Learning Quadrupedal Motion over Challenging Terrain

Authors: Joonho Lee , Jemin Hwangbo, Lorenz Wellhausen,
Vladlen Koltun, and Marco Hutter

Presenter: Alan Van Omen

Legged Motion

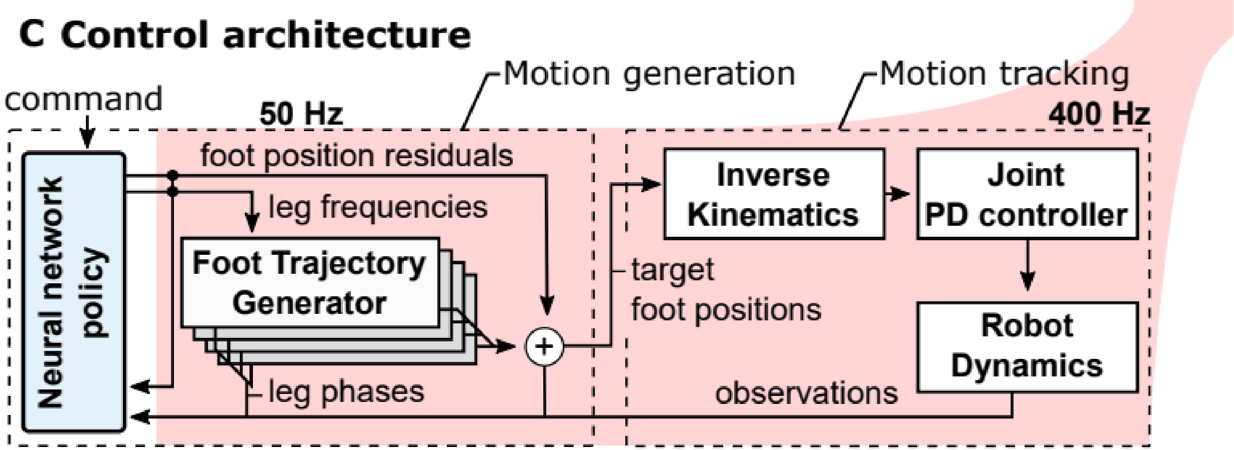
- Allows access to some of the most challenging terrain on earth
- Conventional methods struggle in unexpected/challenging environments:
    1. Rely heavily on exteroceptive sensors (i.e. camera, LiDaR)
    2. Many use carefully-tuned, complex state machines which do not generalize to unexpected conditions

A Novel Proprioceptive Model

- Train a controller on simulated data using only proprioceptive measurements (joint encoders and IMU)
- Requires several additional ingredients to learn robustness
  1. TCN model: use history of proprioceptive states
  2. Privileged learning: pure RL learning approach has sparse rewards, use teacher-student model
  3. Automated learning curriculum: adaptively synthesizes terrain for medium difficulty during training
- These ideas produce a highly-robust controller, which they demonstrate can operate successfully in zero-shot generalization tests

Paper Video Summary

Motion Synthesis



- Each leg moves based on periodic leg phase ($f_0 = 1.25\ Hz$)
- Each leg has a periodic leg phase variable $\phi_i \in [0,2\pi)$ defined at every time step t
$$\phi_i = (\phi_{i,0} + (f_0 + f_i)t)(\mathrm{mod}\ 2\pi)$$
- Step 1 (neural network policy): model outputs f_i and target foot position residuals ($\Delta r_{f\,i,T}$) for each foot
- Step 2 (motion generation): each FTG takes periodic phase variable and gives and outputs a target foot position, $F(\phi_i) \rightarrow \mathbb{R}^3$, the foot targets computed as,
$$r_{f\,i,T} = F(\phi_i) + \Delta r_{f\,i,T}$$
- Step 3 (motion tracking): predicted targets realized as actual joint movements via an IK model and PD joint controllers

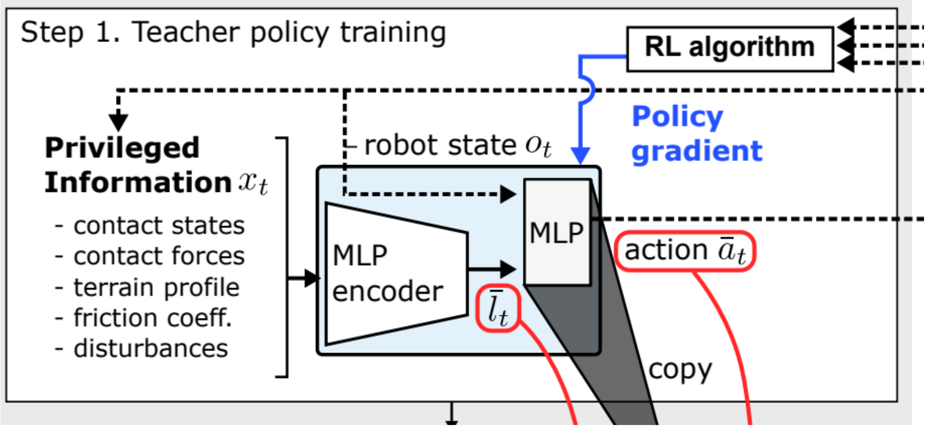
Teacher Policy

- Has access to privileged information in training that would not normally be available to controller
- Receives as input current robot state

$$s_t := \langle o_t, x_t \rangle$$

- Ouputs 16-dimensional action vector, discussed later
- Consists of two MLP blocks
- Uses RL to reward actions which result in moving the robot more quickly to the goal

Data	dimension	x_t	o_t	h_t
Desired direction $(({}^{B}_{IB}\hat{v}_d)_{xy})$	2		✓	✓
Desired turning direction $(({}^{B}_{IB}\hat{\omega}_d)_z)$	1		✓	✓
Gravity vector (e_g)	3		✓	✓
Base angular velocity $({}^{B}_{IB}\omega)$	3		✓	✓
Base linear velocity $({}^{B}_{IB}v)$	3		✓	✓
Joint position/velocity $(\theta_i, \dot{\theta}_i)$	24		✓	✓
FTG phases $(\sin(\phi_i), \cos(\phi_i))$	8		✓	✓
FTG frequencies $(\dot{\phi}_i)$	4		✓	✓
Base frequency (f_o)	1		✓	
Joint position error history	24		✓	
Joint velocity history	24		✓	
Foot target history $((r_{f,d})_{t-1,t-2})$	24		✓	
Terrain normal at each foot	12	✓		
Height scan around each foot	36	✓		
Foot contact forces	4	✓		
Foot contact states	4	✓		
Thigh contact states	4	✓		
Shank contact states	4	✓		
Foot-ground friction coefficients	4	✓		
External force applied to the base	3	✓		



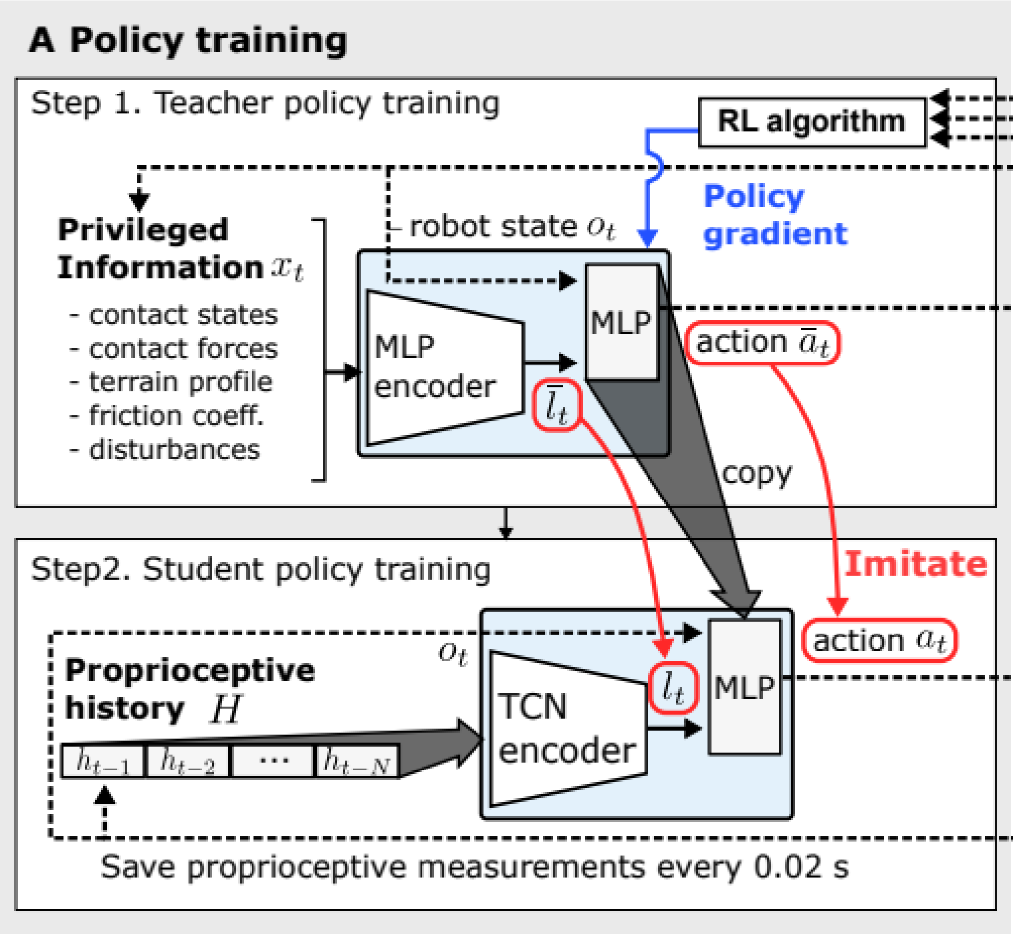
Student Policy

- Student policy learns by imitating teacher policy
- Based on the idea that the latent representation of the priviledged information can be recovered from proprioceptive measurements
- Only has access to proprioceptive measurements (over last 2 seconds)

$$H = \{h_{t-1}, \ldots, h_{t-N-1}\}$$

- Training by minimizing supervised learning objective

$$\mathcal{L} := \left(\bar{a}_t(o_t, x_t) - a_t(o_t, H)\right)^2 + \left(\bar{l}_t(o_t, x_t) - l_t(H)\right)^2$$



Adaptive Terrain Curriculum

- While training using simulation, use a training curriculum that gradually exposes the agent to increasingly more difficult terrain
- Instead of measuring the reward function to measure difficulty, compute the traversability of a given terrain, which they define as the success rate of traversing a terrain
- Three terrain types (hills, steps, stairs) each parameterized by c_T, goal is to pick parameter that gives middle-range traversability, i.e., challenging but still traversable
- Successful traverse is defined as,

$$v(s_t, a_t, s_{t+1}) = \begin{cases} 1 & \text{if} \quad v_{pr}(s_{t+1}) > 0.2 \\ 0 & \text{if} \quad v_{pr}(s_{t+1}) < 0.2 \lor \ \textbf{termination} \end{cases}$$

- Traversability is then defined as,

$$\mathbf{Tr}(c_T, \pi) = \mathbb{E}_{\tilde{\xi} \sim \pi}\{v(s_t, a_t, s_{t+1} \mid c_T)\} \in [0.0, 1.0]$$

Adaptive Terrain Curriculum

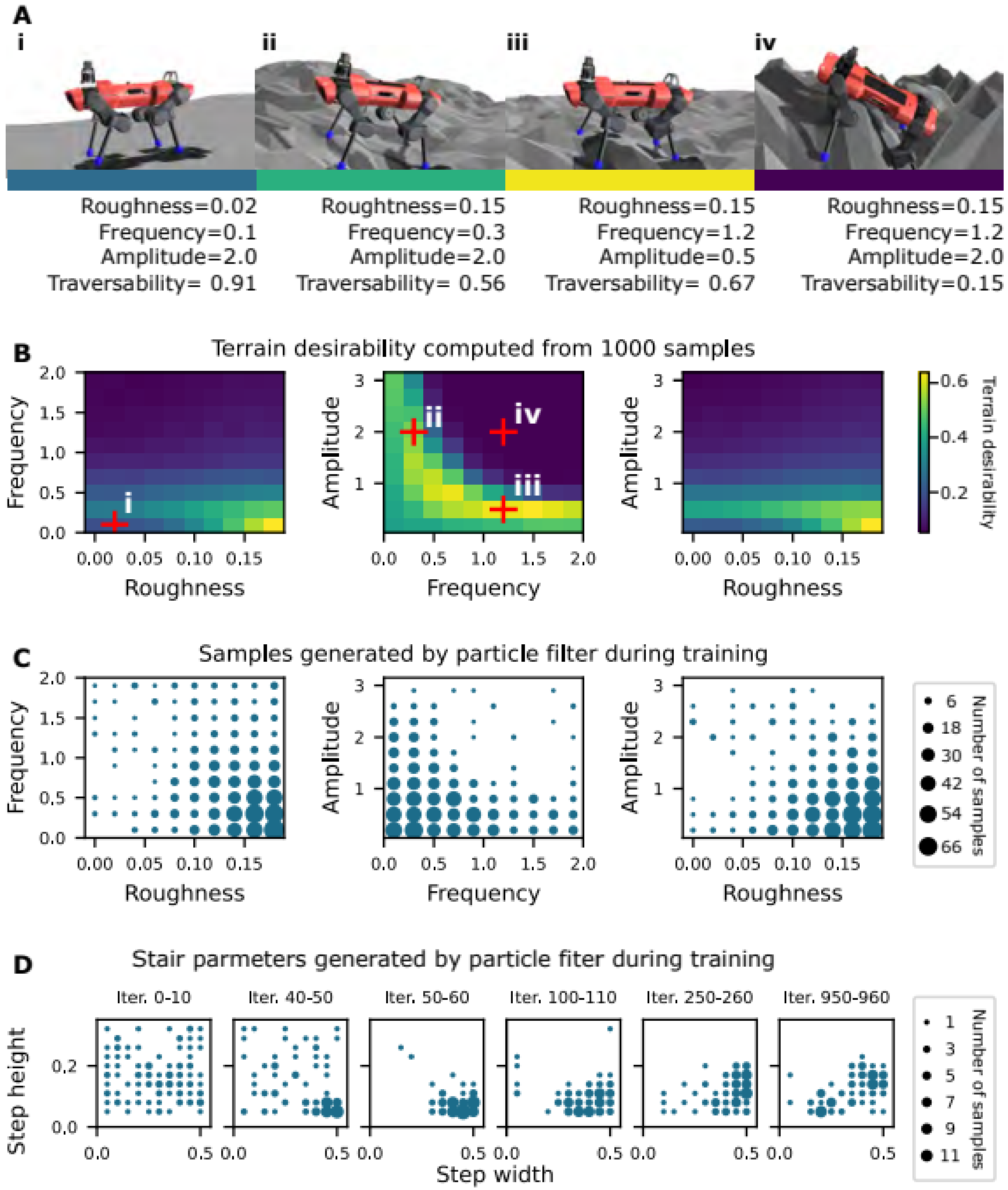
- The goal is to find terrain parameters that give mid-range traversability
- Terrain desirability is defined as,

$$\mathrm{Td}(c_T, \pi) := \Pr(\mathrm{Tr}(c_T, \pi) \in [0.5, 0.9]) \\ = \mathbb{E}_{\xi\sim\pi}\{\mathrm{Tr}(c_T, \pi) \in [0.5, 0.9]\}$$

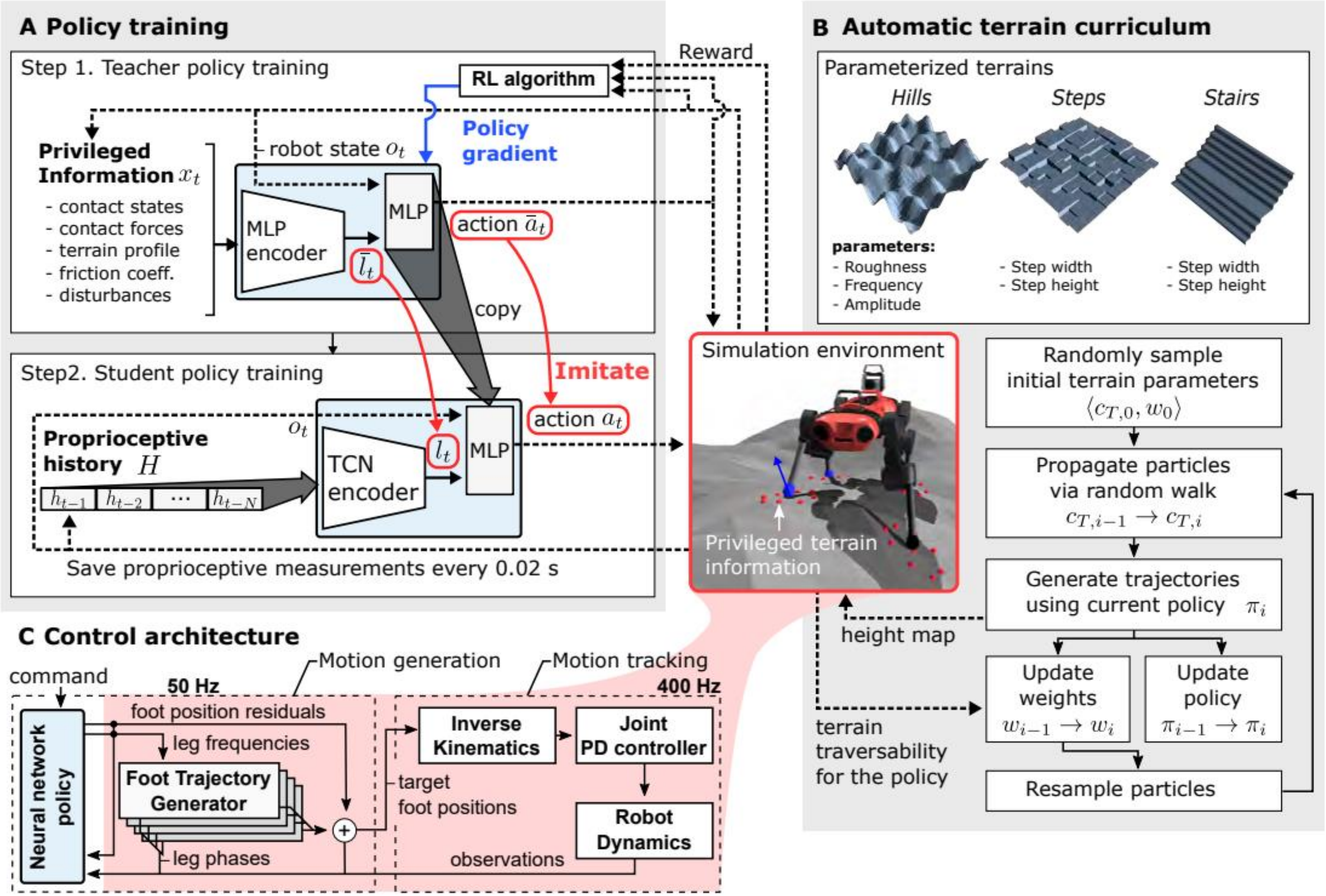
- Use a particle filter to choose desirable distribution of terrain parameters when training
- terrain desirabilities computed empirically during training (start out uniformly distributed)

$$\Pr(y_j^k \mid c_{T,j}^k) \approx \sum^{N_{\text{traj}}} \frac{\mathbf{1}\left(\mathrm{Tr}\left(c_{T,j}^k, \pi\right) \in [0.5, 0.9]\right)}{N_{\text{traj}}}$$

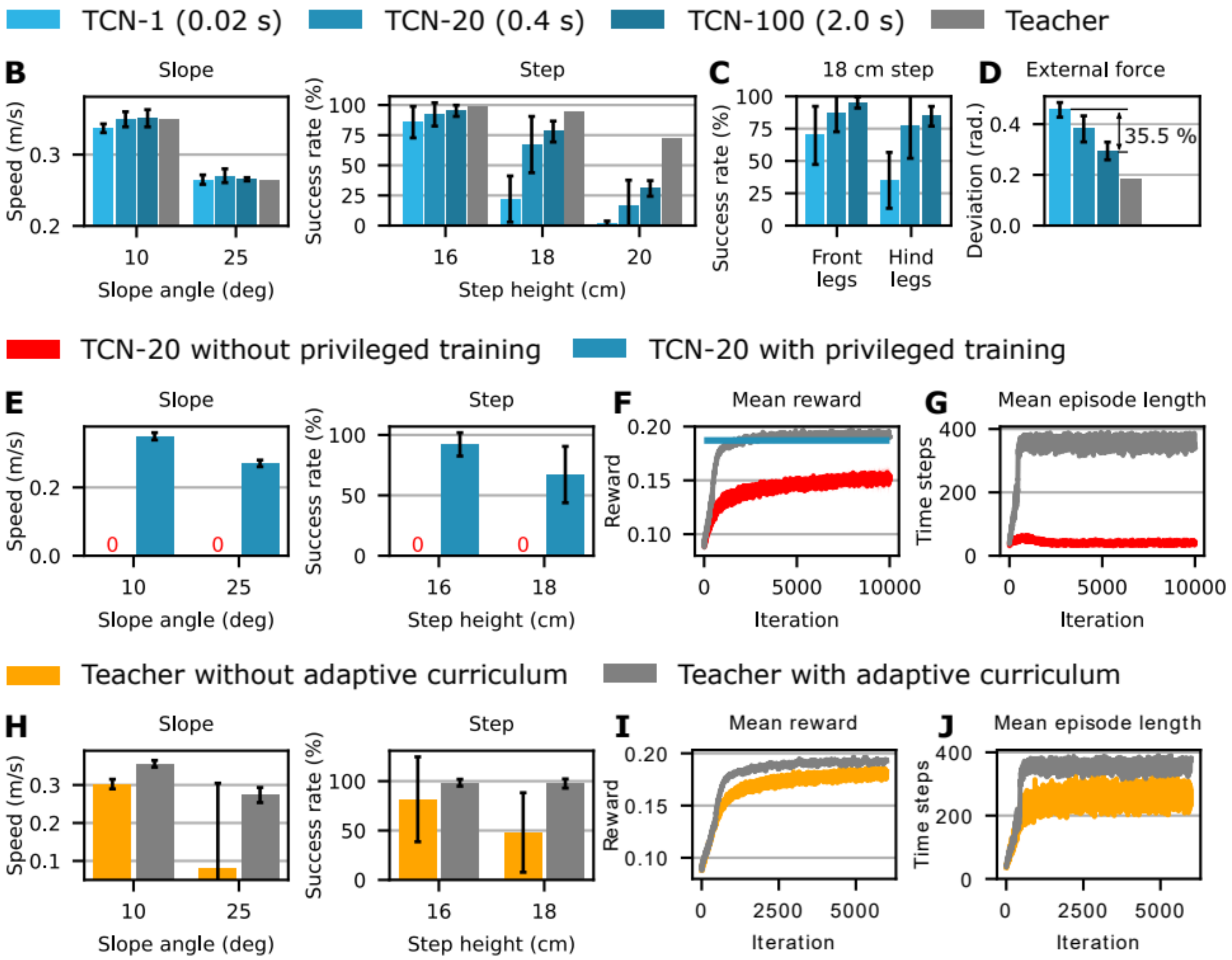
- Discretized to and bounded



Summary of Learning Framework

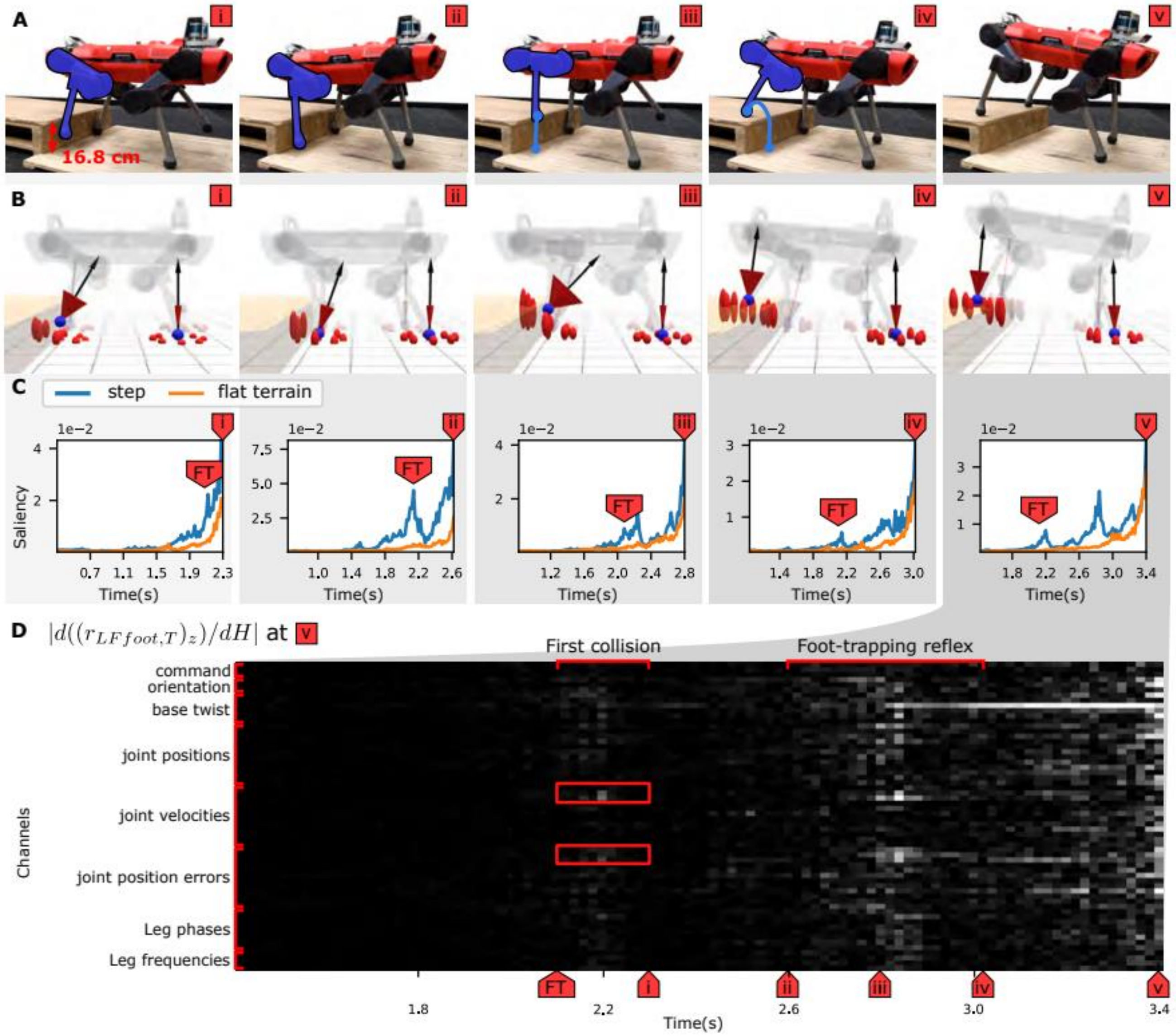


Summary of Results



Emergent Behavior

- Analyze behavior by trying to reconstructing privileged from output of TCN in student policy with a trained decoder
- Minimized with CE and Gaussian log-likelihood (mean and variance) losses
- Reconstructed terrain geometry in B
- D shows sensitivity of output to each measurement



Discussion

Motivated by **@76_f1** and **responses**:

The teacher-student learning framework learned through “cheating”.

- The idea of training a student to imitate a teacher is a clever way to “distill” the teacher policy into a student policy which does not have access to the same information. Are there any alternative ways to “distill” this information or conduct teacher-student learning? (for example, maybe use reconstruction of privileged information as input or something?)
- It is very similar to how children learn by imitation. Are there any other improvements to this learning system that might also be inspired by human behavior?

Discussion Cont.

Also motivated by **@76_f1, @76_f5** and **responses**:

- The author acknowledges that a blind, proprioceptive controller isn’t susceptible to some of the issues with exteroceptive measurements, it is still “inherently limited” (it could easily walk off a cliff). How could a hybrid model be designed that incorporated both proprioceptive and exteroceptive measurements?
- Can we really trust this type of controller for this reason?