

533 A Real Robot Setup

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

543 B HuB Details

544 B.1 State Space Design

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

546 **Teacher Policy.** The teacher policy, trained via RL, has access to the full states required for reference
 547 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.

548 **Student Policy.** The student policy, trained using DAgger with a history of 25 steps, is restricted
 549 to deployment-accessible observations only. Table 4 presents the state space of the student policy.
 550 For the student policy, we select a total of 12 tracking keypoints, corresponding to the left and right
 551 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 | $\exp(-\ \mathbf{p}_{xy}^{\text{com}} - \mathbf{p}_{xy}^{\text{lower-foot}}\ _2^2 / \sigma_{\text{com}}^2) \times \mathbb{1}(\ \hat{\mathbf{p}}_z^{\text{l-foot}} - \hat{\mathbf{p}}_z^{\text{r-foot}}\ _2 > 0.05)$ | 160 | $\sigma_{\text{com}} = 0.1$ |
| Foot contact mismatch | $\mathbf{c}_{\text{feet}} \oplus \hat{\mathbf{c}}_{\text{feet}}^1$ | -250 | |
| Close feet | $\max\{0.16 - \ \mathbf{p}^{\text{l-foot}} - \mathbf{p}^{\text{r-foot}}\ _2, 0\}$ | -1000 | |
| Tracking Rewards | | | |
| Body position | $\exp(-\ \mathbf{p}_t - \hat{\mathbf{p}}_t\ _2^2 / \sigma_{\text{pos}}^2)$ | 30 | $\sigma_{\text{pos}} = 0.6$ |
| Body rotation | $\exp(-\ \boldsymbol{\theta}_t \ominus \hat{\boldsymbol{\theta}}_t\ _2^2 / \sigma_{\text{rot}}^2)$ | 20 | $\sigma_{\text{rot}} = 0.3$ |
| Body velocity | $\exp(-\ \mathbf{v}_t - \hat{\mathbf{v}}_t\ _2^2 / \sigma_{\text{vel}}^2)$ | 8 | $\sigma_{\text{vel}} = 3$ |
| Body angular velocity | $\exp(-\ \boldsymbol{\omega}_t - \hat{\boldsymbol{\omega}}_t\ _2^2 / \sigma_{\text{ang}}^2)$ | 8 | $\sigma_{\text{ang}} = 10$ |
| DoF position | $\exp(-\ \mathbf{d}_t - \hat{\mathbf{d}}_t\ _2^2 / \sigma_{\text{dpos}}^2)$ | 32 | $\sigma_{\text{dpos}} = 0.7$ |
| DoF velocity | $\exp(-\ \dot{\mathbf{d}}_t - \hat{\dot{\mathbf{d}}}_t\ _2^2 / \sigma_{\text{dvel}}^2)$ | 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 | $\mathbb{1}(\mathbf{d}_t \notin [\mathbf{d}_{\min}, \mathbf{d}_{\max}])$ | -30 | |
| DoF velocity limits | $\mathbb{1}(\dot{\mathbf{d}}_t \notin [\dot{\mathbf{d}}_{\min}, \dot{\mathbf{d}}_{\max}])$ | -12 | |
| Termination | $\mathbb{1}_{\text{termination}}$ | -60 | |
| Regularization | | | |
| Torque | $\ \boldsymbol{\tau}_t\ $ | -2.5×10^{-5} | |
| DoF velocity | $\ \dot{\mathbf{d}}_t\ _2^2$ | -1×10^{-3} | |
| DoF acceleration | $\ \ddot{\mathbf{d}}_t\ _2$ | -3×10^{-6} | |
| Action rate | $\ \mathbf{a}_t - \mathbf{a}_{t-1}\ _2^2$ | -1.5 | |
| Feet air time | $T_{\text{air}} - 0.25$ [69] | 250 | |
| Feet contact force | $\ \mathbf{F}_{\text{feet}}\ _2^2$ | -0.2 | |
| Stumble | $\mathbb{1}(F_{\text{feet}}^{xy} > 5 \times F_{\text{feet}}^z)$ | -3×10^{-4} | |
| Slippage | $\ \mathbf{v}_t^{\text{feet}}\ _2^2 \times \mathbb{1}(F_{\text{feet}} \geq 1)$ | -30 | |
| Feet orientation | $\ \mathbf{g}_z^{\text{feet}}\ \times \mathbb{1}(\mathbf{p}_z^{\text{feet}} < 0.05)$ | -62.5 | |
| In the air | $\mathbb{1}(F_{\text{feet}}^{\text{left}}, F_{\text{feet}}^{\text{right}} < 1)$ | -50 | |

¹ \mathbf{c}_{feet} represents the robot's feet contact with the ground, and $\hat{\mathbf{c}}_{\text{feet}}$ the reference's. Whether the robot's feet are in contact is determined by $F_{\text{feet}} \geq 1$ 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.

554 B.3 Domain Randomization

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

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

557 B.4 IMU noise

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

$$\frac{dX_t}{dt} = -\theta X_t + \sigma \epsilon_t$$

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

564 B.5 Hyperparameters

565 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.

| Method | Completion | | Stability | | | Tracking Error | | |
|-----------------------------|---------------------|---------------------|---------------------|-------------|--------------------|--------------------|--------------------|--------------------|
| | 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.