

A Environment Details

A.1 RL environment

We provide a detailed training and test environment setting in this subsection.

Observation For Privileged observation, we use proprioception, including linear velocity, angular velocity, joint position, joint velocity, and last action, and task-relevant observation, including target joint positions, target keypoint positions, target root translations, and target root rotations in the global coordinates. For student policies, we use all proprioception observation, except for linear velocity. For task-relevant information, we only preserve target joint positions, root translation, and root rotations in the local coordinates. For teacher policy, we take observations from 5 timesteps as input, and for student policies, we take observations from 10 timesteps as input.

For the delta action policy, we use the full proprioception of the teacher policy mentioned above, as well as the tracking policy actions. Note that we don't use the global information like root position and keypoint positions.

Action We use roportional derivative (PD) controller to control the 23 DoF of the G1 (totally 29 DoF). And the policy outputs are the target joint position for PD controller.

Termination In addition to falls, we added an additional termination condition during the training and testing process, where the position of the keypoints must not exceed a threshold. During training, the threshold is set from 0.8 down to 0.3 using curriculum learning. During testing, a threshold of 0.8 is used for walking, and 0.4 is used for the other tasks.

Table 7: Reward design for tracking policy.

Term	Expression	Weight
Penalty		
Torque limits	$\mathbf{1}(\tau_t \notin [\tau_{\min}, \tau_{\max}])$	-10
DoF position limits	$\mathbf{1}(d_t \notin [q_{\min}, q_{\max}])$	-10
DoF velocity limits	$\mathbf{1}(\dot{d}_t \notin [\dot{q}_{\min}, \dot{q}_{\max}])$	-10
Regularization		
DoF acceleration	$\ \ddot{d}_t\ _2^2$	-3×10^{-8}
Action rate	$\ a_t - a_{t-1}\ _2^2$	-2
Action smoothness	$\ \dot{a}_t - \dot{a}_{t-1}\ _2^2$	-2
Torque	$\ \tau_t\ $	-0.0001
Stumble	$\mathbf{1}(F_{\text{feet}}^{xy} > 5 \times F_{\text{feet}}^z)$	-0.00125
Feet orientation	$\sum_{\text{feet}} \ \text{gravity}_{xy}\ $	-2.0
Task reward		
Body position	$\exp(-4 \cdot \ \hat{p}_{\text{body}} - p_{\text{ref}}\)$	1.0
Root rotation	$\exp(-4 \cdot \ q_{\text{toot}} - q_{\text{root}}\)$	0.5
Root angular velocity	$\exp(-4 \cdot \ \omega_{\text{body}} - \omega_{\text{ref}}\)$	0.5
Root velocity	$\exp(-4 \cdot \ v_{\text{root}} - v_{\text{ref}}\)$	0.5
DoF position	$\exp(-4 \cdot \ d - d_{\text{ref}}\)$	0.5
DoF velocity	$\exp(-4 \cdot \ \dot{d} - \dot{d}_{\text{ref}}\)$	0.5

A.2 Deployment

The policy runs at an inference frequency of 50 Hz. The low-level interface operates at 200 Hz, ensuring smooth real-time control. Communication between the control policy and the low-level interface is facilitated via Lightweight Communications and Marshalling (LCM).

B Training Details

B.1 Reward Design

We have listed the rewards used for training the WBC policy and the Delta Action model separately in Table 7 and Table 8, respectively. It is worth noting that when training the Delta Action model, compared to the rewards in ASAP, we used the translation of the root position rather than the positions of all body joints. This is because we did not use a motion capture system, but instead relied on odometry.

Table 8: Reward design for delta action model.

Term	Expression	Weight
Penalty		
Torque limits	$\mathbf{1}(\tau_t \notin [\tau_{\min}, \tau_{\max}])$	-10
DoF position limits	$\mathbf{1}(d_t \notin [q_{\min}, q_{\max}])$	-10
DoF velocity limits	$\mathbf{1}(\dot{d}_t \notin [\dot{q}_{\min}, \dot{q}_{\max}])$	-10
Termination	$\mathbf{1}(\text{termination})$	-200.0
Regularization		
DoF acceleration	$\ \ddot{d}_t\ _2^2$	-3×10^{-8}
Action rate	$\ a_t - a_{t-1}\ _2^2$	-2
Action norm	$\ \dot{a}_t\ _2$	-2
Torque	$\ \tau_t\ $	-0.0001
Task reward		
Root position	$\exp(-4 \cdot \ \hat{p}_{\text{root}} - p_{\text{ref}}\ ^2)$	1.0
Root rotation	$\exp(-4 \cdot \ q_{\text{root}} - q_{\text{ref}}\ ^2)$	0.5
Root angular velocity	$\exp(-4 \cdot \ \omega_{\text{root}} - \omega_{\text{ref}}\ ^2)$	0.5
Root velocity	$\exp(-4 \cdot \ v_{\text{root}} - v_{\text{ref}}\ ^2)$	0.5
DoF position	$\exp(-4 \cdot \ d - d_{\text{ref}}\ ^2)$	0.5
DoF velocity	$\exp(-4 \cdot \ \dot{d} - \dot{d}_{\text{ref}}\ ^2)$	0.5

B.2 Domain Randomization

Detailed domain randomization setups are summarized in Table 9.

B.3 RL Hyperparameters

The RL training progress is aligned with standard PPO [Schulman et al., 2017]. We provide the detailed training hyperparameters in Table 10. We also list the hyperparameters used during the distillation process in Table 11.

B.4 Delta Action

For each cluster in each iteration, we randomly sample 20 deployable motions and perform 8 rollouts in the real world. Similar to ASAP, we only train the 4 DoF of the ankles. The average duration per motion is approximately 8 seconds.

B.5 Training Resource

We used two desktop computers for training. Each was equipped with an Intel i9-13900 CPU, an NVIDIA RTX 4090 GPU, and 64 GB of RAM for policy training.

Table 9: The detailed domain randomization implementation. Our types of domain randomization include mechanical properties, external disturbances, and terrain.

Term	Value
Dynamics Randomization	
Friction	$\mathcal{U}(0.5, 1.25)$
Base CoM offset	$\mathcal{U}(-0.1, 0.1)$ m
Link mass	$\mathcal{U}(0.8, 1.2) \times \text{default}$ kg
P Gain	$\mathcal{U}(0.75, 1.25) \times \text{default}$
D Gain	$\mathcal{U}(0.75, 1.25) \times \text{default}$
Control delay	$\mathcal{U}(20, 40)$ ms
External Perturbation	
Push robot	interval = 10s, $v_{xy} = 0.5\text{m/s}$
Randomized Terrain	
Terrain type	flat, rough

Table 10: Hyperparameters for PPO

Hyperparameter	Value
Optimizer	Adam
β_1, β_2	0.9, 0.999
Learning Rate	1×10^{-4}
Batch Size	4096
Discount factor (γ)	0.99
Clip Param	0.2
Entropy Coef	0.001
Max Gradient Norm	1
Learning Epochs	5
Mini Batches	4
Value Loss Coef	1

C Model Details

MLP We use a 3-layer MLP model with hidden layer sizes of 1024, 1024, and 512, respectively. The activation function used is ELU.

Transformer Our general WBC controller is built upon the Gated Transformer-XL architecture, adapted from the open-source implementation at <https://github.com/datvodinh/ppo-transformer>.

The model integrates attention mechanisms with GRU-based gating to enhance memory retention and capture long-range temporal dependencies, making it particularly effective for sequential control tasks in reinforcement learning. The controller takes as input a sequence of 10 consecutive observations and processes them through one Transformer block. Each block employs six attention heads and has a hidden size of 128 and an embedding dimension of 204. The memory length is maintained at 10 to preserve temporal context across sequences.

D Additional Results

Expert Comparison As shown in Figure 6, we visualized the comparison between generalists and specialists across six types of clusters. The same trend can still be observed in the remaining two clusters (*Stand Mid* and *Walk Fast*). In both of these two clusters, the policy of the specialists

Table 11: Hyperparameters for DAgger

Hyperparameter	Value
Optimizer	Adam
Learning Rate	1×10^{-4}
Batch Size	4096
Max Gradient Norm	1
Learning Epochs	2
Mini Batches	4

856 outperforms that of the generalist. However, the final generalist still retains favorable properties and
857 significantly outperforms the initial generalist.

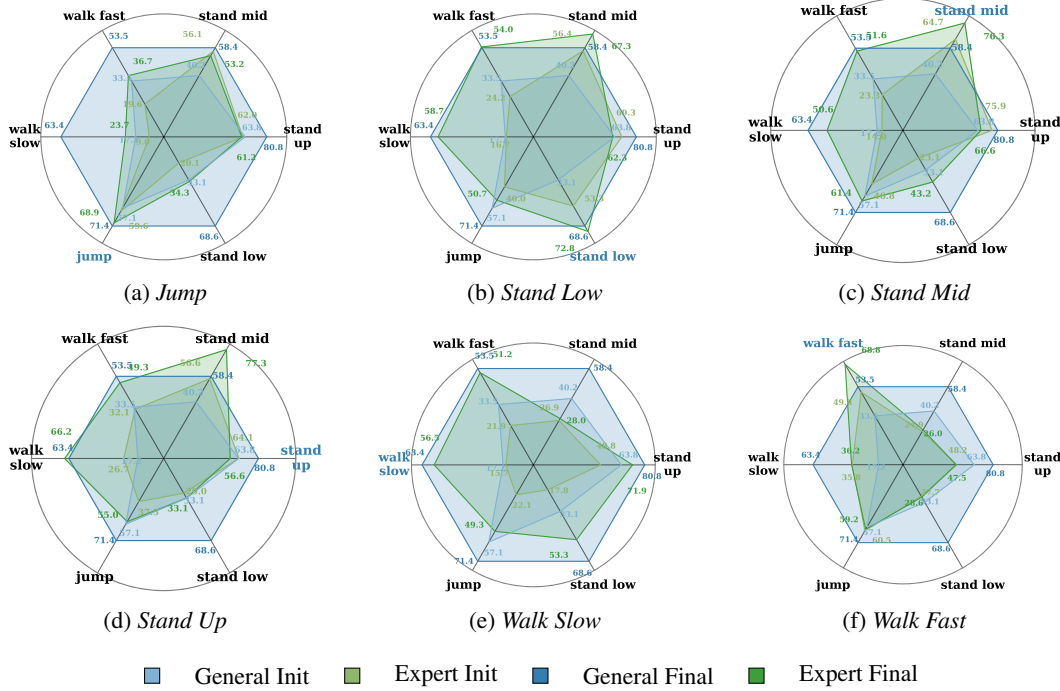


Figure 6: Complete Iterative Comparison.

858 **Clustering** We clustered a total of the following number of data samples for each category: *Jump*
859 – 351, *Stand Low* – 229, *Walk Slow* – 3355, *Stand Mid* – 578, *Stand Up* – 2378, and *Walk Fast* – 307.
860 It is evident that the data distribution across the entire AMASS dataset is imbalanced. To balance
861 the distribution for the general strategy, we ensure that each category is equally represented with a
862 ratio of 1/6 during the distillation process.

863 **Additional Video Results** More detailed real-world policy comparison will be in the link <https://anonymous.4open.science/r/BUMBLEBEE>.
864