

Learning Agile Striker Skills for Humanoid Soccer Robots from Noisy Sensory Input

Zifan Xu¹, Myoungkyu Seo¹, Dongmyeong Lee¹, Hao Fu¹, Jiaheng Hu¹, Jiaxun Cui¹,
Yuqian Jiang¹, Zhihan Wang¹, Anastasiia Brund¹, Joydeep Biswas¹, Peter Stone^{1,2}

¹Department of Computer Science, The University of Texas at Austin, ²Sony AI
Project website: <https://humanoidsoccer.github.io>

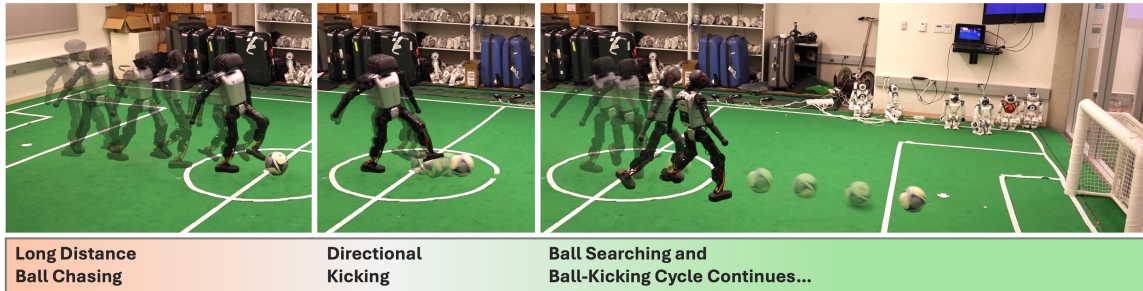


Fig. 1: The illustration of a complete ball-kicking cycle in the robust continual ball-kicking task.

Abstract—Learning fast and robust ball-kicking skills is a critical capability for humanoid soccer robots, yet it remains a challenging problem due to the need for rapid leg swings, postural stability on a single support foot, and robustness under noisy sensory input and external perturbations (e.g., opponents). This paper presents a reinforcement learning (RL)-based training pipeline that enables humanoid robots to execute robust continual ball-kicking with adaptability to different ball-goal configurations. The pipeline extends a typical teacher-student training framework—in which a “teacher” policy is trained with ground truth state information and the “student” learns to mimic it with noisy, imperfect sensing—by including four training stages: (1) long-distance ball chasing (teacher); (2) directional kicking (teacher); (3) teacher policy distillation (student), and (4) student adaptation and refinement (student). Key design elements—including tailored reward functions, realistic noise modeling, and online constrained RL for adaptation and refinement—are critical for closing the sim-to-real gap and sustaining performance under perceptual uncertainty. Extensive evaluations in both simulation and on a real robot demonstrate strong kicking accuracy and goal-scoring success across diverse ball-goal configurations. Ablation studies further highlight the necessity of the constrained RL, noise modeling, and the adaptation stage. This work presents a training pipeline for robust continual humanoid ball-kicking under imperfect perception, establishing a benchmark task for visuomotor skill learning in humanoid whole-body control.

This work has taken place jointly in the Learning Agents Research Group (LARG) and Autonomous Mobile Robotics Laboratory (AMRL) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-24-1-2550), ARO (W911NF-17-2-0181, W911NF-23-2-0004, W911NF-25-1-0065), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork) Lockheed Martin, and UT Austin’s Good Systems grand challenge. AMRL research is supported in part by NSF (CAREER-2046955, OIA-2219236, DGE-2125858, CCF-2319471), ARO (W911NF-23-2-0004), Amazon, and JP Morgan. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors. Peter Stone serves as the Chief Scientist of Sony AI and receives financial compensation for that role. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

I. INTRODUCTION

Humanoid robots are designed with anthropomorphic morphology, allowing them to operate in environments built for humans [1]. This makes them suitable for diverse tasks that potentially requires whole-body control [2], dynamic balance [3], [4], and coordinated interaction with the environment [5], [6]. Recent advances in reinforcement learning (RL) and simulation have enabled progress in humanoid locomotion and manipulation, including robust locomotion [7]–[9], and object interaction skills [10]–[12].

Despite this progress, learning whole-body tasks that combine speed, balance, and perception remains challenging. Such tasks require rapid limb movements, stability on a reduced support polygon, and robustness to perceptual noise and external perturbations. A prime example is the ball-kicking skill: the robot must approach the ball quickly, coordinate a powerful leg swing, maintain balance on one foot, and direct the ball toward the goal with precision. Effective striking further depends on accurate localization of both ball and goal, which is often degraded by sensor noise, delays, and perception errors. These demands make humanoid striking a difficult benchmark for visuomotor whole-body control.

This paper presents an RL-based training pipeline that enables humanoid robots to acquire such a reliable ball-kicking skill with great generalization across diverse ball-goal configurations. The pipeline comprises four consecutive training stages: (1) long-distance ball chasing, (2) directional kicking, (3) teacher policy distillation, and (4) student policy adaptation and refinement. The first two training stages form a simple two-stage curriculum that enables a *privileged teacher*, that has the ground-true ball and goal position information, to acquire ball-chasing skills, and extend them to directional kicking. During this process, aggressive domain

randomization, such as external pushes on the body and ball, is applied to encourage recovery from imperfect states, such as missed kicks and tilted postures. The teacher policy is then distilled into a student policy, that has *imperfect* perception of the ball and goal position, via DAgger [13]. The modeling of such imperfect perception constitutes three components: a velocity-dependent noise model, delayed updates, and frame drops caused by perceptual occlusion. Finally, stage four applies online adaptation using N-P3O [14], a constrained RL algorithm, to refine the kicking skill. This process reduces jittery leg motions and unsafe sharp turning that often arise when using a fixed regularization coefficient, which can impose disproportionately large penalties at certain steps—particularly those immediately preceding the kick, where the agent expects high immediate future reward from the kicking task.

We evaluate the resulting policy extensively in simulation, measuring kick accuracy, success rate, kick strength, and energy cost across diverse ball–goal configurations. Ablation studies highlight the importance of the final adaptation stage and motion refinement through constrained RL. To assess real-world performance, we deploy the policy on a Booster T1 humanoid robot, which achieves an average success rate of 66.7% across five different ball–goal configurations. Overall, this work presents a comprehensive framework for learning robust continual humanoid ball-kicking skill under imperfect perception and establishes a benchmark for agile visuomotor skill learning in humanoid robots.

II. RELATED WORK

A. Robot Soccer

Robot soccer has long served as a benchmark domain for testing integrated perception, control, and decision-making systems [15]–[17]. Early efforts explored rule-based strategies and simple controllers, but recent advances in reinforcement learning (RL) have enabled agile, dynamic skills. On quadrupeds, researchers have demonstrated soccer behaviors such as cooperative play [18] and precise shooting [19] using hierarchical RL, robust dribbling in the wild [20], and dynamic goalkeeping against human players [21], [22]. Quadrupeds naturally benefit from high stability, making perception and balance control easier compared to humanoid platforms.

Humanoid soccer is comparatively less explored, but it presents greater challenges due to bipedal balance, whole-body coordination, and larger perceptual uncertainty [23], [24]. Prior work has focused on dribbling with agile locomotion [25], learning from gameplay recordings, and diffusion-based visuomotor policies [26]. Other efforts target biomechanics-inspired approaches to humanoid kicking [27]. However, most prior humanoid studies emphasize dribbling or generic locomotion skills, while agile striking under noisy perception remains relatively unaddressed. Our work fills this gap by tackling robust and adaptive ball-kicking skills, a critical capability for advancing humanoid soccer.

B. Humanoid Whole-Body Control

Whole-body control of humanoid robots is a central challenge in robotics, requiring dynamic coordination of dozens of degrees of freedom while maintaining balance and stability [28]–[36]. Traditional methods often rely on model predictive control (MPC) or optimization-based controllers that explicitly encode dynamics and contact constraints [28], [29]. While these approaches provide stability guarantees, they can be brittle under perception noise, latency, or unmodeled dynamics.

Reinforcement learning has recently emerged as a powerful alternative, enabling humanoids to learn locomotion [37], acrobatic maneuvers [30], and robust recovery strategies [33] directly from interaction. GPU-accelerated simulation frameworks have further accelerated progress, allowing large-scale training of bipedal policies. Beyond locomotion, researchers have begun to study whole-body visuomotor skills, such as manipulation [34], agile sports behaviors (e.g., badminton [35]), and multi-contact motion tracking [36]. These advances highlight the feasibility of training policies that integrate perception and control for high-dimensional humanoids.

Our work contributes to this growing literature by extending whole-body control to a challenging soccer striker task. Unlike locomotion or manipulation, striking requires rapid leg swings, balance on a single support foot, and accurate timing under noisy perception. By combining robust perception with a staged RL framework, our approach bridges the gap between whole-body control research and the practical demands of humanoid soccer.

III. METHOD

System Overview. The objective of the system is to learn a control policy $a_t = \pi(o_{t-H:t})$ that maps a history of H observations $o_{t-H:t} = [o_{t-H}, \dots, o_t]$ to an action a_t , corresponding to the joint position targets of a humanoid robot at 50 Hz. Each observation at time t is a tuple $o_t = (o_t^p, o_t^b, o_t^g)$ where o_t^p denotes proprioceptive measurements, and o_t^b and o_t^g represent the estimated ball and goal positions, respectively. By executing the learned policy, the robot exhibits continual ball-kicking behaviors. As shown in Fig. 1, each kicking cycle integrates three key phases: (i) approaching the ball from long distance, (ii) performing a kick motion that directs the ball toward the goal, and (iii) reorienting to locate the ball again and seamlessly initiating the next kicking attempt. An illustration of phase (iii) is shown in Fig. 3.

The system is trained through a four-stage training framework, followed by a real-world deployment pipeline:

- 1) **Long-Distance Chasing** – a “teacher” policy, by which we mean a policy with privileged ground-truth ball position, learning a robust walking gait to approach the ball from diverse initial configurations;
- 2) **Directional Kicking** – a teacher policy acquiring precise and robust kick motions using privileged ground-truth ball and goal positions;
- 3) **Teacher Policy Distillation** – transferring privileged-information teacher policies into “student” policies, by

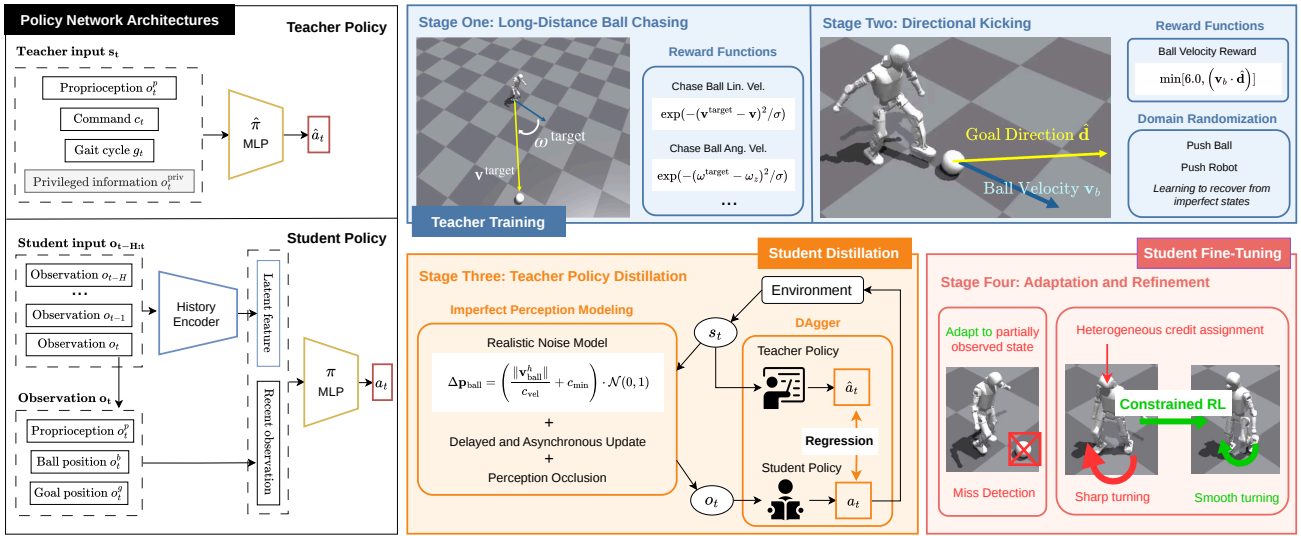


Fig. 2: Left: The network architectures for the teacher and the student network; Right: Multi-stage training framework: (1) Long-distance Chasing; (2) Directional Kicking; (3) Teacher Policy Distillation; (4) Student Adaptation and Refinement.

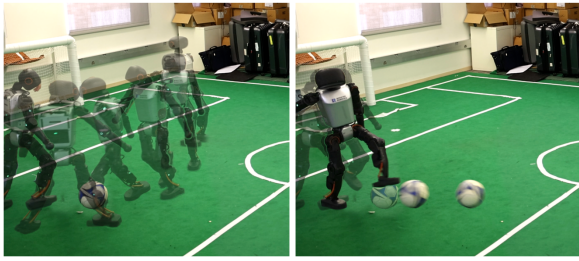


Fig. 3: Kicking cycle phase (iii): reorienting to locate the ball and seamlessly initiating a kicking attempt.

which we mean policies that operate under imperfect perception.

- Student Adaptation and Refinement** – applying an online constrained RL algorithm to adapt to partially observed states created by imperfect perception.

Each stage builds on the previous one, progressively shaping the policy to acquire a ball-kicking skill that is robust to perception noise enabling zero-shot real-world deployment. The following sections detail the four training stages and the deployment pipeline.

A. Long-Distance Chasing

a) Privileged Teacher Policy: The teacher policy is defined as $\hat{a}_t = \hat{\pi}(s_t)$ where the input state s_t at timestep t consists of three components: (i) proprioceptive observations o_t^p , (ii) a velocity command $c_t = (v_t^x, v_t^y, \omega_t^z)$, (iii) a gait cycle g_t , and (iv) privileged observations o_t^{priv} . The design of the proprioceptive and privileged observations varies across the actual morphology of the robot. In this paper, we follow the proprioceptive and privileged observations from Booster Gym [38], with the exception that the privileged input additionally includes ground-truth information about the ball and goal positions, the ball’s linear velocity, and its physical parameters such as mass and friction. The full list of observations is listed in Table I.

TABLE I: Summary of Observation Space.

Components	Dims	Teacher (Critic)	Student
Commands	3	✓	✓
Gait Cycle	2	✓	✓
Gravity Vector	3	✓	✓
Angular Velocity	3	✓	✓
Joint Position	23	✓	✓
Joint Velocity	23	✓	✓
Previous Action	23	✓	✓
Noisy Ball Position	2		✓
Noisy Goal Position	2		✓
Body Mass	1	✓	
Body Center of Mass	3	✓	
Base Linear Velocity	3	✓	
Base Height	1	✓	
Push Force	2	✓	
Push Torque	3	✓	
Real Ball Position	2	✓	
Real Goal Position	2	✓	
Real Ball Velocity	2	✓	
Ball Mass	1	✓	
Ball Friction	1	✓	

b) Environment Design: The environment contains a humanoid robot, a soccer ball, and a point goal representing the kick direction. At the beginning of each episode, the robot’s initial position is sampled uniformly within an annular region centered at the ball, bounded by an inner radius of 0.5 m and an outer radius of 4.0 m. The long-distance chasing task is adapted from the standard flat-ground walking task, with modifications tailored to the ball-chasing task. Instead of tracking a velocity target in the world frame, the commanded linear velocity always aligns with the robot–ball vector. Let $\mathbf{r}_{rb}(t) \in \mathbb{R}^2$ denote the vector from the robot base to the ball at time t , and $\hat{\mathbf{r}}_{rb}(t) = \frac{\mathbf{r}_{rb}(t)}{\|\mathbf{r}_{rb}(t)\|}$ its unit vector. The commanded linear velocity target is defined as $\mathbf{v}_t^{\text{target}} = v_t^x \cdot \hat{\mathbf{r}}_{rb}(t)$. For rotation, let $\theta_t \in [-\pi, \pi]$ be the signed angular difference between the robot heading vector and $\hat{\mathbf{r}}_{rb}(t)$. The target angular velocity is given by $\omega_t^{\text{target}} = \omega_t^z \cdot \text{sign}(\theta_t)$, so that the robot always turns in the direction (left or right) that minimizes $|\theta_t|$. The scalar $\omega_t^z \geq 0$ controls the magnitude of the turning speed.

Components	Equations ¹	Weights
Task Rewards		
Chase ball lin. vel.	$\exp(-\ \mathbf{v}^{\text{target}} - \mathbf{v}\ ^2/\sigma)$	2.0 (0.5)
Chase ball ang. vel.	$\exp(-(\omega^{\text{target}} - \omega_c)^2/\sigma)$	2.0 (0.5)
Feet swing	$1_{\text{feet swing}} \cdot 1_{\text{swing period}}$	3.0 (0.0)
Arm swing	$1_{\text{arm swing}} \cdot 1_{\text{swing period}}$	1.0 (0.0)
Head tracking ball	$1_{\text{ball in FOV}}$	1.0
Ball velocity	$\min[6.0, (\mathbf{v}_b \cdot \hat{\mathbf{d}})]$	0.0 (4.0)
Regularization Rewards		
Survival	1	1.0
Base height	$(h^{\text{des}} - h)^2$	-20.0
Orientation	$\ g\ ^2$	-5.0
Torque	$\ \tau\ ^2$	-2×10^{-4}
Torque tiredness	$\ \tau/\tau_{\text{max}}\ ^2$	-1×10^{-2}
Power	$\max(\tau \cdot \dot{q}, 0)$	-2×10^{-4}
Lin velocity (z)	v_z^2	-2.0
Ang velocity (xy)	$\ \omega_{xy}\ ^2$	-0.2
Joint velocity	$\ \dot{q}\ ^2$	-1×10^{-4}
Joint acceleration	$\ \ddot{q}\ ^2$	-1×10^{-7}
Base acceleration	$\ \ddot{v}\ ^2 + \ \ddot{\omega}\ ^2$	-1×10^{-4}
Action rate	$\ a_t - a_{t-1}\ ^2$	-1.0
Upper body symmetry	$\ q - q_0\ ^2$	-0.5
Joint position limit	$1_{q > q_{\text{max}}} + 1_{q < q_{\text{min}}}$	-1.0
Collision	$n_{\text{collision}}$	-1.0
Feet slip	$1_{\text{feet stance}} \cdot \ \mathbf{v}_{\text{feet}}\ ^2$	-0.1
Feet yaw	$\ \psi_{\text{feet}} - \psi_{\text{base}}\ ^2$	-1.0 (0.0)
Feet roll	$\ \phi_{\text{feet}}\ ^2$	-0.1
Feet distance	$\max(d_{\text{ref}} - d_{\text{feet}}, 0)$	-1.0

TABLE II: Summary of reward functions. Weights are specified for the long-distance chasing task; values in parentheses indicate the weights used for the directional kicking task when they differ.

To improve robustness, we apply domain randomization following prior work on humanoid locomotion [38], [39]. Randomized parameters include joint stiffness, damping, and friction, body masses, and center of mass, as well as external perturbations such as forces and pushes. Additionally, the ball’s mass and surface friction are randomized to encourage robust ball interaction.

c) Reward Design: The reward formulation builds on prior walking-gait designs [38], [39], where the primary task rewards are exponential kernels of linear and angular velocity tracking errors. For the ball-chasing task, the tracking errors are defined relative to $\mathbf{v}_t^{\text{target}}$ and ω_t^{target} grounded by the ball position (detailed in Section III-B (a)). Table II provides the full list of reward functions specific to our hardware platform Booster T1 robot.

B. Directional Kicking

a) Environment Design: The directional kick environment simulates a soccer field consistent with the RoboCup Adult Size Humanoid League rules [40]. The field measures 14 m \times 9 m, with goals placed at $x = -7$ m and $x = +7$ m.

At the beginning of each episode, the ball is uniformly initialized within the region $[-1.5, 6.5] \times [-4.0, 4.0]$, corresponding roughly to the front half of the field. The robot is then placed within an annular region centered at the ball, bounded by an inner radius of 0.5 m and an outer radius of 2.0 m.

If the ball exits the field boundary, it is reset to a new position without terminating the episode. The reset location

¹The notation in this Table is consistent with popular convention, and will be defined in full detail in the Appendix of the online publication.

is sampled from an annular region centered at the original ball position. This setup encourages the robot to repeatedly search for the ball, approach it, and execute a kick, thereby creating continuous striking cycles.

To improve robustness, we introduce additional domain randomization: every four seconds, the ball receives an external disturbance by applying a random linear velocity sampled from $[-0.1, 0.1]$ m/s along both axes. This perturbation produces imperfect states from which the teacher policy must recover. These recovery behaviors can then be distilled into the student policy, which may encounter similar imperfect states due to noisy perception.

b) Reward Design: The directional kick task introduces two extra reward functions: *ball velocity* reward and *head tracking ball* reward detailed as follows.

Ball velocity reward encourages the robot to impart a high ball velocity directed toward the opponent’s goal. Let $\mathbf{p}_g \in \mathbb{R}^2$ denote the goal position, $\mathbf{p}_b \in \mathbb{R}^2$ the ball position, and $\mathbf{v}_b \in \mathbb{R}^2$ the linear velocity of the ball. Define the goal direction unit vector as

$$\hat{\mathbf{d}} = \frac{\mathbf{p}_g - \mathbf{p}_b}{\|\mathbf{p}_g - \mathbf{p}_b\|}. \quad (1)$$

The reward is given by the projection of the ball velocity onto the goal direction, scaled by the velocity magnitude:

$$r_t = \min[6.0, (\mathbf{v}_b \cdot \hat{\mathbf{d}})], \quad (2)$$

The reward is capped at 6.0 to prevent the agent from exploiting the objective by generating excessively large ball velocities.

Head tracking ball reward encourages the robot to maintain visual contact with the ball. At each timestep, a reward of +1 is assigned if the ball lies within the field of view of the robot’s head camera.

C. Teacher Policy Distillation

Once the teacher policy acquires a basic kicking skill with a degree of recovery capability, it is distilled into a student policy that can operate under imperfect perception.

a) Student Policy: The student policy is defined as $a_t = \pi(o_{t-H:t})$, where the input consists of a history of $H = 50$ observations $o_{t-H:t} = [o_{t-H}, \dots, o_t]$. Each observation is represented as $o_t = (o_t^p, o_t^b, o_t^g)$, with o_t^p denoting proprioceptive measurements, and o_t^b and o_t^g corresponding to noisy estimates of the ball and goal positions. The perceptual error modeling for these inputs is described in the next subsection.

To capture temporal dependencies, the observation history $o_{t-H:t}$ is first processed by a 1D convolutional encoder, producing a 64-dimensional latent representation. This latent vector is concatenated with the most recent observation o_t and passed through a multilayer perceptron (MLP) with three hidden layers of 256, 256, and 128 units, respectively, to output the action.

b) Noise Modeling: To simulate realistic perception uncertainty, we employ a velocity-dependent noise model [41] for both the ball and goal positions. The magnitude of the injected noise scales with object velocity in the robot’s head frame, reflecting the fact that rapid-moving objects generally

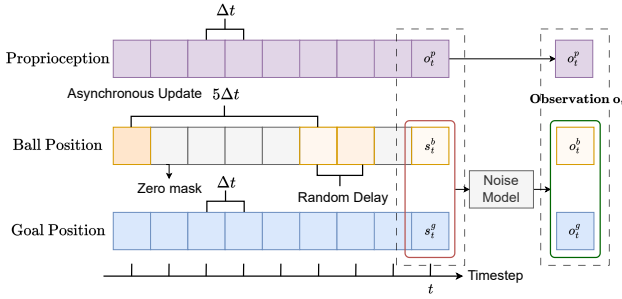


Fig. 4: An illustration of the realistic perception modeling.

lead to larger noise in camera-based localization. Formally, the noise to the ball/goal position $\Delta \mathbf{p}_{\text{ball/goal}}$ is defined as

$$\Delta \mathbf{p}_{\text{ball/goal}} = \left(\frac{\|\mathbf{v}_{\text{ball/goal}}^h\|}{c_{\text{vel}}} + c_{\text{min}} \right) \cdot \mathcal{N}(0, I), \quad (3)$$

where $\mathbf{v}_{\text{ball/goal}}^h$ denotes the ball/goal velocity in the robot’s camera frame, c_{vel} and c_{min} are scaling constants, and $\mathcal{N}(0, I)$ is standard Gaussian vector noise. The student policy is trained using Dataset Aggregation (DAgger) [13], which iteratively collects rollouts under the student policy while querying the teacher for corrective actions. This process ensures that the student learns not only from the teacher’s demonstrations but also from states encountered due to its own errors, improving robustness under noisy perception.

In addition, perception signals are subject to realistic delays and dropouts. Proprioceptive observations are updated at 50 Hz (every 0.02 s), while ball positions are published at around 10 Hz (every 0.1 s) with a random offset to mimic asynchronous updates. If the ball lies outside the camera’s field of view or falls between publish intervals, its position o_t^b is masked with zeros. Goal observations are updated at the same frequency as the proprioception and subject to the same noise injection. Fig. 4 shows an overview of the imperfect perception model.

D. Student Adaptation and Refinement

a) Student Adaptation: Conventional teacher–student frameworks typically involve training a teacher policy and distilling it into a student policy [9], under the assumption that full environment state information can be reconstructed from observation histories. However, recent work shows that certain partially observed states cannot be effectively captured by the teacher training stage, and that online RL-based adaptation can significantly improve performance in such cases [42]. Inspired by this work, we further adapt the student policy with online RL under noisy perception. During adaptation, we apply smaller perturbations to the robot and ball, allowing the policy to concentrate on improving accuracy and robustness to realistic sensory uncertainty.

b) Refinement through constrained RL: While training the adapted student, we observed undesirable behaviors such as sharp rotations during ball search and overly dynamic kicking, which may compromise both safety and stability

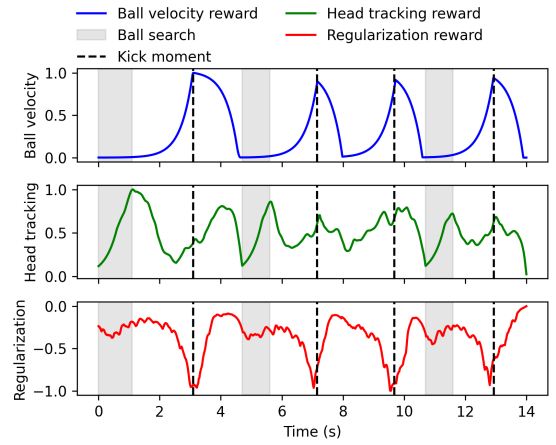


Fig. 5: An illustration of inhomogeneous credit assignment within four complete kicking cycles. y-axis shows the normalized future returns for regularization reward, head tracking reward, and ball velocity reward, respectively. Large dips in regularization reward indicate jittery kick motions.

during the real-world deployment. We attribute these issues to inhomogeneous credit assignment within an episode, where task rewards, e.g., *ball velocity* in Fig. 5 (top) and *head tracking ball* in Fig. 5 (middle), are distributed unevenly across a ball-kicking cycle. In particular, the discounted future reward associated with ball velocity exhibits a pronounced peak at the kick moment (indicated by the dashed vertical lines in Fig. 5). In this case, a fixed regularization coefficient can impose disproportionately large penalties at certain steps—particularly those immediately preceding the kick, where the agent expects high immediate future reward from the kicking task. This effect is reflected in the sharp drops of the regularization reward shown in Fig. 5 (bottom).

To mitigate this problem, we reformulate directional kicking as a constrained RL problem. Let r_{task} denote the aggregated task rewards and r_{regu} denote the aggregated regularization rewards (as defined in Table II). The objective is to maximize task performance while constraining excessive regularization costs:

$$\max_{\pi} J_{\text{task}}(\pi) \quad (4)$$

$$\text{s.t. } r_{\text{regu}}(a_t, s_t) \leq h, \quad \forall (s_t, a_t) \in \tau; \tau \sim p^{\pi}(\tau), \quad (5)$$

where $J_{\text{task}}(\pi) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_{\text{task}}(a_t, s_t) \mid s_0 \sim \mu(s), \tau \sim p^{\pi}(\tau) \right]$ with $\mu(s)$ denotes the initial state distribution, τ a trajectory sampled under policy π , and h is the upper limit on the aggregated regularization reward. In practice, h is set to $1.5 \times$ per-step average regularization cost of a walking gait.

This constrained optimization is solved using N-P3O [14], a constrained policy gradient method shown to be effective for locomotion tasks [35]. Importantly, the role of constrained RL here is not strict hardware safety enforcement, but rather the mitigation of inhomogeneous credit assignment, resulting in smoother, and safer ball-kicking motions.

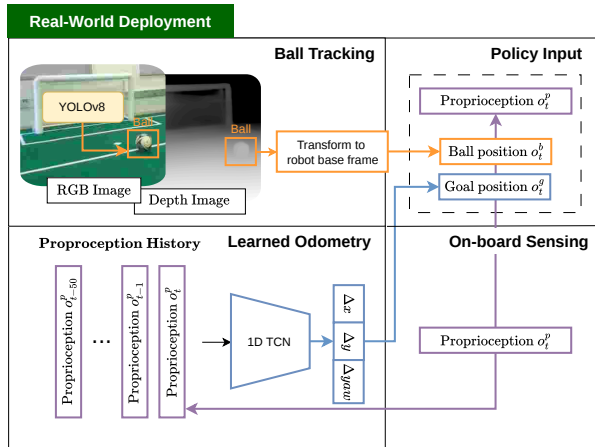


Fig. 6: An overview of the real-world deployment pipeline

IV. REAL-WORLD DEPLOYMENT

This section details the hardware platforms and core components that enables the real-world deployments.

A. Hardware Platform

We deploy our ball-kicking policy on the physical Booster T1 humanoid robot [43]. The robot stands 1.18-meter-tall, features 23 degrees of freedom and is equipped with a rich sensor suite including an IMU and joint encoders for real-time state estimation, as well as a ZED 2i Stereo Camera for visual perception. All onboard computations are performed using an NVIDIA AGX Orin GPU in conjunction with a 14-core high-performance CPU. The ball-kicking policy is executed at a control frequency of 50 Hz.

B. Perception Pipeline

Our perception pipeline provides the policy with real-time estimates of the ball and the opponent’s goal in the robot’s base frame. For ball perception, a YOLOv8 model [44] processes the robot’s RGB-D camera stream to detect the ball’s bounding box. The center of the detected box is unprojected into a 3D point using its corresponding depth value. This point is then transformed into the robot’s base frame using the head’s forward kinematics. This results in the pose of the ball in the robot’s base frame serving as the input to the ball-kicking policy.

C. Odometry System

To enable goal localization, we employ a lightweight data-driven legged-inertial odometry module inspired by Legolas [45]. As shown in Fig. 6 on the bottom left, the system fuses inertial signals—specifically the gravity vector and angular velocity—with leg kinematics such as joint positions and velocities. A neural network processes a 50-frame history of these measurements and predicts the relative pose change ($\Delta x, \Delta y, \Delta \text{yaw}$) in SE(2) space. Robustness of such prediction is further enhanced through multi-step supervision and an SE(2) consistency constraint, which together promote long-term stability and coherent estimation of both translation and

rotation. Across trajectories spanning four complete ball-kicking cycles, the learned odometry achieves an average position error of 0.14 ± 0.09 meters.

V. EVALUATIONS

The learned control policy is evaluated in both simulation and the real world. We detail the evaluation settings and present the results in the following sections.

A. Simulation Results

a) *Evaluation setting*: We conduct evaluation under a diverse set of initial ball positions, represented as a 9×9 matrix evenly spanning the region $[-1.5, 6.5] \times [-4.0, 4.0]$ at 1 m intervals. For each ball position, 50 independent trials are executed with randomly sampled robot initial states.

b) *Kick accuracy and success rate*: Kick accuracy is measured as the cosine similarity between the post-kick ball velocity unit vector and the ground-truth ball–goal direction vector. The *success rate* is defined as the fraction of trials in which the ball crosses the goal line between the the goal posts². Under these metrics, the policy achieves an average success rate of 79.5% and an average kick accuracy of 0.956. Results are visualized in Fig. 7

c) *Maximum ball velocity and energy cost*: We also evaluate the maximum ball velocity and the average energy cost during kicking, where energy cost is computed over timesteps in which the robot is within 0.5 m of the ball. The learned policy achieves an overall maximum ball velocity of 4.13 m/s and an average energy cost of 110.8 J/s. Figure 7 (right) illustrates the distribution of these metrics across different initial ball positions. Notably, an emergent energy-saving strategy is observed: when the ball is positioned closer to the goal, the robot consistently applies lighter kicks and expends less energy, indicating adaptive behavior that balances efficiency with task success.

B. Real-world results

We perform a smaller-scale evaluation on hardware using five different ball positions in a RoboCup Kid-Size soccer field. For each position, three trials are conducted with the robot initialized from the same fixed location (6.5 m in front of the goal). We report averaged kick accuracies and success rates across these trials. The results are presented in Table. III. The policy achieves an overall success rate of 66.7%.

Ball Pos. x (m)	Ball Pos. y (m)	Kick Accuracy	Success Rate
4.5	0.0	0.99 ± 0.01	3/3
3.5	1.8	0.94 ± 0.08	2/3
3.5	-1.8	0.95 ± 0.02	1/3
6.0	2.5	0.99 ± 0.02	2/3
6.0	-2.5	0.99 ± 0.01	2/3

TABLE III: Kick accuracy and success rate for the real-world experiments at different ball positions (relative to goal center). The robot starts from 6.5 meters away from the goal.

²The goal posts distance are specified by RoboCup Adult-Size Humanoid League rules [40].

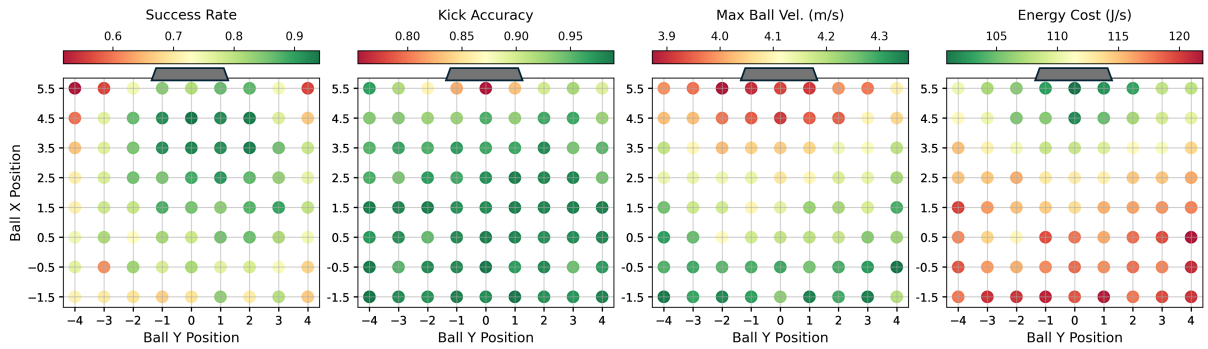


Fig. 7: Visualizations of success rate, kick accuracy, max ball vel., and energy cost, at different initial ball positions. The shadowed trapezoidal regions visualize the goal areas.

C. Ablation Studies

To understand the contribution of individual design components, we perform ablations on the two key aspects of our system:

a) *Constrained RL*: We evaluate the effect of constrained RL by comparing N-P3O against standard PPO with fixed regularization coefficients. Table IV summarizes the results across success rate, peak ball velocities, and energy cost. N-P3O achieves the highest overall performance, with a success rate of 79.5%, while maintaining smooth and stable motions characterized by lower maximum angular velocity and energy consumption at a small cost to peak ball velocity. Energy consumption is also substantially reduced to 108.6 J/s, less than half that of PPO. These results confirm that the constrained optimization in N-P3O mitigates unsafe behaviors and improves both efficiency and stability, leading to the possibility of more reliable real-world deployment.

Method	Succ.	Max Ball Vel.	Max Ang. Vel.	Energy Cost
N-P3O	79.5%	1.05 m/s	2.13 rad/s	108.6 J/s
PPO-1.0	64.8%	1.41 m/s	3.21 rad/s	255.8 J/s
PPO-1.5	43.4%	1.21 m/s	2.86 rad/s	185.7 J/s

TABLE IV: Comparison between N-P3O and PPO with fixed regularization coefficients of 1.0 and 1.5. The bold numbers indicate the best results under each metric.

b) *Student adaptation*: Finally, we evaluate the role of the final student adaptation stage by comparing policies before and after online adaptation. As shown in Table V, the adapted student achieves a success rate of 79.5% with a kick accuracy of 0.956, while also reducing energy consumption to 110.78 J/s. In contrast, the pre-adaptation student attains only 52.3% success and 0.807 accuracy, consuming more than twice as much energy (256.2 J/s). For reference, the privileged teacher policy trained without sensor noise achieves slightly higher success (81.1%). These results highlight the fact that online adaptation is essential for handling partially observed states that cannot be addressed by distillation of a privileged teacher policy.

VI. CONCLUSIONS

This paper introduced a RL-based system for enabling humanoid robots to perform fast and robust ball-kicking under

Method	Succ.	Kick Accuracy	Energy Cost
Student after adaptation	79.5%	0.956	110.8 J/s
Student before adaptation	52.3%	0.807	256.2 J/s
Teacher (w/o sensor noise)	81.1%	0.954	240.1 J/s

TABLE V: Evaluation of student policies before and after adaptation, and a teacher policy with perfect perception. The bold numbers indicate the best results under each metric.

noisy perception. Building on a four-stage teacher–student framework, comprising long-distance chasing, directional kicking, teacher policy distillation, and student adaptation and refinement, the system achieves continuous ball-kicking behaviors that remain stable under perceptual noise. Experiments in both simulation and on real hardware demonstrate high accuracy and goal-scoring success across diverse ball–goal configurations, while ablation studies confirm the necessity of core design choices including the noise modeling, constrained RL, and online adaptation.

A key limitation lies in that the long-distance chasing gaits rely on heavily engineered reward functions adapted from walking gait training, and the success of transferring to ball-kicking depends on their similarity in motion. This approach may not generalize to tasks with different motions, motivating future work on learning from diverse human motions as a more scalable strategy.

REFERENCES

- [1] Z. Gu, J. Li, W. Shen, W. Yu, Z. Xie, S. McCrory, X. Cheng, A. Shamsah, R. Griffin, C. K. Liu, *et al.*, “Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning,” *arXiv preprint arXiv:2501.02116*, 2025.
- [2] J. Dao, H. Duan, and A. Fern, “Sim-to-real learning for humanoid box loco-manipulation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 930–16 936.
- [3] Y. Li, Y. Zhang, W. Xiao, C. Pan, H. Weng, G. He, T. He, and G. Shi, “Learning gentle humanoid locomotion and end-effector stabilization control,” *arXiv preprint arXiv:2505.24198*, 2025.
- [4] T. Zhang, B. Zheng, R. Nai, Y. Hu, Y.-J. Wang, G. Chen, F. Lin, J. Li, C. Hong, K. Sreenath, *et al.*, “Hub: Learning extreme humanoid balance,” *arXiv preprint arXiv:2505.07294*, 2025.
- [5] Z. Wang, X. Yang, J. Zhao, J. Zhou, T. Ma, Z. Gao, A. Ajoudani, and J. Liang, “End-to-end humanoid robot safe and comfortable locomotion policy,” *arXiv preprint arXiv:2508.07611*, 2025.
- [6] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang, “Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control,” *arXiv preprint arXiv:2505.03738*, 2025.

- [7] 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*. PMLR, 2022, pp. 91–100.
- [8] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.
- [9] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Conference on robot learning*. PMLR, 2023, pp. 403–415.
- [10] Z. He, K. Lei, Y. Ze, K. Sreenath, Z. Li, and H. Xu, "Learning visual quadrupedal loco-manipulation from demonstrations," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 9102–9109.
- [11] J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad, M. Itkina, *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025.
- [12] I. Dadiotis, M. Mittal, N. Tsagarakis, and M. Hutter, "Dynamic object goal pushing with mobile manipulators through model-free constrained reinforcement learning," *arXiv preprint arXiv:2502.01546*, 2025.
- [13] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
- [14] J. Lee, L. Schroth, V. Klemm, M. Bjelonic, A. Reske, and M. Hutter, "Evaluation of constrained reinforcement learning algorithms for legged locomotion," *arXiv preprint arXiv:2309.15430*, 2023.
- [15] N. Kohl and P. Stone, "Policy gradient reinforcement learning for fast quadrupedal locomotion," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, vol. 3. IEEE, 2004, pp. 2619–2624.
- [16] M. Hausknecht and P. Stone, "Learning powerful kicks on the aibo ers-7: The quest for a striker," in *Robot Soccer World Cup*. Springer, 2010, pp. 254–265.
- [17] R. Gerndt, D. Seifert, J. H. Baltes, S. Sadeghnejad, and S. Behnke, "Humanoid robots in soccer: Robots versus humans in robocup 2050," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 147–154, 2015.
- [18] Z. Su, Y. Gao, E. Lukas, Y. Li, J. Cai, F. Tulbah, F. Gao, C. Yu, Z. Li, Y. Wu, and K. Sreenath, "Toward real-world cooperative and competitive soccer with quadrupedal robot teams," *arXiv preprint arXiv:2505.13834*, 2025, may 20 2025. [Online]. Available: <https://arxiv.org/abs/2505.13834>
- [19] Y. Ji, G. B. Margolis, and P. Agrawal, "Dribblebot: Dynamic legged manipulation in the wild," *arXiv preprint arXiv:2304.01159*, 2023, april 2023. [Online]. Available: <https://arxiv.org/abs/2304.01159>
- [20] Y. Ji, Z. Li, Y. Sun, X. B. Peng, S. Levine, G. Berseth, and K. Sreenath, "Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2022, arXiv:2208.01160. [Online]. Available: <https://arxiv.org/abs/2208.01160>
- [21] L. Blommers, "Autonomous quadrupedal goalkeeping using hierarchical rl and vision-based localization," Master's thesis, Eindhoven University of Technology, November 2024, master's thesis.
- [22] X. Huang, Z. Li, Y. Xiang, Y. Ni, Y. Chi, Y. Li, L. Yang, X. B. Peng, and K. Sreenath, "Creating a dynamic quadrupedal robotic goalkeeper with reinforcement learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2023, arXiv:2210.04435. [Online]. Available: <https://arxiv.org/abs/2210.04435>
- [23] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, J. Humplik, M. Wulfmeier, *et al.*, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," *Science Robotics*, vol. 9, no. 89, Apr. 2024. [Online]. Available: <http://dx.doi.org/10.1126/scirobotics.adi8022>
- [24] Y. Wang, C. Luo, P. Chen, J. Liu, W. Sun, T. Guo, K. Yang, B. Hu, Y. Zhang, and M. Zhao, "Learning vision-driven reactive soccer skills for humanoid robots," *arXiv preprint arXiv:2511.03996*, 2025.
- [25] Z. Wang, J. Zhou, and Q. Wu, "Dribble master: Learning agile humanoid dribbling through legged locomotion," *arXiv preprint arXiv:2505.12679*, May 2025. [Online]. Available: <https://arxiv.org/abs/2505.12679>
- [26] F. Vahl, J. Griepenburg, J. Gutsche, J. Gldenstern, and J. Zhang, "Soccerdiffusion: Toward learning end-to-end humanoid robot soccer from gameplay recordings," *arXiv preprint arXiv:2504.20808*, Apr 2025. [Online]. Available: <https://arxiv.org/abs/2504.20808>
- [27] D. Marew, N. Perera, S. Yu, S. Roelker, and D. Kim, "A biomechanics-inspired approach to soccer kicking for humanoid robots," in *Proceedings of the 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids)*, 2024, pp. 722–729. [Online]. Available: <https://arxiv.org/abs/2407.14612>
- [28] D. Kim, S. J. Jorgensen, P. Stone, and L. Sentis, "Dynamic behaviors on the nao robot with closed-loop whole body operational space control," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 1121–1128.
- [29] L. Sentis and O. Khatib, "A whole-body control framework for humanoids operating in human environments," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*. IEEE, 2006, pp. 2641–2648.
- [30] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbab, C. Pan, *et al.*, "Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," *arXiv preprint arXiv:2502.01143*, 2025.
- [31] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang, "Expressive whole-body control for humanoid robots," *arXiv preprint arXiv:2402.16796*, 2024.
- [32] J. Hu, P. Stone, and R. Martn-Martn, "Causal policy gradient for whole-body mobile manipulation," *arXiv preprint arXiv:2305.04866*, 2023.
- [33] D. Ferigo, R. Camoriano, P. M. Viceconte, D. Calandriello, S. Traversaro, L. Rosasco, and D. Pucci, "On the emergence of whole-body strategies from humanoid robot push-recovery learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8561–8568, 2021.
- [34] J. Hu, P. Stone, and R. Martn-Martn, "Slac: Simulation-pretrained latent action space for whole-body real-world rl," *arXiv preprint arXiv:2506.04147*, 2025.
- [35] Y. Ma, A. Cramariuc, F. Farshidian, and M. Hutter, "Learning coordinated badminton skills for legged manipulators," *Science Robotics*, vol. 10, no. 102, p. eadu3922, 2025.
- [36] C. Zhang, W. Xiao, T. He, and G. Shi, "Wococo: Learning whole-body humanoid control with sequential contacts," *arXiv preprint arXiv:2406.06005*, 2024.
- [37] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2811–2817.
- [38] Y. Wang, P. Chen, X. Han, F. Wu, and M. Zhao, "Booster gym: An end-to-end reinforcement learning framework for humanoid robot locomotion," *arXiv preprint arXiv:2506.15132*, 2025.
- [39] U. Robotics, "unitree_rl_gym: A reinforcement learning gym for unitree robots (go2, h1, h1_2, g1)," https://github.com/unitreerobotics/unitree_rl_gym, 2025, BSD-3-Clause License; accessed September 15, 2025.
- [40] RoboCup Humanoid League Technical Committee, "Humanoid league laws of the game 2025: Robocup humanoid league adult-size rules," <https://humanoid.robocup.org/wp-content/uploads/RC-HL-2025-Rules.pdf>, Apr. 2025, accessed: 2025-09-13.
- [41] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang, "Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks," *arXiv preprint arXiv:2506.08931*, 2025.
- [42] Q. Zhang, G. Han, J. Sun, W. Zhao, C. Sun, J. Cao, J. Wang, Y. Guo, and R. Xu, "Distillation-ppo: A novel two-stage reinforcement learning framework for humanoid robot perceptive locomotion," *arXiv preprint arXiv:2503.08299*, 2025.
- [43] B. Robotics, "Booster t1," <https://www.boosterobotics.com/robots/>, Booster Robotics, 2023–2024, "Made for Developers", Lightweight, Flexible, Durable. [Online]. Available: <https://www.boosterobotics.com/robots/>
- [44] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*. IEEE, 2024, pp. 1–6.
- [45] J. Wasserman, A. Agarwal, R. Jangir, G. Chowdhary, D. Pathak, and A. Gupta, "Legolas: Deep leg-inertial odometry," in *8th Annual Conference on Robot Learning*, 2024.