

Learning to Walk by Steering: Perceptive Quadrupedal Locomotion in Dynamic Environments

Mingyo Seo¹, Ryan Gupta¹, Yifeng Zhu¹, Alexy Skoutnev², Luis Sentis¹, and Yuke Zhu¹

Abstract—We tackle the problem of perceptive locomotion in dynamic environments. In this problem, a quadrupedal robot must exhibit robust and agile walking behaviors in response to environmental clutter and moving obstacles. We present a hierarchical learning framework, named PRELUDE, which decomposes the problem of perceptive locomotion into high-level decision-making to predict navigation commands and low-level gait generation to realize the target commands. In this framework, we train the high-level navigation controller with imitation learning on human demonstrations collected on a steerable cart and the low-level gait controller with reinforcement learning (RL). Therefore, our method can acquire complex navigation behaviors from human supervision and discover versatile gaits from trial and error. We demonstrate the effectiveness of our approach in simulation and with hardware experiments. Videos and code can be found at the project page: <https://ut-austin-rpl.github.io/PRELUDE>.

I. INTRODUCTION

Legged locomotion in natural environments is an essential step toward unlocking real-world robotic applications, such as autonomous inspection, disaster response, and last-mile delivery. Although recent years have witnessed great strides in low-cost and reliable quadruped hardware, overcoming the large state-action space associated with legged robots in complex environments remains an open challenge. Robots must simultaneously perceive surroundings from onboard sensing and perform agile locomotion to facilitate this advancement. In recent years, a burgeoning body of literature has sought to employ data-driven approaches to improve the robustness and versatility of quadrupedal locomotion. Reinforcement learning, in particular, has led to promising results, including agile behaviors such as balancing, running, jumping, and robust walking, even in the presence of environment uncertainty [4, 26, 43]. However, these methods primarily focus on proprioceptive information while ignoring environment sensing, which hinders their applicability in unstructured human environments.

Pioneering work on perceptive locomotion attempted to incorporate onboard sensory data for motion generation [1, 5, 11, 12, 23, 32, 46]. These methods estimate 3D environment representations (e.g., height maps or voxel grids) for downstream controllers. These representations fall short in capturing rich visual information of semantics and dynamics and instead focus on simple, static environments. Deep reinforcement learning (DRL) approaches [19, 44] aim at training end-to-end policies that directly map sensory data to motor commands in simulation. In practice, these approaches are

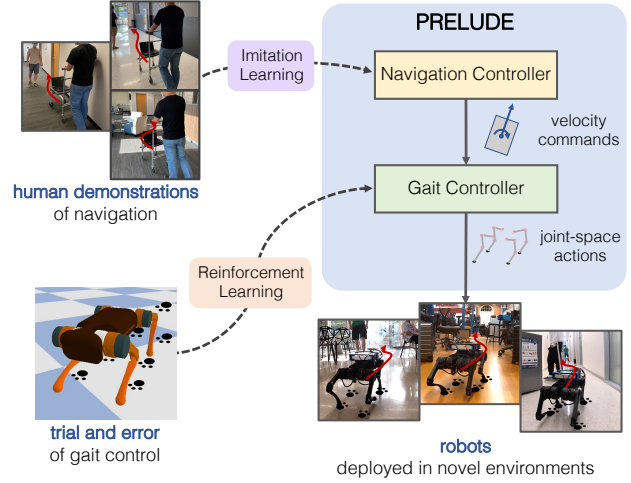


Fig. 1: Method overview. PRELUDE tackles the problem of perceptive locomotion in dynamic environments. We introduce a control hierarchy where the high-level controller, trained with imitation learning, sets navigation commands and the low-level gait controller, trained with reinforcement learning, realizes the target commands through joint-space actuation. This combination enables us to effectively deploy the entire hierarchy on quadrupedal robots in real-world environments.

limited by the sample complexity of DRL, which is further exacerbated in cluttered and dynamic human environments. Furthermore, due to a wide reality gap in rendering RGB images combined with the difficulty of simulating realistic human walking behaviors, these approaches are impractical for real-world deployment.

The goal of this work is to develop a perceptive legged locomotion framework that operates in real-world environments among humans and clutter. Our key idea is to decompose the locomotion problem into a high-level navigation and a low-level gait generation policy to realize target commands. This decomposition allows us to harness the complementary strengths of two different learning mechanisms: 1) we train the high-level navigation controller with *imitation learning*. By learning from human demonstration, the model acquires complex navigation behaviors through moving crowds and dense clutter. Furthermore, imitation learning on real-world demonstration data eliminates the aforementioned reality gap; and 2) we train the low-level gait controller with goal-conditioned *reinforcement learning*, giving flexibility in discovering versatile gait styles which may be difficult for conventional controllers.

We name our method PRELUDE (Perceptive Locomotion Under Dynamic Environments). It consists of a high-level navigation controller and a low-level gait controller (see Figure 1). Both controllers are implemented as deep neural

¹The University of Texas at Austin, ²Vanderbilt University.
Correspondence: mingyo@utexas.edu

networks. The navigation controller, trained with imitation learning, takes in egocentric RGB-D data and generates a desired motion command. To collect data for imitation learning, we design an intuitive steerable cart that can be driven around to collect steering actions in natural environments. A key advantage of this design is that a non-expert human can easily supply demonstrations at scale by pushing the cart around. The navigation controller, trained on data collected on the cart, is able to leverage rich human supervision to perform complex spatiotemporal reasoning and generate dynamic robot motions. Meanwhile, the gait controller, trained with reinforcement learning, learns to achieve variable velocity commands as input. During deployment, it takes as input the target motion command given by the high-level navigation controller and realizes the desired motion.

We evaluate our approach in simulation and with real hardware. In simulation, we demonstrate that our method outperforms state-of-the-art reinforcement learning (RL) methods by 35.4%. Further, we illustrate that our RL locomotion controller exhibits superior performance over conventional model-based methods. The hierarchical design of our method allows us to train the locomotion controller in large-scale simulation and readily transfer it to the real world. Meanwhile, the navigation controller can be trained directly from real-world human demonstration data. By combining these two facets, we can effortlessly deploy the real hardware in a realistic environment.

Our work has three main contributions: 1) We introduce PRELUDE, a hierarchical model for perceptive locomotion in dynamic environments and develop effective strategies to combine human supervision and reinforcement learning; 2) We develop a practical cart design to collect human demonstration data at scale for training our model; 3) We show the effectiveness of our method in simulation and on physical hardware and perform ablative studies to analyze individual model components. We show PRELUDE generalizes to unseen real-world environments with diverse objects and human behaviors.

II. RELATED WORK

a) Quadrupedal Locomotion: The control of legged robots has been studied for decades. One of the main challenges is high computational requirements for optimizing whole-body dynamics. In the last decade, the centrifugal model, which considers the robot as a lumped mass to reduce the computational burden, has been widely used by optimization-based methods [3, 6, 24, 41]. However, such model-based approaches are constrained by computational limits and model mismatch. In response to these limits, the integration of model-based methods and data-driven methods has been actively studied [9, 34, 43, 45]. These methods leverage knowledge of robot dynamics encapsulated in model-based methods to improve training efficiency. However, the boost in training efficiency is hindered by the computational costs of the underlying controller. Alternatively, there has been some recent success when training end-to-end policies for locomotion [14, 25]. These policies boast fast

computation speeds when inferring actions and have been shown to outperform model-based methods. However, these approaches focus on improving legged robots’ agility rather than how legged locomotion is applied to real tasks.

b) Perceptive Locomotion of Quadrupedal Robots:

Legged robots with exteroceptive sensors have been studied for decades. Several model-based approaches exploit exteroceptive sensory modules to reconstruct the shape of surrounding terrains [2, 5, 15, 22, 23, 27, 46]. However, they rely on accurate terrain reconstruction, and the locomotion performance is limited by the precision and speed of reconstruction. Another family of work has employed neural network models to transform raw visual observations into navigation commands [18, 30, 36]. However, they do not account for robot dynamics and instead use a pre-defined gait controller, which is not optimized for the target tasks. There is a growing interest in end-to-end learning of perceptive locomotion [19, 44]. These methods require meticulous simulation training and sim-to-real transfer, which largely limits their applicability to (quasi-)static environments. Inspired by recent advances, our approach uses deep learning models to tackle this problem. Rather than seeking an end-to-end algorithm, we design our method to harness the complementary strengths of imitation and reinforcement learning, allowing the learned model to be deployed to real hardware with minimal domain gaps.

c) Hierarchical Frameworks for Locomotion: Hierarchical frameworks have been used in both model-based and data-driven locomotion. The main advantage of hierarchical frameworks is their ability to break down complicated systems into multiple modules, such that each module can be designed with domain-specific knowledge [3, 6, 24]. In particular, hierarchical approaches to motion generation offers several advantages when using data-driven methods. First, low-level control can be reused with various types of high-level gait planners and vice versa. Second, the overall system can be separated into modules, and individual modules are easier to train and analyze [21, 31, 38]. Our approach also adopts a hierarchical structure. We demonstrate how it gives rise to a practical learning algorithm for perceptive locomotion in dynamic environments.

III. METHOD

We introduce PRELUDE, a hierarchical learning framework for quadrupedal locomotion and navigation in dynamic environments using onboard vision sensing. The key to our approach is to decompose the perceptive locomotion pipeline into a two-level hierarchy consisting of a high-level navigation controller and a low-level gait controller. In this section, we describe our problem formulation and then introduce the hybrid learning scheme used to train the controllers at the two levels of the hierarchy: 1) imitation learning with human supervision for training the navigation controller and 2) reinforcement learning for training the locomotion controller.

A. Problem Formulation

We model the problem of perceptive legged locomotion as a discrete-time Markov Decision Process $\mathcal{M} =$

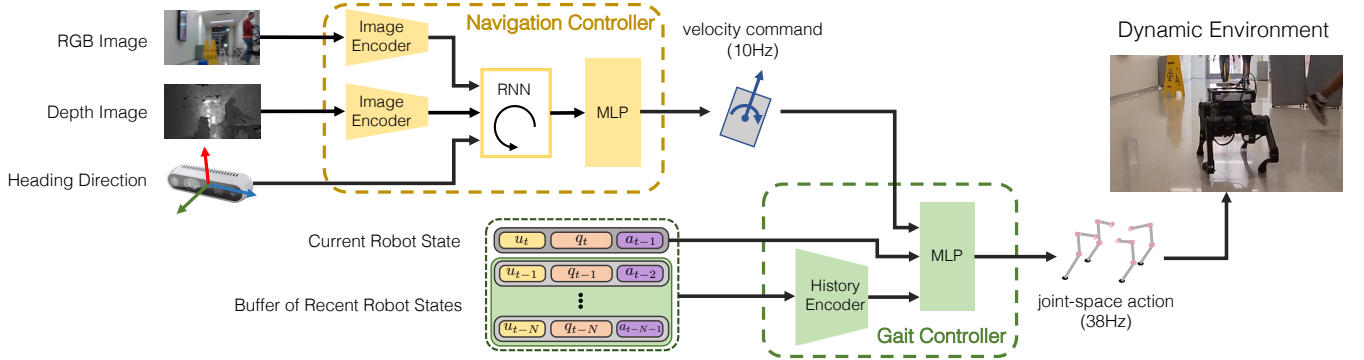


Fig. 2: Model architecture of PRELUDE. The high-level navigation policy generates the target velocity command u_t at 10Hz from the onboard RGB-D camera observation and robot heading. The target velocity command is used as input to the low-level gait controller along with the buffer $\mathcal{B}_t$ of velocity commands u_t , recent robot states q_t and previous joint-space actions a_{t-1} . The low-level gait policy predicts the joint-space actions as the desired joint positions at 38Hz and sends them to the quadrupedal robot for actuation.

$(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, \rho_0)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(\cdot|s, a)$ is the stochastic transition probability, $R(s, a, s')$ is the reward function, $\gamma \in [0, 1]$ is the discount factor, and $\rho_0(\cdot)$ is the initial state distribution. Our goal is to learn a closed-loop visuomotor policy $\pi(a_t|s_t)$ that maximizes the expected return $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})]$. In our context, $\mathcal{S}$ is the space of the robot’s sensor data captured by its egocentric camera and proprioceptive sensors, $\mathcal{A}$ is the space of the robot’s joint-space commands, $R(s, a, s')$ is the reward function designed for the locomotion task, and π is a closed-loop policy that we deploy on the robot.

To handle the complexity of perceptive locomotion and train the policy π effectively, we decompose the policy π into a two-level hierarchy. At the high level is a navigation policy π_H that computes the target motion command $u \in \mathbb{R}^2$ as the target linear and angular velocities. At the low level is a gait policy π_L that computes the robot’s joint-space actions to realize target command u from π_H . With this hierarchical abstraction, we can rewrite the whole policy as, $\pi(a_t|s_t) = \pi_L(a_t|s_t, u_t)\pi_H(u_t|s_t)$.

B. Hierarchical Perceptive Locomotion Policy

PRELUDE consists of a high-level policy π_H for predicting navigation commands and a low-level π_L for generating locomotion gaits (see Figure 2). Here we describe the scheme we used to train both policies and their seamless deployment both in simulation and on the real robotic system.

a) Imitating Navigation Demonstrations in Dynamic Environments: When surrounded by static obstacles and people moving in and out of the scene, the robot must anticipate future motion and subsequently generate high-quality motion plans. By leveraging imitation learning, PRELUDE avoids the difficulty of training an RL policy that must simulate full human walking dynamics and also attempt effective navigation. Instead, we propose to learn the policy π_H for navigation using imitation learning. Through imitating human navigation behaviors, the robot is able to acquire the ability of complex spatial reasoning and motion prediction in dynamic environments. Meanwhile, this form of high-level demonstration is easy to collect on a hardware platform of a much simpler morphology than the quadruped. We describe our hardware platform of the steerable cart in Sec. IV-A.

The collected demonstration dataset $\mathcal{D}$ consists of state-action pairs $\mathcal{D} = \{(s_i, u_i)\}_{i=1}^N$, where the states are the camera observations in a similar view as of the robot’s onboard egocentric camera plus the camera orientation while the actions are the linear and angular velocities estimated by odometry, and N is the total number of data points. With this dataset, we train the navigation policy π_H using a supervised imitation learning algorithm. Specifically, we design a Behavioral Cloning model with recurrent neural networks (RNNs) [28] to capture the temporal information of moving objects in the environments. Furthermore, demonstrations collected from human operators are noisy and multimodal in nature. To capture the multimodality of action distributions while preventing mode collapsing, the policy uses a Gaussian Mixture Model (GMM) to parameterize a distribution of actions [40].

In practice, π_H is implemented as a deep neural network. At each time step, its input consists of the egocentric RGB-D observation and the heading direction of the body. It outputs the target velocity command, i.e., two scalar values for forward-linear velocity and angular velocity. As the RGB images provide rich task-relevant semantic information and the depth images provide geometry information of surrounding environments, we process RGB images and depth images separately through individual *image encoders*. The target velocity command serves as a suitable representation for specifying high-level navigation behaviors, as it can define highly reactive motions in dynamic environments compared to other common representations, such as position-based commands.

b) Learning Robust Locomotion Gaits: A robust gait controller is key for locomotion in real dynamic environments. To this end, a desirable gait controller should meet the following two requirements: 1) it can track constantly varying velocity commands set by π_H to generate agile walking behaviors for dynamic collision avoidance; 2) it can be effortlessly deployed to real hardware in the presence of dynamics uncertainty. Motivated by recent successes in learning gait controllers in simulation and sim2real transfer [9, 25, 26], we use large-scale reinforcement learning in simulation to train our gait controller. Concretely, PRELUDE learns the policy using the distributed Proximal Policy Optimization (PPO)

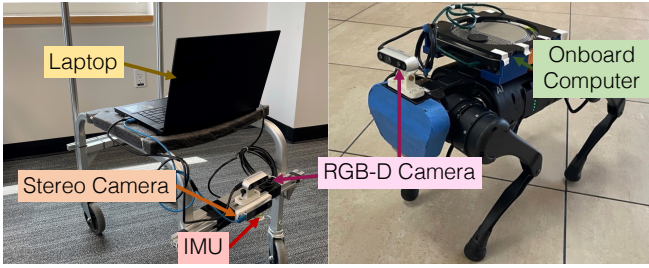


Fig. 3: Hardware platforms. A steerable cart designed for collecting human demonstrations of navigation (left) and the Unitree A1 robot with ego-centric RGB-D camera mounted at cart height (right). It ensures the navigation policy trained on demonstration data can be directly deployed on the robot.

algorithm [35] in parallel copies of simulated environments. We randomize the target velocity commands (as input to the gait policy π_L) to learn a gait policy to track rapidly changing commands. The policy is optimized over the reward function designed for 1) minimizing the tracking command error and energy, 2) stabilizing the robot body, and 3) encouraging concrete foot contacts. In addition, we introduce two key designs to ensure the learned policy to transfer directly to physical hardware: First, we apply aggressive *domain randomization* [37] to physical parameters in simulation to improve the robustness of our controller while bypassing the complicated system identification process on the real robot. We randomize mass, inertia, friction coefficients of robot joints, and external force perturbation. Second, we use a *buffer of recent robot states* as input to the policy π_L rather than raw visual observations. For each time t , the input buffer $\mathcal{B}_t$ contains $T + 1$ tuples $\mathcal{B}_t = \{(u_i, q_i, a_{i-1})\}_{i=t-T, \dots, t}$, where u_i is the navigation command at time i , q_i is the proprioceptive state at time i , and a_{i-1} is the joint-space action output of π_L from the previous time step $i - 1$. This design resonates with recent work RMA [25], which demonstrated that the robot’s recent state-action history can serve as a robust proxy for estimating extrinsics. For perceptive locomotion, using the buffer of robot states as input eliminates the needs for us to handle the reality gap of visual observations from camera sensors.

IV. EXPERIMENTS

We design our experiments to answer the following three questions: 1) How does PRELUDE’s hierarchical design perform compared to existing methods for perceptive locomotion? 2) Is our learned low-level gait controller more robust than conventional model-based controllers in motion generation? 3) Can PRELUDE be deployed on the quadruped robot in real-world environments?

A. Experimental Setup

We validate our proposed method for perceptive locomotion on a track with static obstacles and people who may be moving around. A successful episode is one where the robot walks from one end to the other without falling or colliding with obstacles, walls, or people. We design environments of varying difficulty to analyze our model’s performance in simulation and in real hardware with the Unitree A1 quadrupedal robot [39].

a) Simulation Setup: We develop a simulated corridor environment in Bullet Physics [8] with 3m width and 50m length. We use this environment to systematically compare our approach to baselines under various conditions. We initialize the scenes with different numbers of static obstacles and walking humans, then categorize the conditions into three difficulty levels:

- **Easy Level:** Corridor environments with sparsely placed obstacles (average 5) and zero or one person moving.
- **Medium Level:** Cluttered environments with on average 15 static obstacles and zero or one person. Due to the increase in objects, the robot is more prone to collision.
- **Hard Level:** Challenging environments with on average 15 obstacles and 4 humans. The crowd raises the uncertainty, so the robot must anticipate the human motions.

All the static obstacles are initialized with 3D assets from the ShapeNet [7] and Google Scanned Objects [33] that cover everyday objects. Human motion is governed by a Gaussian Process-based motion generation method. Further, we add random external perturbations to robots to mimic the environmental uncertainties that tend to exist in real-world scenarios.

b) Real-World Setup: We consider our real-world deployment environment as 15m tracks with different numbers of static obstacles and walking humans, then categorize the conditions into four types with two annotations:

- **Static (Sparse):** Open space with on average 10 large pieces of furniture with enough space to bypass them.
- **Static (Complex):** Narrow office area with more than 10 obstacles where the robot may pass through tight spaces.
- **Dynamic (Sparse):** Open space with on average 5 humans and 5 static obstacles. The robot has enough space to bypass humans but needs to anticipate humans’ movements.
- **Dynamic (Complex):** Narrow corridor with 1.6m in length with on average 4 humans and less than 5 obstacles, the most challenging environments where the robot has to pass through narrow space and rapidly react to humans.

Between each trial, obstacles were shuffled around the area and swapped for others. People are instructed to walk in natural ways.

The deployment environments are different from where human demonstration data are collected, with unseen backgrounds, new objects, unexpected human behaviors, and changing light conditions. We mount a RealSense D435 [20] camera on the forehead of the robot for egocentric RGB-D observations. The gait controller predicts desired joint positions at 38Hz, which is transmitted to the internal Unitree SDK for actuation. The controller is deployed on an onboard Apple Mac mini M1 computer for real-time inference.

c) Data Collection: To collect demonstration data using the quadruped would require a human expert to control the robot in complex environments over long periods of time. This approach is hard to scale due to the mechanical limits of the hardware and the difficulty of control. To ease data collection, we design a steerable cart platform (see Figure 3). This cart platform is designed in a non-holonomic morphology, permitting only the forward linear velocity and

the angular velocity. Human operators can easily maneuver this cart among static and dynamic obstacles to generate expert data. We mount an Intel RealSense D435 camera on the cart to record RGB-D observations. We further obtain the cart’s linear velocity and yaw-direction angular velocity from the front-mounted Intel RealSense T265 stereo camera. The heading direction can be estimated by the VectorNav VN100 IMU sensor installed on the cart. This way, we eliminate the need for manual labeling of human actions. We collect 250 real-world trajectories within three hours of wall-clock time. To mimic such a steering maneuver in simulation, we use the 3Dconnexion SpaceMouse to provide velocity commands. We collect 80 demonstration trajectories for each difficulty level. For both domains, our datasets consist of the following information: 1) observation data, including RGB-D images and the heading orientation; 2) the forward linear and angular velocity of the cart.

B. Quantitative Evaluation

a) *Simulation Evaluation:* We compare PRELUDE against the following baselines:

- **Dynamic Window Approach (DWA):** An online collision avoidance method that computes a cost map from the obstacles’ positions and plans navigation commands based on the cost map [13]. It uses the navigation commands with our learned gait controller to actuate the robot. This baseline accesses ground-truth positions of all obstacles in the field of view.
- **Hierarchical RL:** This baseline trains the high-level navigation controller with reinforcement learning (RL) using the same distributed PPO algorithm. It uses the navigation command as the action space and invokes our learned gait controller. The reward is defined as the progress of moving forward at each step. This strategy has been explored in prior work [29, 31, 42], which uses low-level action primitives to level up the action abstraction for RL.
- **PRELUDE (MPC):** This baseline uses a conventional Model Predictive Control (MPC) baseline [10] instead of our learned gait controller. We assume the MPC baseline accesses ground-truth body velocity.
- **PRELUDE (BC):** A variant of our final model, which does not include the recurrent neural networks in the navigation policy architecture. All the other parts remain the same.

To quantitatively evaluate each model, we use the success rate of completing the track and the travel distance, averaged across all test trials. The distance metric allows us to measure partial progress made by the baselines when they struggle to complete the entire task. To evaluate all the models consistently and control the environment setup, we generate 50 randomized scenes for each difficulty level and use the same set of 50 scenes for evaluating each model. We present the average length traversed and success rate in Table I.

Table I suggests that PRELUDE achieves the overall best performance in all three difficulty levels. Results of the DWA baseline imply that all three levels of environments require non-trivial collision avoidance behaviors. Meanwhile, the

TABLE I: Comparison of PRELUDE and baselines in simulation. We report the average length traversed in meters (total track length: 50m) and success rate in percentage as the evaluation metric.

Method	Easy	Medium	Hard
DWA [13]	42.2 ± 12.8 (66%)	34.2 ± 18.2 (52%)	30.2 ± 18.3 (40%)
Hierarchical RL [31]	36.4 ± 14.8 (44%)	24.7 ± 17.6 (26%)	28.4 ± 18.9 (38%)
PRELUDE (BC)	32.0 ± 16.0 (32%)	28.1 ± 15.5 (24%)	24.1 ± 14.9 (16%)
PRELUDE (MPC) [10]	42.3 ± 14.3 (74%)	38.9 ± 15.4 (60%)	34.0 ± 18.1 (50%)
PRELUDE	44.3 ± 12.4 (80%)	41.6 ± 15.2 (74%)	35.3 ± 18.2 (56%)

TABLE II: Evaluation of gait controllers. We report the tracking errors in meters (lower the better) and success rate in percentage (higher the better) on 20 trials of tracking each pre-defined trajectory.

	MPC [10]	Gait Controller (Ours)
<i>x</i> -axis Linear	0.26 ± 0.03 (50%)	0.20 ± 0.02 (100%)
<i>x</i> -axis Sine	0.28 ± 0.05 (60%)	0.24 ± 0.02 (95%)
<i>x</i> -axis Step	0.29 ± 0.04 (45%)	0.26 ± 0.04 (90%)
Sine Trajectory	0.27 ± 0.04 (55%)	0.23 ± 0.03 (90%)
Zig-zag Trajectory	0.26 ± 0.03 (55%)	0.24 ± 0.02 (85%)

Hierarchical RL baseline shows that decoupling gait control and navigation allowed the robot to traverse in simple and static environments. However, RL alone struggles to discover sophisticated navigation strategies, especially for handling dynamic obstacles. Results of PRELUDE (BC) indicate that the lack of temporal modeling leads to inferior performances, especially in the Hard level, where our final model traverses a 44% longer average distance than the BC baseline. We hypothesize that the temporal information, harnessed by our recurrent policies, plays a critical role in reasoning about moving people and anticipating their future motions.

Results of PRELUDE (MPC) indicated that our learned navigation controller is compatible with different types of gait controllers. Even though PRELUDE (MPC) takes advantage of the robot’s dynamic models for MPC, our RL-based gait achieved higher performances overall. These results imply that our learned gait controller exhibits high robustness for agile navigation in dynamic environments.

b) *Analysis of Gait Controllers:* We have shown PRELUDE outperforms baseline approaches in the perceptive locomotion task and further analyze the quality of the learned gait controller. We send the robot velocity commands computed from the set points of a pre-defined position trajectory. At each time step, the velocity command is computed by a simple PD controller based on the next set point and the robot’s current position. We set desired velocity commands with a PD controller because unit testing on the locomotion controllers is performed in a closed-loop manner to mimic the integration with Navigation Controller. Furthermore, realistic navigation tasks require fast convergence speed and stability with respect to velocity error.

We design trajectories with diverse shapes and curvatures to evaluate the agility of our learned gait controller against the MPC baseline [10]. To evaluate that our gait policy handles aggressive changes in linear velocity, we design the below linear trajectories of different set point distances:

- ***x*-axis Linear:** Linear trajectory of set points with constant distances of 0.7m per second.
- ***x*-axis Sine:** Linear trajectory of set points with distances varying as sine functions in the range of 0.4 to 1.0m per

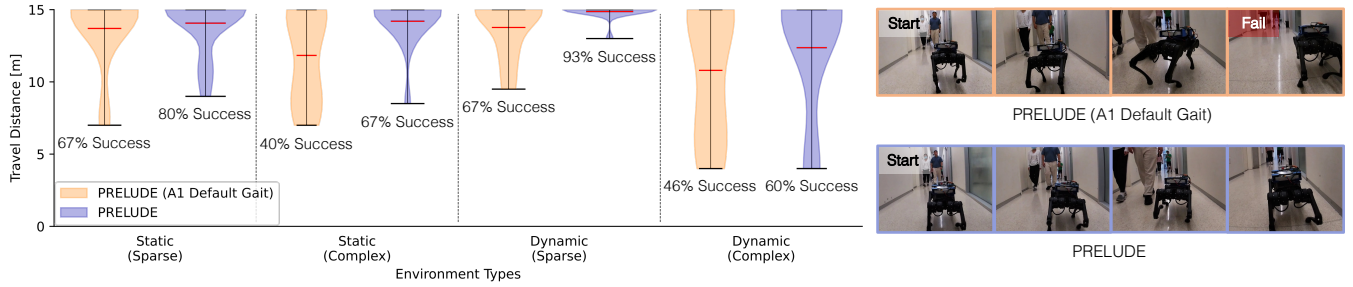


Fig. 4: Real robot experiments. (Left) We perform real-world trials where the robot traverses 15m-length tracks in different configurations. The plot shows distributions of travel distances in meter. The black and red lines indicate the ranges and the means of traverse length, respectively. **(Right)** We observed that PRELUDE (A1 Default Gait) drifts aggressively after a high-speed turning and collides into the wall, while PRELUDE turns rapidly to bypass the walking crowd and completes the trial successfully.

second.

- **x -axis Step:** Linear trajectory of set points with distances varying as step functions with values of 0.5, 0.7, and 1.0m per second.

To show that our gait policy handles large changes in the yaw rate, we designed the additional trajectories below:

- **Sine Trajectory:** Sine-like trajectory of set points with constant distances. The yaw direction of the trajectory changes in the range of 0.0 and 1.4 rad.
- **Zig-zag Trajectory:** Zig-zag trajectory of set points with constant distances where the direction of the path is given as step functions, as -0.4 to 0.4 rad.

We report two evaluation metrics: 1) the quality of trajectory tracking, computed as the average tracking error between the ground-truth trajectories and the robot’s realized trajectories, and 2) the robustness of the controller, computed by evaluating the success rate in tracking the trajectories below the error limit of 1m without failing within 30 seconds. Table III presents the results. Our Gait Controller, learned with reinforcement learning, realizes the trajectories more accurately and robustly than the MPC baseline. In particular, we observe that the MPC method regulates body acceleration and changes the body motion very smoothly when the changes to the velocity command are small. However, when commands change more aggressively, simplifications of inertia terms to convexify the MPC cause failure of the robot. In contrast, we observe that our RL-based controller can safely track aggressively changing commands. These attributes are suitable for performing the given task of navigation amongst dynamic obstacles where changes need to be rapid. Furthermore, our controller has a significantly lower lateral drift than MPC when tracking velocity commands.

c) Real Robot Evaluation: Finally, we demonstrate that PRELUDE can be deployed on the real robot. We compare it with our self-baseline PRELUDE (A1 Default Gait), a variant of our final model, using the robot’s default model-based controller instead. The comparison shows the effectiveness of PRELUDE in real-world deployment. In particular, 1) our navigation controller can be deployed with other types of gait controllers, and 2) our learned gait controller outperforms model-based controllers for locomotion in cluttered and dynamic environments.

To quantitatively evaluate each model, we report the success rate of completing the track and the distance that

the robot traverses before a collision or reaching the end of the track, averaged across all test trials, as shown in Figure 4. To evaluate the models consistently and control the environment setup, we generate 15 randomized scenes for each type of scenario. The result shows that both PRELUDE and PRELUDE (A1 Default Gait) can navigate robustly in static and dynamic environments. It implies that our navigation controller works well with different gait controllers on real robots. Furthermore, PRELUDE achieves overall longer travel distances and higher success rates considering all the four types of environments. We note that the default gait controller suffers from drifts at high-speed turning, often resulting in collisions with obstacles on the side. We hypothesize that model errors from the simplified dynamic model and inaccurate state estimation caused these issues. Our learned gait controller exhibits less drift when turning at high rates, and the robot shows higher agility in cluttered and dynamic environments.

V. CONCLUSION

We introduce PRELUDE, an effective method for learning perceptive locomotion controllers for quadrupedal robots to traverse real-world dynamic environments. Our method combines the complementary strengths of imitation learning and reinforcement learning through a hierarchical design that decomposes the locomotion problem into high-level navigation and low-level gait generation. We designed a steerable cart platform for collecting human navigation demonstrations in complex scenes, and used the collected datasets to train the high-level navigation policy. We used large-scale reinforcement learning to train the low-level gait controller in simulation, demonstrating its effectiveness in transferring to the real world and producing robust and versatile motions. Our work focuses on flat and indoor environments where human steering actions can be conveniently collected with the wheeled platform. For future work, we would like to extend our wheeled carts to more sophisticated mechanical designs to collect human datasets for traversing rough terrains in outdoor environments.

Acknowledgements We would like to thank Zhenyu Jiang and Braham Snyder for providing feedback on this manuscript. We acknowledge the support of the National Science Foundation (1955523, 2145283), the Office of Naval Research (N00014-22-1-2204), and Amazon.

REFERENCES

- [1] C. D. Bellicoso, M. Bjelonic, L. Wellhausen, K. Holtmann, F. Günther, M. Tranzatto, P. Fankhauser, and M. Hutter, "Advances in real-world applications for legged robots," *Journal of Field Robotics*, vol. 35, no. 8, pp. 1311–1326, 2018.
- [2] D. Belter, P. Labecki, P. Fankhauser, and R. Siegwart, "RGB-D terrain perception and dense mapping for legged robots," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 1, pp. 81–97, 2016.
- [3] G. Bledt and S. Kim, "Implementing regularized predictive control for simultaneous real-time footstep and ground reaction force optimization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 6316–6323.
- [4] G. Bledt, P. M. Wensing, and S. Kim, "Policy-regularized model predictive control to stabilize diverse quadrupedal gaits for the mit cheetah," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2017, pp. 4102–4109.
- [5] R. Buchanan, L. Wellhausen, M. Bjelonic, T. Bandyopadhyay, N. Kottege, and M. Hutter, "Perceptive whole-body planning for multilegged robots in confined spaces," *Journal of Field Robotics*, vol. 38, no. 1, pp. 68–84, 2021.
- [6] J. Carius, F. Farshidian, and M. Hutter, "MPC-Net: A first principles guided policy search," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2897–2904, 2020.
- [7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," no. arXiv:1512.03012, 2015.
- [8] E. Coumans and Y. Bai, *Pybullet, a python module for physics simulation in robotics, games and machine learning*, 2017.
- [9] X. Da, Z. Xie, D. Hoeller, B. Boots, A. Anandkumar, Y. Zhu, B. Babich, and A. Garg, "Learning a contact-adaptive controller for robust, efficient legged locomotion," in *Conference on Robot Learning*, 2021, pp. 883–894.
- [10] J. Di Carlo, P. M. Wensing, B. Katz, G. Bledt, and S. Kim, "Dynamic locomotion in the mit cheetah 3 through convex model-predictive control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 1–9.
- [11] P. Fankhauser, M. Bloesch, C. Gehring, M. Hutter, and R. Siegwart, "Robot-centric elevation mapping with uncertainty estimates," in *Mobile Service Robotics*, World Scientific, 2014, pp. 433–440.
- [12] M. Focchi, A. Del Prete, I. Havoutis, R. Featherstone, D. G. Caldwell, and C. Semini, "High-slope terrain locomotion for torque-controlled quadruped robots," *Autonomous Robots*, vol. 41, no. 1, pp. 259–272, 2017.
- [13] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robotics & Automation Magazine*, vol. 4, no. 1, pp. 23–33, 1997.
- [14] Z. Fu, A. Kumar, J. Malik, and D. Pathak, "Minimizing energy consumption leads to the emergence of gaits in legged robots," in *Conference on Robot Learning*, 2021.
- [15] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, "Rloc: Terrain-aware legged locomotion using reinforcement learning and optimal control," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 2908–2927, 2022.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [18] D. Hoeller, L. Wellhausen, F. Farshidian, and M. Hutter, "Learning a state representation and navigation in cluttered and dynamic environments," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5081–5088, 2021.
- [19] C. Imai, M. Zhang, Y. Zhang, M. Kierebinski, R. Yang, Y. Qin, and X. Wang, "Vision-guided quadrupedal locomotion in the wild with multi-modal delay randomization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022, pp. 5556–5563.
- [20] *Intel RealSense SDK*, <https://github.com/IntelRealSense/librealsense>.
- [21] D. Jain, A. Iscen, and K. Caluwaerts, "Hierarchical reinforcement learning for quadruped locomotion," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 7551–7557.
- [22] F. Jenelten, T. Miki, A. E. Vijayan, M. Bjelonic, and M. Hutter, "Perceptive locomotion in rough terrain—online foothold optimization," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5370–5376, 2020.
- [23] D. Kim, D. Carballo, J. Di Carlo, B. Katz, G. Bledt, B. Lim, and S. Kim, "Vision aided dynamic exploration of unstructured terrain with a small-scale quadruped robot," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 2464–2470.
- [24] D. Kim, J. Di Carlo, B. Katz, G. Bledt, and S. Kim, "Highly dynamic quadruped locomotion via whole-body impulse control and model predictive control," *arXiv preprint arXiv:1909.06586*, 2019.
- [25] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," in *Conference on Robot Learning*, 2021.
- [26] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science robotics*, vol. 5, no. 47, 2020.
- [27] O. A. V. Magana, V. Barasuol, M. Camurri, L. Franceschi, M. Focchi, M. Pontil, D. G. Caldwell, and C. Semini, "Fast and continuous foothold adaptation for dynamic locomotion through cnns," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2140–2147, 2019.
- [28] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín, "What matters in learning from offline human demonstrations for robot manipulation," in *Conference on Robot Learning*, 2022, pp. 1678–1690.
- [29] S. Nasiriany, H. Liu, and Y. Zhu, "Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 7477–7484.
- [30] X. Pan, T. Zhang, B. Ichter, A. Faust, J. Tan, and S. Ha, "Zero-shot imitation learning from demonstrations for legged robot visual navigation," in *IEEE International Conference on Robotics and Automation*, 2020, pp. 679–685.
- [31] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne, "Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.
- [32] D. Pongas, M. Mistry, and S. Schaal, "A robust quadruped walking gait for traversing rough terrain," in *IEEE International Conference on Robotics and Automation*, 2007, pp. 1474–1479.
- [33] G. Research. (Aug. 2021). "Google scanned objects," Open Robotics, [Online]. Available: <https://fuel.ignitionrobotics.org/1.0/GoogleResearch/fuel/collections/Google%5C%20Scanned%5C%20Objects>
- [34] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*, 2022, pp. 91–100.
- [35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [36] M. Sorokin, J. Tan, C. K. Liu, and S. Ha, "Learning to navigate sidewalks in outdoor environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3906–3913, 2022.
- [37] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, "Sim-to-real: Learning agile locomotion for quadruped robots," in *Robotics: Science and Systems*, 2018.
- [38] V. Tsounis, M. Alge, J. Lee, F. Farshidian, and M. Hutter, "Deepgait: Planning and control of quadrupedal gaits using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3699–3706, 2020.
- [39] *Unitree-A1*, <https://www.unitree.com/products/a1>
- [40] Z. Wang, A. Novikov, K. Zolna, J. S. Merel, J. T. Springenberg, S. E. Reed, B. Shahriari, N. Siegel, C. Gulcehre, N. Heess, et al., "Critic regularized regression," *Neural Information Processing Systems*, vol. 33, pp. 7768–7778, 2020.
- [41] A. W. Winkler, C. D. Bellicoso, M. Hutter, and J. Buchli, "Gait and trajectory optimization for legged systems through phase-based end-effector parameterization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1560–1567, 2018.
- [42] F. Xia, C. Li, R. Martín-Martín, O. Litany, A. Toshev, and S. Savarese, "Relmogen: Leveraging motion generation in reinforcement learning for mobile manipulation," in *IEEE International Conference on Robotics and Automation*, 2021.
- [43] Z. Xie, X. Da, B. Babich, A. Garg, and M. van de Panne, "Glide: Generalizable quadrupedal locomotion in diverse environments with a centroidal model," *arXiv preprint arXiv:2104.09771*, 2021.
- [44] R. Yang, M. Zhang, N. Hansen, H. Xu, and X. Wang, "Learning vision-guided quadrupedal locomotion end-to-end with cross-modal transformers," in *International Conference on Learning Representations*, 2022.
- [45] Y. Yang, T. Zhang, E. Coumans, J. Tan, and B. Boots, "Fast and efficient locomotion via learned gait transitions," in *Conference on Robot Learning*, 2022, pp. 773–783.
- [46] J. Zico Kolter and A. Y. Ng, "The stanford littledog: A learning and rapid replanning approach to quadruped locomotion," *The International Journal of Robotics Research*, vol. 30, no. 2, pp. 150–174, 2011.

A. Implementation Details

The navigation controller predicts the target velocity commands at 10Hz. The low-level gait controller takes the buffer of recent velocity commands, robot states and previous joint-space actions as input and produces joint-space action commands at 38Hz to actuate the robot.

a) *Navigation Controller*: The navigation policy uses a ResNet18-backbone network [16] as the image encoder. The encoded image features are flattened and concatenated with the 1D heading direction value. The concatenated vector is passed through a two-layer multi-layer perceptron (MLP) with 1024 hidden units in each layer. The input RGB-D images have a size of 212×120 . For imitation learning, we develop our behavioral cloning implementations with the robomimic framework [28]. For the recurrent neural networks, we use LSTM [17] of two layers with 400 hidden units for each layer. The GMM policy output has 5 modes.

b) *Gait Controller*: Each tuple in the buffer $\mathcal{B}_t$ is a 48D vector, containing $u_i \in \mathbb{R}^2$, $q_i \in \mathbb{R}^{34}$, and $a_{i-1} \in \mathbb{R}^{12}$. u_i consists of the forward linear velocity and angular velocity. q_i consists of joint positions and velocities of each leg (24D), binary foot contact states for each foot (4D), the IMU measurement of the robot's body orientation (3D), and angular velocities (3D). a_{i-1} is a 12D joint-space command of the previous time step. We choose $T = 11$ for the size of the history buffer. We use 1D temporal convolution to transform the input $T + 1$ tuples $\mathcal{B}_t$ into a 32D vector. Specifically, the 48D tuple of each time step is first encoded by a linear layer, followed by 3 layers of temporal convolution. The feature from temporal convolution is flattened and projected by another linear layer into the 32D vector. We concatenate the 48D tuple of the most recent state (u_t, q_t, a_{t-1}) and the encoded 32D vector of history states. This concatenated feature is passed through a two-layer MLP of 256 hidden units to produce a Gaussian distribution of joint-space actions.

The reward function for training π_L consists of the terms for tracking commands, balancing the body, minimizing energy consumption, and regulating foot contacts, as below.

- 1) Tracking commands: $K(k_1|e_\psi|^2)e^{k_2|e_{xy}|}$
- 2) Balancing the body: $K(k_3|e_{xy}|^2)K(k_4|\theta|^2)$
- 3) Minimizing energy consumption: $K(k_3|e_{xy}|^2)E$
- 4) Regulating foot contacts: $k_5(\max(n_1 - 3, 0) - n_2)$

where $K(\cdot) := 1 - \tanh(\cdot)$ and $k_1, \dots, k_5$ are given as positive weight coefficients. E is the energy consumption, θ is the vector of roll and pitch, e_{xy} is the linear velocity errors in the x and y axes, and e_ψ is the yaw rate error. n_1, n_2 are the numbers of foot contacts, and non-foot contacts, respectively. The terms of tracking commands and balancing the body are designed to improve the stability of tracking commands, and the other two terms improve the gait patterns. 64 actors of distributed PPO are used for training.

Our learned Gait Controller is trained with the reward design prioritizing tracking rapidly changing commands for reactive locomotion skills. Therefore, this reward design yields

good command-tracking performance, though the behavior of the Gait Controller is jittery and less energy efficient. The learned Gait Controller outperforms the MPC baseline in terms of tracking time-varying velocity commands, as shown in the simulation unit tests (*Analysis of Gait Controllers*), and the impact of better tracking is discussed in Sec. IV-B.

Algorithm 1 PRELUDE

```

// Parameters
N: Number of dataset samples
T: Buffer length of recent robot states
fh: Navigation Controller frequency
fi: Gait Controller frequency

// Training of PRELUDE
D ← {(si, ui)}i=1N           ▷ Human demonstration
πH ← trained by Imitation Learning with D
πL ← trained by Reinforcement Learning in simulation

// Deployment of PRELUDE
ut ← initialized           ▷ velocity command
at ← initialized           ▷ joint-space action
Bt ← {(ui, qi, ai-1)}i=t-T, ..., t ▷ Buffer of recent robot states
th ← t                     ▷ Navigation Controller timer
tl ← t                     ▷ Gait Controller timer
while episode is not done do
  t ← current time
  st ← observed by the robot sensors
  qt ← joint states in st
  if t ≥ th then
    ut ∼ πH(st, ut)
    th ← th +  $\frac{1}{f_h}$ 
  end if
  if t ≥ tl then
    Store (ut, qt, at-1) in Bt
    at ∼ πL(Bt)
    tl ← tl +  $\frac{1}{f_l}$ 
  end if
end while

```

B. Real Robot Evaluation

The location and background are unseen in all of the **Dynamic (Complex)** experiments. Lighting conditions vary for the **Dynamic (Sparse)** episodes compared with the training data in the same location. This location is sensitive to sunlight, and the resulting background is significantly different from the training data. **Static (Sparse)** and **Static (Complex)** are performed over several days in shared spaces; therefore, the background is naturally different. In these two environments, there are 20% and 50% unseen objects, respectively. For sparse scenes where objects are large, we were limited in the number of different obstacles available, while in complex scenes, there was more variation. Also, all humans are unseen obstacles as they differ in size, shape, clothing color, and gait style.