Adaptive Energy Regularization for Autonomous Gait Transition and Energy-Efficient Quadruped Locomotion

Boyuan Liang*, Lingfeng Sun*, Xinghao Zhu*, Bike Zhang, Ziyin Xiong, Yixiao Wang, Chenran Li, Koushil Sreenath, Masayoshi Tomizuka

***Abstract*— We investigate the impact of incorporating an energy-efficient reward term that prioritizes distance-averaged energy consumption into the reinforcement learning framework. Our findings demonstrate that this simple addition enables quadruped robots to autonomously select appropriate gaits—such as four-beat walking at lower speeds and trotting at higher speeds—without the need for explicit gait regularizations. Furthermore, we provide a guideline for tuning the weight of this energy-efficient reward, facilitating its application in real-world scenarios. The effectiveness of our approach is validated through simulations and on a real Unitree Go1 robot. This research highlights the potential of energy-centric reward functions to simplify and enhance the learning of adaptive and efficient locomotion in quadruped robots. Videos and more details are at `https://sites.google.com/berkeley.edu/efficient-locomotion`**

I. Introduction

Using gaits as guidance for locomotion policies is common in reinforcement learning (RL) methods [1], [2], as they provide rich information. However, developing a versatile and robust policy that adapts across speeds and platforms remains challenging, particularly due to the complexity of reward design. Gait references have been used as extended states and as regularization terms to aid policy learning and support low-level MPC controllers [3]. Prior works [2], [4], [5] have successfully trained policies in simulators like IsaacGym [6], [7] and deployed them on hardware, but often rely on intricate reward shaping and weight tuning. In addition to gait information, auxiliary rewards such as feet-air time and contact force penalties [4] have been used to encourage specific behaviors and stabilize training. While intended to induce particular traits, these reward terms ultimately align with the broader goal of minimizing energy consumption [8], [9]. This observation motivates a reconsideration of reward design: could a simpler, energy-centric reward replace complex, behavior-specific terms and generalize across diverse locomotion tasks, including varying speeds, directions, and terrains? Under such a formulation, gaits would naturally emerge as energy-efficient solutions to the locomotion problem.

Building on the correlation between energy-efficient gaits and speed [10], and prior findings that energy minimization at fixed speeds induces specific gaits [9], this study explores a streamlined reward formulation for energy-efficient locomotion. We focus on minimizing energy consumption without relying on intricate reward components, aiming to achieve stable and effective velocity tracking in quadruped robots across a range of speeds. Unlike previous approaches that train separate energy-optimal policies for different speeds [9], we target the development of a single energy-optimal policy capable of handling various linear and angular velocities and diverse terrains through reinforcement learning.

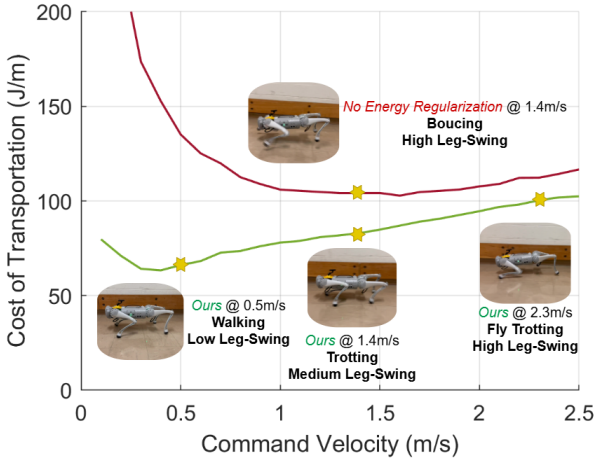


Fig. 1: Compared to the baseline when there is no energy regularization, our single policy (from one-time RL training) autonomously adopted different energy-efficient gaits (walking, trotting and fly trotting).

In this paper, we investigate the impact of incorporating a distance-averaged energy reward term into the reinforcement learning framework. This reward term directly penalizes energy consumption per unit motion traveled, promoting energy-efficient locomotion across various speeds. We explore the effect of different weightings for this energy reward, observing that both excessively low and high weights can lead to undesirable behaviors such as unnatural movements or immobility. By carefully tuning the weight of the distance-averaged energy reward, we demonstrate its effectiveness in facilitating stable velocity tracking and encouraging the emergence of energy-efficient gaits.

Employing this adaptive reward structure within IsaacGym enables the training of robust policies for the Unitree Go1 [11] quadruped robot. As illustrated in section III, our methodology identifies appropriate gaits, such as four-beat walking at lower speeds and trotting at higher speeds, without predefined gait knowledge. The energy-efficient policy also

* Equal contributions.
All authors are with the Department of Mechanical Engineering, University of California, Berkeley, California, USA, 94720.
Corresponding author: Lingfeng Sun (email: lingfengsun@berkeley.edu)

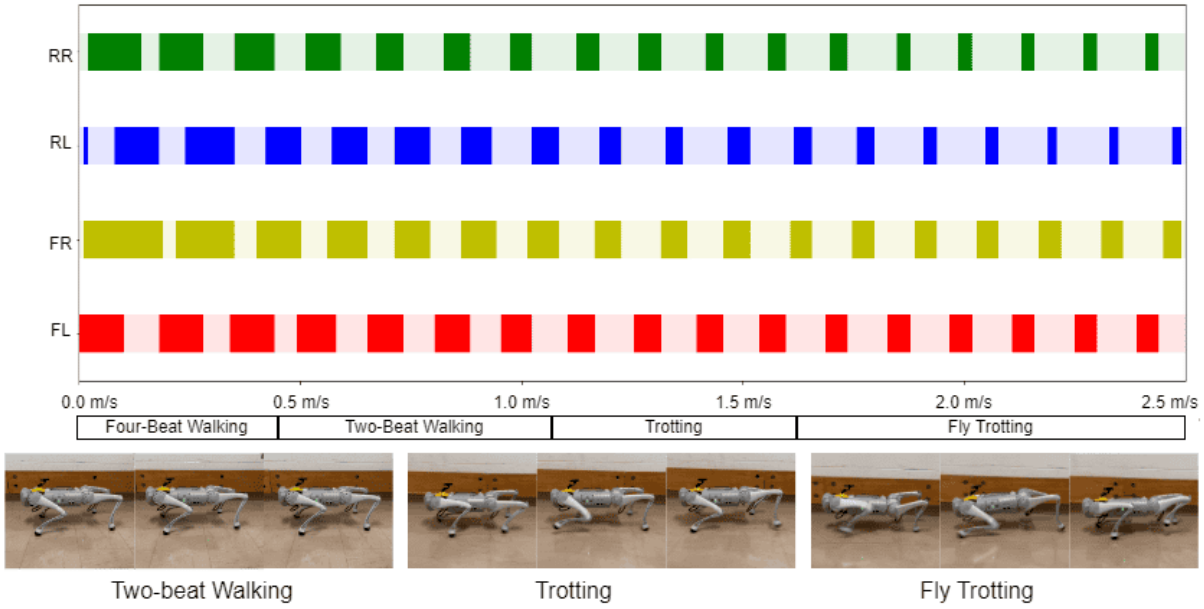


Fig. 2: Gait switching under different command velocities. The policy is generated when $\alpha_{en} = 1.0$. We also demonstrate snapshots of two beat walking at 0.5 m/s, trotting at 1.4 m/s and fly trotting at 2.3 m/s.

shows better performance in circle tracking and terrain clearance tasks. The trained single policy is deployed on a real Go1 robot to verify its stable moving and transition locomotion skills in the real world.

II. Energy Regularization

A general form of energy regularized locomotion reward takes the following form:

$$R = (R_{motion} + R_{energy}) * f(R_{aux}), \tag{1}$$

where R_{motion} encourages accurate velocity tracking, R_{energy} discourages energy consumption and R_{aux} includes other necessary rewards to stabilize training. $*$ and f are respectively an arithmetic operator and functions. Common choices are $* = +$, $f(x) = -x$ [4] and $* = \times$, $f(x) = \exp(-x)$ [2].

In previous work [9], motion rewards include negative squared linear and angular velocity tracking errors; energy rewards include negative time-averaged motor power with a fixed weight $-\tau\dot{q}$; survival bonus is also added. In experiments, we found such a training process unstable across different speeds, primarily for two reasons—*1)* The negatively unbounded nature of tracking and energy rewards. *2)* The scale of energy reward varies across different speeds, and the energy reward weight usually only works within a narrow range of reference speeds. It is hard to find a single energy reward weight value that works for all reference velocities without knowing more simulation settings and training details. As a result, different speeds are trained in different runs separately in [9] to energy-efficient gaits at different speeds.

To overcome this challenge in reward design, we proposed that the energy-related reward function should depend on the robot's velocity to promote an automatic generation of energy-efficient behavior of legged robots under various reference velocities.

$$R = (R_{motion} + \alpha_{en} R_{en}(v_x, \omega_z)) \exp(-R_{aux}) \tag{2}$$

where v_x is the robot moving velocity. The remaining of this section elaborates on each component in (2).

1) Motion Rewards: R_{motion} consists of R_{lin} and R_{ang}, which respectively encourage the legged robot to track the linear reference velocities in two directions $\hat{v}_x, \hat{v}_y$ and angular reference velocities $\hat{\omega}_z$.

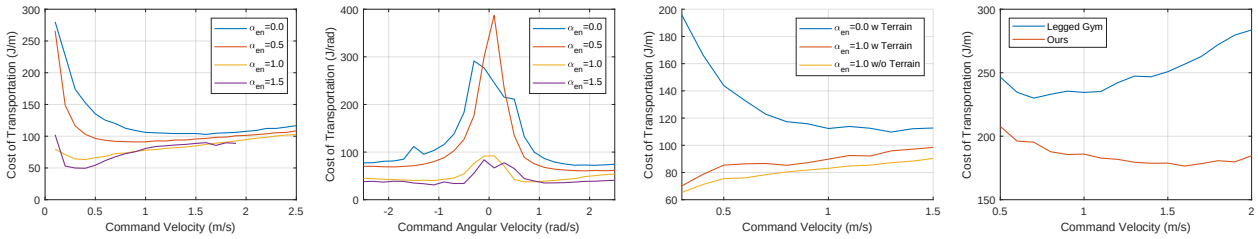
$$\begin{aligned} R_{motion} &= R_{lin} + \alpha_{ang} R_{ang}, \\ R_{lin} &= \exp\Big(-\frac{|v_x - \hat{v}_x|^2 + |v_y - \hat{v}_y|^2}{\sigma_v}\Big), \\ R_{ang} &= \exp\Big(-\frac{|\omega_z - \hat{\omega}_z|^2}{\sigma_\omega}\Big), \end{aligned} \tag{3}$$

where $\hat{v}_y$ and $\hat{\omega}_z$ are not user-specified commands, but randomly sampled during training as explained in section IV-A. σ_v and σ_ω are scaling factors depending on the training velocity range. The structure of motion rewards, the coefficient $\alpha_{ang} = 0.5$ for angular velocity tracking, and the scaling coefficients follow the default setting in legged-gym [4].

2) Energy Rewards: R_{en} rewards the system for consuming less energy while moving.

$$R_{en} = \exp\Big(-\frac{\sum_i |\tau_i||\dot{q}_i|}{\sigma_{en,x}|v_x| + \sigma_{en,z}|\omega_z|}\Big) \tag{4}$$

The energy consumption is averaged against the robot's amount of motion. τ_i's are the actuated torque at each joint while $\dot{q}_i$'s denotes the joint velocities. $\sigma_{en,x}$ and $\sigma_{en,z}$ are energy scaling constants. Components inside the round bracket of (4) is equivalent to the linear-distance-averaged energy



(a) Straight Line Walking (b) Angular Spining (c) Terrain Walking (d) Anymal-C Walking

Fig. 3: Ablation study of energy consumption in Unitree Go1 simulation. For straight line walking, reference linear velocities are chosen from 0.1 to 2.5 m/s with 0.1 common gap. The cost of transportation is measured in J/m. For angular spining, reference angular velocities are chosen from -2.5 to 2.5 rad/s with 0.2 common gap. In both (a) and (b), CoT considerably decreases when α_{en} reaches 1.0. CoT of $\alpha_{en} = 1.5$ when reference velocity of above 1.9 m/s is not plotted because the output velocity drops to zero in this range. This indicates that velocity tracking accuracy will be sacrificed when energy regularization weight α_{en} is too large. For terrain walking, the robot is asked to walk in a straight line on a rough slope terrain with reference linear velocities from 0.3 to 1.5 m/s with 0.1 common gap, because it is hard to walk either too slowly or too quickly on such terrains. (c) shows that the reduced CoT analogously appears on terrains. (d) shows the effect of energy regularization method to ANYmal-C platform with the same parameters $\sigma_{en,x}$, $\sigma_{en,z}$ and α_{en}.

consumption when $\sigma_{en,z} = 0$, which is the conventional definition of the cost of transportation (CoT). Since we are training a linear and angular velocity-dependent policy, we include rotation distance to form a generalized distance while generating the energy reward.

As discussed in [3], directly using the CoT format inside the exponential function in (4) by detecting it from motion reward fails to generate stable policies via end-to-end training. As a result, authors in [3] learn a contact schedule instead and execute it via low-level MPC. In our approach, we deployed the exponential form, which guarantees a positive reward and scale it within $(0, 1]$. We multiply the absolute values of each entry of τ with each entry of $\dot{q}$ and sum them up in (4) to follow the fact that a motor does not get charged back even when the applied torque is opposite to the motion [12]. This adaptive energy regularization term allows us to learn stable locomotion policies across different speeds, tasks, and embodiments.

One may claim that the selection of an appropriate energy reward weight α_{en} in (2) is still nontrivial since a tiny α_{en} diminishes the influence of energy regularization. At the same time, an overly large value can lead to over-enforcement, compromising velocity tracking accuracy. However, the generalized distance design calibrates the energy reward across different speeds, and the exponential function design on motion and energy reward scale both into $(0, 1]$. As a result, a $\alpha_{en} = 1$ works for all cases in our experiments, and the parameter is not that sensitive. This trade-off and detailed ablation study will be shown in Section III.

3) Auxiliary Rewards: Energy regularization alone is usually insufficient for generating proper behaviors. To address this, we add a few auxiliary regularization rewards, including collision avoidance, action rate control and trunk orientation regularization. Higher values of these terms signify less desirable performance, so we deployed a negative exponential function as detailed in (2). Comprehensive information on these auxiliary rewards can be found in the project website.

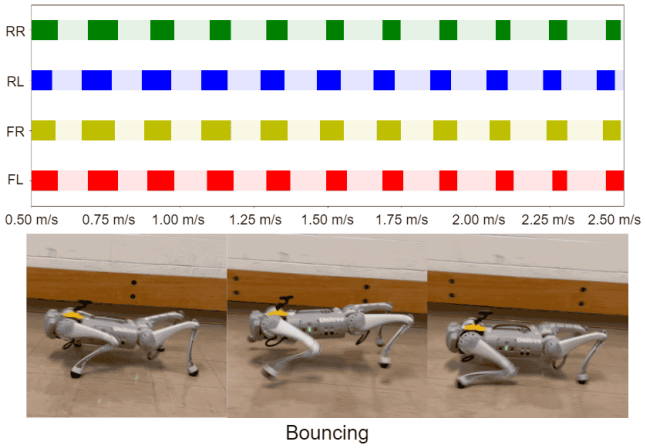


Fig. 4: Gait under different velocities when $\alpha_{en} = 0.0$.

III. Experiments

The experiments are designed to support the following statements after adding the velocity-dependent energy reward as shown in (4) without specifying gait information. Training details are given in the appendices.

- The generated RL policy automatically selects an energy efficient action at different command linear and angular velocity.
- The energy reward weight α_{en} should be comparable to motion rewards in order to get a satisfactory RL policy.
- The energy reward can also be used for locomotion training on terrains with minor amendments on the auxiliary reward R_{aux}.

A. Translation and Rotation

To demonstrate natural emergence of efficient locomotion gaits, we evaluate the trained walking policy under different reference velocities. Fig. 2 demonstrates a trial run where the legged robot was commanded to move from $\hat{v}_x = 0.0$,

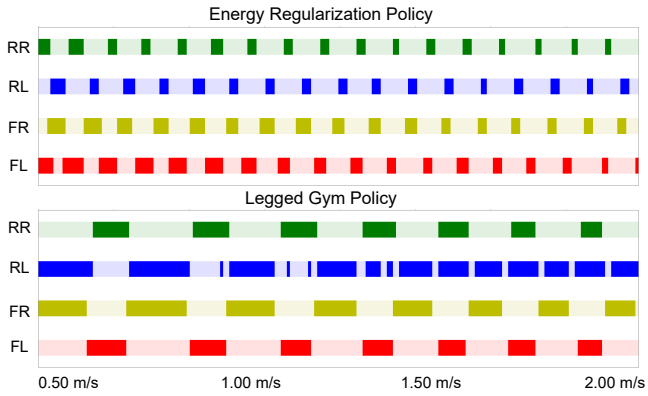
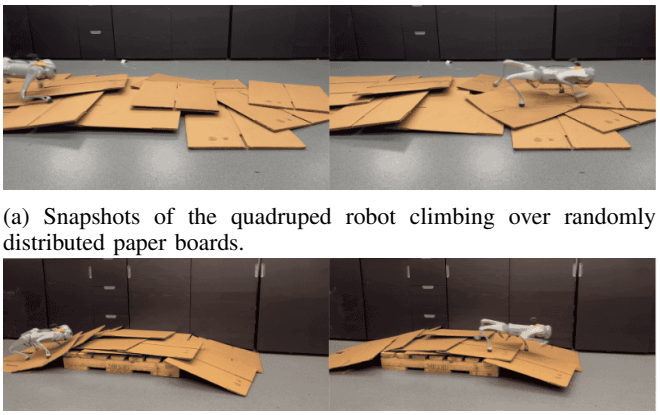


Fig. 5: Gait comparison of ANYmal-C between energy regularization and original legged gym policy [4]. In the original legged gym policy, the lifting height of rear right leg is very low, so it has several unexpected mild contacts with the ground. Videos of ANYmal-C simulation can be found on our project website.

to 2.5 m/s. The robot accelerates at $0.5m/s^2$. We plot the gait recorded from the Go1 real-world deployment. We can see that our trained policy exhibits four-beat walking at low speed (around 0.0 to 0.4 m/s), where four legs touches the ground in front-left, front-right, rear-left, rear-right sequence. Then, the robot shows two-beat walking (around 0.4 to 1.1 m/s) where the four legs touch the ground in diagonal pairs and present observable moments where four legs touches the ground at the same time. At medium speed (around 1.1 to 1.7 m/s), the policy exhibits an trotting gait. At this gait, the four legs still touch the ground in diagonal pairs, but there is neither noticeable moment where the four legs touch the ground at the same time, nor moment where all four legs are in the air. At high speed (around 1.7 m/s and beyond), the trained policy exhibits a fly-trotting gait, which is similar to trotting gait, except there are observable moments where all four legs are in the air. This gait transition is endorsed by previous works [10], [3] that walking and trotting are respectively the most energy-efficient gaits under low and high speeds.

The result in Fig. 2 is generated with $\alpha_{en} = 1.0$. In Fig. 3a and Fig. 3b, we compare CoT and linear velocity tracking results across different α_{en}. When weight is small (like 0.0 or 0.5), the effect of energy regularization is minor, so the generated policy has a much higher CoT. Fig. 4 shows that when $\alpha_{en} = 0.0$, the robot exhibits a bouncing gait across the whole velocity domain, which is not efficient [10].

The proposed energy regularization can be similarly applied to other quadruped platforms. Experiments were conducted on the ANYmal-C simulation environment [4] with the same scaling constants $\sigma_{en,x} = 1000$ and $\sigma_{en,z} = 500$ as well as the regularization weight $\alpha_{en} = 1.0$. The generated policy similarly showed preferred gait transition from walking to trotting, which is show in Fig. 5. As shown in Fig. 3d, it also successfully reduced CoT in comparison to the original legged gym settings [4].



(a) Snapshots of the quadruped robot climbing over randomly distributed paper boards.

(b) Snapshots of the quadruped robot climbing over a 20 cm step covered by paper boards.

Fig. 6: Quadruped robot clearing terrains.

B. Terrain Clearance

The effect of energy regularization on quadruped gait can also be deployed on terrain clearance. Since it is hard to walk either very fast or very slow on a rough slope, the training velocity range is limited to only $[-1.5, 1.5]$ m/s. When no energy regularization presents, the quadruped robot also tends to show a bouncing gait. When $\alpha_{en} = 1.0$, a more natural trotting gait appears. As indicated in Fig. 3c, the policy generated from $\alpha_{en} = 1.0$ is also a more energy efficient gait.

We tested the trained policy on the real Go1 quadruped robot. Fig. 6 shows experiment snapshots. The robot successfully cleared a 20 cm step covered by paper boards, which is a considerable high step compared to the size of Go1.

IV. Discussion and Conclusion

This paper presented a novel approach to energy-efficient locomotion in quadruped robots by implementing a simplified, energy-centric reward strategy within a reinforcement learning framework. Our method demonstrated that quadruped robots, specifically Unitree Go1, could autonomously develop and transition between various gaits across different linear and angular velocites without relying on predefined gait patterns or intricate reward designs. The adaptive energy reward function, adjusted based on velocity, enabled these robots to select the most energy-efficient locomotion strategies naturally.

The energy-centric training framework extends beyond locomotion tasks, where motion rewards move beyond velocity tracking to incorporate more task-specific objectives. This approach requires the system to dynamically adapt its strategy in response to varying reward structures. While this study focuses on locomotion, the underlying principle—using energy efficiency to guide behavior selection—has broader applicability. Future work could explore its extension to other robotic domains, such as manipulation and interaction tasks, where prioritizing energy efficiency may lead to the

emergence of natural and effective behaviors analogous to those seen in biological systems [13], [14]. Such an approach would also promote greater alignment between robotic system design and principles of sustainability and environmental responsibility.

ACKNOWLEDGEMENTS

This project was supported by the Agency of Science, Technology and Research of Singapore via the National Science Scholarship, the FANUC Advanced Research Laboratory, and the InnoHK of the Government of the Hong Kong Special Administrative Region via the Hong Kong Centre for Logistics Robotics.

APPENDICES

A. Training Setup in Experiments

We utilized the robot model and PPO training package in [2]. The system outputs the position command of the 12 joints in the next time step. The system inputs include the projected gravity in the robot frame, the commanded x velocities, the commanded yaw rate, each joint's current position and velocity, as well as the action at previous time step. In addition, the inputs of the previous 30 time steps are also given to the RL system. The training episode will reset after 1000 time steps or if any part of the robot except its feet touches the floor.

1) Details of Reward Parameters: In motion rewards (3), both σ_v and σ_ω were fixed at 0.25. In energy rewards (4), the energy scaling constants are fixed at $\sigma_{en,x} = 1000$ and $\sigma_{en,z} = 500$. As stated in section II, we found that the energy reward R_{en} alone is insufficient to regularize Go1's behavior, which is likely due to the lighter weight compared to its motor power. Thus, following the settings in [2], we add the fixed auxiliary reward R_{aux}. This auxiliary reward is derived mainly from safety concerns, such as penalizing limb-ground collision, out-of-range joint position, and high frequency joint action. The details of R_{aux} can be found on the project website. Compared to [2], we did not include any gait-related rewards.

2) Curriculum and Domain Randomization: We deployed curriculum technique on command velocities. The sampling range of linear and angular velocities start at $[-1, 1]$ m/s and $[-1, 1]$ rad/s. The sampling range increases when the total reward achieves a certain threshold, and the maximal sampling range is set at $[-5, 5]$ m/s and $[-5, 5]$ rad/s. The ground coefficient of friction is randomized between $[0.05, 1.5]$. We also added a disturbance to the robot mass with a uniform random value in $[-0.1, 3.0]$ kg. A uniformly distributed noise was added to the observation. All these randomized domain parameters are renewed every time the training episode resets, except observation noise is resampled after every time step.

3) Amendments for Training on Terrains: All previous techniques are deployed on flat ground trainings. We also tested adding energy regularization while training on terrains. The training terrain shape is rough slope adapted from [4].

REFERENCES

[1] J. Siekmann, Y. Godse, A. Fern, and J. Hurst, "Sim-to-real learning of all common bipedal gaits via periodic reward composition," in *IEEE International Conference on Robotics and Automation*, 2021.

[2] G. B. Margolis and A. Pulkit, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*, 2023.

[3] Y. Yang, T. Zhang, E. Coumans, J. Tan, and B. Boots, "Fast and efficient locomotion via learned gait transitions," in *Conference on Robot Learning*, 2021.

[4] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*, 2021.

[5] G. Feng, H. Zhang, Z. Li, X. B. Peng, B. Basireddy, L. Yue, Z. Song, L. Yang, Y. Liu, K. Sreenath, *et al.*, "Genloco: Generalized locomotion controllers for quadrupedal robots," in *Conference on Robot Learning*, 2023.

[6] J. Liang, V. Makoviychuk, A. Handa, N. Chentanez, M. Macklin, and D. Fox, "Gpu-accelerated robotic simulation for distributed reinforcement learning," in *Conference on Robot Learning*, 2018.

[7] 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," in *Conference on Neural Information Processing Systems*, 2021.

[8] S. Mahankali, C.-C. Lee, G. B. Margolis, Z.-W. Hong, and P. Agrawal, "Maximizing quadruped velocity by minimizing energy," in *IEEE International Conference on Robotics and Automation*, 2024.

[9] Z. Fu, A. Kumar, J. Malik, and D. Pathak, "Minimizing energy consumption leads to the emergence of gaits in legged robots," in *Conference on Robot Learning*, 2021.

[10] W. Xi, Y. Yesilevskiy, and C. D. Remy, "Selecting gaits for economical locomotion of legged robots," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1140–1154, 2016.

[11] "Unitree robotics, go1," https://www.unitree.com/products/go1, online; accessed Jun. 2022.

[12] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and D. Pathak, "Robot parkour learning," in *Conference on Robot Learning*, 2023.

[13] X. Cheng, A. Kumar, and D. Pathak, "Legs as manipulator: Pushing quadrupedal agility beyond locomotion," in *IEEE International Conference on Robotics and Automation*, 2023.

[14] E. Vollenweider, M. Bjelonic, V. Klemm, N. Rudin, J. Lee, and M. Hutter, "Advanced skills through multiple adversarial motion priors in reinforcement learning," in *IEEE International Conference on Robotics and Automation*, 2023.