SkillBlender: Towards Versatile Humanoid Whole-Body Control via Skill Blending

Yuxuan Kuang[1,2,3], **Amine Elhafsi**[2], **Haoran Geng**[4], **Marco Pavone**[2], **Yue Wang**[1]
[1]University of Southern California [2]Stanford University
[3]Peking University [4]University of California, Berkeley

Abstract: Humanoid robots hold significant potential in assisting humans across diverse environments and tasks thanks to their flexibility and human-like morphology. However, whole-body control remains a significant challenge, given the high-dimensional action space and the inherent instability of bipedal systems. Previous works often rely on either precise dynamic models with computationally expensive optimization or task-specific training with extensive reward tuning. In this work, we introduce **SkillBlender**, a hierarchical reinforcement learning framework that first develops a set of primitive skills using pre-designed dense rewards, and then reuses and blends these skills to accomplish more complex new tasks, requiring minimal task-specific reward engineering. Our simulated experiments on two complex loco-manipulation tasks show that our method significantly outperforms all baselines, while naturally regularizing behaviors to avoid reward hacking, resulting in more feasible and human-like movements. Website: https://sites.google.com/view/wcbm-skillblender/.

Keywords: Humanoid, Whole-Body Control, Loco-Manipulation, Bimanual Manipulation, Reinforcement Learning

1 Introduction

Humanoid robots have long held promise to be seamlessly deployed in our daily lives thanks to their flexibility and adaptability, enabled by their human-like morphology. This alignment is crucial since our environments, tasks, and tools are designed around human capabilities [1]. However, controlling humanoids remains extremely challenging due to the high-dimensional nature of their observation and action spaces, as well as the complex dynamics inherent in bipedal locomotion [2]. Even seemingly simple tasks like standing or walking present significant research challenges [1, 3].

Due to the sheer complexity of such problems, previous works focused on building dynamic models for model predictive control (MPC) [4, 5], which have shown robust performance on humanoid locomotion. However, these methods require highly accurate dynamic models tailored to each robot's embodiment, along with time-intensive optimization, limiting their scalability across different environments. Recent model-free reinforcement learning (RL) methods [6, 7, 8, 9, 10, 11] have made significant strides in agile humanoid whole-body control, benefiting from highly parallel simulation training [12, 13]. Nonetheless, those works often require task-specific training with labor-intensive reward engineering to balance terms like gait, contact, curiosity, etc. [6, 7, 8, 14], limiting their scalability to diverse real-world tasks.

In this work, we propose **SkillBlender**, to tackle the humanoid whole-body control problem by leveraging a pretrain-then-blend paradigm. Our approach leverages hierarchical reinforcement learning, where primitive expert skills are first pretrained using goal-conditioned RL. These skills are task-agnostic, reusable, and physically interpretable. Then for each specific high-level task, we train a high-level controller that generates goals for the low-level skills, as well as per-joint weight vectors to blend them. The combination of these specialized low-level skills yields more sophisticated

CoRL 2024 Workshop on Whole-Body Control and Bimanual Manipulation (CoRL 2024 WCBM).

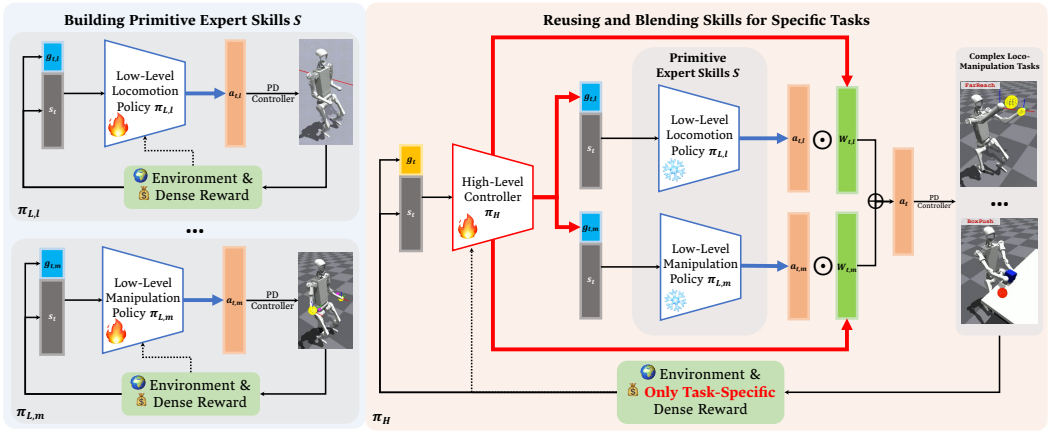

Figure 1: Framework of our proposed **SkillBlender**. We first pretrain primitive expert skills that are task-agnostic, reusable, and physically interpretable, and then reuse and blend these skills to achieve complex whole-body loco-manipulation tasks given only one or few task-specific reward terms.

behaviors, generalizing their use to broader, more complex tasks. Additionally, this architecture reduces the need for extensive reward engineering for those high-level tasks, requiring **only one or few task-specific reward terms per task** [15, 16].

We evaluate our framework on two challenging loco-manipulation tasks, `FarReach` and `BoxPush`. Our experiments in simulation show that **SkillBlender** significantly outperforms baselines in terms of accuracy and feasibility. In addition, our method avoids reward hacking and produces more natural, task-specific behaviors by leveraging pretrained skills and minimal task-related rewards. Videos are available on our website: `https://sites.google.com/view/wcbm-skillblender/`.

2 Related Works

Humanoid Whole-Body Control. Humanoid whole-body control remains extremely difficult due to its high dimensionality and unstable bipedal nature. To tackle this problem, previous non-learning-based methods focused on building dynamic models for MPC [4, 5]. However, these methods require relatively accurate dynamic modeling for each individual embodiment and require time-consuming optimization of cost functions. Recent times witnessed significant progress on learning-based methods leveraging model-free reinforcement learning [6, 7, 8, 9, 10, 11, 17, 18] for their robustness against model mismatch and uncertainties, and capability of real-time agile motions on legged robots [19, 20]. However, most of them only focused on locomotion or motion mimicking tasks and required lots of tedious reward tuning on gait, contact, curiosity, etc. on each setting [14]. Compared to those works, our method neglects the need for tedious reward engineering and only needs one or few task-specific reward terms for each task to train robust, agile and natural policies.

Hierarchical Reinforcement Learning. Hierarchical Reinforcement Learning (HRL) strategies have been used in many works to handle the complex temporal dependencies of long-horizon tasks, which are challenging for conventional RL [21, 22, 23]. HRL has also seen frequent application in quadruped loco-manipulation [24, 25, 26] and physics-based animation [27, 28, 29, 30, 31, 32, 33]. Recently [1, 3] have also shown promising results of HRL on humanoid whole-body control. However, those methods only consider one kind of low-level policy (mimicking or reaching) instead of multiple reusable skills which are more structural for complex whole-body loco-manipulation tasks. Compared to MCP [28] or ASE [30] which consider multiple skills, our method's low-level skills are physically interpretable, which are specialized and generally useful, allowing them to be reused for diverse task and motion planning (TAMP) objectives.

3 Method

3.1 Problem Formulation

We formulate our humanoid whole-body control policy learning problem as a goal-conditioned Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, \mathcal{G} \rangle$ of state $s_t \in \mathcal{S}$, action $a_t \in \mathcal{A}$, transition function $\mathcal{T}$, reward $r_t \in \mathcal{R}$, discount factor γ, and task goal $g \in \mathcal{G}$. The objective is to maximize the expected return $\mathbb{E}\left[\sum_t \gamma^t r_t\right]$ by finding an optimal policy $\pi^*(a_t|g_t, s_t)$.

In our hierarchical pipeline, we divide policies into low-level skills $\{\pi_L(a_{t,i}|g_{t,i}, s_{t,i}\})$ that output humanoid actions based on (sub)goals and states, and a high-level controller $\pi_H(\{g_{t,i}\}, \{W_{t,i}\}|g_t, s_t)$ that outputs subgoals and weight vectors for low-level policies based on the task goal and state.

3.2 Building Primitive Expert Skills

The core concept of our proposed framework is to reuse and blend primitive expert skills for new tasks, with only one or few task-specific reward terms. To achieve this, we first pretrain a set of low-level expert skills as policies $\{\pi_L\}$ using goal-conditioned RL. These policies take the current state and task-specific goal as input, and output actions of all joints for whole-body control. In this work, we consider two primitive expert skills: `Locomotion` $\pi_{L,l}$ and `Manipulation` $\pi_{L,m}$. Note that both expert skills are whole-body policies, which directly actuate all the joints on the humanoid.

For both policies, the input is divided into state s_t and goal g_t, in which s_t contains the proprioception of the humanoid (joint position, velocity, etc.). The goal of the `Locomotion` policy $g_{t,l}$ is specified as the target linear velocities in the xy (i.e., ground) plane and target yaw rate. For `Manipulation`, the goal $g_{t,m}$ is specified as the distance vector between the humanoid's hand positions and their respective targets. The output of both policies $a_t \in \mathbb{R}^d$ is the whole-body joint target positions, which are subsequently converted to torques using a PD controller.

The low-level expert policies are trained with dense rewards, including task-related goal-matching rewards, regularization rewards, gait rewards, and so on. Although reward tuning is required to train these expert skills, they are modular and reusable, which makes them amenable to blending for high-level tasks, minimizing the need for further task-specific reward engineering.

3.3 Reusing and Blending Skills for Specific Tasks

Once the expert skills are constructed, they can be reused either as plug-and-play modules for traditional TAMP, or dynamically blended for novel tasks involving complex whole-body control, guided solely by task-specific rewards. In this blending process, all expert skills are simultaneously activated, and their actions are weighted to accomplish challenging tasks beyond the capability of a single expert policy. Unlike prior multi-expert approaches [24, 28], which apply scalar weights to each expert, we employ vectorized weights, enabling more versatile and flexible skill blending.

Specifically, as shown in Fig. 1, given the task goal g_t and state s_t, we train a high-level controller π_H that takes g_t and s_t, and outputs the goals $g_{t,l}$ and $g_{t,m}$ for the low-level policies, as well as their per-joint weight vectors $W_{t,l}, W_{t,m} \in [0, 1]^d$, which are continuous and match the dimensionality of the actions. The low-level policies then concatenate the current state s_t with the subgoals generated by π_H to produce the actions $a_{t,l}$ and $a_{t,m}$. The overall action a_t is the weighted sum of $a_{t,l}$ and $a_{t,m}$, using the weights provided by π_H, in the following form:

$$a_t = a_{t,l} \odot W_{t,l} + a_{t,m} \odot W_{t,m} \tag{1}$$

where $\odot$ is the Hadamard (element-wise) product operation. In the blending process, only the high-level controller is updated and all the low-level skills remain frozen. Note that similar to [3] both the high-level policy and low-level skills are algorithmically identical but only differ in the dimension of input/output and are both trained with end-to-end RL. Notably, the blending process requires only one or few task-specific reward terms.

Task	FarReach					BoxPush				
Metrics	**Acc.** ↓	$avg(\phi, \theta)$ ↓	h ↑	τ ↓	E ↓	**Acc.** ↓	$avg(\phi, \theta)$ ↓	h ↑	τ ↓	E ↓
PPO	0.033	0.241	0.823	28.7	62.2	0.093	0.609	0.789	52.5	92.5
ExBody	0.109	0.089	0.914	18.8	43.5	0.053	**0.038**	**0.897**	16.6	12.7
Ours	**0.029**	**0.061**	**0.915**	**17.4**	**43.1**	**0.010**	0.044	0.882	**15.9**	**12.4**

Table 1: Quantitative comparison between our method and baseline methods. Our method consistently and significantly outperforms baseline methods on most metrics.

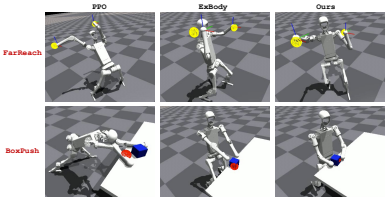

Figure 2: Qualitative results on different methods.

Task	FarReach	BoxPush
w/o Locomotion	0.451	0.039
w/o Manipulation	0.232	0.047
Scalar Weights	0.203	0.025
Ours	**0.029**	**0.010**

Table 2: Accuracy metrics on ablation methods.

4 Experiments

4.1 Experimental Setup

We conduct our experiments in the IsaacGym simulator [12], using the Unitree H1 humanoid robot. For all goal-conditioned RL in this work, we employ Proximal Policy Optimization (PPO) [13] to optimize the policy. All policy networks are implemented as end-to-end MLPs.

We consider two complex whole-body loco-manipulation tasks: `FarReach`, where the humanoid must use both hands to reach randomly placed targets (yellow spots) within a 2-meter range, and `BoxPush`, where the humanoid needs to move forward to push a box on a table to random target positions (red spots). The task rewards are straightforward, incorporating **only one or two** intuitive task-specific terms, such as the distance between the current hand positions and the target positions.

We compare our method with two baselines: vanilla PPO [13], and ExBody [7] which uses [7] as the low-level policy and trains in a Puppeteer [3] fashion. All methods are trained **with the same reward function**. We consider two kinds of metrics: accuracy (**Acc.**) as the main metric and a few proxy metrics [3], measuring task performance and behavior feasibility, respectively. Proxy metrics include average root roll/pitch angle $avg(\phi, \theta)$, root height h, joint torque τ, and joint energy E.

4.2 Results and Analysis

We show our main results in Table 1, where our method significantly outperforms all baselines across most metrics, demonstrating its clear advantages with respect to motion accuracy and naturalness. Qualitative examples are shown in Fig. 2, where our method is more accurate, natural and feasible.

The strength of our framework stems from the structural priors from the low-level expert skills that provide extra robustness and regularization, effectively reducing the RL search space and mitigating reward hacking. Compared to vanilla PPO [13], our method not only achieves better accuracy but also produces more natural and feasible behaviors, as shown by proxy metrics and qualitative results. Compared to ExBody [7], our method is more accurate, thanks to its more structured and versatile action space derived from different primitive skills and their dynamic blending.

4.3 Ablation Studies

To further investigate our framework design, we perform ablation studies on various components. As shown in Table 2, removing either $\pi_{L,l}$ or $\pi_{L,m}$ leads to severe performance degradation due to limited search space. Moreover, we modified our framework to output scalar weights, as in [28, 24], instead of per-joint weight vectors. The decreased performance highlights that vectorized weights enable more flexible skill blending than scalar weights, leading to higher accuracy.

5 Conclusions

In this paper, we introduced **SkillBlender**, a pretrain-then-blend framework for versatile and robust humanoid whole-body control. At the core of SkillBlender is to pretrain primitive skills and blend them for specific tasks, using only one or few task-related reward terms. Extensive experiments demonstrate the effectiveness of our framework. Moreover, our framework has the potential to generalize to even more challenging humanoid tasks beyond those explored in this work.

References

[1] C. Sferrazza, D.-M. Huang, X. Lin, Y. Lee, and P. Abbeel. Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506*, 2024.

[2] Y. Hurmuzlu, F. Génot, and B. Brogliato. Modeling, stability and control of biped robots—a general framework. *Automatica*, 40(10):1647–1664, 2004.

[3] N. Hansen, J. SV, V. Sobal, Y. LeCun, X. Wang, and H. Su. Hierarchical world models as visual whole-body humanoid controllers. *arXiv preprint arXiv:2405.18418*, 2024.

[4] A. Gazar, M. Khadiv, A. Del Prete, and L. Righetti. Stochastic and robust mpc for bipedal locomotion: A comparative study on robustness and performance. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 61–68. IEEE, 2021.

[5] C. Khazoom, S. Hong, M. Chignoli, E. Stanger-Jones, and S. Kim. Tailoring solution accuracy for fast whole-body model predictive control of legged robots. *IEEE Robotics and Automation Letters*, 2024.

[6] 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.

[7] 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.

[8] C. Zhang, W. Xiao, T. He, and G. Shi. Wococo: Learning whole-body humanoid control with sequential contacts. *arXiv preprint arXiv:2406.06005*, 2024.

[9] Z. Fu, Q. Zhao, Q. Wu, G. Wetzstein, and C. Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024.

[10] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024.

[11] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[12] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021.

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

[14] B. van Marum, A. Shrestha, H. Duan, P. Dugar, J. Dao, and A. Fern. Revisiting reward design and evaluation for robust humanoid standing and walking. *arXiv preprint arXiv:2404.19173*, 2024.

[15] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

[16] A. Escontrela, X. B. Peng, W. Yu, T. Zhang, A. Iscen, K. Goldberg, and P. Abbeel. Adversarial motion priors make good substitutes for complex reward functions. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 25–32. IEEE, 2022.

[17] Z. Zhuang, S. Yao, and H. Zhao. Humanoid parkour learning. *arXiv preprint arXiv:2406.10759*, 2024.

[18] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik. Humanoid locomotion as next token prediction. *arXiv preprint arXiv:2402.19469*, 2024.

[19] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

[20] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, 7(62):eabk2822, 2022.

[21] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[22] P.-L. Bacon, J. Harb, and D. Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[23] N. Heess, G. Wayne, Y. Tassa, T. Lillicrap, M. Riedmiller, and D. Silver. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv:1610.05182*, 2016.

[24] C. Yang, K. Yuan, Q. Zhu, W. Yu, and Z. Li. Multi-expert learning of adaptive legged locomotion. sci, 2020.

[25] Y. Ji, Z. Li, Y. Sun, X. B. Peng, S. Levine, G. Berseth, and K. Sreenath. Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1479–1486. IEEE, 2022.

[26] J. Zhang, N. Gireesh, J. Wang, X. Fang, C. Xu, W. Chen, L. Dai, and H. Wang. Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1399–1405. IEEE, 2024.

[27] S. Starke, H. Zhang, T. Komura, and J. Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6):178, 2019.

[28] X. B. Peng, M. Chang, G. Zhang, P. Abbeel, and S. Levine. Mcp: Learning composable hierarchical control with multiplicative compositional policies. *Advances in neural information processing systems*, 32, 2019.

[29] T. Wang, Y. Guo, M. Shugrina, and S. Fidler. Unicon: Universal neural controller for physics-based character motion. *arXiv preprint arXiv:2011.15119*, 2020.

[30] X. B. Peng, Y. Guo, L. Halper, S. Levine, and S. Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

[31] Z. Luo, J. Cao, K. Kitani, W. Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023.

[32] Z. Luo, J. Cao, J. Merel, A. Winkler, J. Huang, K. Kitani, and W. Xu. Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582*, 2023.

[33] Y. Wang, Q. Zhao, R. Yu, A. Zeng, J. Lin, Z. Luo, H. W. Tsui, J. Yu, X. Li, Q. Chen, et al. Skillmimic: Learning reusable basketball skills from demonstrations. *arXiv preprint arXiv:2408.15270*, 2024.