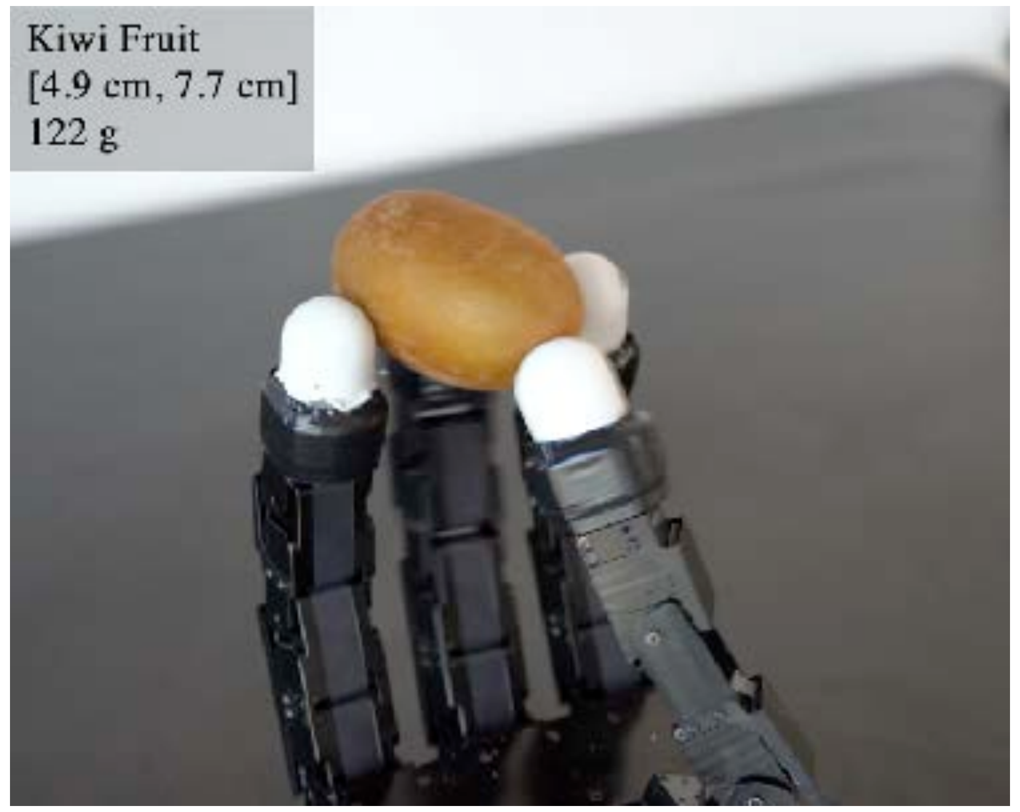


198 g

Coefficient of Friction



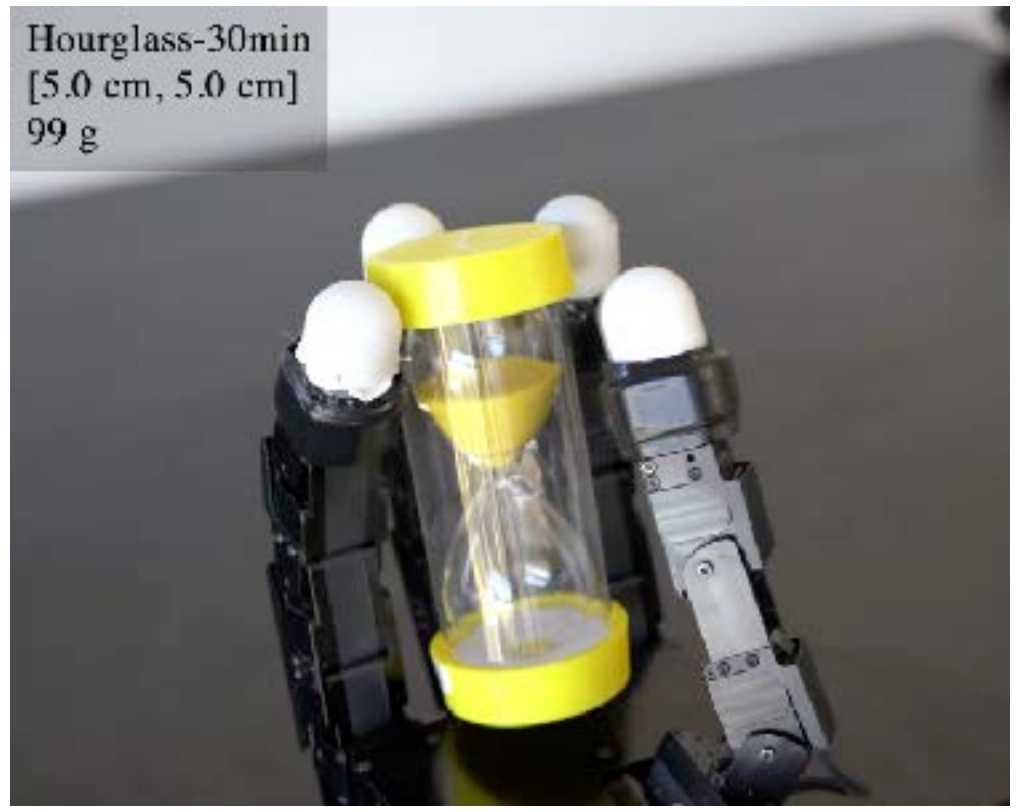
Small



Large

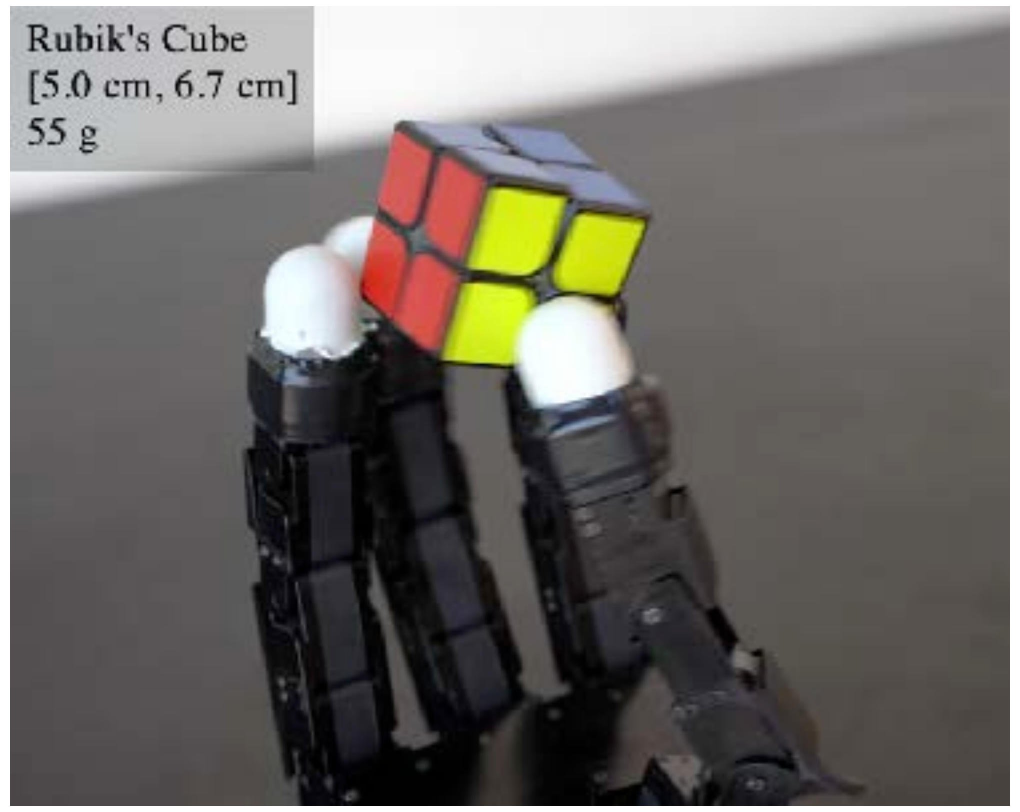
Center of Mass

Higher than Finger

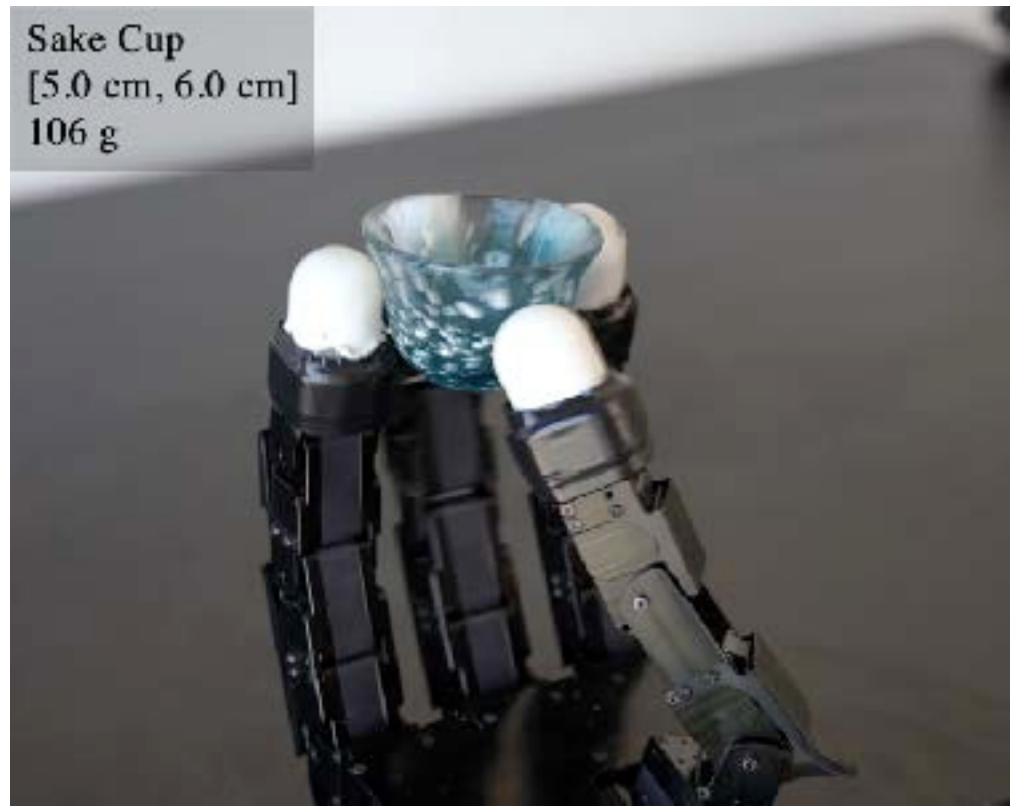


Lower than Finger

Shape

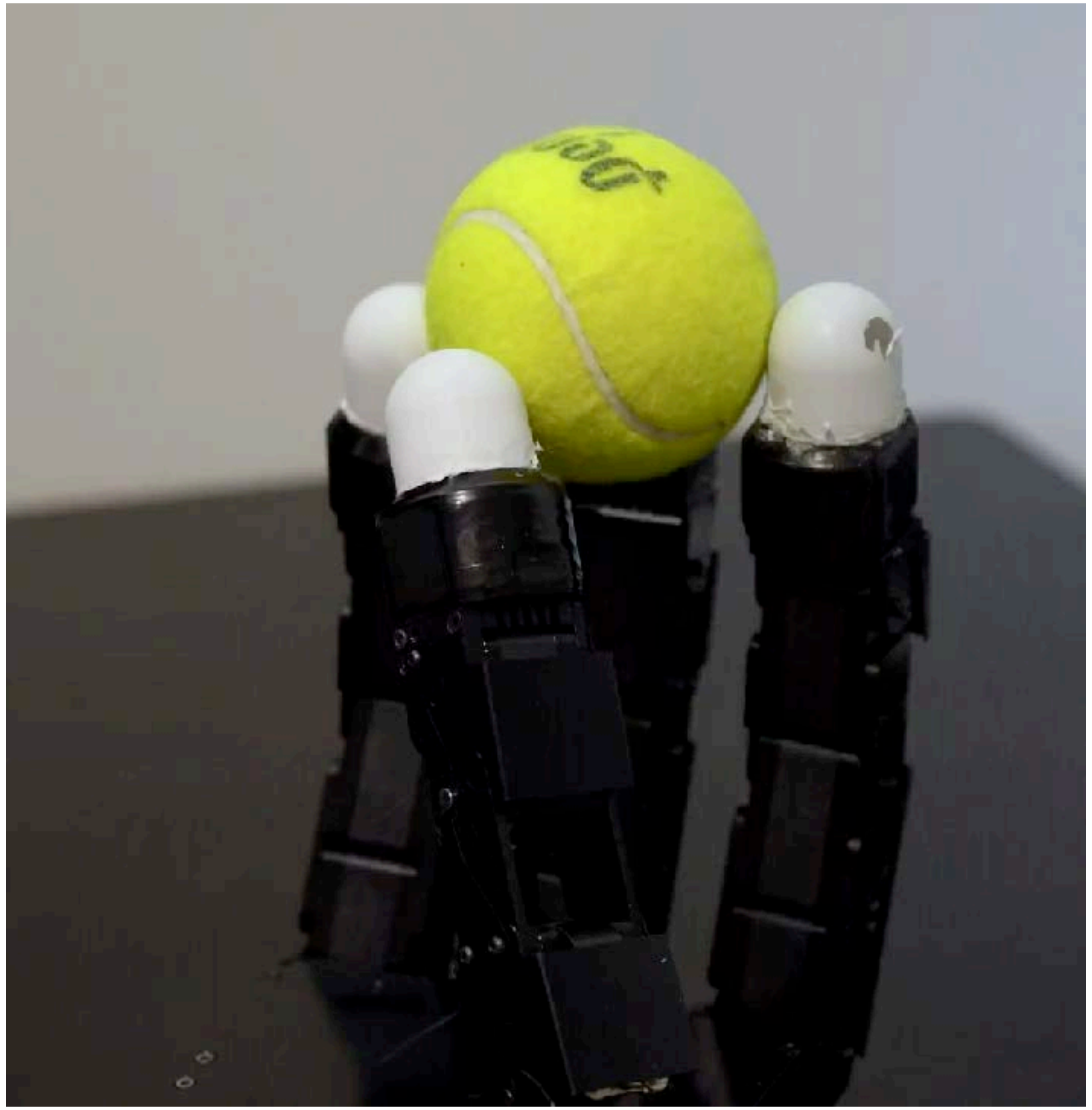


Cube

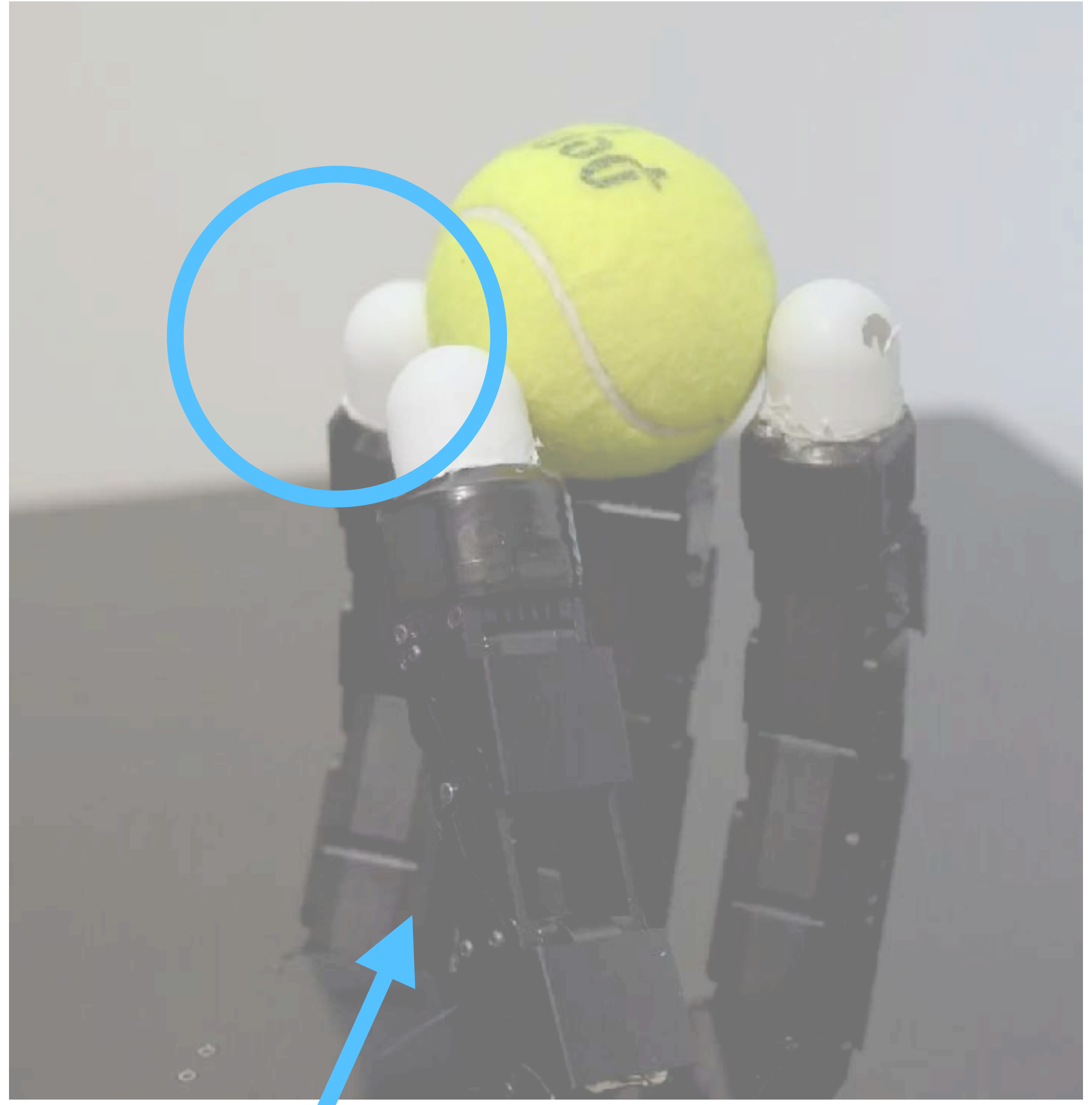


Cup

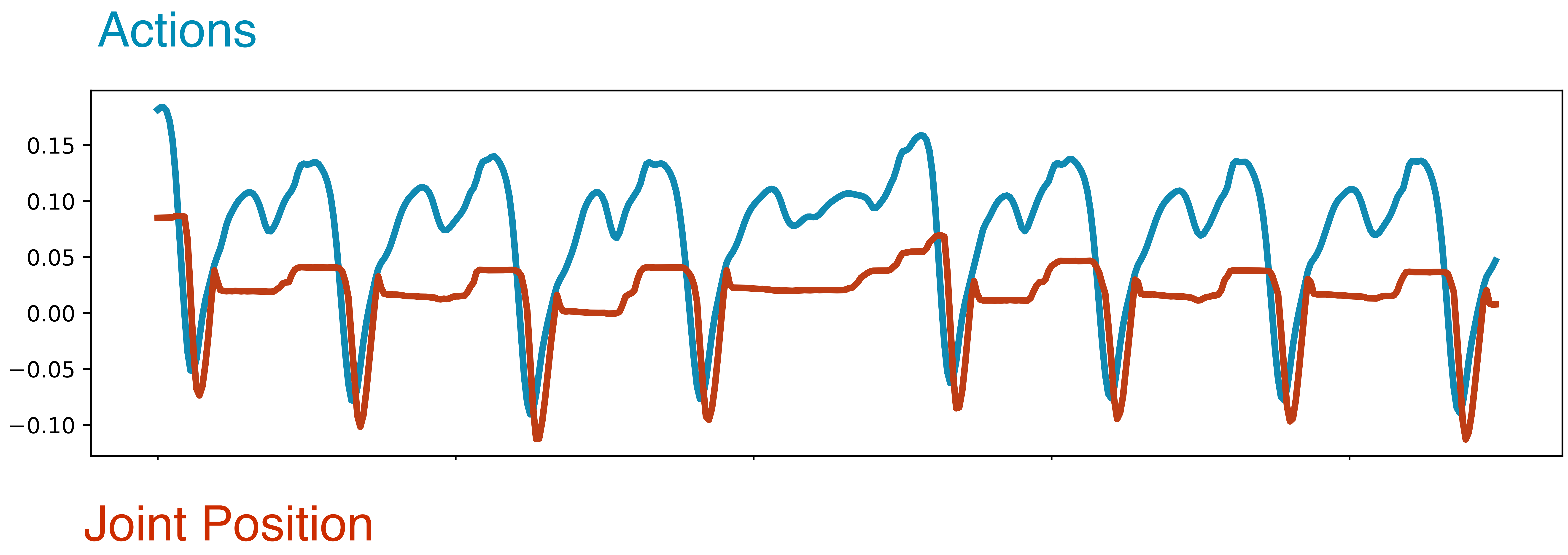
Proprioceptive history for Contact Detection



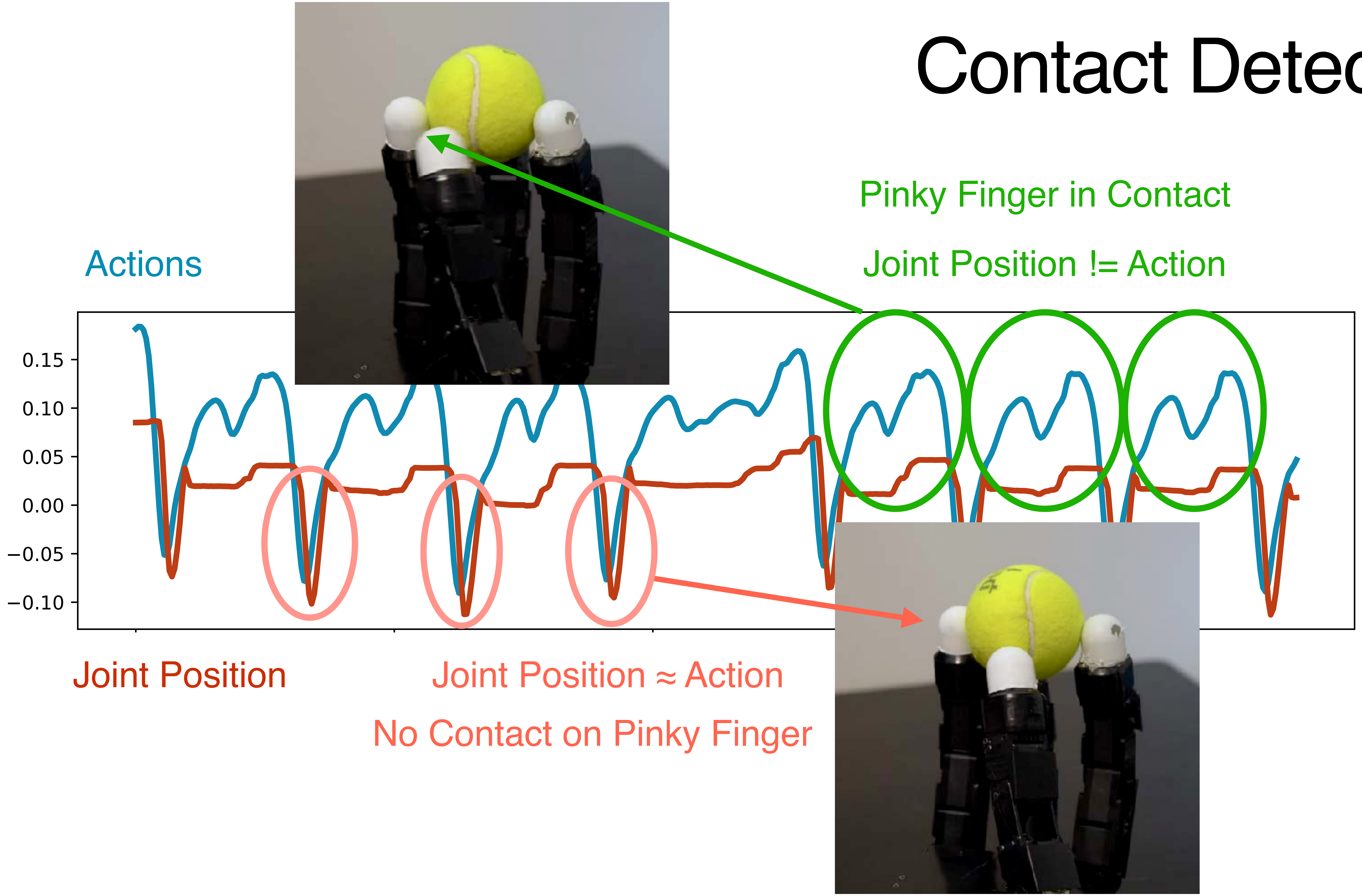
Proprioceptive history for Contact Detection



The Bottom Joint



Contact Detection



Real-world Demos

University of Bristol

CoRL 2022

ICRA 2024

Application to drone flight


Large quad: 792g, 16.6cm Armlength
Small quad: 177g, 5.8cm Armlength

So what’s left to solve?

Humanoids


1x Speed


Fully Autonomous 1.5x
Optimus is learning many new tasks

Getting to reliability and dexterity

Epic Kitchens

- Simulation technology is still very primitive
- Rewards are much harder to specify

Simulation?

- We can, in fact model complex scenarios like deformable objects, etc. However, they are slow.
- Learn from the bitter lesson and bet on compute

"The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin"

\- Rich Sutton

General reward model?

- Most obvious answer are large models trained for general understanding
- Could work, but RL will exploit weaknesses if they exist

Worst case: humans label everything for us!

Superhuman capabilities

Learnings from AlphaGo: Sparse reward + lots of search!

Can we get robots that *exceed* human capability?

Questions?