DARLEI: Deep Accelerated Reinforcement Learning with Evolutionary Intelligence

Saeejith Nair[1] **Mohammad Javad Shafiee**[1,2] **Alexander Wong**[1,2]
[1]University of Waterloo, Waterloo, Ontario, Canada
[2]Waterloo Artificial Intelligence Institute, Waterloo, Ontario, Canada
{smnair, mjshafiee, a28wong}@uwaterloo.ca

Abstract

We present DARLEI, a GPU-accelerated framework to study the interplay between parallel reinforcement learning and morphological evolution in producing emergent ecological complexity. DARLEI harnesses Isaac Gym for scalable multi-agent simulation in rich environments, enabling new research into the dynamics between individual lifetime learning and long-term evolutionary goals. Compared to prior work requiring large distributed CPU clusters, DARLEI achieves over 20x speedup using just a single workstation. We systematically characterize DARLEI's performance under various conditions, revealing factors impacting diversity of evolved morphologies. While current implementations demonstrate limited diversity over generations, we hope future work can build on DARLEI to study mechanisms for open-ended discovery. By bringing scalable accelerated simulation to this domain, DARLEI introduces a new platform to rapidly prototype and evaluate approaches at the intersection of embodied intelligence, reinforcement learning, and evolutionary computation.

1 Introduction

The existence of complex life on Earth demonstrates the creative potential of evolution. However, despite decades of research into evolutionary algorithms, modern implementations still lack the open-endedness of biological evolution [1]. While approaches like genetic programming optimize solutions for specific objectives, a key missing ingredient is the unbounded creativity allowing natural evolution to perpetually invent new solutions. One promising approach is co-evolution of interacting populations, where satisfying minimal criteria relative to a counterpart population enables open-ended discovery [1]. For example, Minimal Criterion Coevolution (MCC) [2] has shown potential by coevolving maze environments along with navigating agents. As mazes grow more complex, agents must develop new navigation strategies, which in turn facilitate further elaboration of the mazes. However, while MCC points toward a fruitful research direction, existing implementations have been limited to simple 2D gridworlds. To fully harness the creative potential of this approach, we need simulation frameworks that allow:

- Procedural generation of realistic, physics-based environments
- Evolution of diverse embodied morphologies
- Scalable distributed execution for computational efficiency
- Multi-agent interactions to study emergent ecological dynamics

Recent tools like Isaac Gym [3] and advancements in sim2real transfer [4] open up new possibilities for such a platform. In particular, the DERL framework [5] introduced a distributed system for automated design and training of embodied agents on challenging locomotion and manipulation tasks. While DERL demonstrated promising results, it requires a distributed CPU cluster, thus making it inaccessible to most researchers.

To overcome these limitations, we present the Deep Accelerated Reinforcement Learning with Evolutionary Intelligence (DARLEI) framework. DARLEI adapts DERL's core ideas into a GPU-accelerated platform using Isaac Gym, achieving over 20x speedup on a single workstation. Beyond computational acceleration, DARLEI's integration with Isaac Gym also enables future work on multi-agent coevolution in rich simulated environments.

While our current experiments only showcase locomotion tasks in simple planes, DARLEI lays the groundwork to study open-ended discovery at the intersection of evolution, embodied intelligence, and multi-agent coevolution. By bringing scalable accelerated simulation to this domain, DARLEI introduces a new platform to rapidly prototype and evaluate approaches that harness the creative potential of interacting co-evolving populations.

2 Methods

DARLEI enables large-scale evolutionary learning by combining a distributed asynchronous architecture with GPU-accelerated simulation. It builds upon the UNIMAL [5] design space and tournament selection approach of DERL, while harnessing the parallelism and speed of Isaac Gym for agent training.

2.1 System Architecture

DARLEI employs a distributed asynchronous architecture similar to DERL, with separate worker processes for population initialization, agent training, and tournament evolution. This decouples the different stages, allowing them to be parallelized across CPU and GPU resources.

The core element borrowed from DERL is the UNIMAL (UNIversal aniMAL) design space, enabling the learning of locomotion and manipulation skills in stochastic environments without needing an accurate model of the agent or environment. UNIMAL agents are hierarchical rigid-body structures, generated procedurally through mutation operations starting from a root node. This genotype generation is conceptually similar to the morphological generation proposed in Evolved Virtual Creatures [6], with the key distinction that agents are limited to 10 limbs and cyclic graphs are forbidden. Population initialization runs on the CPU, leveraging multiple processes to generate P topologically unique UNIMALs from an initial pool of $10P$ candidate morphologies. Proprioceptive force sensors are then added to "foot" limbs before serializing to a MuJoCo-based XML representation [7] on a filesystem that all nodes and workers have access to.

2.2 Agent Training

Each UNIMAL agent is trained with PPO [8] using reinforcement learning through a process called lifetime learning, where the agent learns to perform locomotion tasks over 30 million simulation steps. These steps are parallelized across M Isaac Gym environments on the GPU. We utilize Isaac Gym's default hyperparameters that were tuned for the Ant demo task. While it may be possible to reduce the number of steps required or get better results through further tuning, we leave this as future work and adopt the default setting for now.

During training, agents receive only proprioceptive observations (joint positions/velocities and force sensor data) and ego-centric exteroceptive observations (head position/velocity relative to the target). Currently, we evaluate agents on the simple environment shown in Figure 1 where the task is to move towards a fixed target at the end of a flat terrain. While adding more environments is straightforward, we focus all our current experiments on the flat terrain task and defer additional environments to future work.

The reward formulation (Equation 1) encourages forward locomotion towards the target, staying upright, and avoiding early termination. It is identical to the formulation used in Isaac Gym's Ant and Humanoid demos [3], with one key difference: the termination height is dynamically set to 50% of the agent's initial head height. This prevents excessive crawling behaviors, as noted in DERL's work [5]. The final fitness is the mean reward over the last 100,000 lifetime steps.

$$\begin{aligned} R = R_{\text{progress}} + R_{\text{alive}} \times 1(&\text{head\_height} \\ \geq &\text{ termination\_height}) + R_{\text{upright}} + R_{\text{heading}} \\ &+ R_{\text{effort}} + R_{\text{act}} + R_{\text{dof}} + R_{\text{death}} \times 1(\text{head\_height} \\ \leq &\text{ termination\_height}) \end{aligned} \tag{1}$$

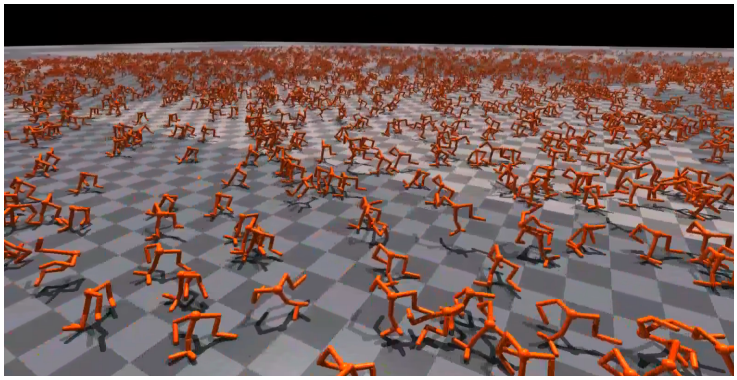

Fig. 1: Overhead view of 8192 agents in Isaac Gym simulation.

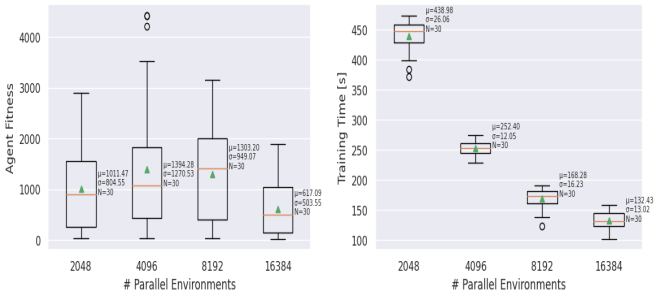

Fig. 2: Impact of parallel environments on agent fitness and training time. To enable a fair comparison [3], the horizon length is decreased proportionally with more environments to ensure that the overall experience an RL agent observes is constant. Specifically, we use horizon lengths of 64, 32, 16, and 8 for 2048, 4096, 8192, and 16384 environments respectively. Results were collected based on lifetime learning across 30 agents from the initial population.

2.3 Tournament Evolution

Once the initial population has been trained, tournament evolution begins asynchronously across W parallel worker processes. Each worker repeatedly samples 4 agents uniformly at random from the range $[T \cdot G, Q]$, where G is the current generation number, T is the number of tournaments per generation, P is the initial population size, and Q is the current total number of evolved agents. We compute $G = \lfloor (Q-P)/T \rfloor$ to determine the generation number. The 4 sampled agents then participate in a tournament where the agent with the highest fitness wins. Fitness values for each agent are computed once during its initial lifetime learning phase. The winning agent undergoes mutation by randomly selecting and applying a modification from the UNIMAL design space, including deleting a limb, adding new limbs, or altering limb parameters like length, angle, density, etc. The mutated child agent is added back into the population for future tournaments. This evolutionary loop continues until a maximum of 10 generations, imposed due to time constraints.

Similar to DERL, we employ an aging criteria based on the range R to maintain population diversity and ensure robustness to initially lucky genotypes. Aging provides a more egalitarian approach compared to directly culling low fitness agents, since all agents eventually succumb to old age regardless of fitness. Basing aging on completed generations rather than raw population size also improves fault tolerance. If a worker fails, new workers can be added without affecting the current population. Consequently, our population size temporarily exceeds P until the next generation completes. With fewer workers than DERL in our experiments, postponing aging until after full generations are completed allows more mutations per agent compared to aging prematurely based on raw population size alone.

We conducted all experiments on a workstation with 2x NVIDIA A6000 GPUs and a 32-core AMD Ryzen Threadripper PRO 3955WX CPU. To ensure accurate benchmarking, no other applications were active during the experiments.

3 Results

To evaluate DARLEI's capabilities, we conducted experiments analyzing its performance, scalability, and the quality of evolved solutions.

3.1 Scalability via Parallel Environments

A key advantage of DARLEI is its ability to leverage large numbers of parallel environments during training to achieve significant speedups. As shown in Figure 2, increasing the number of environments reduces training time, with 16384 environments providing over 3.3x faster training than 2048. However, using too many environments can negatively impact final agent fitness if the horizon is not sufficiently long, as the RL objective becomes short-term focused [3]. Based on these results, we select 8192 parallel environments for the remainder of our experiments as it provides the best compromise between training time and agent fitness.

Furthermore, we measure the total time for a complete evolutionary run ($P = 100, T = 50, W = 10$) evolving 600 morphologies. Across 4 trials, DARLEI takes (205±8) minutes, or 3.41 minutes per agent per worker. In contrast, DERL requires 16 hours for 4000 morphologies using 288 workers, equating to 69.12 minutes per agent per worker. Thus, DARLEI provides a significant 20.3x speedup over DERL using just a single workstation. Additional compute nodes can further reduce the total time exponentially.

3.2 Impact of Simulation Parameters

To understand how simulation parameters can impact learning, we investigated the effect of varying the environment radius (Figure 4). As shown in Figure 3, larger radii improve median fitness by allowing further exploration before termination, which usually happens when the agent loses balance or collides with another agent. However, moderately smaller radii introduce collisions earlier, driving more robust policies. For instance, agents drawn from the peak in outlier fitness at a 2m radius exhibit agile behaviors like high-jumping and cartwheeling to avoid collisions, suggesting this "sweet spot" radius promotes such strategies. However, the tradeoff is that smaller radii increase reset frequency and training time due to more frequent termination. Overall, the optimal radius balances robustness gains from collisions against exploration benefits of larger areas. While larger radii maximize median fitness, moderate radii may better discover diverse survival strategies.

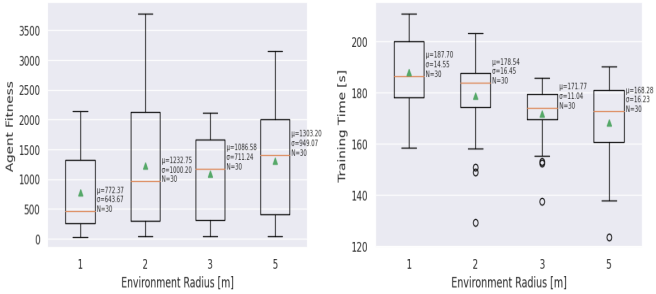

Fig. 3: Impact of environment radius on agent fitness (left) and training time (right). Results based on 30 agents from the initial population in a simulation with 8192 parallel environments and horizon length of 16.

3.3 Quality of Generated Solutions

The quality of evolved solutions is analyzed by examining mutation cycles, defined as the number of times an agent has been mutated. Four experiments are conducted with varying population sizes, tournament counts, and asynchronous worker processes.

The results reveal two key insights. First, mutations are generally harmful rather than beneficial. The top plot in Figure 7 shows

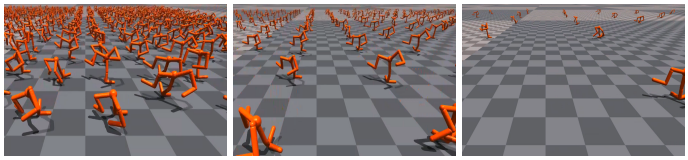

Fig. 4: Environments with radii of 1m, 2m, and 5m.

Fig. 5: Population diversity is seen to decrease over generations. Lines trace rewards of agents' lineages.

agent fitness increasing with more mutations, falsely implying mutations improve fitness. However, the center plot shows that in most experiments, mutations actually reduce fitness between ancestors and descendants on average. The fitness increase in the top plot stems from selection bias - fitter ancestors reproduce more, accumulating additional mutations.

Second, diversity collapses rapidly over generations as shown in Figure 5. All final agents descend from just two initial ancestors, despite starting with a diverse population. Even with more workers, convergence emerges quickly.

In essence, while selection propagates fit solutions, mutations degrade fitness without mechanisms to maintain diversity. The lack of sustained open-ended evolution implies further techniques are needed to drive ongoing innovation. Potential directions include speciation, fitness sharing, or novelty search criteria. By rewarding novel behaviors instead of raw task performance, agents may continue discovering new strategies.

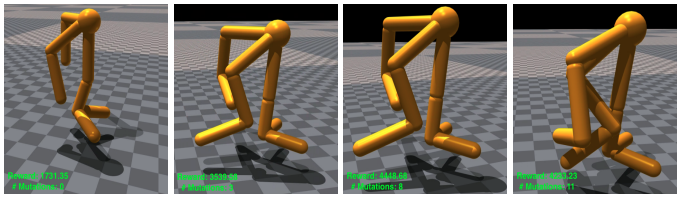

Fig. 6: Morphological changes in the best agent from experiment configuration P=100, T=50, W=20 over successive mutations. Despite visual similarity, invisible modifications to limb parameters (joint angle and density) occurred between mutation 3 and 8 improving fitness. However, additional mutations were harmful causing regressions.

4 Discussion and Future Work

Our results reveal limitations in DARLEI's current approach for maintaining diversity and enabling ongoing innovation. While selection propagates high-performing solutions, mutations degrade fitness without mechanisms to actively promote diversity. As a result, the population rapidly converges to human-like forms despite efforts like more workers and tournaments. This lack of sustained diversity indicates that key elements are missing for achieving open-ended discovery.

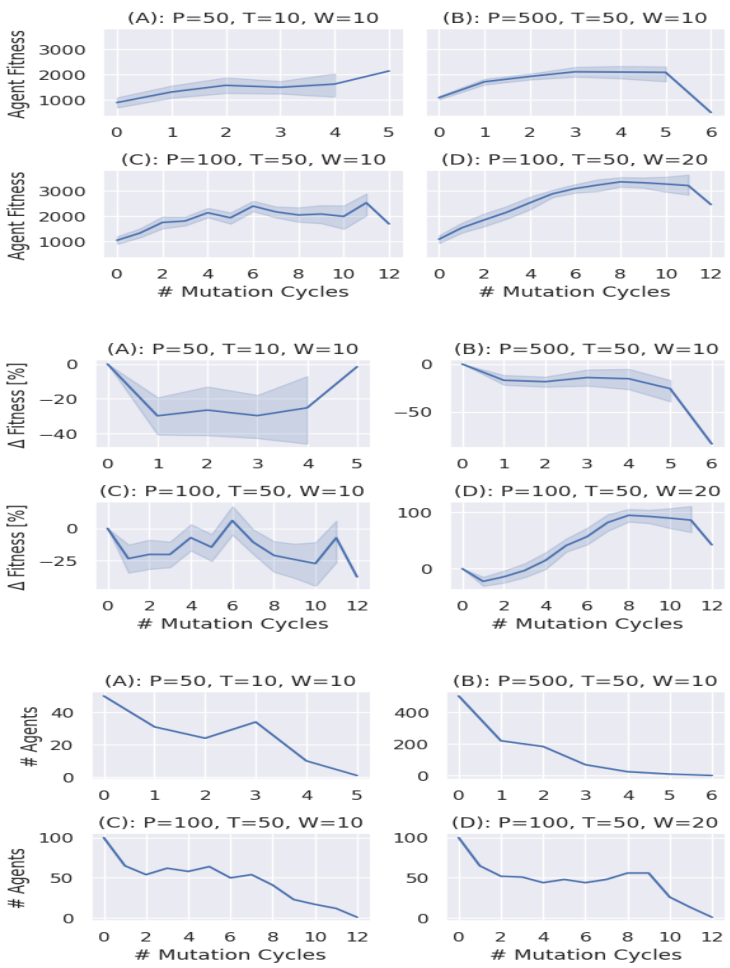

Fig. 7: Impact of mutation cycles on: agent fitness (top), percentage improvement in fitness between the youngest child and oldest ancestor (middle), and number of agents (bottom). Results are shown for 4 experiments (A,B,C,D), each with a unique configuration of initial population size P, number of tournaments T, and number of parallel worker processes W.

To promote greater open-endedness, future work could modify the fitness criteria to reward novelty over raw task performance. Approaches like Minimal Criteria Novelty Search [9] or novelty search in coevolution [10] may help drive morphological and behavioral diversity by evaluating agents based on how differently they accomplish the task rather than absolute performance. Balancing extrinsic rewards like task completion with intrinsic rewards for novelty could prevent premature convergence.

Additionally, complex procedurally generated environments satisfying a Minimal Criterion Coevolution could use multi-objective rewards enabling diverse agents to succeed in orthogonal ways. The environment itself could also coevolve to challenge the capabilities of the current population, driving the emergence of new strategies. By co-evolving agents and environments through continual elaboration on both sides, open-ended innovation may be sustained.

By decoupling individual lifetime learning from long-term evolution, DARLEI provides a platform to rapidly prototype and evaluate methods combining reinforcement learning and evolution. While our current experiments show limited diversity, we hope to extend DARLEI by adding support for mechanisms like coevolving populations, multi-objective rewards, and novelty-driven objectives to harness the creative potential of open-ended evolution. The accelerated simulation lowers the barrier for future work at the intersection of evolution, multi-agent interactions, and embodied intelligence.

References

[1] L. K. O. S. Soros, Joel Lehman. Open-endedness: The last grand challenge you've never heard of. [Online]. Available: https://www.oreilly.com/radar/open-endedness-the-last-grand-challenge-youve-never-heard-of/

[2] J. C. Brant and K. O. Stanley, "Minimal criterion coevolution: a new approach to open-ended search," in *Proceedings of the Genetic and Evolutionary Computation Conference*. ACM, pp. 67–74. [Online]. Available: https://dl.acm.org/doi/10.1145/3071178.3071186

[3] 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." [Online]. Available: http://arxiv.org/abs/2108.10470

[4] A. Kadian, J. Truong, A. Gokaslan, A. Clegg, E. Wijmans, S. Lee, M. Savva, S. Chernova, and D. Batra, "Sim2real predictivity: Does evaluation in simulation predict real-world performance?" vol. 5, no. 4, pp. 6670–6677. [Online]. Available: http://arxiv.org/abs/1912.06321

[5] A. Gupta, S. Savarese, S. Ganguli, and L. Fei-Fei, "Embodied intelligence via learning and evolution," vol. 12, no. 1, p. 5721, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-021-25874-z

[6] K. Sims, "Evolving virtual creatures," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques - SIGGRAPH '94*. ACM Press, pp. 15–22. [Online]. Available: http://portal.acm.org/citation.cfm?doid=192161.192167

[7] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, ISSN: 2153-0866.

[8] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms." [Online]. Available: http://arxiv.org/abs/1707.06347

[9] J. Gomes, P. Mariano, and A. L. Christensen, "Novelty search in competitive coevolution," vol. 8672, pp. 233–242. [Online]. Available: http://arxiv.org/abs/1407.0576

[10] J. Lehman and K. O. Stanley, "Revising the evolutionary computation abstraction: minimal criteria novelty search," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, ser. GECCO '10. Association for Computing Machinery, pp. 103–110. [Online]. Available: https://doi.org/10.1145/1830483.1830503