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

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

***Abstract*— In reinforcement learning for legged robot locomotion, crafting effective reward strategies is crucial. Predefined gait patterns and complex reward systems are widely used to stabilize policy training. Drawing from the natural locomotion behaviors of humans and animals, which adapt their gaits to minimize energy consumption, we propose a simplified, energy-centric reward strategy to foster the development of energy-efficient locomotion across various speeds in quadruped robots. By implementing an adaptive energy reward function and adjusting the weights based on velocity, we demonstrate that our approach enables ANYmal-C and Unitree Go1 robots to autonomously select appropriate gaits—such as four-beat walking at lower speeds and trotting at higher speeds—resulting in improved energy efficiency and stable velocity tracking compared to previous methods using complex reward designs and prior gait knowledge. The effectiveness of our policy is validated through simulations in the IsaacGym simulation environment and on real robots, demonstrating its potential to facilitate stable and adaptive locomotion. Videos and more details are at `https://sites.google.com/berkeley.edu/efficient-locomotion`**

I. Introduction

Humans and animals exhibit various locomotion behaviors at different speeds, optimizing for their energy efficiency. For instance, humans typically walk at low speeds and run at higher speeds, rarely opting for jumping. Prior research demonstrated through optimal control on planar models the correlation between speed and optimal gait choices concerning the cost of transport (CoT). For quadrupeds, the optimal gaits were four-beat walking[1] at low speeds, trotting at intermediate speeds, and trotting/galloping at high speeds [1].

Due to the rich information in the gaits, using a gait as guidance for locomotion policies is popular among lots of reinforcement learning (RL) based methods [2], [3]. However, crafting a versatile and robust locomotion policy that can adapt to and transition between multiple speeds while generalizing across different platforms poses substantial challenges. One of the main challenges here is the reward design. Gait reference can be used as extended state or extra regularization terms in reward functions to provide more supervision. Previous works [3], [4], [5] trained on different quadruped robots within simulation environments like IsaacGym [6], [7] and successfully transferred these policies to physical hardware. However, they often necessitate intricate reward designs and weight tuning. Apart from gait information, reward terms like feet-air time and contact force penalizing [4] were also used to encourage specific behaviors and help stabilize the training. While these additional reward components are aimed at inducing or preventing specific behavioral traits, they inadvertently align with the broader objective of reducing energy costs. This convergence prompts a reconsideration of our reward strategy: could a more straightforward, energy-centric reward term be used to replace the specifically designed terms used in prior policy training? Such a term would encapsulate the core objective of reducing energy consumption and fostering stable and efficient locomotion.

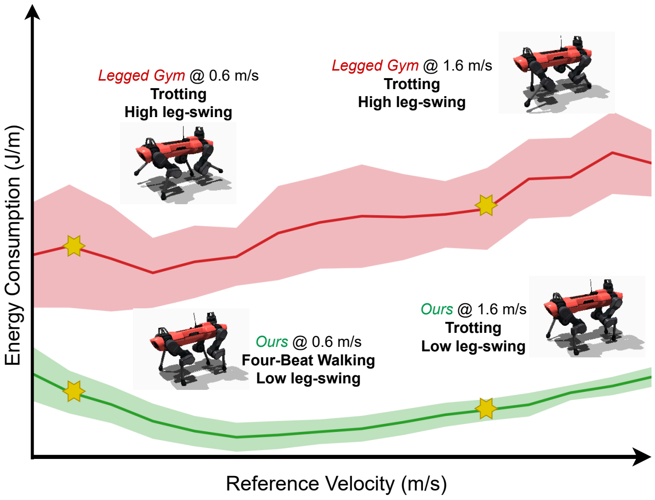


Fig. 1: Compared to legged-gym baseline [4], our single policy (from one-time RL training) autonomously adopted different energy-efficient gaits (four-beat walking and trotting). It achieved lower energy consumption (low leg-swing) at varying speeds ($0.1m/s$ interval; mean and variance from 50 independent runs).

Building on the concept that energy-efficient gaits correlate with speed [1] and aligning with prior work emphasizing that energy minimization at pre-selected speed results in the emergence of specific gaits [8], this study investigates a more streamlined reward formulation for energy-efficient locomotion. By focusing on energy minimization without intricately designed reward components, we aim to verify if such a simplified approach can yield stable and effective

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

[1] Four-beat walking, two-beat walking, trotting and galloping are typical gaits for quadruped robots defined in [1] based on feet contact schedule.

velocity-tracking in quadruped robots across various speeds. Instead of generating multiple velocity-specific energy optimal policies [8], we focus on getting a single energy-optimal policy across all target velocities via RL training.

In our research, we examine the influence of energy regularization weights on policy performance, identifying that both excessively low and high weights can lead to unnatural movements or immobility. Recognizing that energy terms have different scales across velocities and require adaptive velocity-conditioned weights, we first design a non-negative energy reward function and then find an adaptive reward form by interpolating the maximum energy weights at selected speeds to facilitate effective velocity tracking.

Employing this adaptive reward structure within IsaacGym enables the training of robust policies for the ANYmal-C [9] and Unitree Go1 [10] quadruped robots. Our methodology, illustrated in Figure 1, identifies appropriate gaits, such as four-beat walking at lower speeds and trotting at higher speeds, without predefined gait knowledge—outperforming baseline approaches [4] in energy efficiency. Our policy also significantly improves velocity tracking and energy consumption performance compared to a policy trained with fixed-weight energy rewards. The trained single policy is deployed on a real Go1 robot to verify its stable moving and transition locomotion skills in the real world.

The main contribution of this paper includes:

- Introduction of a streamlined reward formula integrating basic velocity-tracking and adaptive energy minimization to foster stable, velocity-sensitive locomotion policies.
- Demonstration that the derived policies autonomously adopt different energy-efficient gaits at varying speeds without preset gait knowledge.
- Evaluation of the velocity-tracking and energy efficiency across reward structures and weight tunings for ANYmal-C and Unitree Go1, culminating in the real-world application of these policies on a Go1 robot, affirming their efficacy in stable locomotion and gait transition.

II. Related Works

A. Reinforcement Learning for Locomotion Skills

After deep RL demonstrated its capability to fit general policies in an unsupervised manner, researchers actively sought its potential to be deployed on legged locomotion. Hwangbo et al. [11] achieved RL-generated walking policies on a plane ground using a pre-trained actuator net to reduce the sim-to-real gap. Further research also achieved training the policy with adaptive actuator net [12] or motor control parameters [13], [14]. In Hwangbo's follow-up works, locomotion on complicated terrains using a similar approach was also accomplished [15], [16], [17]. With modern GPU-accelerated simulators [6], [7], more time-efficient training frameworks were proposed [4], [18]. Due to the intensive reward engineering in these approaches, the model-free RL tends to converge to a single gait, usually trotting gait, which may not be the most efficient for all terrains and target velocities [19].

Many efforts were made to overcome this limitation by promoting the behavioral diversity of legged robots. Researchers in [20] proposed a hierarchical framework that prespecifies a set of gait primitives and allows an RL model to choose from them. In [21], [22], gait primitives were parameterized using the contact schedule and RL policy was trained to select these parameters. These works used model predictive control (MPC) as the lower-level controller, which demands preliminary knowledge of locomotion and contact modeling [23]. In [3], behavioral-related arguments such as body height, step frequency, and phase are directly added to the RL model input for end-to-end training. Although various two-beat gaits (where legs touch the ground in pairs) were realized, they cannot generate four-beat gaits (where legs touch the ground in orders) and require manual gait specification in different scenarios. It is more desirable if the quadruped robot can select the most suitable behavior by itself. In [8], a pipeline was proposed to output different gaits under different velocities via minimizing energy consumption. However, this pipeline relies on training and distillation of several velocity-specific RL policies.

B. Energy Studies on Locomotion

The energetic economy of legged robots has always been an important concern for researchers. The cost of transport, namely the amount of energy used per distance traveled, was introduced in [24], [25], where the minimization of CoT was connected to the choice of speeds [26] and step lengths [27] in human locomotion behaviors. The optimal velocity varies for different gaits, allowing animals to transit between different gaits to move at various speeds.

Conceptual legged models of bipedal [28], [29], and quadruped [30], [31], [1] robots are later designed by researchers to search for energetically optimal motions on various gaits. For quadruped robots, genetic algorithms were used in [30] to study potential gaits at different speeds. Our research is mainly inspired by [1], where optimal control problems are formulated on realistic robot models considering the effects of leg mass, plastic collisions, and damping losses. This work uses an unbiased search on energy-efficient locomotion patterns at various velocities and demonstrates the energy vs. velocity curve for different gaits. Inspired by the results, we believe there exists a policy that can transit between gaits at different velocities with an energy-optimization reward. Among previous works, [8] is close to ours in utilizing relations between energy and gaits. It generated multiple policies, with each policy velocity-specific and energy-optimal. In contrast, our work focuses on generating a single energy-optimal policy across various speeds, and it aims to replace the complex-designed reward terms in RL.

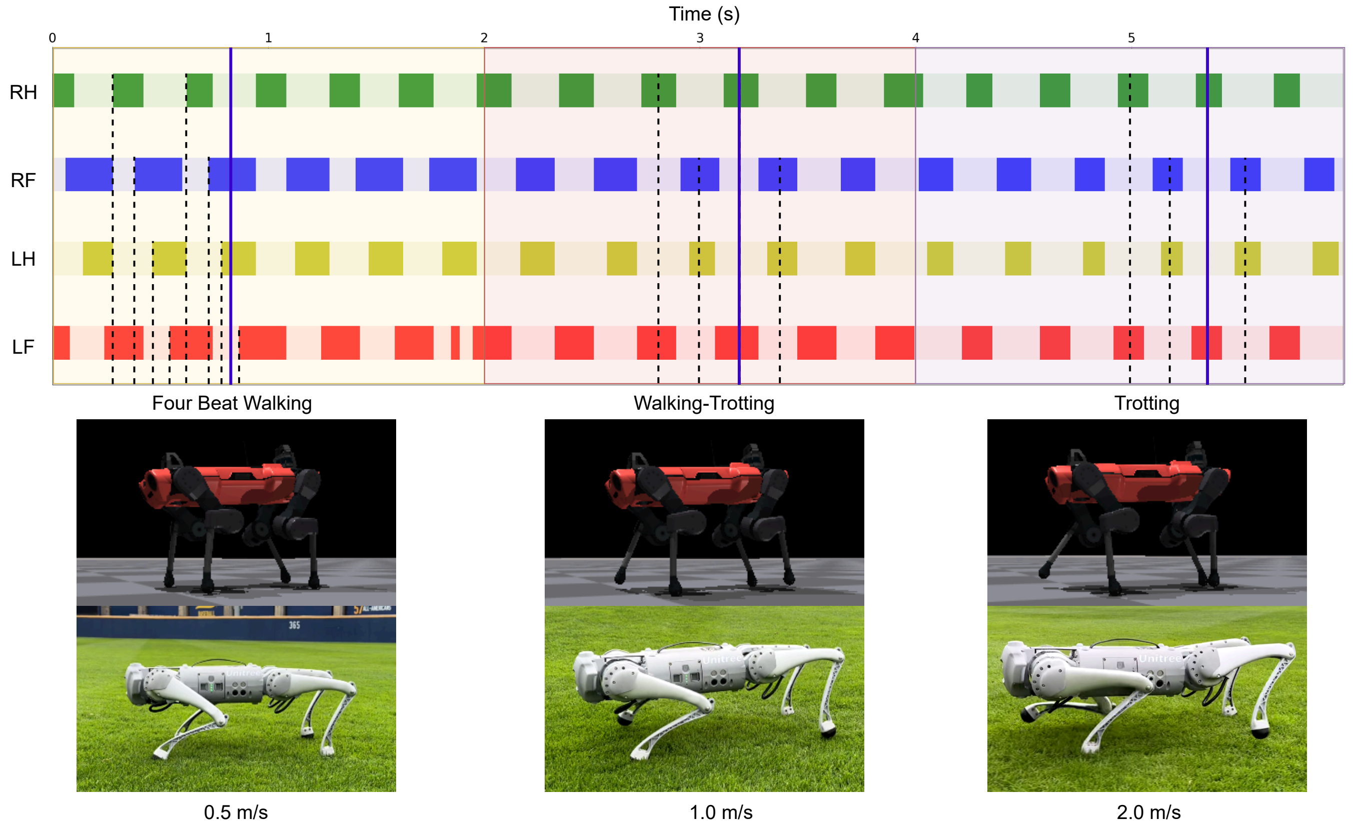


Fig. 2: **Top row**: Gait plot generated from the ANYmal-C simulation where the legged robot is first commanded to move at 0.5 m/s from 0 to 2 seconds and demonstrated four-beat walking (four legs touch the ground one by one in order, refer to [1] for details), then at 1.0 m/s from 2 to 4 seconds and demonstrated a transition gait between walking and trotting, finally at 2.0 m/s for 4 to 6 seconds and demonstrated trotting (legs touch the ground in pairs). **Middle row**: Snapshots from the ANYmal-C simulation taken approximately at the vertical purple lines in the gait plot. **Bottom row**: Snapshots of Go1 moving on a playground. Its feet contact state approximately corresponds to the vertical purple line in the gait plot.

III. Locomotion Reward Design

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

$$R = R_{motion} + R_{energy} + R_{others} \tag{1}$$

where R_{motion} encourages accurate velocity tracking, R_{energy} discourages energy consumption and R_{others} includes other necessary rewards to stabilize training. In previous work [8], motion rewards include penalty on linear and angular velocity tracking errors; energy rewards include penalty on motor power with a fixed weight; survival bonus is also added. In experiments, we found the training process unstable potentially due to the negative nature of tracking and energy rewards. Besides, each energy reward weight usually only works within a very 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, we proposed the following reward function to promote the automatic generation of energy-efficient behavior of legged robots under various reference velocities.

$$R = \frac{1}{Z(\hat{v}_x)}\Big[R_{lin} + \alpha_{ang}R_{ang} + \alpha(\hat{v}_x)R_{en}\Big] \tag{2}$$

where $\alpha_{ang} = 0.5$, $\hat{v}_x$ is the user-specified reference velocity, $\alpha(\hat{v}_x)$ is the adaptive energy reward weight, $Z(\hat{v}_x)$ is the normalizing index for total reward, R_{lin}, R_{ang} are velocity tracking rewards, and R_{en} is the energy reward. The remaining section elaborates on each component in (2).

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

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

$\hat{v}_y$ and $\hat{\omega}_z$ are not user-specified commands, but randomly sampled during training as explained in section IV. 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.

$$R_{en} = \exp\Big(- \frac{\sum_i |\tau_i||\dot{q}_i|}{\sigma_{en}} \Big) \tag{4}$$

The exponential form guarantees a positive reward. τ are each joint's actuated torques, and $\dot{q}$ are joint velocities. 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 [32]. σ_{en} is an energy scaling constant.

3) Adaptive Energy Weight: $\alpha(\hat{v}_x)$ is the weight of the energy reward terms. Previous works take this value as a constant [8], [32], but we argue that it should be adaptive to the reference velocity $\hat{v}_x$ to achieve suitable behaviors directly from RL. To verify this, we first run RL under fixed sampled $\hat{v}_x$ using pre-selected σ_{en} and try various values of $\alpha(\hat{v}_x)$. When $\alpha(\hat{v}_x)$ is too large, the robot tends to stay unmoved to save energy, neglecting the velocity tracking task. When $\alpha(\hat{v}_x)$ is too small, the robot tends to use a highly inefficient and unnatural way to walk.

To find an adaptive weight, we first collect the largest $\alpha(\hat{v}_x)$ when the velocity tracking error is smaller than a pre-defined small threshold δ for a set of speeds $\hat{v}_x$. After collecting velocity-weight sample pairs $(\hat{v}_{x,j}, \alpha(\hat{v}_{x,j}))_{j=1}^{M}$ for a range of velocities, we see a clear trend (see Figure 3): with velocity increasing, the maximum allowed α decreases. This trend corresponds to the fact that the kinetic energy increases quadratically with the velocity. While the motion rewards are expected to converge close to zero, the energy reward has variant optimal values across velocities. Therefore, we can set a larger weight for lower-speed training but need to decrease it as the velocity increases. When training velocity-conditioned locomotion policies, we linearly interpolate between selected pairs to acquire the velocity-conditioned energy weight $\alpha(\hat{v}_x)$ for the current sampled velocity. We will show in experiments that this is a simple but effective way of training energy-effective policies across different velocities.

4) Normalization Index: $Z(\hat{v}_x)$ is an adaptive normalization index. Due to the adaptive $\alpha(\hat{v}_x)$, the reward scale under different command velocities varies. This often leads to instability in RL training. As such, for each velocity-weight sample pair $(\hat{v}_{x,j}, \alpha(\hat{v}_{x,j}))$, we also record the final achieved reward $Z(\hat{v}_{x,j})$ and analogously use linear interpolation to get the normalization index curve $Z(\hat{v}_x)$ to stabilize training.

IV. Locomotion Skill Training Details

Section III focused on the reward design of the proposed method, which constitutes the most essential part of RL training. However, the energy reward is not a stand-alone one that can be used directly for RL training, and we have to use it together with the basic locomotion rewards. Across quadruped locomotion baselines, the default training settings and locomotion rewards slightly differ. This section explains the basic RL training and simulation settings other than energy regularization. These basic settings can also generate a naive locomotion policy without energy regularization, but these policies usually have low energy efficiency and might have undeployable abnormal behavior. The two baselines we use in this research are *legged-gym* [4] for ANYmal-C and *walk-these-ways* [3] for Unitree Go1.

A. ANYmal-C Settings

We utilized the robot model and PPO training package in [4]. The system outputs the position command of the 12 joints in the next time step. The system inputs include the linear and angular velocities of the trunk, projected gravity in the robot frame, the commanded x-y velocities, the commanded yaw rate, each joint's position and velocity, as well as the action at previous time step. The commanded y-velocity and yaw rate are fixed at zero here; only the commanded x-velocity will be set. The training episode will reset after 1000 time steps or if any part of the robot except its feet touches the floor.

The policy was trained on a flat ground with the coefficient of friction randomized between $[0.0, 1.5]$. We also disturbed the mass of the robot with a uniform random value in $[-5.0, 5.0]$ kg. A uniformly distributed noise was added to the observation. A random push with x-y velocity uniformly sampled between $[-1, 1]$ m/s lasting 15 seconds was exerted on the robot. All these randomized domain parameters are renewed every time the training episode resets, except observation noise is resampled after every time step.

In motion rewards (3), during fixed velocity training (to get results in Figure 3) and when the $\hat{v}_x$ is large, the policy may fail to converge to a desirable tracking accuracy. Hence, we use $\sigma_v(\hat{v}_x) = |\hat{v}_x|^2/3$ to overcome this difficulty and σ_ω is fixed at 0.25. In energy rewards (4), the energy scaling constant σ_{en} is fixed at 800. To obtain the energy weight curve $\alpha(\hat{v}_x)$ and the normalization index $Z(\hat{v}_x)$, we trained fixed-velocity policies by setting $\hat{v}_x$ at 0.5 to 2.0 m/s, with 0.1 common difference. For each fixed reference velocity, we trained 11 policies using different $\alpha(\hat{v}_x)$ values ranging from 0.5 to 4.5. Figure 3 summarizes the corresponding $\alpha(\hat{v}_x)$ and $Z(\hat{v}_x)$ for each reference velocity. These dense parameter pairs are exhibited to visualize the α-$\hat{v}_x$ relation; in real experiments, only a few pairs of parameters at representative speeds are required.

B. Go1 Settings

We mainly inherited the training methods released by [3]. Similar to [4], the system also outputs the position command of the 12 joints, but its input excludes the linear and angular velocities of the trunk. In addition, the inputs of the previous 30 time steps are also given to the RL system. Compared to ANYmal-C, 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 [3], we further add a fixed auxiliary reward R_{aux} to (2) as an amendment to the normalization index $Z(\hat{v}_x)$. 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.

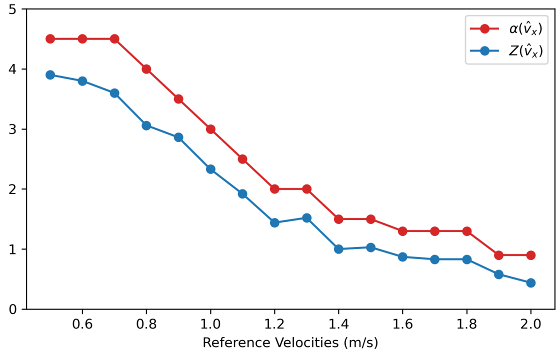


Fig. 3: $\alpha(\hat{v}_x)$ and $Z(\hat{v}_x)$ corresponding to each reference velocities. These two values are defined in (2). For each reference velocity, fixed-velocity trainings were conducted across various $\alpha(\hat{v}_x)$, then we select the largest $\alpha(\hat{v}_x)$ whose velocity tracking error is smaller than a threshold.

The details of R_{aux} can be found on the project website. The training episode will reset after 1000 time steps or if the trunk touches the floor. Similar domain randomization in section IV-A is applied here. Compared to [3], we did not include any gait-related rewards.

We also discovered that curriculum technique is essential even for fixed velocity training. Given a reference velocity $\hat{v}_x$, the sampling range of x-velocity starts with $[-\min\{\hat{v}_x, 1\}, \min\{\hat{v}_x, 1\}]$ m/s. the sampling range increases when the total reward achieves a certain threshold, and the maximal sampling range is set at $[-\hat{v}_x - 0.1, \hat{v}_x + 0.1]$ m/s. The fixed velocity training is considered a success if the x-velocity sampling range expanded to $[-\hat{v}_x - 0.1, \hat{v}_x + 0.1]$ m/s in the end and the tracking error for speed $\hat{v}_x$ is smaller than a threshold δ.

In motion rewards (3), both $\sigma_v(\hat{v}_x)$ and σ_ω were fixed at 0.25. In energy rewards (4), the energy scaling constant σ_{en} is fixed at 300. We trained fixed-velocity policies by setting $\hat{v}_x$ from 0.5 to 2.5 m/s, with a 0.5 common difference. For each reference velocity, we trained eight policies using different α values ranging from 0.7 to 2.1, then fit the $\alpha(\hat{v}_x)$ curve with the method in section III.

V. Experiments

The experiments were designed to show the following after the legged robot was trained using the adaptive energy-regularized reward shown in Equation (2).

- With adaptive $\alpha(\hat{v}_x)$ and $Z(\hat{v}_x)$, the legged robot automatically selects suitable behaviors to move with the reference velocity. We also demonstrate that the trained policy is deployable to a real quadruped robot.
- If $\alpha(\hat{v}_x)$ and $Z(\hat{v}_x)$ are constants, the legged robot may not be able to find an energy-efficient walking policy for all target velocities.
- Our method can generate a more energy-efficient walking policy compared to established baselines.

A. *One Policy with Different Gaits at Different Velocities*

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 at $\hat{v}_x = 0.5$, 1.0 and then 2.0 m/s. Each commanded velocity lasts for two seconds. We plot the gait recorded from the ANYmal-C simulation and show the snapshots for simulated ANYmal-C and real world Go1. We can see that our trained policy exhibits four-beat walking at low speed (0.5 m/s) by moving one leg each time in the order of right hind, right front, left hind and left front. At medium speed (1.0 m/s), the policy exhibits an intermediate gait between walking and trotting. At this gait, the right hind and left front legs move at the same time, while the right front and left hind legs have a displacement in their motions. At high speed (1.0 m/s), the trained policy exhibits a standard trotting gaits, where the right hind and left front legs move together, and the right front and left hind legs move together. This gait transition is endorsed by previous works [1], [21] that four-beat walking and trotting are respectively the most energy-efficient gaits under low and high speeds.

A closer look at the snapshots in Figure (2) uncovers that the swing ratio was also regularized. The legged robot only lifts up its leg to a height necessary to reach the reference velocity $\hat{v}_x$ to avoid wasting energy. At low speed, it only mildly lifts up its feet. As the reference velocity increases, the robot also lifts its feet higher, but only to a necessary height. Without considering energy-efficiency, the trained policy usually tends to lift the feet redundantly high. This will be argued in more detail in section V-C.

Finally, Figure 5 also showcases a successful hardware deployment of our policy. Videos can be found in our project website stated in the abstract.

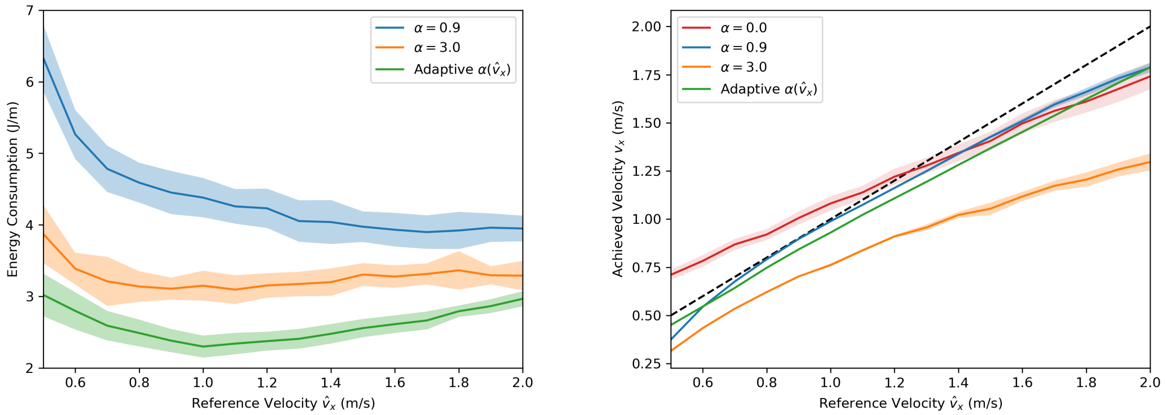
B. *Ablation Studies with Fixed Energy Rewards*

In ANYmal-C simulation, when the reference velocity is fixed at 1.0 and 2.0 m/s, we found that velocity tracking error was reasonably small when α was set as 3.0 and 0.9. Therefore, we run ablation experiments by fixing α at these values and 0.0 to compare with varying $\alpha(\hat{v}_x)$.

Figure 4a shows the energy consumption for adaptive α, $\alpha = 0.9$ and $\alpha = 3.0$. The policy with $\alpha = 0.0$ is dropped because it consumes multiple orders more energy. We observe that adaptive α reaches the lowest energy consumption. Figure 4b shows the velocity tracking error. Only the $\alpha = 0.9$ policy has comparable tracking accuracy with adaptive α, but Figure 4a shows that $\alpha = 0.9$ policy consumes considerable more energy. These observations conclude the non-triviality of finding a constant energy reward scale α and elaborates the usefulness of $\alpha(\hat{v}_x)$ varying with reference velocity.

C. *Comparison to Other Methods on Go1*

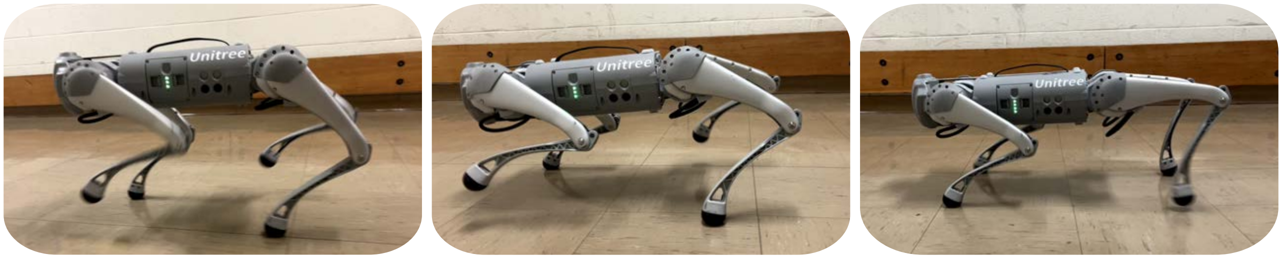
We also compare our method with built-in MPC and *walk-these-ways* on real world Go1 robot. All policies are commanded to move at 0.5 m/s. Such low velocity requires only mild movement of each leg for Go1. Figure 5 shows



(a) Comparison of energy consumption

(b) Comparison of velocity tracking error

Fig. 4: Ablation study in ANYmal-C simulation. Reference velocities are chosen from 0.5 to 2.0 m/s with 0.1 common gap. Results of each velocity are obtained from 50 independent runs. The solid line indicates the average and the shadow indicates the variance. **Left**: Energy consumption per unit moving distance under different reference velocities. We can see that adaptive energy regularization generates the most energy-efficient policy compared to fixed energy regularization. The $\alpha = 0$ policy is neglected because it consumes an uncontrollable amount of energy. **Right**: Velocity tracking error under different reference velocities. Among all fixed energy regularization policies, only $\alpha = 0.9$ has a comparable tracking error. However, the left figure shows that $\alpha = 0.9$ policy is considerably less energy efficient.



MPC @ 0.5 m/s
High CoM High Swing

WTW @ 0.5 m/s
Low CoM High Swing

Ours @ 0.5 m/s
Low CoM Low Swing

Fig. 5: Three policies walking at 0.5 m/s. Built-in MPC and *walk-these-ways* (WTW) [3] swing the legs to a redundant height, while our policy only makes necessary lifting of the leg. This demonstrates the energy-efficiency of our policy.

snapshots of each policy. It can be seen that both built-in MPC and *walk-these-ways* over-swing the legs, which squanders energy. Our policy swings the leg only to a necessary height, which can be visually recognized as the most energy-efficient.

VI. Discussions

This paper focuses on developing energy-efficient locomotion strategies for quadruped robots, employing a simple yet effective reinforcement learning approach. However, there are some inherent limitations which provide avenues for future research.

1) Limitations: The core limitation of the presented approach lies in the requirement for pre-running experiments to determine appropriate energy regularization weights. This necessity stems from our method's reliance on empirical observations to calibrate the weights, which, while effective, does not afford the flexibility in a fully adaptive reinforcement learning system. Although the policy developed is applicable across different speeds post-training, it cannot be obtained with only one training for a new quadruped platform. Moreover, we only tested its generalizability across speeds. We did not verify the adaptive capability of energy rewards in different environments, which would be crucial for deploying these robots in real-world scenarios where they might encounter multiple operational challenges.

2) Future extensions: Future research could address these limitations to develop methodologies for automatically tuning energy regularization weights within one single reinforcement learning training. This would enable the system to dynamically adjust its strategy in response to multi-task RL [33] or cross-embodiment settings [34], [35], [36].

Moreover, while this study concentrated on locomotion tasks, the underlying principle of leveraging energy efficiency to drive behavior selection holds broader potential. Future work could explore applying this energy-centric approach across different robotic tasks. For instance, manipulation and interaction tasks could also benefit from strategies prioritizing energy efficiency, potentially finding natural, efficient behaviors analogous to those observed in biological systems [37], [38]. Such a framework would align robotic systems more closely with sustainability principles and environmental consciousness.

VII. Conclusions

This paper presented a novel approach to energy-efficient locomotion in quadruped robots through the implementation of a simplified, energy-centric reward strategy within a reinforcement learning framework. Our method demonstrated that quadruped robots, specifically ANYmal-C and Unitree Go1, could autonomously develop and transition between various gaits across different velocities 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. Our policy showed energy-efficient behaviors and gait transitions in both simulation experiments (ANYmal-C) and hardware experiments (Go1). We also demonstrated the usefulness of adaptive energy regularization via ablation studies.

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.

References

[1] 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.

[2] 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.

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

[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*. PMLR, 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] 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.

[9] M. Hutter, C. Gehring, A. Lauber, F. Gunther, C. D. Bellicoso, V. Tsounis, P. Fankhauser, R. Diethelm, S. Bachmann, M. Bloesch, H. Kolvenbach, M. Bjelonic, L. Isler, and K. Meyer, "Anymal - toward legged robots for harsh environments," *Advanced Robotics*, vol. 31, no. 17, pp. 918–931, 2017.

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

[11] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Science Robotics*, vol. 4, no. 26, p. eaau5872, 2019.

[12] G. Ji, J. Mun, H. Kim, and J. Hwangbo, "Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4630–4637, 2022.

[13] Z. Li, X. Cheng, X. B. Peng, P. Abbeel, S. Levine, G. Berseth, and K. Sreenath, "Reinforcement learning for robust parameterized locomotion control of bipedal robots," in *IEEE International Conference on Robotics and Automation*, 2021.

[14] S. Chen, B. Zhang, M. W. Mueller, A. Rai, and K. Sreenath, "Learning torque control for quadrupedal locomotion," in *IEEE-RAS International Conference on Humanoid Robots*, 2023.

[15] 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*, vol. 7, no. 62, p. eabk2822, 2022.

[16] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[17] S. Choi, G. Ji, J. Park, H. Kim, J. Mun, J. H. Lee, and J. Hwangbo, "Learning quadrupedal locomotion on deformable terrain," *Science Robotics*, vol. 8, no. 74, p. eade2256, 2023.

[18] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 572–587, 2024.

[19] M. H. Raibert, "Trotting, pacing and bounding by a quadruped robot," *Journal of Biomechanics*, vol. 23, pp. 79–98, 1990.

[20] X. Da, Z. Xie, D. Hoeller, B. Boots, A. Anandkumar, Y. Zhu, B. Babich, and A. Garg, "Learning a contact-adaptive controller for robust, efficient legged locomotion," in *Conference on Robot Learning*, 2020.

[21] 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.

[22] H. Duan, A. Malik, J. Dao, A. Saxena, K. Green, J. Siekmann, A. Fern, and J. Hurst, "Sim-to-real learning of footstep-constrained bipedal dynamic walking," in *IEEE International Conference on Robotics and Automation*, 2022.

[23] S. Le Cleac'h, T. A. Howell, S. Yang, C.-Y. Lee, J. Zhang, A. Bishop, M. Schwager, and Z. Manchester, "Fast contact-implicit model predictive control," *IEEE Transactions on Robotics*, vol. 40, pp. 1617–1629, 2024.

[24] G. Gabrielli, "What price speed? specific power required for propulsion of vehicles," *Mechanical Engineering-CIME*, vol. 133, no. 10, pp. 4–5, 2011.

[25] V. Tucker, "The energetic cost of moving about: Walking and running are extremely inefficient forms of locomotion. much greater efficiency is achieved by birds, fish, and bicyclists." *American Scientist*, vol. 63, no. 4, pp. 413–419, 1975.

[26] H. Ralston, "Energy-speed relation and optimal speed during level walking." *Internationale Zeitschrift für angewandte Physiologie einschließlich Arbeitsphysiologie*, vol. 17, no. 4, pp. 277–283, 1958.

[27] B. Umberger and P. Martin, "Mechanical power and efficiency of level walking with different stride rates." *Journal of Experimental Biology*, vol. 210, no. 18, pp. 3255–3265, 2007.

[28] C. Chevallereau and Y. Aoustin, "Optimal reference trajectories for walking and running of a biped robot." *Robotica*, vol. 19, no. 5, pp. 557–569, 2001.

[29] C. D. Remy, "Optimal exploitation of natural dynamics in legged locomotion." Ph.D. dissertation, Eidgenossische Technische Hochschule., 2011.

[30] K. Kiguchi, Y. Kusumoto, K. Watanabe, I. K, and F. T, "Energy-optimal gait analysis of quadruped robots." *Artificial Life and Robotics.*, vol. 6, no. 3, pp. 120–125, 2002.

[31] A. Muraro, C. Chevallereau, and Y. Aoustin, "Optimal trajectories for a quadruped robot with trot, amble and curvet gaits for two energetic criteria." *Multibody System Dynamics*, vol. 9, no. 1, pp. 39–62, 2003.

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

[33] L. Sun, H. Zhang, W. Xu, and M. Tomizuka, "Paco: Parameter-

compositional multi-task reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 21 495–21 507, 2022.

[34] K. Zakka, A. Zeng, P. Florence, J. Tompson, J. Bohg, and D. Dwibedi, "Xirl: Cross-embodiment inverse reinforcement learning," in *Conference on Robot Learning*, 2022, pp. 537–546.

[35] L. Sun, H. Zhang, W. Xu, and M. Tomizuka, "Efficient multi-task and transfer reinforcement learning with parameter-compositional framework," *Robotics and Automation Letters*, 2023.

[36] Z.-H. Yin, L. Sun, H. Ma, M. Tomizuka, and W.-J. Li, "Cross domain robot imitation with invariant representation," in *IEEE International Conference on Robotics and Automation*, 2022, pp. 455–461.

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

[38] 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.