2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Agile Hauler Curriculum: Learning High-Speed Locomotion for Robots under Demanding Payloads

Yawen Zhou[1,2], Haopeng Tang[3,4,6], Huiqiao Fu[5], Peng Li*[1,2,3,6]

***Abstract*— Dynamic control for legged robots confronts a fundamental trilemma, where concurrent demands for high agility, substantial payload capacity and energy efficiency impose deeply coupled and often conflicting constraints. We introduce the Agile Hauler Curriculum (AHC), a learning-based method that bypasses complex mathematical modeling to address this problem. The core of AHC is an Elo-based dual-axis dynamic sampling curriculum that continuously focuses training on the agent's performance frontier, systematically pushing the robot's agility-payload performance envelope while an energy-aware gating mechanism ensures efficiency. In real-world deployment on the Go2 robot, the AHC-trained policy achieved a max speed of 2.42 m/s with a 12 kg payload, representing a 46.7% increase in speed and a 20.5% average reduction in energy consumption compared to standard grid adaptive curriculum.**

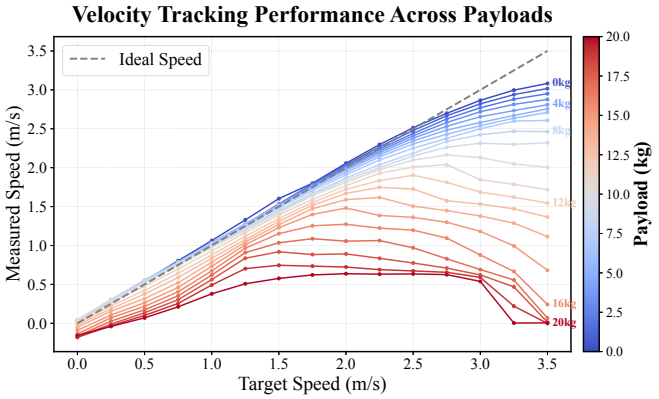

Fig. 1: Velocity Tracking Performance across Payloads.

I. INTRODUCTION

The burgeoning demand for autonomous robotic systems capable of performing dynamic interactions with the physical world is driving innovation across various fields. A critical requirement emerging from this trend is the ability for mobile robots to move under high payload capacities while maintaining significant agility and speed. Higher speeds directly enhance task execution efficiency, while lower velocity tracking errors enable more precise control. This precision is paramount for ensuring safety and stability, allowing the robot to avoid collisions and overturning while strictly adhering to its trajectory. Furthermore, it is crucial for counteracting payload-induced disturbances in high-precision tasks and minimizing unnecessary energy consumption. These advanced capabilities are poised to transform application in autonomous logistics, heavy material handling, disaster response, facilitate meticulous inspection, and defense in security scenarios.

However, achieving agile locomotion under substantial payloads remains a formidable challenge. Recent advancements have shown impressive robotic agility in tasks such as parkour [1–3], play football [4], backflip [5] or high-speed running [6] etc. However, their controllers are predominantly trained and validated under zero or low-load conditions, hindering their generalization to high-payload scenarios. Just as shown in Figure 1, higher payloads often lead to a significant rise in velocity tracking error, a problem which is exacerbated during high-speed maneuvers.

Author affiliations: [1]Institute of Software Chinese Academy of Sciences, [2]Hangzhou Institute for Advanced Study, UCAS, [3]University of Chinese Academy of Sciences, Nanjing, [4]Hohai University, [5]Nanjing University, [6]Nanjing Institute of Software Technology.
* Corresponding author: lipeng@iscas.ac.cn

Prior work on enhancing the capacity of legged robots has pursued both hardware and software innovations. Hardware-centric efforts have focused on advancing actuation systems [7–12], optimizing materials and structural designs [13–17]. Algorithmically, researchers have sought to enhance stability by refining foot-terrain interaction models and developing specialized gait generation techniques [18–21]. However, these approaches suffer from significant drawbacks. Hardware-level optimizations are often platform-specific, relying on custom designs that are difficult to generalize across different robots. Meanwhile, the few existing algorithmic solutions for payload conditions are predominantly model-based, demanding substantial engineering effort for tuning and exhibiting limited robustness to external disturbances and agility under high load.

To overcome these limitations, we introduce the Agile Hauler Curriculum (AHC), which is designed to systematically expand a quadruped's agility-payload envelope without complex mathematical modeling. Our approach is founded on the principle that efficient learning requires dynamically concentrating training effort on the agent's evolving performance boundary. By framing the curriculum as a dynamic and adaptive challenge, AHC creates a holistic training regimen that pushes the limits of speed and load capacity while simultaneously promoting energy efficiency.

Our primary contributions are summarized as follows.

- We introduce a novel curriculum that integrates game-theoretic principles into curriculum learning, framing the training process as a multiplayer game between the policy and a suite of velocity tasks. It leverages an Elo-based rating system for continuous evaluation and a sampling strategy inspired by Prioritized Fictitious Self-

Play to intelligently concentrate training on the agent's performance frontier, systematically pushing the robot's agility limits.
- We complement this core mechanism with a staged payload curriculum to systematically build load-bearing capacity, and an energy-aware reward gating function that promotes the discovery of highly efficient gaits without sacrificing peak performance.
- We demonstrate successful zero-shot sim-to-real transfer on a Go2 quadruped. The resulting policy achieves a maximum speed of 2.42 m/s with a 12 kg payload, representing a 46.7% improvement in speed and a 20.5% reduction in energy consumption over baseline.

II. RELATED WORK

A. Curriculum Learning in Robot Control

Curriculum Learning (CL) [22] incorporates a progressive training structure that enables models to gradually transition from simple tasks to challenging ones. This strategy has proven highly effective in robotics, particularly for locomotion control and sim-to-real transfer. These approaches range from automatic [23], [24] to manual [25–29] terrain curricula on geometric or physical parameters. Some works also add extra disturbances [28] while others fade external guidance, gradually removing initial assistive forces to build robustness [30] on different terrain. Curricula have also been applied to internal system parameters, utilizing Automatic Domain Randomization (ADR) to progressively expand the distribution of randomized simulation parameters [31], adapting the robot's command space [32] or transitioning from soft to hard physical constraints [3]. Among these, the work of Margolis et al. [32] is the most closely related to ours. Our approach distinguishes itself by not only introducing payload as a new curriculum dimension but also by developing a more flexible curriculum framework, which ensures that the difficulty level of training is more dynamically and precisely tailored to the agent's evolving capabilities.

B. Locomotion Control With Payload

Controlling legged robots under varying payloads presents a significant challenge, some progress has been made through biologically-inspired Central Pattern Generators (CPGs) and model-based methods like Model Predictive Control (MPC). Traditional CPGs with fixed parameters exhibit limited adaptation to varying dynamics and struggle with intricate parameter tuning. Recent efforts have sought to solve this by incorporating force feedback and online Bayesian optimization to adjust CPG parameters [33] or leveraging reinforcement learning to directly modulate oscillator parameters [34]. Similarly, traditional MPC performance degrades under unknown payloads due to model mismatch, motivating a research thrust towards adaptive MPC frameworks, some centered on online system identification: Jin et al. [35] recursively estimate payload mass and inertia to update the MPC model, while Amanzadeh et al. [36] use gradient descent for parameter estimation with guaranteed stability. Other works integrate robust control techniques: Sombolestan et al. [37] combine MPC with L1 adaptive control to compensate for uncertainties, Minniti et al. [38] embed adaptive Control Lyapunov Functions within the MPC formulation to ensure tracking stability. While these methods achieve higher payload capacity or robustness, most reported results exhibit limited agility. In contrast, our fully learning-based approach not only bypasses complex mathematical modeling but also enables more agile locomotion under heavy payloads.

C. Elo Rating System

The Elo rating system [39] is one of the most widely used approaches for evaluating the relative skill levels of players in competitive environments. Several extensions have been proposed to address Elo's limitations. The Glicko [40] and Glicko-2 systems [41] incorporate rating volatility, enabling faster adaptation to changes in player strength. Whole-History Rating (WHR) [42] treats ratings within a probabilistic framework and account for time-varying performance. Alternative ranking models, including Bradley–Terry [43] and Plackett–Luce [44], [45], also share the assumption of modeling pairwise or sequential comparisons within a probabilistic framework. Despite these advances, a common characteristic of Elo and its variants is their fundamental design around pairwise matches, which does not fit the setting of our work, where the reinforcement learning training process is modeled as a multi-player game. Therefore, we propose an adapted Elo algorithm that preserves the traditional zero-sum update while better reflecting task difficulty and policy performance.

III. METHODS

To enable agile locomotion under heavy payloads, we propose a novel curriculum learning mechanism, the Agile Hauler Curriculum (AHC). As depicted in Figure 2, AHC features a dual-axis curriculum that jointly manages velocity and payload targets through three key components: a). An Elo-based dynamic velocity curriculum that continuously adapts the task sampling distribution to keep the agent focused on its performance frontier. b). A staged payload curriculum that incrementally increases the robot's load-bearing capacity. c). An energy-aware reward gating that incentivizes the learning of energy-efficient actions without compromising agility.

A. Elo-Based Dynamic Velocity Curriculum

For a legged robot tasked with learning omnidirectional locomotion, the longitudinal and yaw velocity commands $\mathbf{v}_x^{\text{cmd}}$, ω_z^{cmd} during episode k are sampled from a probability distribution $p^k_{\mathbf{v}_x,\omega_z}(\cdot,\cdot)$. A static and uniform distribution over a wide range is usually inefficient due to the sparse rewards resulting from frequent, infeasible velocity commands. To solve this, prior work [6] proposed a grid adaptive curriculum, which enabled the velocity curriculum to expand from a limited initial range to high-speed domain automatically and progressively. The evolution of the sampling probability

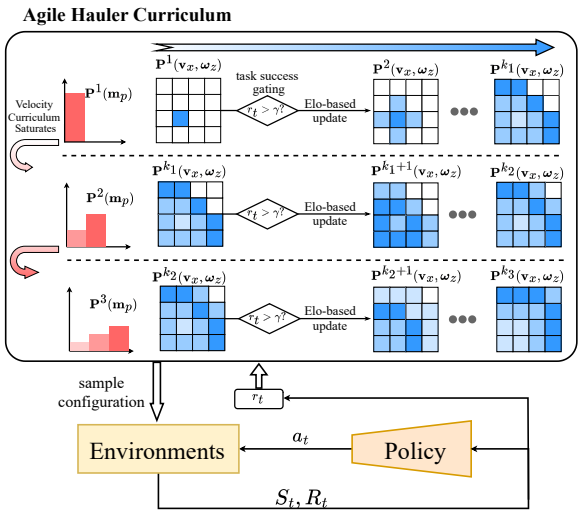

Fig. 2: Agile Hauler Curriculum Framework. This figure illustrates the core principle of AHC, visualizing the evolution of the velocity and payload curricula as they advance from left to right and subsequently from top to bottom.

distribution is governed by the following expression.

$$p^{k+1}_{\mathbf{v}_x,\omega_z}(\mathbf{v}^{\mathrm{n}}_x,\omega^{\mathrm{n}}_z) \leftarrow \begin{cases} p^{k}_{\mathbf{v}_x,\omega_z}(\mathbf{v}^{\mathrm{n}}_x,\omega^{\mathrm{n}}_z) & r_{v^{\mathrm{cmd}}_x} < \gamma \text{ or } r_{\omega^{\mathrm{cmd}}_z} < \gamma, \\ 1 & \text{otherwise.} \end{cases} \tag{1}$$

Where γ is the success threshold, with constant value between 0 and 1. $r_{v^{\mathrm{cmd}}_x}$ and $r_{\omega^{\mathrm{cmd}}_z}$ are the rewards the agent received. $(\mathbf{v}^{\mathrm{n}}_x,\omega^{\mathrm{n}}_z)$ are neighbors of $(\mathbf{v}^{\mathrm{cmd}}_x,\omega^{\mathrm{cmd}}_z)$. We retain two mechanisms from [6]: a). the velocity command space is discretized into a grid with a resolution of $[0.5\ m/s, 0.5\ rad/s]$; b). the sampling domain should expand to neighboring regions once the robot succeeds in a command area.

However, Equation 1 also reveals a key limitation of this method: the velocity sampling distribution eventually converges to be approximately uniform, forcing the agent to redundantly train on already mastered low-speed tracking tasks (as visualized in the second row of Figure 3). Although this may mitigate catastrophic forgetting, we argue that this redundancy creates a performance bottleneck that inhibits breakthroughs into more agile behaviors. Therefore, our core objective is to design a dynamic curriculum that preserves proficiency on simpler command tacking while concentrating training efforts on the most challenging tasks that offer the greatest potential for improvement.

Therefore, we define each cell in velocity grid as a unique task. Inspired by game theory, we frame the training process as a multiplayer competition between the current policy and the N active tasks (i.e., tasks with non-zero sampling probabilities), forming an $N+1$ players system. We then employ the Elo rating system to quantify both the policy's proficiency and each task's difficulty. Although tasks do not compete with each other directly, their Elo ratings are updated based on outcomes against the same policy, thus

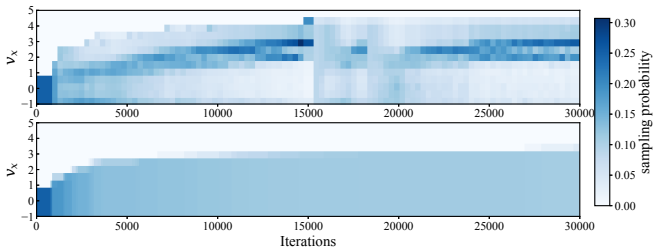

Fig. 3: Marginal distribution of sampled linear velocities during training. **Row 1**: Elo-based Dynamic Velocity Curriculum (AHC); **Row 2**: Grid Adaptive Curriculum (WTW).

reflecting their difficulty relative to the agent's skill. The policy's rating, in turn, captures its aggregate performance across all tasks, serving as a proxy for the robot's overall locomotion capability. Finally, each robot's velocity commands are sampled based on its updated Elo rating.

Elo updating begins by computing the smoothed win rate of the current policy against task i.

$$s_i = \frac{m_i + \delta}{n_i + 2\delta}, \quad i = 1,2,...,N \tag{2}$$

Where n_i denotes total sample count for task i in the current batch and m_i is success count, δ serves as a smoothing parameter. These rates are used to define a difficulty metric for the tasks (q_i) and a proficiency metric for the policy (q_p).

$$q_i = 1 - s_i, \quad q_p = q_{N+1} = \frac{1}{N}\sum_{i=1}^{N} q_i \tag{3}$$

Each player's actual score is then computed via softmax normalization.

$$S_i = \frac{exp(q_i/\tau)}{\sum_{j=1}^{N+1} exp(q_j/\tau)}, \quad i = 1,2,...,N+1 \tag{4}$$

This scoring formulation preserves the relative magnitudes of the win rates q_i, while the temperature coefficient τ controls the smoothness of the resulting scores.

For each player i, we can compute its expected score E_i with Elo score R_i before update.

$$E_i = \frac{\sum_{j\neq i} \frac{1}{1+10^{(R_j-R_i)/400}}}{(N+1)N/2}, \quad i = 1,2,...,N+1 \tag{5}$$

Because there are $(N+1)N/2$ pairs of players in total, the scaling factor ensures the scores for all $N+1$ players sum to one, which is analogous to the property of the E_i.

Finally, we aggregate the matchup data from the entire batch to perform a single, holistic update of the Elo ratings.

$$\begin{cases} R_i \leftarrow R_i + \Delta R_i \\ \Delta R_i = K \cdot N(S_i - E_i) \end{cases} \quad i = 1,\ldots,N+1 \tag{6}$$

It is straightforward to verify that when $N+1=2$ (i.e., a single task against the policy), the above calculation converges to the conventional pairwise Elo update.

The task sampling strategy is inspired by the principles of Prioritized Fictitious Self-Play (PFSP) [46], which suggests

that optimal training opponents are those challenging yet beatable. To this end, we compute expected win rate of task i against policy and get $e = \{e_1, e_2, ..., e_N\}$.

$$e_i = \frac{1}{1 + 10^{(R_p - R_i)/400}} \tag{7}$$

We then define a sampling distribution on a finite set e with weights for each task i governed by the probability density function of the Beta distribution $B(\alpha, \beta)$.

$$P(X = i) = \frac{f(e_i; \alpha, \beta)}{\sum_{j=1}^{N} f(e_j; \alpha, \beta)} \tag{8}$$

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad x \in [0, 1] \tag{9}$$

We set $\alpha = 1.8$ and $\beta = 5.0$ to create a skewed sampling distribution to encourage the agent to select moderately difficult tasks. Figure 3 illustrates the resulting difference in sampling distributions between our method and the baseline. AHC dynamically adapts the sampling distribution to consistently focus on the agent's performance frontier, thereby avoiding redundant training on mastered tasks. AHC's curriculum expands faster and broader than baseline on velocity space, which respectively indicates a greater exploration efficiency and a higher performance ceiling.

B. Staged Payload Curriculum

The payload curriculum, analogous to the velocity curriculum, is designed to follow a structure of progressively increasing difficulty. We employ a simple heuristic to divide the training process into distinct stages. Within each stage, the payload is sampled from a fixed distribution via domain randomization, allowing the agent to focus on mastering the velocity curriculum while becoming robust to the current set of payloads. The specific parameters for each stage are listed in Table I. To prevent catastrophic forgetting, the distribution of each new stage is designed to encompass the range of the preceding one. At the same time, heavier payloads are given a predominant role in the curriculum to continually push the agent's capabilities. Once the velocity curriculum saturates, the payload curriculum advances. The saturation of the velocity curriculum is defined by two criteria: the mean win rate across all tasks surpassing a threshold c, and the growth in maximum velocity sampled Δv_{max} over a time window ΔT falling below a threshold ε.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{m_i}{n_i} > c, \quad \frac{\Delta v_{max}}{\Delta T} < \varepsilon \tag{10}$$

The introduction of a new payload stage may significantly alter the robot's contact dynamics and actuator saturation limits, potentially invalidating the current velocity curriculum. However, our Elo-based mechanism adapts autonomously without manual reset. Figure 3 provides a clear example of this self-correction: the increase of payload at about 15,000 iterations causes a temporary performance drop, and thus AHC refocuses on the low-velocity region in response until the agent readapts. After about 20,000 iterations, the focus naturally shifts back to the high-speed frontier.

TABLE I: Payload Curriculum Setting

	Stage 1	**Stage 2**		**Stage 3**		
Payload	[0, 4]	[0, 4]	[4, 8]	[0, 4]	[4, 8]	[8, 12]
Percentage	1	1/3	2/3	1/6	2/6	3/6

C. Energy-Aware Gating for Efficient Agility

Whether the policy succeeds in a 'matchup' against task i is determined by its performance quantified as the mean reward obtained over an episode (r_t in Figure 2). This determination is gated by two performance metrics designed to promote efficient agility: tracking accuracy and energy efficiency. The former is derived from the linear and angular velocity tracking errors and is defined as follows:

$$\begin{aligned} r_{v_x^{cmd}} &= exp(-\frac{|\mathbf{v}_x - \mathbf{v}_x^{\mathbf{cmd}}|}{\sigma_x}) \\ r_{\omega_z^{cmd}} &= exp(-\frac{|\omega_z - \omega_z^{\mathbf{cmd}}|}{\sigma_z}) \end{aligned} \tag{11}$$

Where σ_x and σ_z are scaling parameters.

In previous work [47], minimizing energy consumption is achieved by adding an energy penalty term $-\tau^{\mathbf{T}}\dot{\mathbf{q}}$ with a fixed weight, a simple but effective technique that ensures the stability of training. However, this formulation is ill-suited as a success gate: as the loads or speed command raises, a corresponding increase in energy consumption is physically unavoidable. Consequently, a fixed threshold cannot reliably evaluate the energy efficiency. Furthermore, a static penalty weight may excessively punish high-power actions, forcing the agent into a suboptimal trade-off that sacrifices peak performance for energy conservation [26].

To address this limitation, we adopt a CoT-like reward (Equation 12) that effectively normalizes energy consumption by the robot's current velocity and payload, rather than applying a blanket penalty on high energy expenditure:

$$r_{\text{en}} = exp(-\frac{\sum_{i=1}^{12}\max(\tau_i \cdot \dot{\mathbf{q}}_i, 0)}{(m+m_p)(\lambda_x|\mathbf{v_x}| + \lambda_z|\omega_\mathbf{z}|)g}) \tag{12}$$

where τ_i and $\dot{\mathbf{q}}_i$ are the torque and velocity of the i-th joint; m and m_p are the masses of the robot body and the payload respectively; g is gravitational constant; λ_x and λ_z are scale factors. Collectively, a policy gets success on a sampled task i only when it satisfies the specified thresholds for all three reward functions defined in Equations 11-12.

D. Control Policy and Training

Although AHC is agnostic to the specific control framework, we integrate it into the training framework from [32] to learn a gait-based policy for experiments. A quadruped learn to walk and run with trotting gait for its significant advantage in high-speed locomotion.

Observation and Action Space The main objective of the controller is to track velocity commands under blind settings, using only proprioception information without exteroceptive

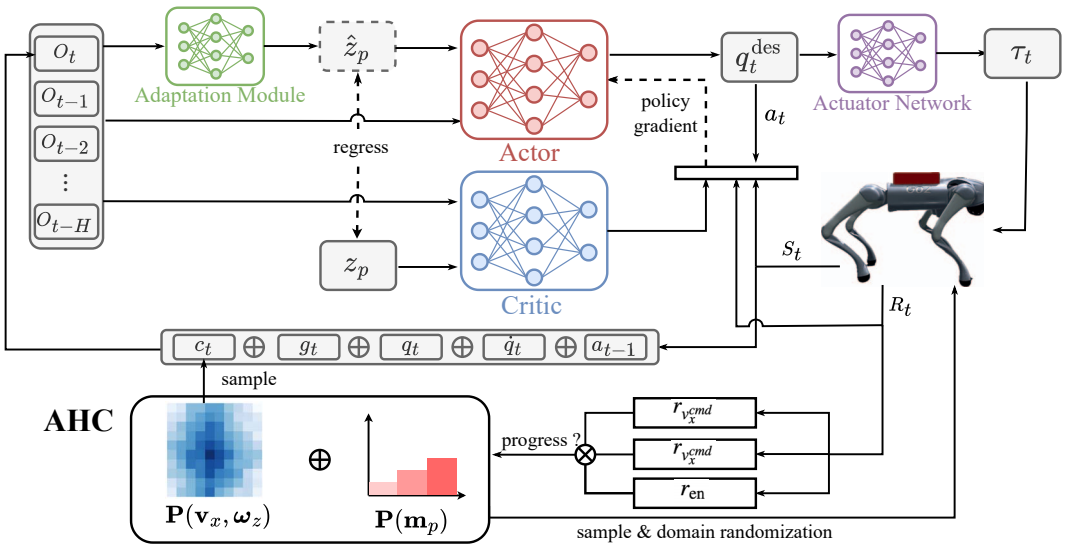

Fig. 4: Locomotion Policy Training Pipeline. Both adaptation module and actuator network are used to bridge sim-to-real gap. Velocity sampling distribution and payload randomization are governed by AHC throughout the entire training process.

sensors. The observation of every single time consists of joint angles $\mathbf{q}_t \in \mathbb{R}^{12}$, joint velocity $\dot{\mathbf{q}}_t \in \mathbb{R}^{12}$, orientation of the gravity vector $\mathbf{g}_t$, last actions $\mathbf{a}_{t-1} \in \mathbb{R}^{12}$, velocity command $\mathbf{c}_t = [\mathbf{v}_x, \mathbf{v}_y, \boldsymbol{\omega}_z] \in \mathbb{R}^3$, and behavior command $\mathbf{b}_t \in \mathbb{R}^7$ which includes body height, step frequency, pitch, roll, swing height, stance length, stance width. So $\mathbf{o}_t = [\mathbf{g}_t, \mathbf{c}_t, \mathbf{b}_t, \mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{a}_{t-1}]$. The input of the policy consists of the observation history $\mathbf{o}_{H-t:H}$ and an estimate of privileged information $\hat{z}_p \in \mathbb{R}^5$ derived from $\mathbf{o}_{H-t:H}$. z_p encompasses key physical parameters, including the ground's coefficient of friction and restitution, payload mass, and the current locomotion velocity. The policy outputs target joint angles $\mathbf{a}_t = \mathbf{q}_t^{des} \in \mathbb{R}^{12}$, which then the actuator network maps to the final motor torques $\tau_t \in \mathbb{R}^{12}$.

Policy Training An overview of our training pipeline is depicted in Figure 4. Four networks are implemented as MLPs with two or three hidden layers. The actor and critic networks are optimized using the PPO algorithm [48]. Following the methodology in [49], the Adaptation Module is trained concurrently with the policy within a single-stage learning framework.

IV. Experiments

We designed a series of experiments in simulation and the real world to validate our key design of curriculum, aiming to answer the following questions:

- **Capability**: Does our curriculum method enable the agent to surpass the performance bottlenecks of the grid adaptive curriculum, achieving superior locomotion speeds and payload capacities?
- **Energy Efficiency**: Is the energy-aware gating effective for learning energy-efficient actions?
- **Sim-to-Real Transfer**: Can the performance advantage demonstrated in simulation successfully transfer to the physical world?

A. Materials

Software & Hardware. We trained our policy in the Isaac Gym Simulator and deployed the controller on a Unitree Go2 EDU quadrupedal robot. This platform has a mass of approximately 15 kg and is equipped with 12 joint motors capable of a maximum output torque of 45 N·m. Unitree manufacturer officially reports a maximum payload capacity of approximately 8 kg and an extreme payload capacity of about 12 kg.

Baseline. We benchmark our approach against the method proposed in [6], hereafter referred to as WTW. We selected WTW as the baseline because its grid adaptive curriculum currently stands as a state-of-the-art method for training high-speed locomotion policies in quadrupedal robots.

B. Metrics

Power Power consumption (W) is measured by summing the product of joint velocity and torque at each of the 12 motors, following the methodology in [32].

$$P = \sum_{i=1}^{12} \max(\tau_i \cdot \dot{q}_i, 0) \tag{13}$$

Cost of Transport (CoT) The energy consumed by the robot per unit weight and distance, which is dimensionless.

$$\text{CoT} = \frac{\int_0^T P(t)dt}{m \cdot g \cdot d} \tag{14}$$

Velocity Tracking Error The Root Mean Square Error (RMSE) between the commanded and the actual robot velocities, measured after the system has reached a steady state.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{v}_{x,t}^{\text{cmd}} - \mathbf{v}_{x,t})^2} \tag{15}$$

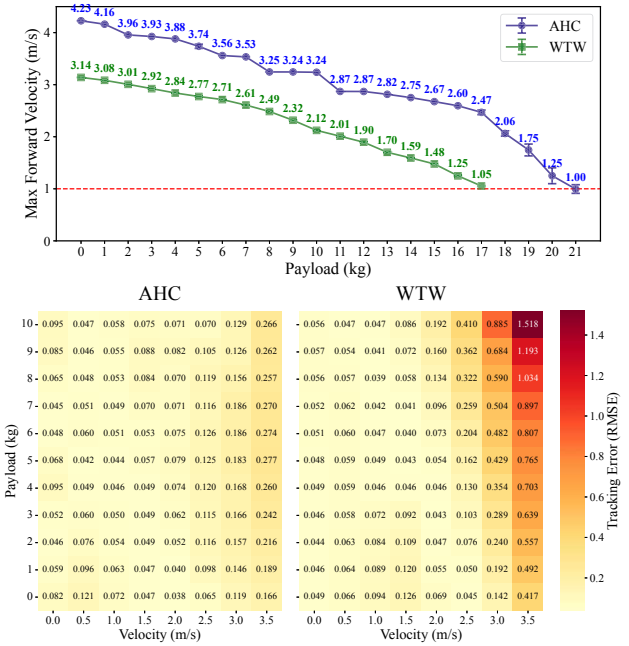

Fig. 5: Agility Metrics under Varying Payloads and Velocity Commands. **Top**: maximum forward velocity; **Bottom**: tracking error.

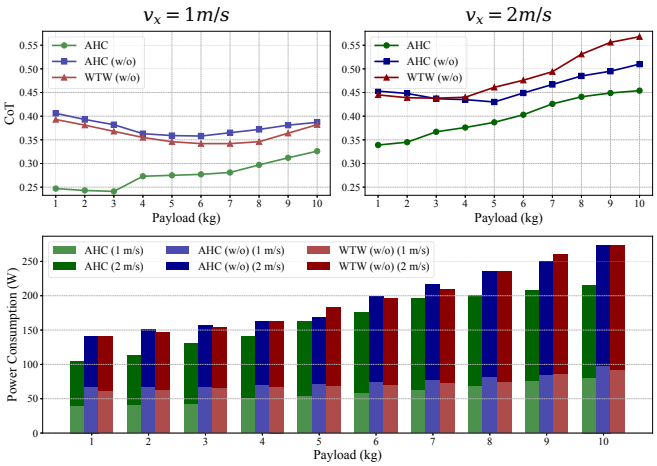

Fig. 6: The ablation study about energy reward gating, where the w/o denotes the variant excluding Equation 12 as reward gating.

V. Results

A. Agility

We evaluated the maximum achievable speed of the AHC-trained policy versus the baseline method across a range of payload conditions in simulation. To account for stochasticity, the reported simulation results are averaged over 100 independent trials. The results in Figure 5 Top demonstrate that the AHC policy achieves a higher maximum velocity than WTW across all payload conditions. Furthermore, despite both methods being trained with the same payload randomization range, the AHC policy generalizes to a maximum payload of 21 kg, surpassing WTW's 17 kg limit.

We also compare the AHC and WTW policies across a comprehensive range of commanded velocities and payloads. Figure 5 Bottom shows that both policies exhibit low tracking error in low-speed scenarios. However, as conditions become more challenging, WTW's error increases sharply, whereas AHC maintains relatively robust tracking across the full spectrum of tasks. This result confirms AHC's ability to push performance boundaries in demanding regimes while avoiding catastrophic forgetting of fundamental skills and ensuring stability where the baseline fails.

B. Energy Efficiency (Simulated)

To evaluate the effectiveness of our energy-aware reward gating, we compared the Cost of Transport (CoT) and power consumption of our full AHC method against the baseline and an ablation variant of AHC without the gating mechanism (AHC w/o). The results in Figure 6 clearly demonstrate the benefits of our approach. The top row shows that the AHC policy achieves a significantly lower CoT across all tested payloads at both 1 m/s and 2 m/s, indicating superior energy efficiency. This finding is corroborated by the power consumption data in the bottom row, where the AHC policy consistently consumes less absolute power, achieving an average energy saving of 20.5% and a peak saving of 41.16% relative to the AHC (w/o). Crucially, the energy performance of the AHC (w/o) is nearly identical to that of the WTW baseline. This ablation study confirms that the observed improvement in energy efficiency is a direct result of the proposed reward gating mechanism, rather than a side effect of the dynamic curriculum itself, validating its role in discovering energy-efficient actions.

C. Speed up Training

Finally, we analyze the exploration efficiency in training of AHC compared to baseline. Figure 7 illustrates two key aspects of the training dynamics: the policy performance (left) and the command area (right). The right panel shows that the AHC policy expands its command area at a significantly faster rate and converges to a larger area than the baseline. This indicates a more rapid and comprehensive exploration of the task space. Crucially, this accelerated exploration does not come at the expense of learning quality. As shown in the left panel, the AHC policy maintains a consistently higher average reward throughout the training process. Taken together, these results demonstrate that by intelligently focusing on tasks at the agent's performance boundary, our method enhances exploration efficiency, leading to a policy that not only learns faster but also achieves a superior final performance.

D. Sim-to-real Deployment and Analysis

We successfully deployed the zero-shot policies (AHC and WTW) on a Unitree Go2 EDU quadrupedal robot without fine-tuning and conducted real-world agility evaluation experiments on synthetic racetracks. Physical experiments in-

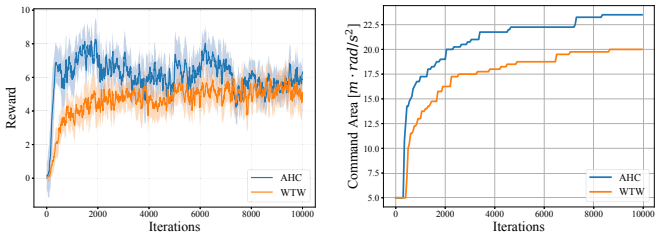

Fig. 7: Performance and Explore Efficiency during Training. Command Area is defined as area of the sampling region in the $\mathbf{v_x^{cmd}} - \boldsymbol{\omega_z^{cmd}}$ grid during training. Intuitively, a larger command area indicates controller has achieved a wider range of velocities.

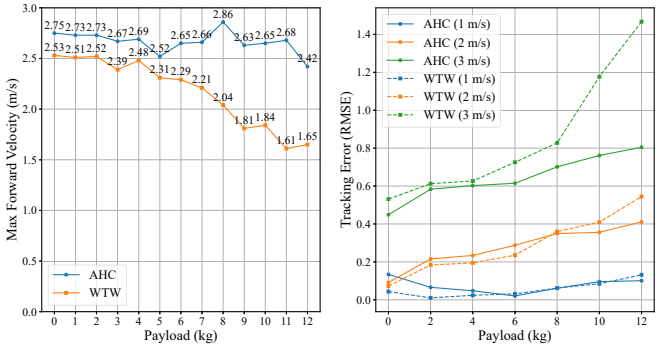

Fig. 8: Results of Real-world Experiments. **Left**: the comparison of policies' available maximum forward velocity; **Right**: velocity tracking error under relatively lower commands.

clude the achievable maximum forward velocity, the velocity tracking performance, and the maximum payload capacity.

As shown in the left part of Figure 8, the maximum forward velocity results indicate that our policy exhibited enhanced locomotion capability compared to baseline under all payload conditions, increasing by approximately 25.7% on average. In particular, AHC obtained a 46.7% improvement in maximum velocity in the 12 kg payload scenario.

We noticed that these results did not demonstrate the same continuous downward trend as the simulation results. Aside from the impact of the sim-to-real gap, we believe the main reason lies in the non-uniform task sampling distribution in training caused by AHC's mechanism, which enables the robot to perform better under greater payloads.

In velocity tracking experiments, shown in the right part of Figure 8, AHC and WTW demonstrated similar locomotion performance under low-speed or low-payload scenarios; when velocity and payload increased, AHC exhibited more accurate and stable velocity tracking capabilities than WTW. Given the target speed of 3 m/s and the payload of 12 kg, AHC achieved an average speed of 2.20 m/s (73.3% of the target) and 1.53 m/s (51.3%) for WTW. These physical experimental results are consistent with the performance in the simulator (Figure 5).

We also tested the maximum payload capacity of the policies, with the results shown in Table II. The reasons why the policies failed were different. AHC maintained its agility and stability until the payload increased to 20 kg, when the robot stopped for protection due to the joint motors exceeded their torque limits. However, WTW gradually lost its ability to steer and experienced a severe lateral deviation after the payload reached 16 kg. This difference indicates that the AHC-trained policy could successfully achieve the limit of locomotor performance while exhibiting higher payload capacity than baseline.

TABLE II: Max Payload Capacity

Average Speeds	**Payloads (kg)** 13	14	15	16	17	18	19	20
AHC	0.88	0.88	0.90	0.87	0.90	0.84	0.88	×
WTW	0.86	0.87	0.83	0.77	0.79	×	×	×

Average speeds (m/s) are measured under the command of 1 m/s.
× denotes the failure of completing the experiments.

VI. CONCLUSIONS

This work introduces AHC, a framework whose core lies in leveraging the Elo rating system to transform curriculum learning into an adaptive competition, thereby providing a robust mechanism for quantifying task difficulty relative to the agent's real-time proficiency. Our findings demonstrate that this dynamic focusing principle is vital for heavy-duty agile locomotion. Furthermore, especially in high-dimensional task spaces with deeply coupled constraints, this performance-aware method exhibits significant potential for the structured and adaptive exploration of the robot's physical limits. However, as the current framework relies on discretized task grids, its scalability to continuous, high-dimensional task representations requires further validation. Future research will explore the generalizability of the AHC paradigm across diverse robotic platforms and more intricate environments, advancing toward truly autonomous, high-performance robotic mobility in the physical world.

ACKNOWLEDGMENT

This work was supported by Science and Technology Major Project of Jiangsu Province (No.BG2024041) and Nanjing Science and Technology Plan (No.Y23002ZX01).

REFERENCES

[1] S. Luo, S. Li, R. Yu, Z. Wang, J. Wu, and Q. Zhu, "Pie: Parkour with implicit-explicit learning framework for legged robots," *IEEE Robotics and Automation Letters*, 2024.

[2] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 443–11 450.

[3] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," *arXiv preprint arXiv:2309.05665*, 2023.

[4] Y. Ji, G. B. Margolis, and P. Agrawal, "Dribblebot: Dynamic legged manipulation in the wild," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 5155–5162.

[5] C. Li, M. Vlastelica, S. Blaes, J. Frey, F. Grimminger, and G. Martius, "Learning agile skills via adversarial imitation of rough partial demonstrations," in *Conference on Robot Learning*. PMLR, 2023, pp. 342–352.

[6] 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.

[7] M. Raibert, K. Blankespoor, G. Nelson, and R. Playter, "Bigdog, the rough-terrain quadruped robot," *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 10 822–10 825, 2008.

[8] C. Semini, N. G. Tsagarakis, E. Guglielmino, M. Focchi, F. Cannella, and D. G. Caldwell, "Design of hyq–a hydraulically and electrically actuated quadruped robot," *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, vol. 225, no. 6, pp. 831–849, 2011.

[9] H.-W. Park, P. M. Wensing, and S. Kim, "High-speed bounding with the mit cheetah 2: Control design and experiments," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 167–192, 2017.

[10] B. Katz, J. Di Carlo, and S. Kim, "Mini cheetah: A platform for pushing the limits of dynamic quadruped control," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 6295–6301.

[11] J. Luo, S. Ye, J. Su, and B. Jin, "Prismatic quasi-direct-drives for dynamic quadruped locomotion with high payload capacity," *International Journal of Mechanical Sciences*, vol. 235, p. 107698, 2022.

[12] M. Hutter, C. Gehring, D. Jud, A. Lauber, C. D. Bellicoso, V. Tsounis, J. Hwangbo, K. Bodie, P. Fankhauser, M. Bloesch *et al.*, "Anymal-a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 38–44.

[13] K. Merckaert, A. De Beir, N. Adriaens, I. El Makrini, R. Van Ham, and B. Vanderborght, "Independent load carrying and measurement manipulator robot arm for improved payload to mass ratio," *Robotics and Computer-Integrated Manufacturing*, vol. 53, pp. 135–140, 2018.

[14] Z. Xu, H. Yi, D. Liu, R. Zhang, and X. Luo, "Design a hybrid energy-supply for the electrically driven heavy-duty hexapod vehicle," *Journal of Bionic Engineering*, vol. 20, no. 4, pp. 1434–1448, 2023.

[15] X. Chen, F. Gao, C. Qi, and X. Zhao, "Spring parameters design to increase the loading capability of a hydraulic quadruped robot," in *Proceedings of the 2013 international conference on advanced mechatronic systems*. IEEE, 2013, pp. 535–540.

[16] S. Feng, Y. Gu, W. Guo, Y. Guo, F. Wan, J. Pan, and C. Song, "An overconstrained robotic leg with coaxial quasi-direct drives for omni-directional ground mobility," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 477–11 484.

[17] Y. Kamikawa, M. Kinoshita, N. Takasugi, K. Sugimoto, T. Kai, T. Kito, A. Sakamoto, K. Nagasaka, and Y. Kawanami, "Tachyon: Design and control of high payload, robust, and dynamic quadruped robot with series-parallel elastic actuators," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 894–901.

[18] C. Yang, L. Ding, D. Tang, H. Gao, Z. Deng, and G. Wang, "Analysis of the normal bearing capacity of the terrain in case of foot-terrain interaction based on terzaghi theory," in *2016 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2016, pp. 443–448.

[19] C. Yang, L. Ding, D. Tang, H. Gao, L. Niu, Q. Lan, C. Li, and Z. Deng, "Improved terzaghi-theory-based interaction modeling of rotary robotic locomotors with granular substrates," *Mechanism and Machine Theory*, vol. 152, p. 103901, 2020.

[20] X. Zhao, Y. You, A. Laurenzi, N. Kashiri, and N. Tsagarakis, "Locomotion adaptation in heavy payload transportation tasks with the quadruped robot centauro," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 5028–5034.

[21] I. Dadiotis, A. Laurenzi, and N. Tsagarakis, "Trajectory optimization for quadruped mobile manipulators that carry heavy payload," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp. 291–298.

[22] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.

[23] 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.

[24] Z. Xie, H. Y. Ling, N. H. Kim, and M. van de Panne, "Allsteps: curriculum-driven learning of stepping stone skills," in *Computer Graphics Forum*, vol. 39, no. 8. Wiley Online Library, 2020, pp. 213–224.

[25] N. Heess, D. Tb, S. Sriram, J. Lemmon, J. Merel, G. Wayne, Y. Tassa, T. Erez, Z. Wang, S. Eslami *et al.*, "Emergence of locomotion behaviours in rich environments," *arXiv preprint arXiv:1707.02286*, 2017.

[26] 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*. PMLR, 2022, pp. 91–100.

[27] 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.

[28] 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.

[29] Y. Wang, Z. Jiang, and J. Chen, "Learning robust, agile, natural legged locomotion skills in the wild," *arXiv preprint arXiv:2304.10888*, 2023.

[30] B. Tidd, N. Hudson, and A. Cosgun, "Guided curriculum learning for walking over complex terrain," *arXiv preprint arXiv:2010.03848*, 2020.

[31] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas *et al.*, "Solving rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[32] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*. PMLR, 2023, pp. 22–31.

[33] G. Bellegarda and A. Ijspeert, "Cpg-rl: Learning central pattern generators for quadruped locomotion," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 547–12 554, 2022.

[34] Z. Zhang, G. Bellegarda, M. Shafiee, and A. Ijspeert, "Online optimization of central pattern generators for quadruped locomotion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 547–13 554.

[35] B. Jin, S. Ye, J. Su, and J. Luo, "Unknown payload adaptive control for quadruped locomotion with proprioceptive linear legs," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 4, pp. 1891–1899, 2022.

[36] L. Amanzadeh, T. Chunawala, R. T. Fawcett, A. Leonessa, and K. A. Hamed, "Predictive control with indirect adaptive laws for payload transportation by quadrupedal robots," *IEEE Robotics and Automation Letters*, 2024.

[37] M. Sombolestan and Q. Nguyen, "Adaptive force-based control of dynamic legged locomotion over uneven terrain," *IEEE Transactions on Robotics*, 2024.

[38] M. V. Minniti, R. Grandia, F. Farshidian, and M. Hutter, "Adaptive clf-mpc with application to quadrupedal robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 565–572, 2021.

[39] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York: Arco Publishing, 1978.

[40] M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 48, no. 3, pp. 377–394, 1999.

[41] ——, "The glicko-2 system," Boston University, Tech. Rep., 2001, technical Report.

[42] R. Coulom, "Whole-history rating: A bayesian rating system for players of time-varying strength," in *International conference on computers and games*. Springer, 2008, pp. 113–124.

[43] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[44] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 24, no. 2, pp. 193–202, 1975.

[45] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. New York: John Wiley & Sons, 1959.

[46] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[47] Z. Fu, A. Kumar, J. Malik, and D. Pathak, "Minimizing energy consumption leads to the emergence of gaits in legged robots," *arXiv preprint arXiv:2111.01674*, 2021.

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

[49] 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.