HuB: Learning Extreme Humanoid Balance

Anonymous Author(s)
Affiliation
Address
email

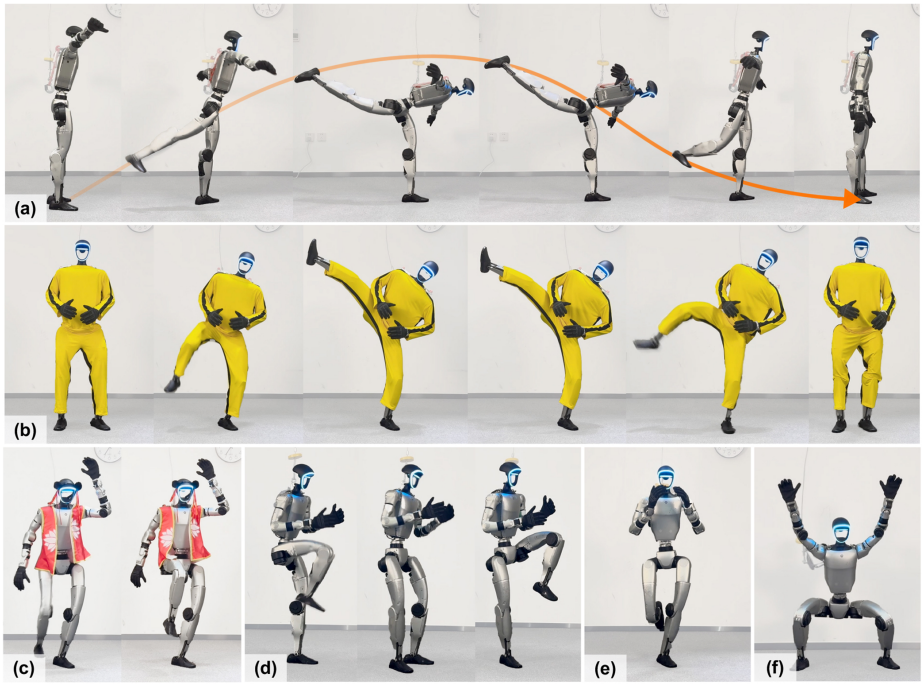

Figure 1: **Extreme Balance Tasks.** HuB enables humanoids to perform extreme quasi-static balance tasks with high stability. (a) `Swallow Balance`: holding a challenging T-shaped pose with the torso extended horizontally; (b) `Bruce Lee's Kick`: executing a high kick with full leg extension while balancing on one foot; (c) `Ne Zha Pose`: a martial arts-inspired one-legged stance with a raised arm; (d) `High Knees`; (e) `Single-Leg Stand`; (f) `Deep Squat`. Videos are available at: hub-corl.github.io

Abstract: The human body demonstrates exceptional motor capabilities—such as standing steadily on one foot or performing a high kick with the leg raised over 1.5 meters—both requiring precise balance control. While recent research on humanoid control has leveraged reinforcement learning to track human motions for skill acquisition, applying this paradigm to balance-intensive tasks remains challenging. In this work, we identify three key obstacles: instability from reference motion errors, learning difficulties due to morphological mismatch, and the sim-to-real gap caused by sensor noise and unmodeled dynamics. To address these challenges, we propose **HuB** (**Hu**manoid **B**alance), a unified framework that integrates *reference motion refinement*, *balance-aware policy learning*, and *sim-to-real robustness training*, with each component targeting a specific challenge. We validate our approach on the Unitree G1 humanoid robot across challenging quasi-static balance tasks, including extreme single-legged poses such as `Swallow Balance` and `Bruce Lee's Kick`. Our policy remains stable even under strong physical disturbances—such as a forceful soccer strike—while baseline methods consistently fail to complete these tasks. Project website: hub-corl.github.io.

Keywords: Humanoid Whole-body Control, Balance Control, Reinforcement Learning

Submitted to the 9th Conference on Robot Learning (CoRL 2025). Do not distribute.

1 Introduction

Developing humanoid robots that can emulate the versatility, agility, and robustness of human movement in complex, unstructured environments has long been a central pursuit in robotics research [1, 2, 3, 4, 5, 6]. Achieving this vision requires not only the ability to execute diverse motor skills, but also the capacity to maintain balance under challenging conditions. Studies in neuroscience and motor control suggest that human balance relies on intricate sensorimotor loops involving the vestibular system, proprioception, and high-level planning [7, 8], making it a particularly demanding aspect of motor control to replicate in robotics. This difficulty is exemplified by the `Swallow Balance` task shown in Figure 1, in which a humanoid must maintain stability in an extreme single-legged pose with the upper body extended horizontally. Such movements require full-body coordination, precise control of the center of mass, and robustness to perturbations—highlighting the demanding nature of humanoid balance.

In recent work on learning-based humanoid control [4, 5, 9, 10, 11, 12, 6], a common approach for enabling humanoids to perform diverse motions is to train a control policy to track reference poses. The standard pipeline typically begins by obtaining human poses either from video-based motion capture algorithms [13, 14] or marker-based motion capture systems. These poses are then retargeted to humanoid-specific reference motions. Next, a control policy is trained in simulation to track these reference motions, and finally, the trained policy is deployed to real-world hardware. However, this pipeline faces significant challenges when applied to complex balancing tasks. In the following, we identify these challenges and present our proposed solutions to address them.

Challenges due to Reference Motion Errors. For tracking-based methods, the successful execution of high-precision balancing critically depends on the accuracy of the reference motion. However, video-based motion capture algorithms [13] often introduce significant errors, and although marker-based motion capture systems offer better precision, they are impractical for Internet videos or consumer-grade recordings. Moreover, optimization-based retargeting [4] can further degrade reference quality due to non-convex optimization, imperfect model alignment, and a lack of temporal continuity constraints. These issues can lead to artifacts such as foot sliding even during stationary phases, which cannot be tolerated in demanding balance tasks. These inaccuracies pose substantial challenges for humanoids attempting to perform complex balancing tasks. To address this, we develop a pipeline that leverages carefully designed initialization to accelerate retargeting convergence and incorporates post-processing techniques to enhance physical plausibility and transition stability.

Challenges in Training Balance Policies. Even with relatively accurate reference data, training balance policies still presents significant challenges. Due to morphological differences between the human body and the humanoid, their centers of mass (COM) do not necessarily align. As a result, strictly tracking the reference motion does not always lead to stable equilibrium for the humanoid. To address this, we relax the policy's tracking objective, allowing it to explore more stable behaviors near the reference trajectory. In addition, to regulate the policy's motion and encourage physically plausible behavior, we introduce a set of shaping rewards. These design choices enable the policy to discover balance strategies better suited to the humanoid's own dynamics.

Challenges in Sim-to-Real Transfer. The sim-to-real gap is a fundamental challenge in simulation-based robot learning, and becomes particularly problematic in complex balance tasks. In the real world, sensors—especially IMUs and visual-inertial odometry (VIO) systems—are often noisy, which leads to inaccurate policy inputs. This subsequently causes jitter in the action outputs and can trigger a vicious feedback loop of instability. Moreover, our experiments show that prior tracking-based methods [4, 9] often cause humanoids to wobble or jitter during real-world balance tasks. These phenomena primarily arise from modeling discrepancies between simulation and reality, particularly in the simulation of ground contact and frictional interactions. To improve robustness under the sim-to-real gap, we adopt localized reference tracking to eliminate VIO dependence, introduce IMU-centric observation perturbation to model realistic sensor noise, and apply high-frequency external pushes to approximate real-world jitter effects, thereby bridging the sim-to-real gap and enhancing deployment robustness.

The strategies outlined above constitute a comprehensive framework for addressing the inherently complex challenge of balance maintenance in humanoid robots. We validated HuB on the Unitree G1 humanoid robot, and the experiments demonstrate that our method enables the robot to perform highly challenging balance tasks, such as `Swallow Balance` and `Bruce Lee's Kick`, as illustrated in Figure 1. In contrast, tracking-based baselines consistently fail to accomplish these tasks, either losing balance and falling, or abandoning single-leg motions. Ablation studies further validate the necessity of each component of our approach. Furthermore, HuB exhibits rapid adaptation and strong robustness against external disturbances, such as a forceful strike from a soccer ball, and enables humanoids to successfully complete 10 consecutive executions within a single continuous rollout without any intervention or resets.

2 Related Work

Humanoid Balance Control. Maintaining balance is a fundamental capability for humanoids. Classical approaches typically adopt model-based control, including feedback-based [15, 16, 17, 18] and optimization-based methods [19, 20, 21]. While effective in structured settings, they often rely on accurate dynamics modeling and struggle under uncertainty. More recently, learning-based approaches have leveraged reinforcement learning (RL) for dynamic stepping [22, 6], push recovery [23], standing-up motions [24, 25, 26], and balancing with uncertain contacts [27]. However, prior works primarily focus on locomotion or transient stabilization rather than sustained quasi-static balance under extreme conditions. In contrast, we address the challenge of sustained balancing, requiring precise whole-body coordination and strong disturbance resilience.

Learning-based Humanoid Control. Recent years have seen rapid progress in learning-based methods for humanoid control. These approaches have demonstrated impressive success in humanoid locomotion [28, 29, 30, 31, 32, 33, 34, 35, 36, 37]. More recent work expands beyond basic walking to include agile and expressive behaviors such as running [38, 39], jumping [40, 6], dancing [11, 12], parkour [41, 42], and loco-manipulation [4, 5, 9, 43]. Despite achieving diverse whole-body motion, most focus on dynamic stabilization and do not address the precise balance control required for quasi-static poses. In contrast, our work introduces a balance-centric learning framework that emphasizes sustained stability in extreme configurations such as single-leg support.

Sim-to-Real Transfer in Robot Learning. Transferring policies from simulation to the real world remains a fundamental challenge. Common approaches include system identification, which calibrates simulation parameters using real-world data [44, 45, 46, 47, 48, 49, 50, 51]; real-to-sim feedback, which corrects simulator behavior by incorporating real-world observations or learned residuals [52, 53, 54, 6, 55]; and domain randomization, which enhances robustness by training over a distribution of randomized dynamics and sensory conditions [56, 57, 58, 59, 60]. While these strategies have shown success in locomotion and manipulation, their effectiveness in humanoid quasi-static balance remains underexplored, where even small sensor or contact inconsistencies can lead to instability. Building upon domain randomization, our method introduces balance-specific perturbations to improve real-world robustness.

3 Learning Framework for Extreme Humanoid Balance

As discussed in Section 1, we identify three key challenges in learning extreme quasi-static balance tasks for humanoids. In this section, we first introduce the necessary background, then present a detailed description of the components of our proposed framework to address these challenges. An overview of the challenges and their corresponding solutions is illustrated in Figure 2.

3.1 Background

Problem Formulation. Our balance learning framework adopts a tracking-based control paradigm, and we formulate the balance task as a goal-conditioned RL problem, modeled as a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $\mathcal{T}$ the

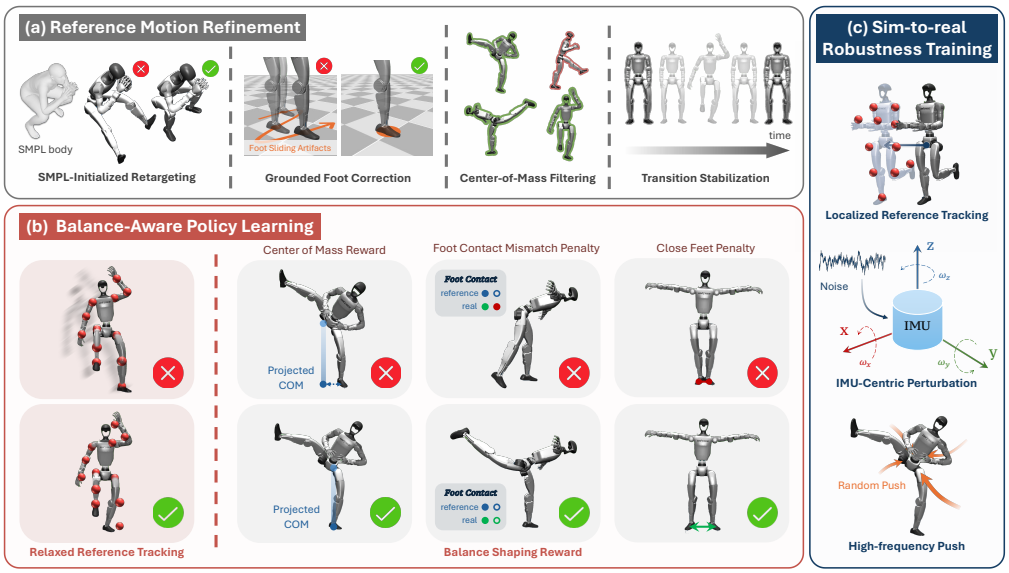

Figure 2: **HuB Overview.** To tackle the challenges of extreme balance tasks on humanoids, HuB integrates three components: (a) a motion refinement process that improves the quality and feasibility of reference motions; (b) a balance-aware policy learning strategy that enables stable execution of challenging balance motions; and (c) a robustness training mechanism to improve sim-to-real consistency and deployment stability.

transition dynamics, $\mathcal{R}$ the reward function, and γ the discount factor. Each state $\boldsymbol{s}_t \in \mathcal{S}$ includes the agent's proprioceptive observation $\boldsymbol{s}_t^{\mathrm{p}}$ and a goal state $\boldsymbol{s}_t^{\mathrm{g}}$ from the reference motion. The agent outputs actions $\boldsymbol{a}_t$ specifying desired joint angles, executed by a low-level proportional-derivative (PD) controller. The reward $r_t = \mathcal{R}(\boldsymbol{s}_t^{\mathrm{p}}, \boldsymbol{s}_t^{\mathrm{g}}, \boldsymbol{a}_t)$ encourages accurate tracking and stable control.

Overall Pipeline. To enable tracking-based humanoid balance control, we first collect video clips and extract human poses using video-based motion capture algorithms such as WHAM [13], representing them in the SMPL format [61]. These poses are then retargeted to humanoid reference motions. Based on the generated reference motions, we adopt a teacher-student learning paradigm [9] to train balance policies. A teacher policy is first trained using Proximal Policy Optimization (PPO) [62] with access to privileged observations. Then, a student policy is distilled via DAgger [63] using only onboard observations. All policies are trained in simulation [64], and the final student policy is deployed on the real humanoid robot. Additional details are provided in Appendix B.

3.2 Reference Motion Refinement

Tracking performance in balance tasks is highly sensitive to the quality of reference motions, yet these often contain artifacts that hinder physical feasibility (see Figure 2 (a)). To mitigate this, we introduce a set of motion refinement strategies.

SMPL-Initialized Retargeting. Some prior humanoid retargeting approaches [4, 9] initialize joint angles to a zero pose and perform optimization by minimizing the positional differences between corresponding joints. However, in non-convex settings, such initialization can yield suboptimal results (see Figure 2(a)), as the zero pose can be far from the optimal solution. To address this, we propose a more effective initialization strategy based on the human SMPL pose. Given the humanoid's joint degrees of freedom form a subset of those in SMPL, we initialize each humanoid joint using the corresponding Euler angles from the SMPL pose. This SMPL initialization provides a starting point closer to the optimal solution, leading to faster convergence and improved accuracy.

Grounded Foot Correction. To further improve reference motion quality after retargeting, we introduce a data post-processing step to address foot instability caused by motion capture noise and retargeting errors. We assume that in the original human motion, the support foot should remain stationary without slipping. Based on this assumption, during single-legged phases, we adjust the global root position while keeping all local joint angles unchanged, ensuring that the grounded foot remains static across consecutive frames. This correction enhances foot stability and mitigates unrealistic foot-sliding artifacts introduced by noisy pose estimation and imperfect retargeting.

Center-of-Mass Filtering. Due to motion capture errors and human-humanoid mass mismatch, reference motions may exhibit large center-of-mass (COM) shifts, leading to physically infeasible trajectories, especially during single-leg phases. To address this, we compute the COM from the URDF-defined [65] body masses and positions, and discard trajectories where the ground-projected COM deviates from the center of the support foot by more than $0.2\,\mathrm{m}$, ensuring feasible references.

Transition Stabilization. Challenging balance motions are often sensitive to the initial pose, and even slight instability in the double-support stance prior to execution can adversely affect performance. To address this issue, we propose a simple yet effective strategy: extending the double-support phase before and after the balance phase. Specifically, we duplicate the first and last frames of the reference motion so that their total duration equals the balance phase. This not only increases the proportion of training time spent in stable double-foot stance—facilitating policy learning of standing balance—but also gives the humanoid more time to settle before transitioning into the extreme balance motion during deployment, ensuring a more reliable transition into the target pose.

3.3 Balance-Aware Policy Learning

Even with refined reference motions, humanoids face inherent challenges in maintaining balance due to morphological mismatch and the lack of structured guidance for balance behaviors. We overcome these issues through relaxed reference tracking and balance shaping rewards.

Relaxed Reference Tracking. Due to structural differences and the resulting center of mass misalignment, closely tracking human motions on a humanoid often leads to instability. To mitigate this, we leverage the exploratory nature of RL and allow the policy to make fine-grained adjustments to the center of mass. Specifically, we relax the tracking objective by setting a relatively large tolerance ($\sigma = 0.6\,\mathrm{m}$) in the reward function (see Appendix B.2 for details), enabling the policy to deviate from the reference when strict adherence would compromise balance. This flexibility promotes the emergence of more stable behaviors, facilitating successful execution of extreme balance tasks.

Balance Shaping Rewards. Merely relaxing the tracking constraint does not guarantee that the policy will learn physically feasible behaviors. Without structured guidance, reinforcement learning can converge to suboptimal solutions that violate task intent or physical plausibility. To handle this challenge, we design a set of shaping rewards. (i) **Center of mass (COM) reward** encourages the horizontal projection of the COM to remain within the support polygon, thereby helping the policy learn to adjust its pose to satisfy balance constraints. (ii) **Foot contact mismatch penalty** categorizes the landing state of each foot as either in contact with the ground or not, and penalizes discrepancies between the humanoid's and the reference's contact states. For example, it discourages the non-supporting foot from making unintended ground contact during single-leg balancing. While such contact may offer momentary stability, it violates task constraints and compromises the integrity of the intended balance behavior. (iii) **Close feet penalty** prevents the feet from getting too close to one another, reducing the risk of inter-foot collisions and encouraging more stable lower-body poses. Collectively, these shaping rewards promote the emergence of balance-aware, physically plausible motion policies. More reward details are described in Appendix B.2.

3.4 Sim-to-Real Robustness Training

Sensor noise—particularly from IMU and visual-inertial odometry (VIO) systems—and unmodeled real-world dynamics pose significant challenges for sim-to-real transfer in balance control. To enhance real-world robustness, we adopt localized reference tracking and IMU-centric observation perturbation to mitigate the issues caused by VIO and IMU noise, respectively, and apply high-frequency push disturbances to improve resilience against simulation modeling inaccuracies.

Localized Reference Tracking Training. To address the noise issues in VIO, we discard odometry information during both student policy training and deployment, and align the reference root with the humanoid's current root pose, expressing all tracking targets in the local coordinate frame. Prior work, such as ExBody2 [12], discards odometry only at deployment but relies on global tracking during training, leading to a mismatch that prevents the policy from accurately perceiving its own

motion. As a result, the robot often fails to correct balance loss, persistently falling or jumping in a particular direction. In contrast, our approach maintains consistency between training and deployment while avoiding the adverse effects of VIO noise.

IMU-Centric Observation Perturbation. Prior approaches [4, 9] inject uniform noise into observations to improve robustness, but this fails to capture the specific characteristics of IMU noise. Since the IMU-provided root orientation defines the local coordinate frame, many observation quantities—such as projected gravity, localized angular velocity, and localized reference targets—are intrinsically coupled. Simply adding independent uniform noise overlooks these dependencies. Moreover, real-world IMU noise exhibits significant temporal correlation. To address these issues, we perturb the observed root orientation—represented in Euler angles—with Ornstein-Uhlenbeck (OU) noise [66], a temporally correlated stochastic process, during student training. All observations are then computed based on the perturbed orientation observation, ensuring that the resulting observation noise reflects both the temporal dynamics and the structural dependencies induced by IMU errors, thereby yielding a more realistic simulation of real-world sensor behavior.

High-Frequency Push Disturbance. Tracking-based policies often fail during real-world deployment of single-leg tasks, as minor initial oscillations can progressively amplify due to unmodeled dynamics. To better approximate this failure mode, we apply random external pushes during teacher policy training by injecting small, high-frequency velocity offsets into the root (push every 1s at up to 0.5 m/s). This strategy effectively incorporates real-world instability into simulation, significantly enhancing sim-to-real transferability and disturbance robustness. In contrast, prior work [9] introduced low-frequency, high-magnitude pushes to train recovery from sudden external forces. However, such large perturbations are unsuitable for single-leg balance, where the feasible region is extremely narrow, and fail to capture the subtle instability dynamics critical for maintaining balance.

4 Experiments

Our experiments aim to answer the following questions: (1) How well does HuB perform on extreme balance tasks compared to prior tracking-based approaches? (2) What are the contributions of each key component in HuB's design to its overall performance? (3) Can HuB transfer successfully to the real world, and how robust is it to external perturbations?

4.1 Experiments Setup

Environment and Tasks. We conduct our experiments on the Unitree G1 humanoid robot, evaluating HuB across a set of balance tasks with varying difficulty levels (visualized in Figure 1). Simulation experiments are performed in the IsaacGym environment [64]. To better simulate real-world jitter and sensor noise, we introduce two perturbations during testing: (i) random external pushes every 1s by perturbing root velocity up to 0.1 m/s, and (ii) IMU noise via Ornstein-Uhlenbeck (OU) noise [66] added to the root orientation in Euler angles. Each policy is evaluated over 100 episodes under these perturbed simulation conditions. For real robot setup, please refer to Appendix A.

Metrics. We design a set of metrics to comprehensively evaluate policy performance on balance tasks, organized into three categories: (1) **Task Completion.** *Contact Mismatch (frame)* counts frames where foot contact states are incorrect (e.g., the non-supporting foot touches the ground during single-leg balancing). *Success Rate (%)* is the percentage of episodes where the humanoid maintains balance without (i) falling, (ii) foot contact mismatch, or (iii) an average tracking error exceeding 0.5 meters. (2) **Stability.** *Slippage (mm/s)* measures the support foot's ground-relative velocity, where higher values indicate unstable foot contact; *Air (frame)* counts frames where both feet are airborne, typically indicating a loss of ground contact due to instability; *Action Rate (rad/frame)* measures the action change magnitude between consecutive steps, where higher rates may suggest abrupt, unstable control behaviors. (3) **Tracking Error.** We report average global errors in keypoint position E_{pos} *(mm)*, velocity E_{vel} *(mm/frame)*, and acceleration E_{acc} *(mm/frame^2)*. In real-world experiments, due to the absence of odometry, we instead compute local errors relative to the robot base frame, denoted as $E_{pos\text{-}l}$, $E_{vel\text{-}l}$, and $E_{acc\text{-}l}$.

	Swallow Balance								Bruce Lee's Kick							
	Completion		Stability			Tracking Error			Completion		Stability			Tracking Error		
Method	Succ[1]↑	Cont[2]↓	Slip[3]↓	Air ↓	Act[4]↓	E_{pos} ↓	E_{vel} ↓	E_{acc} ↓	Succ ↑	Cont ↓	Slip ↓	Air ↓	Act ↓	E_{pos} ↓	E_{vel} ↓	E_{acc} ↓
H2O [4]	0	181.85	203.24	2.29	8.02	572.82	8.18	4.39	4	4.57	368.65	17.11	9.53	328.75	7.79	5.13
OmniH2O [9]	0	237.09	149.48	2.02	5.87	155.45	4.05	2.54	3	15.54	191.67	9.13	2.47	116.21	6.27	3.41
HuB (ours)	**100**	**0.00**	**80.81**	**0.89**	**0.51**	**83.15**	**3.50**	**1.89**	**100**	**0.00**	**76.44**	**1.28**	**0.46**	**67.18**	**3.31**	**2.33**
(a) Ablation on Relaxed Tracking																
HuB-track-sigma-0.15m	97	0.01	103.89	3.33	0.81	**67.74**	**3.20**	1.99	**100**	**0.00**	96.73	2.59	0.80	69.43	**3.29**	2.43
HuB-track-sigma-0.3m	99	**0.00**	107.12	3.76	0.63	89.50	3.29	1.95	99	**0.00**	104.67	5.46	0.58	89.09	3.48	2.45
HuB-track-sigma-1.2m	73	0.41	**54.64**	**0.14**	**0.47**	223.96	6.52	2.11	99	0.03	**72.34**	**1.02**	**0.43**	80.61	3.37	**2.33**
HuB-track-sigma-0.6m (ours)	**100**	**0.00**	80.81	0.89	0.51	83.15	3.50	**1.89**	**100**	**0.00**	76.44	1.28	0.46	**67.18**	3.31	**2.33**
(b) Ablation on Balance Shaping Rewards																
HuB-w/o-COM-reward	99	**0.00**	91.29	1.09	0.53	91.74	3.58	**1.89**	98	**0.00**	76.86	3.43	**0.40**	67.71	**3.18**	**2.28**
HuB-w/o-contact-penalty	74	0.49	93.39	1.53	0.58	104.14	3.96	2.03	62	0.87	78.43	2.01	0.56	67.66	3.31	2.33
HuB-w/o-close-feet-penalty	96	0.06	102.76	**0.75**	0.67	123.00	3.97	2.02	**100**	**0.00**	83.47	1.77	0.53	81.42	3.52	2.37
HuB	**100**	**0.00**	**80.81**	0.89	**0.51**	**83.15**	**3.50**	**1.89**	**100**	**0.00**	**76.44**	**1.28**	0.46	**67.18**	3.31	2.33
(c) Ablation on Sim-to-Real Robustness Training																
HuB-w/o-localized-tracking	92	5.65	152.62	11.14	1.75	311.56	5.24	2.55	99	0.01	116.86	2.74	1.24	187.27	4.69	2.93
HuB-w/o-imu-noise	92	2.13	233.40	17.10	2.53	253.77	6.56	3.13	93	0.62	287.32	14.53	5.14	268.26	7.14	4.19
HuB-w/o-push	90	3.65	102.50	4.75	0.74	98.06	3.75	1.97	89	1.19	156.69	7.67	6.40	134.01	4.63	3.04
HuB-push (5s interval, 1 m/s)	97	0.33	99.09	2.78	0.83	141.74	**3.49**	1.91	**100**	**0.00**	**74.37**	1.51	**0.43**	69.10	3.38	2.34
HuB	**100**	**0.00**	**80.81**	**0.89**	**0.51**	**83.15**	3.50	**1.89**	**100**	**0.00**	76.44	**1.28**	0.46	**67.18**	**3.31**	**2.33**

Abbreviation for [1] *Success Rate* [2] *Contact Mismatch* [3] *Slippage* [4] *Action Rate*

Table 1: **Simulation Results.** We compare HuB against baselines and ablations. The results demonstrate that HuB successfully completes extreme balance tasks, whereas baselines consistently fail. Ablation studies further highlight the critical contributions of each component of HuB to the overall balance performance.

Baselines. To evaluate the performance of HuB compared to standard tracking-based approaches, we consider the following baselines: (1) **H2O** [4]: a tracking-based humanoid control framework that retargets human motion data to the humanoid and trains a RL policy to track the reference motion. (2) **OmniH2O** [9]: an extension of H2O that introduces a teacher-student learning paradigm, where a teacher policy is trained with privileged information using RL, and a student policy is distilled from it via DAgger [63] with only deployment-accessible observations. For a fair comparison, the baselines are adapted to our localized tracking framework and trained from scratch using the same balance motion data as HuB, tracking the same set of keypoints.

4.2 Simulation Results

As shown in Table 1, we present the quantitative results of HuB, the baselines, and the ablations on two challenging tasks, with additional results provided in Appendix C.

HuB and Baselines Performance. The results demonstrate that HuB is the only method that completes these extreme balance tasks with a 100% success rate, while the baselines almost always fail due to large contact mismatches—specifically, unintended ground contacts by the non-supporting foot, which should remain airborne during single-leg balancing. Moreover, HuB exhibits smaller ground slippage, shorter airborne durations, and lower action variability, indicating stronger stability and smoother motion execution. In addition, HuB achieves lower tracking errors, suggesting more accurate task completion.

Ablations on Relaxed Tracking. We experiment with different tracking tolerance σ values. The results show that a smaller σ reduces the tracking error but simultaneously increases slippage, airborne time, and action rate, indicating degraded overall stability. In contrast, excessively large σ values lead to improved stability but at the cost of significantly higher tracking errors, more frequent contact mismatches, and lower task success rates. Based on these observations, we select a moderately large tolerance of $\sigma = 0.6\,\mathrm{m}$, which achieves a favorable balance between tracking fidelity and policy stability compared to prior tracking-based methods.

Ablations on Shaping Rewards. Removing the COM reward leads to increased slippage and airborne time, underscoring its critical role in stabilizing the center of mass to maintain balance. Eliminating the foot contact mismatch penalty results in a significant rise in contact mismatches and a notable drop in success rates, demonstrating its importance for ensuring successful single-foot balancing. Additionally, removing the close-feet penalty degrades performance, primarily because the humanoid's feet can come excessively close, which reduces overall standing stability.

Ablations on Robustness Training. First, replacing localized tracking with global tracking during training—while still deploying with localized tracking, as in ExBody2 [12]—introduces a mismatch between training and deployment. This significantly degrades performance on tasks like `Swallow Balance`, where successful execution depends on the precise completion of preceding motions. Second, removing IMU noise injection leads to performance deterioration across all metrics, indicating policies not exposed to sensor noise during training are highly sensitive to deployment

errors. Third, omitting high-frequency push perturbations increases contact mismatches, lowers success rates, and degrades stability, indicating that perturbation exposure during training is critical for successful task execution and overall stability. Finally, replacing high-frequency pushes with low-frequency ones (push every 5s at up to 1 m/s), as in prior tracking-based methods [9], also degrades stability, especially on tasks with narrow feasible balance regions like `Swallow Balance`. This is likely because infrequent large perturbations are too hard for the humanoid to withstand during balance tasks.

Retargeting Results. To assess the impact of SMPL-initialization on optimization-based retargeting, we compare the retargeting loss after 500 optimization steps between solutions optimized with and without SMPL-initialization. As shown in Figure 3, solutions with SMPL-initialization consistently achieve lower losses across all tasks, with notably large reductions in `Deep Squat`.

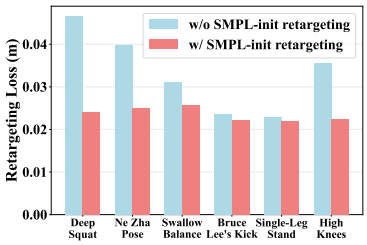

Figure 3: Comparison of Retargeting Loss w/ and w/o SMPL-Initialization.

4.3 Real-World Results

Balance Performance. We evaluate the performance of HuB and the baseline OmniH2O on real-world balance tasks. Figure 1 visualizes real-world executions of HuB, and Table 2 quantitatively compares HuB and OmniH2O across evaluation metrics. Videos are available on the project website. The results show that HuB successfully completes challenging balance tasks in the real world, including `Swallow Balance` and `Bruce Lee's Kick`, which are difficult even for humans. The humanoid holds these extreme poses with stability and fluidity, maintaining steady balance and making only minor foot adjustments when necessary. These results highlight the strong balance capabilities of HuB. In contrast, the baseline struggles to complete the tasks: for example, it fails to maintain balance and falls during `Swallow Balance`, exhibits excessive shaking during `Bruce Lee's Kick`, and abandons leg lifting completely in `Single-Leg Stand` and `Ne Zha Pose` to reduce the risk of falling.

Method	Succ ↑	$E_{\text{pos-l}}$ ↓	$E_{\text{vel-l}}$ ↓	$E_{\text{acc-l}}$ ↓
(a) Swallow Balance				
OmniH2O	0/5	119.73	2.20	1.86
HuB	**4/5**	**38.31**	**1.73**	**1.13**
(b) Bruce Lee's Kick				
OmniH2O	0/5	80.69	5.80	4.80
HuB	**5/5**	**27.87**	**1.58**	**1.14**
(c) Ne Zha Pose				
OmniH2O	0/5	50.48	1.29	1.37
HuB	**5/5**	**30.91**	**0.73**	**0.37**
(d) Single-Leg Stand				
OmniH2O	0/5	32.10	**0.49**	**0.25**
HuB	**5/5**	**29.58**	0.96	0.47
(e) Deep Squat				
OmniH2O	4/5	42.97	3.21	4.21
HuB	**5/5**	**29.90**	**2.08**	**1.08**

Table 2: Real-World Results.

Robustness Evaluation. We evaluate the robustness of HuB across two aspects. (1) **External Perturbations.** As illustrated in Figure 4, we apply external disturbances by striking the humanoid with a forcefully kicked soccer ball during balance tasks. Despite significant disruptions to the robot's balance, HuB rapidly reacts and recovers with minimal corrective motion, returning to a stable state within a short period, demonstrating strong disturbance tolerance. (2) **Long-Horizon Task Execution.** We conduct repeated trials of the `Bruce Lee's Kick` task without resetting the humanoid between trials. HuB successfully completes 10 consecutive executions in a single take without any failures or external intervention. This demonstrates HuB's strong reliability, balance consistency, and control stability in real-world deployment. These two evaluations highlight the robustness of our framework and validate the effectiveness of its design.

Figure 4: External Perturbations.

5 Conclusion

We present HuB, a unified learning-based framework for humanoid control in extreme balance tasks. By systematically addressing challenges such as reference motion inaccuracies, balance policy learning difficulties, and the sim-to-real gap, HuB enables humanoid robots to stably execute

challenging balance poses that baseline methods consistently fail to complete. It further demonstrates strong robustness to disturbances and consistency over long-horizon deployments.

6 Limitations

One limitation of our method is that certain components are specifically designed for balance tasks and rely on task-specific assumptions. As a result, they may not be directly applicable to other task categories, such as jumping or parkour. Moreover, although the trained policies are capable of accomplishing complex balancing behaviors, they exhibit limited generalization: adapting to novel and substantially different motions typically necessitates retraining. Developing policies capable of acquiring versatile motor skills that can be reliably deployed in the real world remains an important research direction.

References

[1] K. Hirai, M. Hirose, Y. Haikawa, and T. Takenaka. The development of honda humanoid robot. In *Proceedings. 1998 IEEE international conference on robotics and automation (Cat. No. 98CH36146)*, volume 2, pages 1321–1326. IEEE, 1998.

[2] M. Johnson, B. Shrewsbury, S. Bertrand, T. Wu, D. Duran, M. Floyd, P. Abeles, D. Stephen, N. Mertins, A. Lesman, et al. Team ihmc's lessons learned from the darpa robotics challenge trials. *Journal of Field Robotics*, 32(2):192–208, 2015.

[3] S. Kuindersma, R. Deits, M. Fallon, A. Valenzuela, H. Dai, F. Permenter, T. Koolen, P. Marion, and R. Tedrake. Optimization-based locomotion planning, estimation, and control design for the atlas humanoid robot. *Autonomous robots*, 40:429–455, 2016.

[4] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8944–8951. IEEE, 2024.

[5] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. In *Conference on Robot Learning (CoRL)*, 2024.

[6] 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. Hodgins, L. J. Fan, Y. Zhu, C. Liu, and G. Shi. Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. *arXiv preprint arXiv:2502.01143*, 2025.

[7] R. C. Fitzpatrick and B. L. Day. Probing the human vestibular system with galvanic stimulation. *Journal of applied physiology*, 96(6):2301–2316, 2004.

[8] R. J. Peterka. Sensorimotor integration in human postural control. *Journal of neurophysiology*, 88(3):1097–1118, 2002.

[9] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. M. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *8th Annual Conference on Robot Learning*.

[10] T. He, W. Xiao, T. Lin, Z. Luo, Z. Xu, Z. Jiang, J. Kautz, C. Liu, G. Shi, X. Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. *arXiv preprint arXiv:2410.21229*, 2024.

[11] 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.

[12] M. Ji, X. Peng, F. Liu, J. Li, G. Yang, X. Cheng, and X. Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.

[13] S. Shin, J. Kim, E. Halilaj, and M. J. Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024.

[14] Y. Wang, Z. Wang, L. Liu, and K. Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487. Springer, 2024.

[15] S. Kajita, F. Kanehiro, K. Kaneko, K. Fujiwara, K. Harada, K. Yokoi, and H. Hirukawa. Biped walking pattern generation by using preview control of zero-moment point. In *2003 IEEE international conference on robotics and automation (Cat. No. 03CH37422)*, volume 2, pages 1620–1626. IEEE, 2003.

[16] J. Pratt, J. Carff, S. Drakunov, and A. Goswami. Capture point: A step toward humanoid push recovery. In *2006 6th IEEE-RAS international conference on humanoid robots*, pages 200–207. Ieee, 2006.

[17] B. Stephens. Integral control of humanoid balance. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4020–4027. IEEE, 2007.

[18] T. Koolen, T. De Boer, J. Rebula, A. Goswami, and J. Pratt. Capturability-based analysis and control of legged locomotion, part 1: Theory and application to three simple gait models. *The international journal of robotics research*, 31(9):1094–1113, 2012.

[19] C. G. Atkeson and B. Stephens. Multiple balance strategies from one optimization criterion. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, pages 57–64. IEEE, 2007.

[20] B. J. Stephens and C. G. Atkeson. Dynamic balance force control for compliant humanoid robots. In *2010 IEEE/RSJ international conference on intelligent robots and systems*, pages 1248–1255. IEEE, 2010.

[21] C. Ott, M. A. Roa, and G. Hirzinger. Posture and balance control for biped robots based on contact force optimization. In *2011 11th IEEE-RAS International Conference on Humanoid Robots*, pages 26–33. IEEE, 2011.

[22] X. B. Peng, G. Berseth, K. Yin, and M. Van De Panne. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Acm transactions on graphics (tog)*, 36(4): 1–13, 2017.

[23] A. Duburcq, F. Schramm, G. Boéris, N. Bredeche, and Y. Chevaleyre. Reactive stepping for humanoid robots using reinforcement learning: Application to standing push recovery on the exoskeleton atalante. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9302–9309. IEEE, 2022.

[24] X. He, R. Dong, Z. Chen, and S. Gupta. Learning getting-up policies for real-world humanoid robots. *arXiv preprint arXiv:2502.12152*, 2025.

[25] T. Huang, J. Ren, H. Wang, Z. Wang, Q. Ben, M. Wen, X. Chen, J. Li, and J. Pang. Learning humanoid standing-up control across diverse postures. *arXiv preprint arXiv:2502.08378*, 2025.

[26] Z. Zhuang and H. Zhao. Embrace collisions: Humanoid shadowing for deployable contact-agnostics motions. *arXiv preprint arXiv:2502.01465*, 2025.

[27] C. Yang, K. Yuan, W. Merkt, T. Komura, S. Vijayakumar, and Z. Li. Learning whole-body motor skills for humanoids. In *2018 IEEE-RAS 18th international conference on humanoid robots (humanoids)*, pages 270–276. IEEE, 2018.

[28] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato. Learning from demonstration and adaptation of biped locomotion. *Robotics and autonomous systems*, 47 (2-3):79–91, 2004.

[29] R. Calandra, A. Seyfarth, J. Peters, and M. P. Deisenroth. Bayesian optimization for learning gaits under uncertainty: An experimental comparison on a dynamic bipedal walker. *Annals of Mathematics and Artificial Intelligence*, 76:5–23, 2016.

[30] T. Li, H. Geyer, C. G. Atkeson, and A. Rai. Using deep reinforcement learning to learn high-level policies on the atrias biped. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 263–269. IEEE, 2019.

[31] 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)*, pages 2811–2817. IEEE, 2021.

[32] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen. Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning. *arXiv preprint arXiv:2408.14472*, 2024.

[33] Q. Liao, B. Zhang, X. Huang, X. Huang, Z. Li, and K. Sreenath. Berkeley humanoid: A research platform for learning-based control. *arXiv preprint arXiv:2407.21781*, 2024.

[34] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik. Humanoid locomotion as next token prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[35] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath. Real-world humanoid locomotion with reinforcement learning. *Science Robotics*, 9(89):eadi9579, 2024.

[36] Q. Zhang, P. Cui, D. Yan, J. Sun, Y. Duan, G. Han, W. Zhao, W. Zhang, Y. Guo, A. Zhang, et al. Whole-body humanoid robot locomotion with human reference. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11225–11231. IEEE, 2024.

[37] Z. Chen, X. He, Y.-J. Wang, Q. Liao, Y. Ze, Z. Li, S. S. Sastry, J. Wu, K. Sreenath, S. Gupta, et al. Learning smooth humanoid locomotion through lipschitz-constrained policies. *arXiv preprint arXiv:2410.11825*, 2024.

[38] D. Crowley, J. Dao, H. Duan, K. Green, J. Hurst, and A. Fern. Optimizing bipedal locomotion for the 100m dash with comparison to human running. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12205–12211. IEEE, 2023.

[39] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *The International Journal of Robotics Research*, page 02783649241285161, 2024.

[40] Z. Li, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath. Robust and versatile bipedal jumping control through reinforcement learning. In *Robotics science and systems*. RSS, 2023.

[41] Z. Zhuang, S. Yao, and H. Zhao. Humanoid parkour learning. In *8th Annual Conference on Robot Learning*.

[42] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang. Learning humanoid locomotion with perceptive internal model. *arXiv preprint arXiv:2411.14386*, 2024.

[43] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

[44] P. K. Khosla and T. Kanade. Parameter identification of robot dynamics. In *1985 24th IEEE conference on decision and control*, pages 1754–1760. IEEE, 1985.

[45] M. Gautier and W. Khalil. On the identification of the inertial parameters of robots. In *Proceedings of the 27th IEEE Conference on Decision and Control*, volume 3, pages 2264–2269. IEEE Piscataway, NJ, 1988.

[46] S. Zhu, A. Kimmel, K. E. Bekris, and A. Boularias. Fast model identification via physics engines for data-efficient policy search. In *27th International Joint Conference on Artificial Intelligence, IJCAI 2018*, pages 3249–3256. International Joint Conferences on Artificial Intelligence, 2018.

[47] J. Tan, Z. Xie, B. Boots, and C. K. Liu. Simulation-based design of dynamic controllers for humanoid balancing. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2729–2736. IEEE, 2016.

[48] S. Kolev and E. Todorov. Physically consistent state estimation and system identification for contacts. In *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pages 1036–1043. IEEE, 2015.

[49] W. Yu, J. Tan, C. K. Liu, and G. Turk. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453*, 2017.

[50] W. Yu, V. C. Kumar, G. Turk, and C. K. Liu. Sim-to-real transfer for biped locomotion. In *2019 ieee/rsj international conference on intelligent robots and systems (iros)*, pages 3503–3510. IEEE, 2019.

[51] A. Allevato, E. S. Short, M. Pryor, and A. Thomaz. Tunenet: One-shot residual tuning for system identification and sim-to-real robot task transfer. In *Conference on Robot Learning*, pages 445–455. PMLR, 2020.

[52] N. Liu, Y. Cai, T. Lu, R. Wang, and S. Wang. Real–sim–real transfer for real-world robot control policy learning with deep reinforcement learning. *Applied Sciences*, 10(5):1555, 2020.

[53] L. Wang, R. Guo, Q. Vuong, Y. Qin, H. Su, and H. Christensen. A real2sim2real method for robust object grasping with neural surface reconstruction. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–8. IEEE, 2023.

[54] M. Torne, A. Simeonov, Z. Li, A. Chan, T. Chen, A. Gupta, and P. Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. *arXiv preprint arXiv:2403.03949*, 2024.

[55] T. Lin, K. Sachdev, L. Fan, J. Malik, and Y. Zhu. Sim-to-real reinforcement learning for vision-based dexterous manipulation on humanoids. *arXiv preprint arXiv:2502.20396*, 2025.

[56] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017.

[57] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[58] J. Tobin, L. Biewald, R. Duan, M. Andrychowicz, A. Handa, V. Kumar, B. McGrew, A. Ray, J. Schneider, P. Welinder, et al. Domain randomization and generative models for robotic grasping. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3482–3489. IEEE, 2018.

[59] B. Mehta, M. Diaz, F. Golemo, C. J. Pal, and L. Paull. Active domain randomization. In *Conference on Robot Learning*, pages 1162–1176. PMLR, 2020.

[60] J. Huber, F. Hélénon, H. Watrelot, F. B. Amar, and S. Doncieux. Domain randomization for sim2real transfer of automatically generated grasping datasets. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4112–4118. IEEE, 2024.

[61] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

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

[63] 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*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[64] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021.

[65] ROS Wiki. URDF - Unified Robot Description Format. http://wiki.ros.org/urdf, 2025. [Online; accessed 24-April-2025].

[66] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Physical review*, 36(5):823, 1930.

[67] Unitree Robotics. G1 developer documentation: Dds services interface. https://support.unitree.com/home/en/G1_developer/dds_services_interface, 2025. Accessed: 2025-04-28.

[68] Unitree Robotics. unitree_sdk2_python: Python interface for unitree sdk2. https://github.com/unitreerobotics/unitree_sdk2_python, 2025. Accessed: 2025-04-28.

[69] N. Rudin, D. Hoeller, P. Reist, and M. Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning, 2022.

[70] L. Campanaro, S. Gangapurwala, W. Merkt, and I. Havoutis. Learning and deploying robust locomotion policies with minimal dynamics randomization, 2023.

A Real Robot Setup

We conduct our experiments on the Unitree G1 humanoid robot, which features 29 degrees of freedom (DoF), including two 7-DoF arms, two 6-DoF legs, and a 3-DoF waist. For real-world deployment, we use the robot's onboard IMU to obtain root orientation and angular velocity, and joint encoders to obtain joint positions and velocities. The control policy receives keypoint tracking targets and proprioceptive information as input, computes the desired joint positions for each actuator, and sends commands to the robot's low-level interface. Policy inference is executed in real time on the onboard NVIDIA Jetson Orin NX, with a control frequency of 50 Hz. Observations, including keypoint tracking information and proprioceptive data, are transmitted to the control policy via DDS [67], using the `unitree_sdk2_python` implementation [68].

B HuB Details

B.1 State Space Design

This subsection details the state space design for both the teacher and student policies in HuB.

Teacher Policy. The teacher policy, trained via RL, has access to the full states required for reference tracking. Table 3 presents the state space of the teacher policy.

State term	Dimensions
Rigid body position	87
Rigid body rotation	180
Rigid body velocity	90
Rigid body angular velocity	90
Rigid body position difference	90
Rigid body rotation difference	180
Rigid body velocity difference	90
Rigid body angular velocity difference	90
Local reference rigid body position	90
Local reference rigid body rotation	180
Actions	29
Total dim	1196

Table 3: State space information of the teacher policy.

Student Policy. The student policy, trained using DAgger with a history of 25 steps, is restricted to deployment-accessible observations only. Table 4 presents the state space of the student policy. For the student policy, we select a total of 12 tracking keypoints, corresponding to the left and right sides of the hips, knees, ankles, shoulders, elbows, and wrists.

State term	Dimensions
DoF position	29
DoF velocity	29
Base angular velocity	3
Projected gravity	3
Localized reference keypoints position	36
Keypoints position difference	36
Keypoints velocity difference	36
Actions	29
Single step total dim	201
History state term	Dimensions
DoF position	29
DoF velocity	29
Base angular velocity	3
Projected gravity	3
Actions	29
History single step total dim	93
Total dim	2526 (201 + 93×25)

Table 4: State space information of the student policy.

B.2 Rewards

Table 5 provides a summary of the detailed reward components.

Term	Expression	Weight	Remarks
	Balance Shaping Rewards		
Center of mass	$\exp(-\lVert \boldsymbol{p}_{\text{xy}}^{\text{com}} - \boldsymbol{p}_{\text{xy}}^{\text{lower-foot}} \rVert_2^2 / \sigma_{\text{com}}^2) \times \mathbb{1}(\lVert \hat{\boldsymbol{p}}_{\text{z}}^{\text{l-foot}} - \hat{\boldsymbol{p}}_{\text{z}}^{\text{r-foot}} \rVert_2 > 0.05)$	160	$\sigma_{\text{com}} = 0.1$
Foot contact mismatch	$\boldsymbol{c}_{\text{feet}} \oplus \hat{\boldsymbol{c}}_{\text{feet}}$ [1]	-250	
Close feet	$\max\{0.16 - \lVert \boldsymbol{p}^{\text{l-foot}} - \boldsymbol{p}^{\text{r-foot}} \rVert_2, 0\}$	-1000	
	Tracking Rewards		
Body position	$\exp(-\lVert \boldsymbol{p}_t - \hat{\boldsymbol{p}}_t \rVert_2^2 / \sigma_{\text{pos}}^2)$	30	$\sigma_{\text{pos}} = 0.6$
Body rotation	$\exp(-\lVert \boldsymbol{\theta}_t \ominus \hat{\boldsymbol{\theta}}_t \rVert_2^2 / \sigma_{\text{rot}}^2)$	20	$\sigma_{\text{rot}} = 0.3$
Body velocity	$\exp(-\lVert \boldsymbol{v}_t - \hat{\boldsymbol{v}}_t \rVert_2^2 / \sigma_{\text{vel}}^2)$	8	$\sigma_{\text{vel}} = 3$
Body angular velocity	$\exp(-\lVert \boldsymbol{\omega}_t - \hat{\boldsymbol{\omega}}_t \rVert_2^2 / \sigma_{\text{ang}}^2)$	8	$\sigma_{\text{ang}} = 10$
DoF position	$\exp(-\lVert \boldsymbol{d}_t - \hat{\boldsymbol{d}}_t \rVert_2^2 / \sigma_{\text{dpos}}^2)$	32	$\sigma_{\text{dpos}} = 0.7$
DoF velocity	$\exp(-\lVert \dot{\boldsymbol{d}}_t - \hat{\dot{\boldsymbol{d}}}_t \rVert_2^2 / \sigma_{\text{dvel}}^2)$	16	$\sigma_{\text{dvel}} = 10$
	Penalty		
Torque limits	$\mathbb{1}(\boldsymbol{\tau}_t \notin [\boldsymbol{\tau}_{\min}, \boldsymbol{\tau}_{\max}])$	-0.5	
DoF position limits	$\mathbb{1}(\boldsymbol{d}_t \notin [\boldsymbol{d}_{\min}, \boldsymbol{d}_{\max}])$	-30	
DoF velocity limits	$\mathbb{1}(\dot{\boldsymbol{d}}_t \notin [\dot{\boldsymbol{d}}_{\min}, \dot{\boldsymbol{d}}_{\max}])$	-12	
Termination	$\mathbb{1}_{\text{termination}}$	-60	
	Regularization		
Torque	$\lVert \boldsymbol{\tau}_t \rVert$	-2.5×10^{-5}	
DoF velocity	$\lVert \dot{\boldsymbol{d}}_t \rVert_2^2$	-1×10^{-3}	
DoF acceleration	$\lVert \ddot{\boldsymbol{d}}_t \rVert_2$	-3×10^{-6}	
Action rate	$\lVert \boldsymbol{a}_t - \boldsymbol{a}_{t-1} \rVert_2^2$	-1.5	
Feet air time	$T_{\text{air}} - 0.25$ [69]	250	
Feet contact force	$\lVert F_{\text{feet}} \rVert_2^2$	-0.2	
Stumble	$\mathbb{1}(F_{\text{feet}}^{xy} > 5 \times F_{\text{feet}}^{z})$	-3×10^{-4}	
Slippage	$\lVert \boldsymbol{v}_t^{\text{feet}} \rVert_2^2 \times \mathbb{1}(F_{\text{feet}} \geq 1)$	-30	
Feet orientation	$\lVert \boldsymbol{g}_{\text{z}}^{\text{feet}} \rVert \times \mathbb{1}(\boldsymbol{p}_{\text{z}}^{\text{feet}} < 0.05)$	-62.5	
In the air	$\mathbb{1}(F_{\text{feet}}^{\text{left}}, F_{\text{feet}}^{\text{right}} < 1)$	-50	

[1] $\boldsymbol{c}_{\text{feet}}$ represents the robot's feet contact with the ground, and $\hat{\boldsymbol{c}}_{\text{feet}}$ the reference's. Whether the robot's feet are in contact is determined by $F_{\text{feet}} \geq 1\,\text{N}$. For the reference, both feet are considered grounded if their height difference is below 0.05m; otherwise, the lower foot is considered grounded.

Table 5: Reward components and weights. Quantities with the hat symbol ($\hat{\cdot}$) represent reference motion variables, while unmarked terms refer to the humanoid's own state variables.

B.3 Domain Randomization

Table 6 summarizes the domain randomization strategies used in HuB, including high-frequency push disturbances designed to bridge the sim-to-real gap and improve balance robustness.

Term	Value
High-Frequency Push Disturbance	
Push robot	interval $= 1\,\mathrm{s}$, $v_{\mathrm{xy}} \in \mathcal{U}(0, 0.5)$ m/s
Dynamics Randomization	
Friction	$\mathcal{U}(2.5, 3.5)$
Torso COM offset	$\mathcal{U}(-0.1, 0.1)$ m
Link mass	$\mathcal{U}(0.7, 1.3) \times$ default kg
PD gains	$\mathcal{U}(0.75, 1.25) \times$ default
Torque RFI [70]	$0.1 \times$ torque limit N $\cdot$ m
Control delay	$\mathcal{U}(20, 60)$ms
Motion reference offset	$\mathcal{U}([-0.02, 0.02], [-0.02, 0.02], [-0.1, 0.1])$ m

Table 6: Domain randomizations for HuB.

B.4 IMU noise

As illustrated in Section 3.4, we introduce Ornstein-Uhlenbeck (OU) noise [66] to the IMU's Euler angles observation (in degree). OU noise is modeled by the following differential equation:

$$\frac{\mathrm{d}X_t}{\mathrm{d}t} = -\theta X_t + \sigma \epsilon_t$$

where X_t represents the OU noise, θ is the mean reversion rate, σ is the noise intensity, and ϵ_t is a standard Gaussian noise term ($\epsilon_t \sim \mathcal{N}(0, 1)$) at each time step. The noise term introduces random fluctuations, while the mean reversion term prevents excessive drift. For our experiments, we set the parameters to $\theta = 25$ and $\sigma = 250$.

B.5 Hyperparameters

Table 7 presents the hyperparameters used for training HuB.

Hyperparameters	Values
Optimizer	Adam
β_1, β_2	0.9, 0.999
Learning rate	1×10^{-3}
Batch size	64
Discount factor (γ)	0.99
Clip param	0.2
Entropy coef	0.005
Max grad norm	0.2
Value loss coef	1
Entropy coef	0.005
Init noise std (RL)	1.0
Init noise std (DAgger)	0.001
Num learning epochs	5
MLP size	[512, 256, 128]

Table 7: Hyperparameters.

C Experiments Details

C.1 Experiments Setup Details

It is worth noting that, for a fair comparison, all baselines (OmniH2O and H2O) are trained from scratch using the same set of balance motion data as HuB, and are tasked with tracking the same set of keypoints.

To better approximate real-world conditions, we apply the same domain randomization during both training and evaluation, except for the random external pushes. As described in Section 3.4 and Section 4.1, different push magnitudes are used for training and evaluation—larger magnitudes (0.5 m/s) are applied during training to ensure the policy learns robustness under stronger disturbances, while smaller perturbations (0.1 m/s) are used in evaluation to more closely reflect realistic deployment scenarios.

C.2 Additional Results

Table 8 shows the performance of HuB and baselines across additional three tasks. HuB consistently outperforms the baselines in completion, stability, and tracking errors, demonstrating superior performance.

	Completion		**Stability**			**Tracking Error**		
Method	Succ[1]↑	Cont[2]↓	Slip[3]↓	Air ↓	Act[4]↓	E_{pos} ↓	E_{vel} ↓	E_{acc} ↓
(a) Ne Zha Pose								
H2O	0	129.27	227.06	2.72	6.59	257.31	6.11	3.77
OmniH2O	0	146.19	219.04	5.03	4.60	102.38	4.70	3.41
HuB	**97**	**0.02**	**72.76**	**0.69**	**0.46**	**74.13**	**2.94**	**1.65**
(b) Single-leg Stand								
H2O	0	172.71	236.28	3.05	8.76	478.23	7.23	4.25
OmniH2O	0	196.74	309.68	27.01	5.95	219.73	6.45	3.67
HuB	**97**	**0.56**	**78.16**	**2.45**	**0.62**	**70.03**	**3.03**	**1.80**
(c) Deep Squat								
H2O	**100**	**0.00**	236.48	2.35	6.65	371.76	14.24	5.08
OmniH2O	99	**0.00**	141.20	0.94	1.46	101.40	7.04	2.84
HuB	**100**	**0.00**	**77.93**	**0.12**	**0.77**	**62.28**	**5.58**	**2.31**

Abbreviation for [1] *Success Rate* [2] *Contact Mismatch* [3] *Slippage* [4] *Action Rate*

Table 8: Simulation results of HuB and baselines on additional 3 tasks.