Appendix

1. Detailed Observation Space

In this section, we provide more details about different components of observation space for both the privileged teacher policy and visual policies. The Unitree A1 robot we used has 12 joints, corresponding to 12 Degrees of Freedom (DoF), and we use positional control for the 12 DOF. Specifically, the proprioceptive input contains:

- **Joint Rotations -** $\mathbb{R}^{12\times 2}$ contains joint rotations for all joints (12D) for the past two control step.
- **Joint Velocity -** $\mathbb{R}^{12}$ contains joint velocities for all joints (12D).
- **Previous Action -** $\mathbb{R}^{12\times 2}$ contains positional command for all joints (12D) for the past two control step.
- **Projected Gravity -** $\mathbb{R}^3$ contains the projected gravity in the robot frame, representing the orientation of the robot.

The privileged information for privileged teacher policy training contains:

- **Linear Velocity -** $\mathbb{R}^3$ contains the linear velocity of the robot in the world frame.
- **Angular Velocity -** $\mathbb{R}^3$ contains the linear velocity of the robot in the world frame.
- **Environment Parameters -** $\mathbb{R}^8$ contains randomized environment parameters.

For visual observation, we provide $N = 5$ frames of the depth image (64×64) to construct the perception history. To simulate the noisy visual observation in the real world, for each time step, we randomly sample 1% pixels in $(64, 64)$ depth image and set the reading for these pixels to be the maximum reading.

2. Reward for Privileged Teacher Training

For the training of privileged teacher policy in all environments we use the same reward function as follows:

$$R = \alpha_{\text{forward}} * R_{\text{forward}} + \alpha_{\text{energy}} * R_{\text{energy}} \\ + \ \alpha_{\text{height}} * R_{\text{height}} + \alpha_{\text{amp}} * R_{\text{amp}}$$

In our experiment, we use $\alpha_{\text{forward}} = 1, \alpha_{\text{energe}} = -0.005, \alpha_{\text{height}} = -2.0, \alpha_{\text{amp}} = 1.0$

we provide specific formulations of different reward terms in our reward function

$$R_{\text{forward}} = 1 + |v_{\text{robot}} - v_{\text{target}}|/v_{\text{target}}$$

where v_{robot} is the current robot speed along the forward direction, and the $v_{\text{target}} = 0.4$ is the target moving forward speed.

$$R_{\text{energy}} = \sum_i |\tau_i \times \dot{q}_i|$$

where τ_i is the the motor torques applied to the ith joint, and the $\dot{q}_i$ is the joint velocity for the ith joint.

$$R_{\text{height}} = ||h_{\text{robot}} - h_{\text{target}}||$$

where h_{robot} is the current relative height of the robot with respect to the terrain, and the $h_{\text{target}} = 0.265$ is the target height to track.

For AMP reward, we follow the setting in Peng et al [1], where a gait discriminator is trained to distinguish the gait in the reference motions and the gait produced by the RL policy. The score (between 0 and 1) from the discriminator is used as the AMP reward. The gait closer to the reference motions is assigned a higher reward.

3. Commitment on Releasing Code

We commit to release the code of our work covering the simulated environment, policy training, and policy deployment upon the acceptance of the work. We believe our open-sourced code will be an important contribution to the community.

References

[1] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Trans. Graph.*, 40(4), July 2021. 1