arXiv:2610.40003v1 [cs.RO] 29 Oct 2026

Humanoid Dodgeball: Learning Fast Dynamic Obstacle Avoidance Motions for Humanoids

Aoqian Zhang[1,*], Chunzheng Wang[1,*], Haoran Chen[1,*], Annan Tang[2], Cheng Xiang[3], Fan Shi[1], and Shuzhi Sam Ge[1]

***Abstract*— Humanoids have become increasingly agile and capable in real-world environments. To operate safely in dynamic scenarios, they must rapidly evade fast-moving obstacles while maintaining stability and physically feasible whole-body motion. Learning such behaviors remains challenging due to the high-dimensional dynamics of humanoid systems and the need to jointly balance reactive obstacle avoidance, motion naturalness, and control robustness. In this work, we present *Humanoid Dodgeball*, a reinforcement learning (RL) framework for agile dynamic obstacle avoidance in humanoids. Our method enables highly dynamic whole-body evasive maneuvers against incoming projectiles through threat-aware reward shaping, while multiple direction-specific adversarial motion priors regularize the learned behaviors toward stable and diverse human-like motions. Extensive simulation and hardware experiments demonstrate that the proposed framework substantially improves dynamic obstacle avoidance performance across diverse obstacle speeds and launch distances, while preserving robust and physically plausible humanoid motions. Supplementary videos are available at https://humanoid-dodge.github.io/.**

I. INTRODUCTION

Recent advances in reinforcement learning for humanoids have significantly improved humanoid whole-body motion tracking and control through training on large human motion datasets [1]–[3]. These developments have enabled humanoids to acquire a wide range of dynamic and human-like skills, including athletic sports behaviors such as soccer [4] and tennis [5], highly dynamic parkour locomotion [6], and complex whole-body tasks such as skateboarding [7]. To further improve the safety and robustness of humanoid deployment in complex and human-centered environments, recent research has explored protective falling strategies that reduce damage to both the robot and its surroundings during unavoidable failures [8], as well as collision-free navigation and traversal in cluttered environments with static obstacles [9]. These advances represent important steps toward deploying humanoids in real-world environments where safety, robustness, and physical interaction are critical.

Despite these remarkable advances, humanoids still lack the ability to safely and robustly react to fast-moving dynamic objects. The capability to evade incoming obstacles is a fundamental and natural behavior observed in legged animals, including humans, and is essential for agile operation in real-world settings. However, learning such reactive whole-body avoidance behaviors for humanoids remains highly challenging due to the need for rapid reaction, dynamic balance maintenance, and coordinated whole-body motion generation.

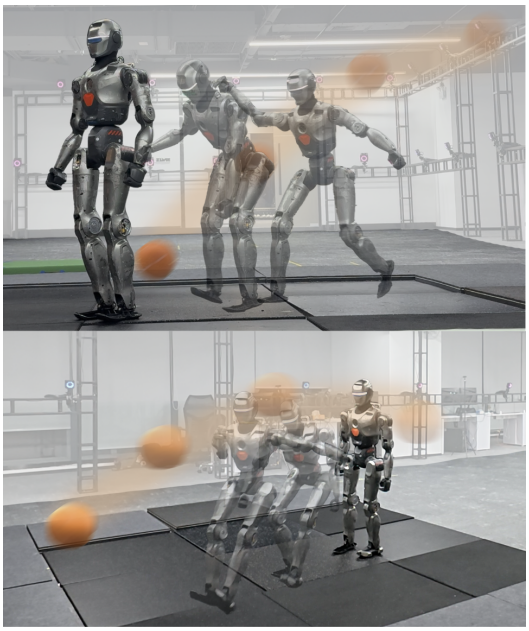


Fig. 1. Our dynamic obstacle avoidance policy deployed on the EngineAI PM01 humanoid. (Top) The robot dodges by stepping forward. (Bottom) The robot dodges by jumping backward.

To address these challenges, we present *Humanoid Dodgeball*, a novel humanoid whole-body control framework for dynamic obstacle avoidance. The main contributions of this work are as follows:

- We design a reinforcement learning framework that enables humanoids to evade fast-moving projectiles through coordinated whole-body evasive maneuvers while maintaining stability.
- We combine threat-aware rewards, direction-specific motion priors, and capture-point guidance to promote effective avoidance, diverse natural motions, and stable recovery.
- We validate our proposed method through simulation experiments on two humanoid platforms and hardware experiments on the EngineAI PM01 humanoid.

II. RELATED WORK

Dynamic obstacle avoidance has been widely studied across different robotic platforms. In aerial robotics, Falanga

*Equal contribution.
[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore.
[2]Shenzhen EngineAI Robotics Technology Co., Ltd., China.
[3]School of Information and Communication Engineering, Dalian University of Technology, China.

et al. proposed an event-camera-based framework for quadrotor obstacle avoidance that leverages the microsecond-level response of event cameras to generate low-latency reactive avoidance behaviors using artificial potential fields [10]. More recently, Fan et al. introduced an end-to-end reinforcement learning framework that uses multi-frame LiDAR observations to directly generate control commands for avoiding high-speed dynamic obstacles [11]. In quadrupedal robots, He et al. proposed Agile But Safe (ABS), a dual-policy reinforcement learning framework that combines agile locomotion and recovery policies through a learned reach-avoid value network for high-speed collision-free navigation [12]. Wang et al. further proposed Omni-Perception, an end-to-end reinforcement learning framework that directly processes raw LiDAR point clouds using a hierarchical perception network for omnidirectional obstacle avoidance [13]. To address the slow response of navigation-based avoidance strategies, Xu et al. proposed the Reflexive Evasion Robot (REBot) framework, which introduces a biologically inspired finite-state-machine architecture for reflexive evasion behaviors [14]. REBot enables rapid obstacle avoidance through shorter reaction times and faster whole-body quadruped maneuvers.

Dynamic obstacle avoidance is significantly more challenging for humanoids than for aerial and quadruped robots due to humanoids' high-dimensional whole-body dynamics, ground contact limitations, and strict balance constraints. When faced with fast-moving incoming obstacles, humanoids cannot simply evade threats through low-dimensional velocity modulation or direct body translation. Instead, they must coordinate each body part and foot placement in a fully integrated manner to perform agile evasive maneuvers while maintaining locomotion stability and physically plausible motion. Concurrently with our work, Yang et al. proposed a perception-aware reinforcement learning framework for humanoid dynamic obstacle avoidance that combines control barrier function (CBF)-inspired reward shaping with adversarial motion priors (AMP)-based motion regularization [15]. However, it considers a limited set of obstacle trajectories, which limits the policy's ability to generalize to more diverse projectile conditions. Compared with existing approaches, our method enables more effective dynamic obstacle avoidance for humanoids by achieving higher avoidance success rates and robust evasive behaviors against obstacles approaching from diverse directions and speeds.

III. METHODS

A. Problem formulation

We formulate dynamic obstacle avoidance for humanoid robots as a discrete-time reinforcement learning problem. We model the task as a Markov decision process (MDP) $\mathcal{M}_{\mathrm{MDP}} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{T}(\boldsymbol{s}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t)$ the transition dynamics induced jointly by the robot dynamics and the obstacle motion, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, and $\gamma \in [0, 1)$ the discount factor. The objective is to learn a policy $\pi_\theta(\boldsymbol{a}_t \mid \boldsymbol{o}_t)$ that maximizes the expected discounted return $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{\infty} \gamma^t r_t]$. In this work, the policy's objective is to avoid dynamic obstacles while maintaining stability without failure. A failure is defined as any non-foot body part of the humanoid making contact with the ground.

Observation space. At each control step t, the policy receives an observation $\boldsymbol{o}_t = [\boldsymbol{s}_t^p, \boldsymbol{s}_t^o]$, comprising proprioceptive and obstacle states. The proprioceptive state $\boldsymbol{s}_t^p = [\boldsymbol{\omega}_t, \phi_t, \theta_t, \boldsymbol{q}_t, \dot{\boldsymbol{q}}_t, \boldsymbol{a}_{t-1}]$ contains the base angular velocity $\boldsymbol{\omega}_t \in \mathbb{R}^3$, the base roll ϕ_t and pitch θ_t, the joint positions $\boldsymbol{q}_t \in \mathbb{R}^N$ and velocities $\dot{\boldsymbol{q}}_t \in \mathbb{R}^N$, and the previous action $\boldsymbol{a}_{t-1} \in \mathbb{R}^N$, where N is the number of actuated joints. The obstacle state $\boldsymbol{s}_t^o = [x_t^o, y_t^o, z_t^o] \in \mathbb{R}^3$ gives the obstacle's 3D position in the robot base frame.

Action space. The policy outputs an action $\boldsymbol{a}_t \in \mathbb{R}^N$ specifying joint position offsets relative to a nominal standing configuration. The desired joint targets are $\boldsymbol{q}_t^{\mathrm{target}} = \alpha\, \boldsymbol{a}_t + \boldsymbol{q}^{\mathrm{default}}$, where α is the action scaling factor and $\boldsymbol{q}^{\mathrm{default}}$ the default standing configuration. These targets are tracked by joint-level proportional–derivative (PD) controllers producing torques $\boldsymbol{\tau}_t = K_p(\boldsymbol{q}_t^{\mathrm{target}} - \boldsymbol{q}_t) - K_d \dot{\boldsymbol{q}}_t$, with stiffness and damping gains K_p and K_d.

B. Threat-aware avoidance policy training

We approximate the dynamic obstacle as a sphere with center position $\boldsymbol{p}_t^o \in \mathbb{R}^3$, velocity $\boldsymbol{v}_t^o \in \mathbb{R}^3$, and radius R^o. The humanoid is represented by $K = 10$ key bodies $\mathcal{B} = \{1, \ldots, K\}$ (the head, torso, and upper and lower links of both arms and legs). Each body i is represented by a sphere with center position $\boldsymbol{p}_t^i \in \mathbb{R}^3$, velocity $\boldsymbol{v}_t^i \in \mathbb{R}^3$, radius R^i, and safety margin $m = 0.5$ m. This representation permits efficient per-body threat estimation using sphere–sphere geometry. Let $R_{\mathrm{safe}}^i = R^o + R^i + m$ be the radius of the safety region.

Predictive threat. We first estimate whether the obstacle will enter the safety region of each body within a prediction horizon $H = 3$ s. Under a constant-velocity approximation, we compute the time of closest approach and its validity mask as

$$\tau_t^{i*} = -\frac{(\Delta \boldsymbol{p}_t^i)^\top \boldsymbol{w}_t^i}{\|\boldsymbol{w}_t^i\|^2}, \qquad M_t^i = \mathbb{I}(0 < \tau_t^{i*} < H), \tag{1}$$

where $\Delta \boldsymbol{p}_t^i = \boldsymbol{p}_t^o - \boldsymbol{p}_t^i$ and $\boldsymbol{w}_t^i = \boldsymbol{v}_t^o - \boldsymbol{v}_t^i$ are the relative position and velocity from body i to the obstacle, respectively. In this formulation, obstacles that are already receding or whose closest approach lies beyond H do not activate the predictive threat. The additional displacement required to keep body i outside the region is

$$\delta_t^i = M_t^i \max(0, R_i^{\mathrm{safe}} - d_t^{i*}), \tag{2}$$

where $d_t^{i*} = \|\Delta \boldsymbol{p}_t^i + \tau_t^{i*} \boldsymbol{w}_t^i\|$ is the distance of closest approach. The required acceleration magnitude and the resulting predictive threat are computed as

$$a_t^{\mathrm{req},i} = \frac{2\delta_t^i}{{\tau_t^{i*}}^2}, \qquad S_t^i = 1 - \exp\left[-\left(\frac{a_t^{\mathrm{req},i}}{a_{\mathrm{ref}}}\right)^2\right], \tag{3}$$

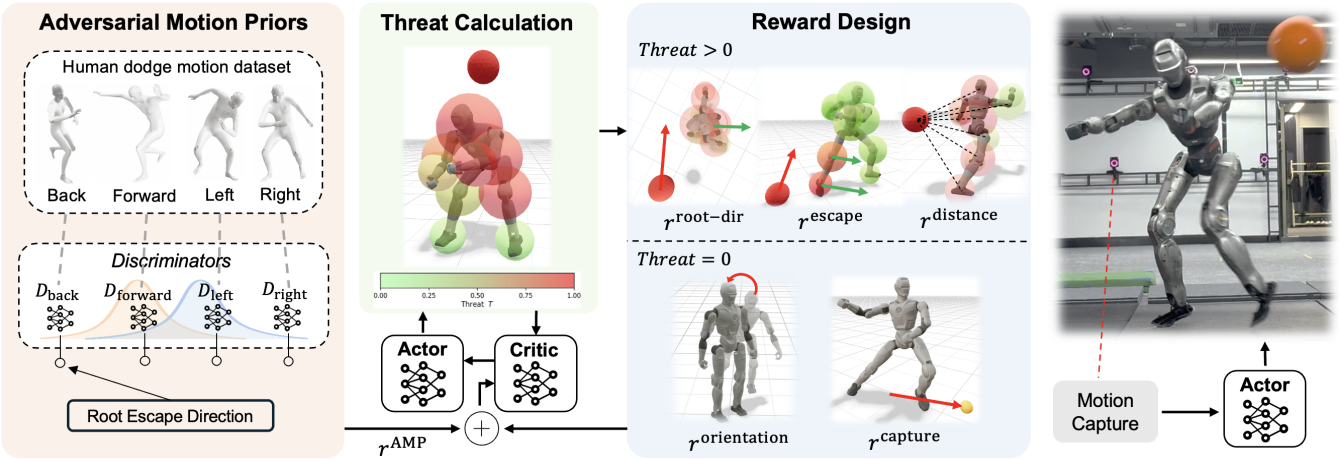


Fig. 2. Overview of *Humanoid Dodgeball*. During simulation training, predictive and near-field threat for key bodies are computed to characterize incoming projectile hazards and guide threat-aware reward shaping for whole-body evasion. Direction-specific adversarial motion priors, selected according to the root escape direction, regularize the learned behaviors toward diverse and natural human-like dodge motions. When the threat subsides, recovery rewards, including orientation and capture-point objectives, encourage rapid stabilization. During hardware deployment, the trained actor is executed using obstacle positions estimated from the motion-capture system.

where $a_{\text{ref}} = 2\ \text{m/s}^2$ is a reference acceleration. Consequently, $S_t^i \in [0,1]$ jointly captures the future collision geometry, remaining reaction time, and required evasive effort.

Near-field threat. The predictive threat is complemented by a bounded near-field threat that depends only on the current surface clearance $c_t^i = \|\Delta \boldsymbol{p}_t^i\| - R_{\text{safe}}^i$. We map this clearance to a bounded near-field threat using a one-sided Gaussian kernel,

$$D_t^i = \exp\left[-\left(\frac{\max(c_t^i, 0)}{\sigma_{\text{near}}}\right)^2\right], \tag{4}$$

where $\sigma_{\text{near}} = 0.5\ \text{m}$ controls the spatial decay. The threat equals one when the obstacle enters the safety region and smoothly decreases as the obstacle moves away.

Threat-aware avoidance reward. To encourage the humanoid to dodge dynamic obstacles, we design rewards based on the predictive and near-field threats. We define the body escape reward as

$$r_t^{\text{escape}} = \sum_{i \in \mathcal{B}} \left[S_t^i \tanh\left(\frac{\max(0, \boldsymbol{v}_t^i \cdot \boldsymbol{e}_t^{\text{pred},i})}{1.5}\right)\right], \tag{5}$$

where

$$\boldsymbol{e}_t^{\text{pred},i} = -\frac{\Delta \boldsymbol{p}_t^i + \tau_t^{i*} \boldsymbol{w}_t^i}{\|\Delta \boldsymbol{p}_t^i + \tau_t^{i*} \boldsymbol{w}_t^i\|}, \tag{6}$$

is the desired body escape direction, which points away from the predicted closest-approach point. The reward r_t^{escape} encourages threatened bodies to move away from the predicted obstacle trajectory, with each body's contribution weighted by S_t^i. We also penalize obstacle proximity to the robot's bodies using the near-field threat measure, with a larger penalty for obstacles within the safety region. The body distance reward is calculated as

$$r_t^{\text{distance}} = \sum_{i \in \mathcal{B}} \begin{cases} -D_t^i, & c_t^i \geq 0, \\ -\left(1 + \dfrac{|c_t^i|}{\sigma_{\text{near}}}\right), & c_t^i < 0. \end{cases} \tag{7}$$

While the body escape and distance rewards encourage dodging behavior, they do not explicitly guide the robot's overall movement. Therefore, we further define a root-body escape-direction reward,

$$r_t^{\text{root-dir}} = \mathbb{I}(S_t^{\text{root}} > 0)\phi\left(\hat{\boldsymbol{v}}_t^{\text{root}} \cdot \boldsymbol{e}_{t,xy}^{\text{pred,root}}\right), \tag{8}$$

where $\hat{\boldsymbol{v}}_t^{\text{root}} = \boldsymbol{v}_{t,xy}^{\text{root}} / \|\boldsymbol{v}_{t,xy}^{\text{root}}\|$ is the root velocity direction in the xy plane. The asymmetric direction-scoring function is

$$\phi(x) = \max(x, 0) - \lambda_{\text{wrong}} \max(-x, 0), \tag{9}$$

where $\lambda_{\text{wrong}} = 5$ penalizes root motion in the wrong direction. Regularization rewards are included to help the robot maintain a stable posture and produce smooth joint-level actions. When the overall threat $T = \max_{i \in \mathcal{B}}[1 - (1 - S_t^i)(1 - D_t^i)]$ is zero (in practice, we use $T < 10^{-3}$ to disregard small nonzero tails in D_t^i), we apply recovery rewards that help the robot return to a stable standing posture. We further include a capture-point reward to encourage stable recovery after evasive maneuvers:

$$r^{\text{capture}} = \exp\left(-\left\|\boldsymbol{p}^{\text{capture}} - \boldsymbol{p}^{\text{supp}}\right\|_2^2 / \sigma_{\text{cp}}^2\right), \tag{10}$$

where $\boldsymbol{p}^{\text{capture}} = \boldsymbol{p}_{xy}^{\text{root}} + \boldsymbol{v}_{xy}^{\text{root}} / \omega_0$ is the instantaneous capture point, $\omega_0 = \sqrt{g / z^{\text{root}}}$ is the natural frequency of the linear inverted pendulum model, g is the gravitational acceleration, and z^{root} is the robot root height. The support position $\boldsymbol{p}^{\text{supp}} = (\boldsymbol{p}_{xy}^L + \boldsymbol{p}_{xy}^R)/2$ is approximated by the midpoint of the left and right feet, and $\sigma_{\text{cp}} = 0.12\ \text{m}$ controls the sensitivity of the reward to the capture-point tracking error.

TABLE I

REWARD FUNCTIONS FOR POLICY TRAINING AND THEIR CORRESPONDING WEIGHTS.

Name	Expression	Weight
Dodge rewards ($T > 0$)		
Body escape	r^{escape} (Eq. 5)	40.0
Body distance	r^{distance} (Eq. 7)	50.0
Root-body escape direction	$r^{\text{root}-\text{dir}}$ (Eq. 8)	60.0
Recovery rewards in the absence of threats ($T = 0$)		
Root velocity	$\exp\left(-\left\|\boldsymbol{v}_{xy}^{\text{root}}\right\|_2^2 / 0.5^2\right)$	10.0
Torso orientation	$\left(\phi^{\text{torso}}\right)^2 + \left(\theta^{\text{torso}}\right)^2$	-20.0
Joint deviation	$\sum_{i=1}^{N}\left(q_i - q_{i,\text{default}}\right)^2$	-40.0
Capture point	r^{capture} (Eq. 10)	10.0
Regularization rewards		
AMP reward	r^{AMP}	10.0
Angular velocity penalty	$\left\|\omega_{xy}^{\text{root}}\right\|^2$	-0.2
Joint velocities	$\|\dot{\boldsymbol{q}}\|_2^2$	-0.001
Joint accelerations	$\|(\dot{\boldsymbol{q}}_t - \dot{\boldsymbol{q}}_{t-1})/dt\|^2$	-2.5e-7
Action rate	$\|(\boldsymbol{a}_t - \boldsymbol{a}_{t-1})/dt\|^2$	-0.1
Action acceleration	$\|\boldsymbol{a}_t - 2\boldsymbol{a}_{t-1} + \boldsymbol{a}_{t-2}\|_2^2$	-0.05
Joint position limit	$\sum_{i=1}^{N}\max\left(\lvert q_i\rvert - 0.9 \cdot q_{i,\text{lim}},\, 0\right)$	-10.0
Joint velocity limit	$\sum_{i=1}^{N}\max\left(\lvert\dot{q}_i\rvert - 0.9 \cdot \dot{q}_{i,\text{lim}},\, 0\right)$	-5.0
Joint torque limit	$\sum_{i=1}^{N}\max\left(\lvert\tau_i\rvert - 0.85 \cdot \tau_{i,\text{lim}},\, 0\right)$	-5.0
Foot slip	$c_{\text{L}}\left\|\boldsymbol{v}_{xy}^{\text{L}}\right\|_2^2 + c_{\text{R}}\left\|\boldsymbol{v}_{xy}^{\text{R}}\right\|_2^2$	-10.0
Foot clearance	$\sum_{i\in\{\text{L,R}\}}(1 - c_i)\ \text{clip}(h_i, 0, 0.25)$	20.0

c_{L} and c_{R} denote the left and right foot contact states, respectively, and $\boldsymbol{v}^{\text{L}}$ and $\boldsymbol{v}^{\text{R}}$ denote the foot velocities. h_i denotes the height of foot i above the ground.

The detailed reward terms and corresponding weights are shown in Table I.

C. Adversarial motion priors

We use adversarial motion priors (AMP) [16] to regularize the learned dodging behaviors toward dynamically feasible and natural human-like motions. We use 90 human evasive motion clips from the dataset in [17], spanning four dodge directions: left, right, backward, and forward. For each direction, multiple motion styles are included to capture diverse realizations of the evasive behavior, such as dodging with an upright torso or combining the dodge with downward motion of the head and torso. All motion clips are retargeted to the robot's morphology using the method in [18].

Discriminator training. We employ four direction-specific AMP discriminators, each associated with one cardinal dodge direction: left, right, forward, or backward. Each discriminator is trained using joint-state transitions from the reference motion dataset corresponding to its associated dodge direction and transitions generated by the policy. At each timestep, the discriminator is selected according to the desired root escape direction $\boldsymbol{e}_{t,xy}^{\text{pred,root}}$. If $\boldsymbol{e}_{t,xy}^{\text{pred,root}} = \mathbf{0}$, one of the four discriminators is selected uniformly at random. Otherwise, $\boldsymbol{e}_{t,xy}^{\text{pred,root}}$ is mapped to the nearest cardinal direction, and the corresponding discriminator is activated. For discriminator D_k, where $k \in \{\text{left}, \text{right}, \text{forward}, \text{backward}\}$, let $\boldsymbol{x}_t = (\boldsymbol{q}_t, \boldsymbol{q}_{t+1})$ denote a joint-state transition. The discriminator is trained using the least-squares adversarial objective

$$\begin{aligned}\min_{D_k} \quad & \mathbb{E}_{\boldsymbol{x}_t \sim d_M^k}\left[\left(D_k(\boldsymbol{x}_t) - 1\right)^2\right] \\ & + \mathbb{E}_{\boldsymbol{x}_t \sim d_\pi^k}\left[\left(D_k(\boldsymbol{x}_t) + 1\right)^2\right],\end{aligned} \tag{11}$$

where d_M^k denotes the reference transition distribution for dodge direction k, and d_π^k denotes the policy transition distribution assigned to discriminator D_k according to the selected root escape direction. The policy receives the AMP reward from the selected discriminator D_k as $r_t^{\text{AMP}} = \max\left[0,\, 1 - 0.25\left(D_k(\boldsymbol{q}_t, \boldsymbol{q}_{t+1}) - 1\right)^2\right]$.

D. Training details

Training environment setup. We train the policies in the mjlab simulator [19] using 4096 parallel environments and evaluate our framework on two humanoid platforms, Unitree G1 and EngineAI PM01. To expose the policies to diverse dynamic obstacles, an obstacle is initialized at a randomly sampled position relative to the robot. Its initial velocity is generated using one of two schemes: 1) the obstacle is directed toward a randomly selected body part with a randomly sampled speed; or 2) the obstacle is aimed toward the robot root in the horizontal plane while an additional vertical velocity component is applied, producing a parabolic projectile trajectory. We further randomize the obstacle radius and the gravitational acceleration acting on the obstacle to increase the diversity of projectile dynamics. The detailed domain randomization settings are summarized in Table II.

Policy network. The policy actor and critic are parameterized by multilayer perceptrons (MLPs) with hidden-layer sizes $[512, 256, 128]$. Both networks receive the current observation $\boldsymbol{o}_t$ together with a 20-step observation history. The history is processed by a temporal encoder consisting of a per-frame linear embedding, followed by a two-layer 1-D temporal convolutional network and a final linear projection that produces a 128-dimensional latent representation. Notably, the obstacle velocity is not provided to the actor; instead, projectile motion is inferred implicitly from the history of obstacle positions. In addition, the critic receives the privileged observation $\boldsymbol{o}_t' = [\boldsymbol{D}_t, \boldsymbol{S}_t, c_{\text{L}}, c_{\text{R}}, \boldsymbol{v}_t, \boldsymbol{v}_t^o, \boldsymbol{p}_t^b, r^o]$, where $\boldsymbol{D}_t$ and $\boldsymbol{S}_t$ denote the near-field and predictive threat values, respectively, for all key bodies; c_{L} and c_{R} indicate the left and right foot contact states; $\boldsymbol{v}_t$ is the robot root velocity; $\boldsymbol{v}_t^o$ is the obstacle velocity; $\boldsymbol{p}_t^b$ contains the positions of the robot's key bodies expressed in the robot base frame; and r^o denotes the obstacle radius. We use the proximal policy optimization (PPO) algorithm for policy optimization [20]. The policy is trained for 30,000 iterations.

IV. SIMULATION EXPERIMENTS

In this section, we present simulation results, analyze our method, and compare it with baseline methods. Specifically, we answer the following questions: **Q1** Does our method achieve better avoidance performance than the baselines across a wide range of dynamic obstacle states? **Q2** Does

TABLE II

DOMAIN RANDOMIZATION PARAMETERS. ALL QUANTITIES ARE SAMPLED UNIFORMLY WITHIN THE SPECIFIED RANGES.

Category	Term	Range
Dynamics	Friction coefficient	$[0.1, 2.0]$
	Joint K_p, K_d scaling	$[0.8, 1.2]\times$
	Joint position bias	$[-0.05, 0.05]$ rad
	Torso COM offset	$[-0.05, 0.05]$ m
Sensor noise	Base angular velocity $\boldsymbol{\omega}$	$[-0.2, 0.2]$ rad/s
	Base roll/pitch (ϕ, θ)	$[-0.25, 0.25]$ rad
	Joint positions $\boldsymbol{q}$	$[-0.02, 0.02]$ rad
	Joint velocities $\dot{\boldsymbol{q}}$	$[-1.5, 1.5]$ rad/s
	Previous action $\boldsymbol{a}_{t-1}$	$[-0.2, 0.2]$
	Obstacle position $\boldsymbol{s}^o$	$[-0.2, 0.2]$ m
Disturbances	Linear velocity push $\boldsymbol{v}_{xy}^{\text{push}}$	$[-1.0, 1.0]$ m/s
	Roll/pitch velocity push $\boldsymbol{\omega}_{\phi\theta}^{\text{push}}$	$[-0.6, 0.6]$ rad/s
	Yaw velocity push $\omega_{\psi}^{\text{push}}$	$[-0.8, 0.8]$ rad/s
Obstacle initial state	Horizontal position (x^o, y^o)	$[-6, 6]$ m
	Height z^o	$[0, 1.8]$ m
	Launch speed $\|\boldsymbol{v}_{xy}^o\|$	$[2, 16]$ m/s
	Launch speed $\|\boldsymbol{v}_{z}^o\|$	$[1, 6]$ m/s
	Obstacle radius R^o	$[0.05, 0.125]$ m
	Gravity scale	$[0.0, 1.0]g$

our policy exhibit more diverse avoidance motions, and does this diversity lead to better obstacle avoidance performance? **Q3** How do our reward design and the use of multiple AMP discriminators affect performance?

A. Metrics and baselines

We consider the following performance metrics:

1) Success rate (%): $SR = \frac{N_{\text{success}}}{N_{\text{obstacle}}} \times 100\%$, where N_{success} is the number of times the robot dodges the obstacle and maintains stability without falling, and N_{obstacle} is the total number of dynamic obstacle occurrences.

2) Recovery time (s): $T_{\text{recovery}} = t_{\text{stable}} - t_{\text{onset}}$, where t_{onset} denotes the timestep at which the robot initiates an evasive maneuver, identified when the root velocity satisfies $\|\boldsymbol{v}^{\text{root}}\| > 0.15$ m/s. The timestep t_{stable} is defined as the beginning of the first continuous 1-s interval after the threat becomes zero during which $\|\boldsymbol{v}^{\text{root}}\| < 0.15$ m/s, $|\theta^{\text{torso}}| < 0.2$ rad, and $|\phi^{\text{torso}}| < 0.2$ rad are all satisfied.

3) Average power (W): The average motor power consumed by all actuated joints during a successful evasive maneuver is computed as

$$P_{\text{motor}} = \frac{1}{t_{\text{stable}} - t_{\text{onset}}} \int_{t_{\text{onset}}}^{t_{\text{stable}}} \sum_i |\tau_i(t)\dot{q}_i(t)| \, dt.$$

We compare our approach against two whole-body dynamic obstacle avoidance baselines:

1) REBot [14]: A reinforcement learning framework for reflexive dynamic obstacle avoidance in quadruped robots using curriculum learning and policy switching. We reimplement its training pipeline for humanoid robots.

2) PAC-MAN [15]: Developed concurrently with our work, it presents a perception-aware reinforcement learning framework for humanoid dynamic obstacle avoidance that combines CBF-inspired reward shaping with AMP-based motion regularization. For a fair comparison, we retrain PAC-MAN using our human motion dataset and use the state-oracle perception option in which the ground-truth obstacle states are observable to the actor.

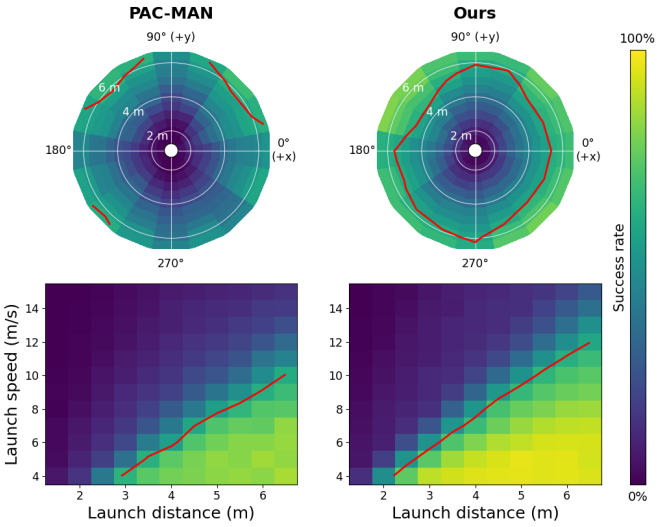


Fig. 3. Top: Polar heatmaps of obstacle avoidance success rates across different launch distances and approach directions relative to the robot's heading. Bottom: Success-rate heatmaps across different obstacle launch speeds and distances. The red contour indicates the 60% success-rate boundary.

For a fair comparison, both baselines are trained using the same obstacle initialization settings as our method.

B. Evaluation of avoidance performance

To evaluate avoidance performance across a broad range of obstacle states, we systematically vary the obstacle launch speed $v^o \in [4, 5, \ldots, 15]$ m/s, the approach direction relative to the robot heading $\theta^o \in [0, \frac{\pi}{8}, \ldots, \frac{15\pi}{8}]$ rad, and the horizontal launch distance $d^o \in [1.0, 1.5, \ldots, 6.5]$ m. This results in $12 \times 16 \times 12 = 2304$ distinct obstacle configurations. Each configuration is evaluated over 100 trials. In every trial, the targeted body part is sampled uniformly from the robot's key bodies, while the obstacle gravity scale is independently randomized within $[0, 1]g$, covering trajectories from straight-line motion to fully ballistic flight.

We compute the success rate (SR) for each obstacle configuration. Fig. 3 (top) reports SR averaged over launch speeds v^o for each launch distance-direction pair (d^o, θ^o). PAC-MAN exhibits pronounced directional sensitivity, with noticeable performance degradation for several frontal, rear, and lateral approach directions. In contrast, our method maintains a consistently higher success rate over nearly the entire angular range, indicating stronger robustness to obstacle approach direction and more uniform whole-body evasive capability. Fig. 3 (bottom) reports SR averaged over approach directions θ^o for each launch speed-distance pair (v^o, d^o). As expected, both methods become less successful as the obstacle speed increases or the launch distance decreases, reducing the available reaction time. However, our method consistently achieves higher success rates across the speed-distance space. The red contour denotes the 60% success-rate boundary. Compared with PAC-MAN, our method shifts this boundary toward substantially more challenging obstacle

TABLE III

DODGE PERFORMANCE UNDER DIFFERENT TIME-TO-HIT (TTH) RANGES IN SIMULATION. RECOVERY TIME AND AVERAGE MOTOR POWER ARE REPORTED AS MEAN ± STANDARD DEVIATION. BOLD INDICATES THE BEST RESULT FOR EACH ROBOT AND TTH RANGE.

		Unitree G1			EngineAI PM01		
TTH (s)	**Method**	$\boldsymbol{SR}$ (%)	$\boldsymbol{T}_{\mathbf{recovery}}$ (s)	$\boldsymbol{P}_{\mathbf{motor}}$ (W)	$\boldsymbol{SR}$ (%)	$\boldsymbol{T}_{\mathbf{recovery}}$ (s)	$\boldsymbol{P}_{\mathbf{motor}}$ (W)
< 0.4	Ours	**11.8**	1.194 ± 0.356	569.6 ± 124.5	**14.2**	$\mathbf{1.197 \pm 0.439}$	586.3 ± 145.9
	REBot [14]	4.8	1.772 ± 0.468	645.3 ± 93.5	8.1	2.293 ± 0.972	830.9 ± 266.6
	PAC-MAN [15]	5.9	$\mathbf{0.825 \pm 0.467}$	$\mathbf{316.7 \pm 77.2}$	3.9	1.788 ± 1.382	$\mathbf{335.9 \pm 139.7}$
0.4–0.8	Ours	**64.6**	1.438 ± 0.900	587.7 ± 105.7	**55.5**	1.369 ± 0.666	590.0 ± 129.3
	REBot [14]	10.0	1.735 ± 0.443	618.7 ± 86.1	9.8	2.162 ± 0.575	834.1 ± 228.0
	PAC-MAN [15]	36.8	$\mathbf{0.570 \pm 0.337}$	$\mathbf{319.8 \pm 61.0}$	33.6	$\mathbf{1.097 \pm 0.939}$	$\mathbf{396.8 \pm 104.8}$
0.8–1.2	Ours	**95.1**	1.119 ± 0.300	507.9 ± 95.8	**94.2**	1.298 ± 0.343	580.3 ± 117.2
	REBot [14]	34.0	1.826 ± 0.396	553.0 ± 67.6	17.3	2.416 ± 0.813	777.3 ± 182.9
	PAC-MAN [15]	79.2	$\mathbf{0.244 \pm 0.210}$	$\mathbf{312.8 \pm 58.5}$	79.5	$\mathbf{0.589 \pm 0.319}$	$\mathbf{395.3 \pm 76.8}$
≥ 1.2	Ours	**95.6**	1.190 ± 0.363	399.3 ± 71.2	**97.2**	1.303 ± 0.341	486.0 ± 104.8
	REBot [14]	53.1	2.008 ± 0.565	522.4 ± 102.7	43.9	2.190 ± 0.825	752.8 ± 174.4
	PAC-MAN [15]	81.9	$\mathbf{0.262 \pm 0.192}$	$\mathbf{305.2 \pm 58.7}$	86.6	$\mathbf{0.237 \pm 0.150}$	$\mathbf{386.0 \pm 73.7}$

configurations, demonstrating improved tolerance to faster projectiles and shorter reaction distances.

We group the obstacle launch states into four nominal time-to-hit (TTH) categories, where $\text{TTH} = d^o/v^o$, and summarize the results in Table III. Our method consistently achieves the highest success rate (SR) across all TTH ranges and on both humanoid platforms. The advantage becomes particularly pronounced as more reaction time becomes available. This indicates that the learned policy effectively exploits additional reaction time and generalizes consistently across two humanoid platforms. REBot performs substantially worse, reaching only around 50% for $\text{TTH} \geq 1.2$ s, while its success rate drops rapidly as TTH decreases. REBot also generally exhibits longer recovery times and considerably higher motor power consumption, particularly on PM01, suggesting that its reflexive evasion training pipeline, originally developed for quadrupeds without AMP-based motion regularization, transfers poorly to humanoid whole-body avoidance. PAC-MAN performs significantly better than REBot and has the lowest average mechanical power and generally the shortest recovery time. However, its avoidance success remains below ours in every TTH range. Overall, our method provides the strongest obstacle avoidance performance at the cost of higher motor power consumption and slower post-evasion recovery than PAC-MAN.

Fig. 4 visualizes the collision frequency of individual body parts for the G1 and PM01 humanoids during evaluation. For PAC-MAN, collisions are concentrated primarily around the head, torso, and lower limbs, while the pelvis and upper legs experience comparatively fewer contacts. In contrast, our method substantially reduces collision frequency across most upper-body regions, with the remaining contacts occurring mainly around the lower legs and feet. This concentration is likely due to the functional constraints of the lower limbs during dynamic evasion: the feet must prioritize support and balance maintenance, leaving less freedom for aggressive collision-avoidance motions than the upper body.

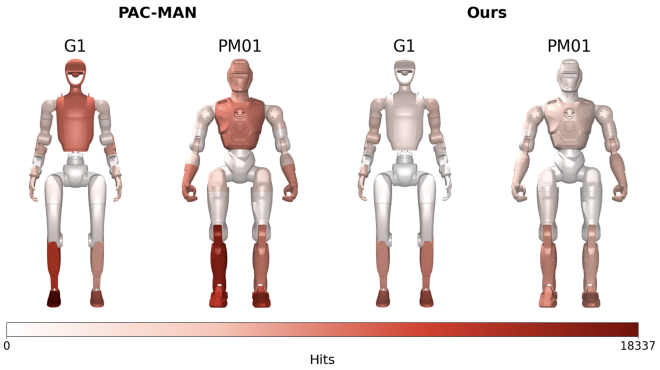


Fig. 4. Visualization of obstacle collision frequency on the G1 and PM01 humanoids during the simulation evaluation.

C. Analysis of avoidance motions

To characterize the diversity of learned evasive behaviors, we collect motion windows generated during evaluation across a wide range of obstacle configurations. Each window contains the 20 most recent frames, where each frame is represented by $\boldsymbol{f}_t = [\boldsymbol{q}_t, \boldsymbol{v}_t^{\text{root},h}, \boldsymbol{\omega}_t^{\text{root}}, z_t^{\text{root}}]$. Each motion window is labeled according to the robot's movement direction during the window. Fig. 5 shows the t-SNE embeddings of 10000 motion windows randomly sampled during evaluation for our method, our method with a single AMP discriminator, and PAC-MAN. Our method exhibits the broadest and most balanced distribution across forward, backward, leftward, and rightward evasive motions, indicating greater behavioral diversity. In contrast, PAC-MAN is strongly dominated by backward dodging, with comparatively few forward and lateral motions. Using only a single AMP discriminator also leads to clear mode imbalance, particularly with very limited forward-dodge behaviors. Introducing direction-specific AMP discriminators produces more distinct and better represented motion modes across all four dodge directions, demonstrating their effectiveness in promoting diverse evasive behaviors.

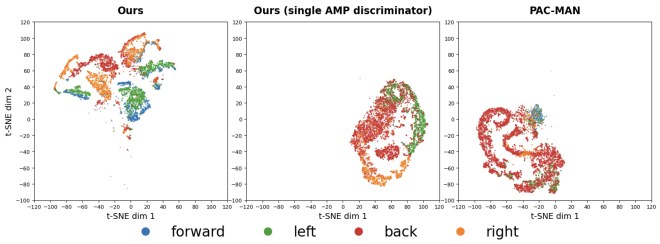


Fig. 5. t-SNE visualization of avoidance motions from the evaluation rollouts.

D. Ablation studies

To assess the contributions of individual reward terms and direction-specific motion priors, we compare the full framework with the following five ablated variants: **1) w/o body escape reward** r^{escape}, **2) w/o body distance reward** r^{distance}, **3) w/o root-direction reward** $r^{\text{root-dir}}$, **4) w/o capture-point reward** r^{capture}, and **5) single AMP discriminator**, which replaces the four direction-specific discriminators with one shared discriminator to evaluate the benefit of direction-specific motion regularization.

Fig. 6(a) evaluates the contribution of the individual reward terms and the multi-discriminator AMP design to obstacle avoidance performance. Removing r^{distance} consistently reduces the success rate, with the largest degradation observed for short reaction times, indicating that the near-field distance penalty is important for maintaining sufficient clearance from rapidly approaching obstacles. Removing r^{escape} also leads to a substantial performance drop for $\text{TTH} < 1.4$ s, particularly in the more challenging short-TTH regime, demonstrating that the escape-velocity reward is critical for eliciting timely evasive motions. A similar trend is observed when removing $r^{\text{root-dir}}$: the success rate decreases noticeably for $\text{TTH} < 1.0$ s, showing that the root-direction reward helps resolve conflicting local body-part escape directions and promotes coherent whole-body displacement under imminent threats. Replacing the direction-specific AMP discriminators with a single discriminator causes a consistent reduction in success rate, suggesting that direction-specific motion priors improve the effectiveness of the learned evasive behaviors. This result further suggests that promoting a more diverse repertoire of avoidance motions contributes to better obstacle avoidance performance, consistent with the findings of [14], which show that greater gait diversity improves dynamic obstacle avoidance.

Fig. 6(b) further examines the effect of stabilization and motion-prior design on fall rate. Removing r^{capture} substantially increases the probability of falling across all TTH ranges, confirming that the capture-point reward is important for recovering balance after highly dynamic evasive maneuvers. Using a single AMP discriminator leads to an even higher fall rate, indicating that direction-specific AMP regularization not only improves avoidance performance but also contributes to more dynamically feasible and stable motions. In contrast, the full system maintains a consistently low fall rate across all TTH ranges.

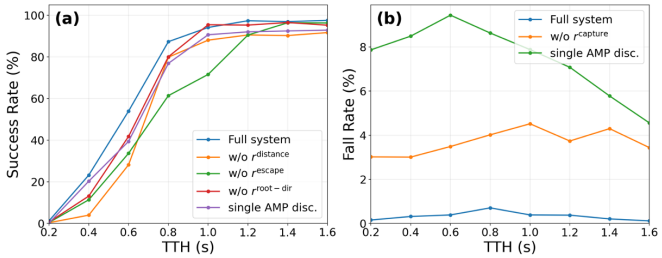


Fig. 6. Ablation studies under different time-to-hit (TTH) conditions. (a) Success rate after removing individual reward terms or replacing the direction-specific AMP discriminators with a single discriminator. (b) Fall rate after removing the capture-point recovery reward or using a single AMP discriminator.

V. Hardware Deployment

We deploy the trained policy on the EngineAI PM01 humanoid to evaluate its performance under real-world conditions. A motion capture system tracks the robot and the incoming obstacle. The obstacle position is transformed into the robot's coordinate frame and provided to the policy. We evaluate obstacle avoidance by throwing a ball with a radius of 0.12 m toward the robot from different directions and at different speeds. The ball speed recorded by the motion capture system ranges from 3.0 m/s to 10 m/s.

As shown in Fig. 7, the robot responds with diverse whole-body evasive maneuvers, including forward and lateral steps combined with ducking. Out of the 30 trials, the robot successfully dodges 22 times, achieving an overall success rate of approximately 73%. Out of the 8 failure cases, the robot gets hit by the ball 6 times and falls to the ground while dodging twice. Most hit failures occur when fast projectiles (typically $> 8\,\text{m/s}$) are directed toward the torso; although the threat successfully triggers an evasive response, the remaining reaction time is sometimes insufficient to avoid contact. Another failure mode occurs when a fast-rolling ball approaches the feet: although the robot steps laterally to evade it, the swinging foot inadvertently contacts the ball. These results demonstrate that the learned policy transfers to hardware and generates coordinated evasive behaviors in response to incoming projectiles, with a modest performance gap consistent with expected sim-to-real differences.

VI. Conclusion and Future Work

We presented *Humanoid Dodgeball*, a reinforcement learning framework for reactive whole-body avoidance of fast-moving projectiles. By combining threat-aware reward shaping, direction-specific adversarial motion priors, and capture-point guidance, the framework learns diverse evasive maneuvers while promoting balance and post-evasion recovery. Simulation experiments on Unitree G1 and EngineAI PM01 demonstrate higher avoidance success rates than the evaluated baselines across all tested time-to-hit ranges. Ablation studies further support the contributions of the reward design and direction-specific motion priors to avoidance performance and stability. Hardware experiments on PM01 demonstrate coordinated avoidance behaviors.

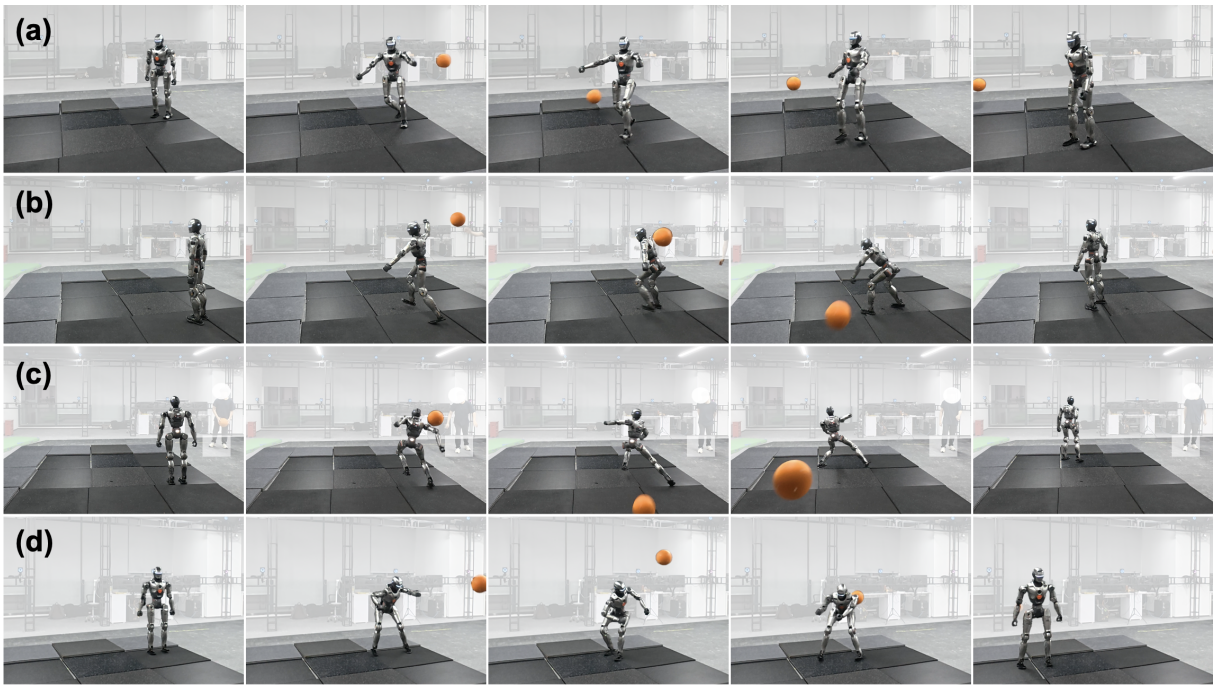


Fig. 7. Hardware demonstrations on the EngineAI PM01 humanoid under diverse projectile conditions. (a) A ball aimed toward the robot's center is avoided by stepping forward. (b) A projectile directed toward the head is avoided through a coordinated forward step and ducking motion. (c) A ball aimed at the right shoulder is avoided by stepping laterally to the left. (d) A projectile targeting the head is avoided through a rightward step combined with a ducking maneuver. These examples demonstrate that the learned policy produces diverse whole-body evasive strategies in response to different obstacle trajectories and incoming directions.

Despite these results, avoidance remains challenging under short reaction times, and improved success comes at the cost of higher motor power and longer recovery times than PAC-MAN. Future work will explore policies that better balance dodge success rate and power consumption, replace motion capture with onboard perception, and reduce the sim-to-real gap through improved actuator modeling. We also plan to extend the framework to handle multiple obstacles and integrate avoidance with locomotion and manipulation policies.

REFERENCES

[1] T. He, J. Gao, W. Xiao, Y. Zhang, Z. Wang, J. Wang, Z. Luo, G. He, N. Sobanbabu, C. Pan, Z. Yi, G. Qu, K. Kitani, J. K. Hodgins, L. Fan, Y. Zhu, C. Liu, and G. Shi, "ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills," in *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, Jun. 2025.

[2] Z. Luo, Y. Yuan, T. Wang *et al.*, "Sonic: Supersizing motion tracking for natural humanoid whole-body control," *arXiv preprint arXiv:2511.07820*, 2025.

[3] Y. Ze, S. Zhao, W. Wang *et al.*, "Twist2: Scalable, portable, and holistic humanoid data collection system," in *Proceedings of the 2026 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2026, originally released as arXiv preprint arXiv:2511.02832.

[4] 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," *Science Robotics*, vol. 11, no. 117, p. eaed1152, 2026.

[5] T. Huang, R. Liu, X. Tang, X. Zhang, J. Ren, H. Wang, F. Jia, Y. Qi, K. Yin, W. Zeng, L. Chen, X. Li, T. Wu, K. Li, R. Dai, J. Wang, L. Han, and J. Pang, "Towards professional tennis styles for humanoid robots with adaptive motion planning and tracking," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2026.

[6] Z. Wu, X. Huang, L. Yang, Y. Zhang, K. Sreenath, X. Chen, P. Abbeel *et al.*, "Perceptive humanoid parkour: Chaining dynamic human skills via motion matching," in *Proceedings of Robotics: Science and Systems (RSS)*, 2026, originally released as arXiv preprint arXiv:2602.15827.

[7] J. Han, D. Wang, C. Zhang *et al.*, "Husky: Humanoid skateboarding system via physics-aware whole-body control," in *Proceedings of Robotics: Science and Systems (RSS)*, 2026, originally released as arXiv preprint arXiv:2602.03205.

[8] Z. Meng, T. Liu, L. Ma, Y. Wu, R. Song, W. Zhang, and S. Huang, "Safefall: Learning protective control for humanoid robots," *arXiv preprint arXiv:2511.18509*, 2025.

[9] H. Xue, S. Liang, Z. Zhang, Z. Zeng, Y. Liu, Y. Lian, J. Wang, Q. Liu, X. Shi, and L. Yi, "Collision-free humanoid traversal in cluttered indoor scenes," *arXiv preprint arXiv:2601.16035*, 2026.

[10] D. Falanga, K. Kleber, and D. Scaramuzza, "Dynamic obstacle avoidance for quadrotors with event cameras," *Science Robotics*, vol. 5, no. 40, p. eaaz9712, 2020.

[11] X. Fan, M. Lu, B. Xu, and P. Lu, "Flying in highly dynamic environments with end-to-end learning approach," *IEEE Robotics and Automation Letters*, 2025.

[12] T. He, C. Zhang, W. Xiao, G. He, C. Liu, and G. Shi, "Agile but safe: Learning collision-free high-speed legged locomotion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2024, originally released as arXiv preprint arXiv:2401.17583.

[13] Z. Wang, T. Ma, Y. Jia, X. Yang, J. Zhou, W. Ouyang, Q. Zhang, and J. Liang, "Omni-perception: Omnidirectional collision avoidance for legged locomotion in dynamic environments," in *Proceedings of the Conference on Robot Learning (CoRL)*, 2025, originally released as arXiv preprint arXiv:2505.19214.

[14] Z. Xu, C. Hao, C. Wang, K. Sima, F. Shi, and J. S. Dong, "Rebot: Reflexive evasion robot for instantaneous dynamic obstacle avoidance," *IEEE Robotics and Automation Letters*, vol. 11, no. 2, pp. 1746–1753, 2025.

[15] L. Yang, J. Li, and A. D. Ames, "Pac-man: Perception-aware cbf-rl for whole-body safety in humanoid dodgeball," *arXiv preprint arXiv:2607.28623*, 2026.

[16] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa, "Amp: Adversarial motion priors for stylized physics-based character con-

trol," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–20, 2021.

[17] Bones Studio, "BONES-SEED: Skeletal everyday embodiment dataset," https://huggingface.co/datasets/bones-studio/seed, 2026, human motion capture dataset.

[18] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[19] K. Zakka, Q. Liao, B. Yi, L. Le Lay, K. Sreenath, and P. Abbeel, "mjlab: A lightweight framework for GPU-accelerated robot learning," *arXiv preprint arXiv:2601.22074*, 2026. [Online]. Available: https://arxiv.org/abs/2601.22074

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.