Article

CCL: Collaborative Curriculum Learning for Sparse-Reward Multi-Agent Reinforcement Learning via Co-evolutionary Task Evolution

Yufei Lin , Chengwei Ye , Huanzhen Zhang , Kangsheng Wang * , Linuo Xu , Shuyan Liu , Zeyu Zhang

Posted Date: 9 May 2025

doi: 10.20944/preprints202505.0657.v1

Keywords: Multi-Agent Reinforcement Learning (MARL); Sparse Reward Environments; Curriculum Learning; Co-evolutionary Algorithms; Task Generation; Evolutionary Reinforcement Learning; Cooperative Problem Solving

Article

CCL: Collaborative Curriculum Learning for Sparse-Reward Multi-Agent Reinforcement Learning via Co-Evolutionary Task Evolution

Yufei Lin [1], Chengwei Ye [1], Huanzhen Zhang [2], Kangsheng Wang [3,*], Linuo Xu [4], Shuyan Liu [5] and Zeyu Zhang [6]

[1] Homesite Group Inc; 0009-0005-0941-3316 (Y.L.); 0009-0004-2593-3621 (C.Y.)
[2] Chewy Inc; 0009-0008-3051-5642
[3] University of Science and Technology Beijing
[4] Yunnan University of Finance and Economics; 0009-0004-2901-6193
[5] Yunnan University; 0009-0004-7641-2623
[6] The Australian National University
* Correspondence: jackie@ieee.org; 0009-0009-8392-4148

Abstract: Sparse reward environments pose significant challenges in reinforcement learning, especially within multi-agent systems (MAS) where feedback is delayed and shared across agents, leading to suboptimal learning. We propose Collaborative Multi-dimensional Course Learning (CCL), a novel curriculum learning framework that addresses this by (1) refining intermediate tasks for individual agents, (2) using a variational evolutionary algorithm to generate informative subtasks, and (3) co-evolving agents with their environment to enhance training stability. Experiments on five cooperative tasks in the MPE and Hide-and-Seek environments show that CCL outperforms existing methods in sparse reward settings.

Keywords: multi-agent reinforcement learning (MARL); sparse reward environments; curriculum learning; Co-evolutionary Algorithms; task generation; Evolutionary Reinforcement Learning; Cooperative Problem Solving

1. Introduction

Deep Reinforcement Learning (DRL) has shown substantial success in Multi-Agent Systems (MAS), with notable applications in robotics [1,2], gaming [3], and autonomous driving [4]. Despite this progress, sparse reward environments continue to hinder learning efficiency, as agents often receive feedback only after completing complex tasks. This delayed reward signal limits exploration and makes policy optimization difficult.

To improve exploration under sparse rewards, several strategies have been proposed, including reward shaping [5,6], imitation learning [7], policy transfer [8], and curriculum learning [9,10]. These methods aim to strengthen the reward signal and guide agents toward effective behaviors. While effective in single-agent environments, their performance often degrades in MAS, where multiple interacting agents exacerbate environmental dynamics and expand the joint state-action space [11–13].

In response, we propose **Collaborative Multi-dimensional Course Learning (CCL)**, a co-evolutionary curriculum learning framework tailored for sparse-reward cooperative MAS. CCL introduces three core innovations:

(1) It generates agent-specific intermediate tasks using a variational evolutionary algorithm, enabling balanced strategy development.

(2) It models co-evolution between agents and their environment [14], aligning task complexity with agents' learning progress.

(3) It improves training stability by dynamically adapting task difficulty to match agent skill levels.

Through extensive experiments across five tasks in the MPE and Hide-and-Seek (HnS) environments, CCL consistently outperforms existing baselines, demonstrating enhanced learning efficiency and robustness in sparse-reward multi-agent scenarios.

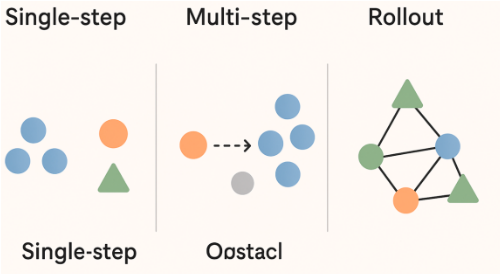


Figure 1. MPE is validated with three different collaborative task scenarios.

2. Problem Statement

In reinforcement learning, the reward signal is a critical feedback mechanism guiding agents to assess their actions and learn optimal policies via the Bellman equation [15]. While a well-designed reward function defines the task objective and measures agent behavior, agents may still pursue suboptimal strategies. Nonetheless, carefully crafted rewards greatly enhance learning efficiency and policy convergence [16].

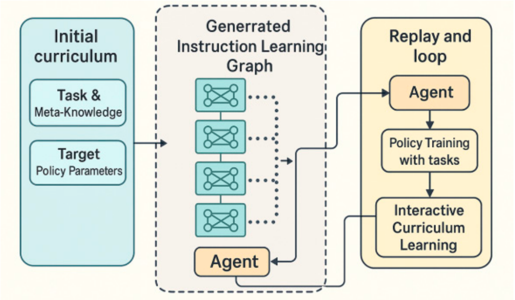

Initial curriculum
Task & Meta-Knowledge
Target Policy Parameters
Generrated Instruction Learning Graph
Agent
Replay and loop
Agent
Policy Training with tasks
Interactive Curriculum Learning

Figure 2. Intermediate task generation in MAS is more complex than in single-agent settings due to the need to account for agent-specific subtasks. In sparse reward environments where rewards are shared, incorporating an individual perspective mechanism becomes essential to ensure effective task decomposition and learning.

Designing dense rewards in complex MAS is challenging due to reliance on prior knowledge, which often fails to capture all interaction dynamics. Sparse rewards offer a more flexible alternative by providing feedback only upon reaching a critical goal state [17], reducing dependence on manual reward design and improving generalization.

In non-sparse reward settings, at each time step $\boldsymbol{t}$, the agent observes its current state $s_t \in S$ and selects an action $a_t \in A$ based on its policy $\pi(a_t|s_t)$. The chosen action results in a transition to a new state s_{t+1}, determined by the environment's transition dynamics $p(s_{t+1}|s_t, a_t)$, and an associated reward r_t is obtained from the reward function $r(s_t, a_t, s_{t+1})$. The sequence of states, actions, following states, and rewards over an episode of T time steps form the trajectory $\tau = (s_t, a_t, s_{t+1}, r_t)_{t=0}^{T-1}$, where T is either determined by the maximum episode length or specific task termination conditions. This outlines the process of reinforcement learning for a single agent.

The goal of this individual agent is to learn and maximize its expected cumulative rewarded policy:

$$J = \mathbb{E}_{\pi}\left[\sum_{t} \gamma^{t} r_{t}\right] \tag{1}$$

where γ is the discount factor, representing future rewards' diminishing value refinement degree of the optimization process is carried out by each time step inside the trajectory, that is, the optimization granularity is accurate to each time step.

However, the system dynamics significantly intensify when extending this general framework to MAS under sparse reward conditions. In this system, there are N decision-making agents, where each agent i takes an action a_i at time step t based on the observed state information and following its dedicated policy $\pi_i(a_i|s_i)$. The global state st of the system is composed of the joint states of all individual agents, denoted as $s_t = (s_1, s_2, \ldots, s_n)$. Correspondingly, the joint action a_t at each time step is also formed by the combination of actions from all agents, i.e., $a_t = (a_1 a_2, \ldots, a_n)$. In the sparse reward environment, reward signals only emerge when the system achieves specific predefined goal states, posing more significant challenges for agent collaboration and strategy optimization.

In cooperative multi-agent tasks, the goal of each agent is no longer focused on maximizing its reward but instead shifts toward optimizing the cumulative reward of the entire system. This requires agents to collaborate effectively, coordinating their actions to achieve the shared objective, thereby improving the overall performance of the multi-agent system. Consequently, the objective function J for each agent i is transformed into $J_i(\pi_i) = \mathbb{E}_{\pi_i}\left[\sum_t \gamma^t r_i(s_t, a_t)\right]$, where $r_i(s_t, a_t)$ represents the reward received by agent i at time step t given the state st and joint action a_t. The overall goal of the multi-agent system (MAS) then becomes the sum of the individual objectives, denoted as $J = \sum_i J_i(\pi_i)$.

At this point, it becomes evident that the essence of a multi-agent reinforcement learning algorithm lies in utilizing the rewards earned by all agents to optimize the overall collaborative strategy. However, this challenge is significantly heightened in a sparse reward environment, where agents receive limited feedback, making it difficult to effectively guide their actions and improve coordination toward the collective goal. In the case that there are only very few 0-1 reward signals, the total reward of the system can be simplified to a binary function:

$$r(s_t, g) = \begin{cases} 1, & s_t = g \\ 0, & \text{Otherwise} \end{cases} \tag{2}$$

As the number of agents increases, training variance in MAS grows exponentially. In sparse reward settings, agents must achieve sub-goals aligned with a shared objective, yet often receive little to no feedback, making learning difficult. This lack of guidance hampers exploration and destabilizes training, rendering many single-agent methods ineffective. To address these challenges, we propose Collaborative Multi-dimensional Course Learning (CCL) for more stable and efficient multi-agent training.

$$s_t = g \iff \forall i \in n, s_i = g_i \tag{3}$$

3. Related Work

3.1. Curriculum Learning

Sparse reward environments have driven the development of various exploration strategies in reinforcement learning, including reward shaping [18], intrinsic motivation [19], and curriculum learning [20]. While the first two enhance learning by densifying rewards, curriculum learning adopts a divide-and-conquer approach—decomposing complex tasks into simpler subtasks arranged in increasing difficulty [1,21,22].

In reinforcement learning, curriculum learning involves three main components: task generation, task ranking, and transfer learning [23]. These can be guided by automated methods [9] or expert knowledge, though the latter often introduces biases [24,25]. Adaptive Automatic Curriculum Learning (ACL) addresses this by dynamically tailoring task sequences to agent progress, without manual intervention.

Despite its promise, ACL faces challenges in defining effective evaluation metrics and managing computational cost [26–28]. Current approaches often rely on coarse performance metrics or costly replay mechanisms [29,30], making it difficult to scale in complex multi-agent settings.

3.2. Evolutionary Reinforcement Learning

Evolutionary Algorithms (EAs) optimize policies through selection, mutation, and recombination of candidate solutions based on fitness scores [31]. Their integration with reinforcement learning aims to address issues like sparse rewards and limited policy diversity [29,30,34].

Though promising [32,33], combining EAs with RL introduces challenges, notably the computational overhead from large populations [24] and the difficulty of retaining informative environmental features during evolutionary encoding. Effective integration requires balancing exploration benefits with computational feasibility[44–51].

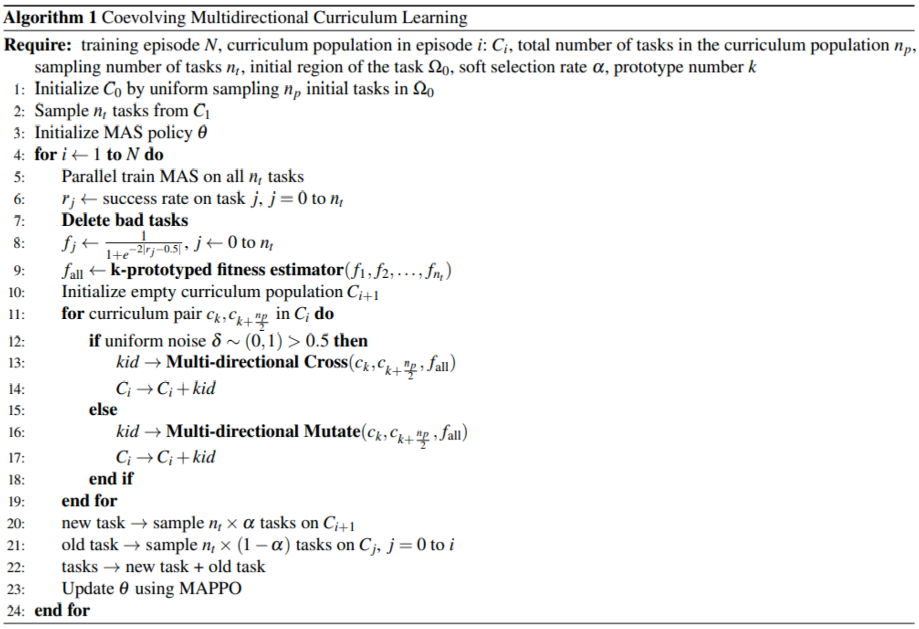
4. Methodology

4.1. The Variational Individual-Perspective Evolutionary Operator

In this section, we provide a detailed explanation of all the components of CMCL. As a coevolutionary system with two primary parts, the agents are trained using the existing Multi-Agent Proximal Policy Optimization (MAPPO) algorithm[35], which will not be elaborated on here. The complete workflow of the CMCL algorithm is outlined in Algorithm 1.

Evolutionary Curriculum Initialization Due to the low initial policy performance of a MAS at the start of training, agents struggle to accomplish complex tasks. Therefore, minimizing the norm of task individuals within the initial population is essential. Assuming the initial task domain is Ω_0, the randomly initialized task population should meet the following conditions, where d represents the initial Euclidean norm between the agent and the task, and δ is a robust hyperparameter, typically set to be about one percent of the total task space size.

$$\frac{1}{|\Omega_0|} \sum_{t_i \in \Omega_0} d(s_i, g_i) < \delta \tag{4}$$



Algorithm 1 Coevolving Multidirectional Curriculum Learning

```
Require: training episode $N$, curriculum population in episode $i$: $C_i$, total number of tasks in the curriculum population $n_p$,
    sampling number of tasks $n_t$, initial region of the task $\Omega_0$, soft selection rate $\alpha$, prototype number $k$
Initialize $C_0$ by uniform sampling $n_p$ initial tasks in $\Omega_0$
Sample $n_t$ tasks from $C_1$
Initialize MAS policy $\theta$
for $i \leftarrow 1$ to $N$ do
    Parallel train MAS on all $n_t$ tasks
    $r_j \leftarrow$ success rate on task $j$, $j = 0$ to $n_t$
    Delete bad tasks
    $f_j \leftarrow \frac{1}{1+e^{-2|r_j-0.5|}}$, $j \leftarrow 0$ to $n_t$
    $f_{\text{all}} \leftarrow$ k-prototyped fitness estimator$(f_1, f_2, \ldots, f_{n_t})$
    Initialize empty curriculum population $C_{i+1}$
    for curriculum pair $c_k, c_{k+\frac{np}{2}}$ in $C_i$ do
        if uniform noise $\delta \sim (0,1) > 0.5$ then
            $kid \rightarrow$ Multi-directional Cross$(c_k, c_{k+\frac{np}{2}}, f_{\text{all}})$
            $C_i \rightarrow C_i + kid$
        else
            $kid \rightarrow$ Multi-directional Mutate$(c_k, c_{k+\frac{np}{2}}, f_{\text{all}})$
            $C_i \rightarrow C_i + kid$
        end if
    end for
    new task $\rightarrow$ sample $n_t \times \alpha$ tasks on $C_{i+1}$
    old task $\rightarrow$ sample $n_t \times (1-\alpha)$ tasks on $C_j$, $j = 0$ to $i$
    tasks $\rightarrow$ new task + old task
    Update $\theta$ using MAPPO
end for
```

Task Fitness Definition: Previous methods often assessed intermediate tasks using agent performance metrics [24,30,36] or simple binary filters [37], which fail to capture the non-linear nature of task difficulty. Tasks with success rates near 0 or 1 offer little training value, while those closer to the midpoint present a more suitable challenge. To address this, we model fitness as a non-linear function, favoring tasks of moderate difficulty that best support learning progression. To capture this non-linear relationship, we establish a sigmoid-shaped fitness function to describe the adaptability of tasks to the current level of agent performance, where r represents the average success rate of the agents on task t.

$$f = \frac{1}{1 + e^{-2|r-0.5|}} \tag{5}$$

Variational Individual-perspective Crossover In a MAS, the single reward signal is distributed across multiple dimensions, especially from the perspective of different agents, leading to imbalances in the progression of individual strategies. Therefore, based on the encoding method mentioned earlier, operating on intermediate tasks at the individual level within the MAS is necessary. Assuming that in a particular round of intermediate task generation, N individuals from the previous task generation $\{T = t_1, t_2, \ldots, t_N\}$ are randomly divided into two groups T_A and T_B. Then, we take $N/2$ task pairs t^A, t^B from T_A and T_B to produce new children in the population.

$$\begin{cases} t_i^{A*} \leftarrow t_i^A + S_i \overrightarrow{D_i}, \\ t_i^{B*} \leftarrow t_i^B + S_i \overrightarrow{D_i} \end{cases}, \quad i = 1, 2, \ldots, \frac{N}{2} \tag{6}$$

In the above formula, s_i represents the crossover step size for pair i, and $\overrightarrow{D_i}$ represents the crossover direction for pair i. The calculations of s_i and $\overrightarrow{D_i}$ are shown below:

$$s_{c,i} = \frac{|f(t_i^A) - f(t_i^B)|}{\max(f(T)) - \min(f(T))} \tag{7}$$

$$\overrightarrow{D_i} = [D_{i,1}, D_{i,2}, \ldots, D_{i,n}]$$

$D_{i,j}$ denotes the direction of the j-th agent in pair i, obtained by uniform random sampling.

$$D_{i,j} = \begin{cases} 0, & \text{if random variable } \delta_j < 0.5 \\ \theta_{i,j}^A - \theta_{i,j}^B, & \text{if random variable } \delta_j \geq 0.5 \end{cases} \tag{8}$$

The proposed variational individual-perspective crossover ensures each agent's subtask direction contributes equally to curriculum evolution, enabling broader exploration compared to traditional methods. In a MAS with nnn entities, this results in 2n2^n2n possible direction combinations, enhancing diversity in intermediate task generation.

To address catastrophic forgetting [38,39], we adopt a soft selection strategy. Rather than discarding low-fitness individuals, the entire population is retained, and a fraction (α\alphaα, typically 0.2–0.4) of historical individuals is reintroduced each iteration. This maintains task diversity, preserves challenging tasks for future stages, and helps avoid local optima.

4.2. Elite Prototype Fitness Evaluation

Evolutionary algorithms often require maintaining a sufficiently large population to ensure diversity and prevent being trapped in local optima or influenced by randomness. However, evaluating the fitness of intermediate tasks in a large curriculum population significantly increases computational cost. To mitigate this issue, we propose a prototype-based fitness estimation method.

First, we uniformly sample tasks in each iteration and measure their success rate r and fitness f . These sampled tasks, called prototypes, are used in actual training. Next, we employ a K-Nearest Neighbor (KNN) approach to estimate the fitness of tasks not directly used in training.

Assume there are m individuals in the prototype task set P, with fitness values f_i for each i, and n individuals in the query task set Q, represented as vectors q_j for each j. For any individual q_j in the query set Q, its fitness value f_j can be calculated as shown below

$$f_j = \frac{1}{k}\sum_{i\in N_j} e_{f_i} \tag{9}$$

In the formula, N_j represents the set of indices of the k-closest individuals in the prototype task set P to the vector q_j, based on the Euclidean distance. This can be expressed as follows:

$$N_j = \arg\min_{S\subseteq P,|S|=k} \sum_{i\in S} \|q_j - p_i\|^2 \tag{10}$$

5. Experiment

5.1. Main Result

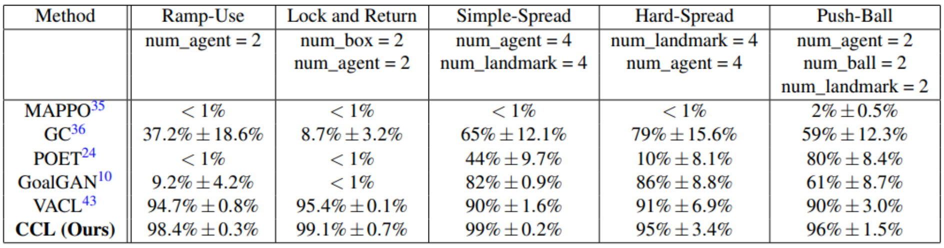
We evaluate CCL on five cooperative tasks across two environments: simple/complex propagation and Push-ball from the MPE benchmark [40], and ramp-passing and lock-back from the MuJoCo-based HnS environment [41]. All tasks use a binary (0-1) sparse reward structure, with results averaged over three random seeds. Training is conducted using MAPPO [35] on a system with an Nvidia RTX 3090 GPU and a 14-core CPU. Attention mechanisms [42] are also integrated to improve agent coordination.

We compare CCL with five baselines:

1. Vanilla MAPPO [35] – Direct training on the target task without intermediate tasks.
2. POET [24] – Uses task evolution; implemented with the same setup as CCL for fairness.
3. GC [36] – An improved version of POET with enhanced task generation.
4. GoalGAN [10] – Combines curriculum learning with attention-based enhancements.
5. VACL [43] – Applies variational methods to create robust intermediate tasks.

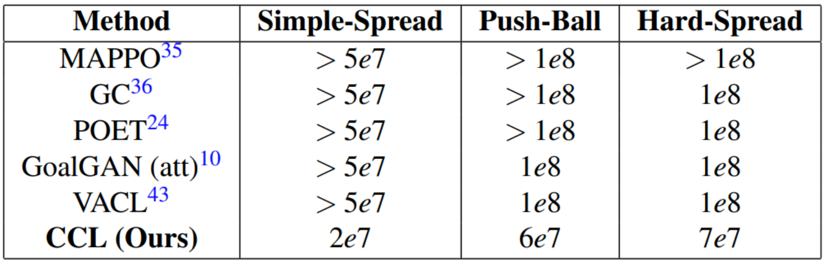
Across all environments, baseline methods struggle under sparse rewards, especially in HnS. CCL consistently outperforms them in both learning speed and final performance, achieving over 95% success in the most complex tasks (see Tables 1 and 2).

Table 1. The Performance Comparison of CCL and Other Baselines on Simulated Environments.



Method	Ramp-Use	Lock and Return	Simple-Spread	Hard-Spread	Push-Ball
	num_agent = 2	num_box = 2 num_agent = 2	num_agent = 4 num_landmark = 4	num_landmark = 4 num_agent = 4	num_agent = 2 num_ball = 2 num_landmark = 2
MAPPO[35]	$< 1\%$	$< 1\%$	$< 1\%$	$< 1\%$	$2\% \pm 0.5\%$
GC[36]	$37.2\% \pm 18.6\%$	$8.7\% \pm 3.2\%$	$65\% \pm 12.1\%$	$79\% \pm 15.6\%$	$59\% \pm 12.3\%$
POET[24]	$< 1\%$	$< 1\%$	$44\% \pm 9.7\%$	$10\% \pm 8.1\%$	$80\% \pm 8.4\%$
GoalGAN[10]	$9.2\% \pm 4.2\%$	$< 1\%$	$82\% \pm 0.9\%$	$86\% \pm 8.8\%$	$61\% \pm 8.7\%$
VACL[43]	$94.7\% \pm 0.8\%$	$95.4\% \pm 0.1\%$	$90\% \pm 1.6\%$	$91\% \pm 6.9\%$	$90\% \pm 3.0\%$
CCL (Ours)	$98.4\% \pm 0.3\%$	$99.1\% \pm 0.7\%$	$99\% \pm 0.2\%$	$95\% \pm 3.4\%$	$96\% \pm 1.5\%$

Table 2. Performance Metrics for Various Methods across Different Tasks.



Method	**Simple-Spread**	**Push-Ball**	**Hard-Spread**
MAPPO[35]	$> 5e7$	$> 1e8$	$> 1e8$
GC[36]	$> 5e7$	$> 1e8$	$1e8$
POET[24]	$> 5e7$	$> 1e8$	$1e8$
GoalGAN (att)[10]	$> 5e7$	$1e8$	$1e8$
VACL[43]	$> 5e7$	$1e8$	$1e8$
CCL (Ours)	$2e7$	$6e7$	$7e7$

5.2. Ablation Studies

Adaptive Mutation Step: Ablation studies show that using an adaptive mutation step size enhances flexibility and performance in sparse reward environments compared to fixed or no mutation. While mutation promotes strategy diversity, improper step sizes can degrade learning. Notably, adaptive mutation proves as effective as crossover and individual-perspective variation in improving CCL's performance (see Figure 3).

Non-linear Factor in Fitness Function: As shown in Figure 4, the sigmoid fitness function delivers better performance than the linear form $f = -k|r - 0.5|$. This improvement stems from the sigmoid function's properties: as the agent's success rate approaches 0 or 1, the task's suitability to the agent's abilities decreases exponentially. Specifically, when the success rate is exactly 0.5, the fitness value remains consistently at 0.5. This approach effectively integrates nonlinear elements into the success rate distribution, enabling the fitness function to more accurately represent the relationship between task difficulty and the agent's skill level.

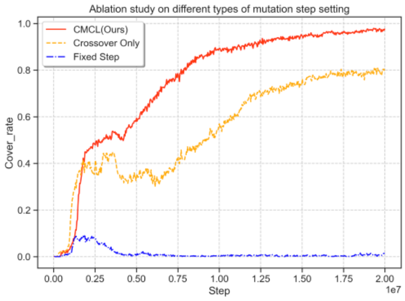


Figure 3. The adaptive step usage ablation experiments which shows its effect.

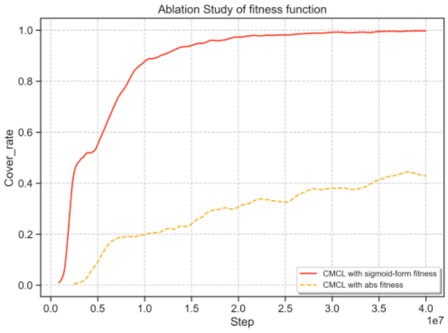


Figure 4. The comparison of using absolute value and sigmoid-shaped fitness function.

6. Conclusion

This paper presents CCL, a co-evolutionary curriculum learning framework designed to improve training stability and performance in sparse-reward multi-agent systems (MAS). By generating a population of intermediate tasks, using a variational individual-perspective crossover, and employing elite prototype-based fitness evaluation, CCL enhances exploration and coordination. Experiments in MPE and HnS environments show that CCL consistently outperforms existing baselines. Ablation studies further validate the importance of each component.

Despite its strengths, CCL's current design is focused on cooperative MAS. Future work should explore its applicability in competitive or mixed-behavior settings, where coordination and conflict coexist. Moreover, the storage of historical tasks for soft selection increases memory usage; optimizing this via compression or selective retention is a promising direction for reducing overhead.

References

1. Abbass, H., Petraki, E., Hussein, A., McCall, F. & Elsawah, S. A model of symbiomemesis: machine education and communication as pillars for human-autonomy symbiosis. Philos. Transactions Royal Soc. A 379, 20200364 (2021).
2. Perrusquía, A., Yu, W. & Li, X. Multi-agent reinforcement learning for redundant robot control in task-space. Int. J. Mach. Learn. Cybern. 12, 231–241 (2021).
3. Rashid, T. et al. Monotonic value function factorisation for deep multi-agent reinforcement learning. J. Mach. Learn. Res. 21, 1–51 (2020).
4. Shalev-Shwartz, S., Shammah, S. & Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295 (2016).
5. Ng, A. Y., Harada, D. & Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In Icml, vol. 99, 278–287 (1999).
6. Hu, Y. et al. Learning to utilize shaping rewards: A new approach of reward shaping. Adv. Neural Inf. Process. Syst. 33, 15931–15941 (2020).
7. Ross, S., Gordon, G. & Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics, 627–635 (JMLR Workshop and Conference Proceedings, 2011).
8. Duan, Y. et al. One-shot imitation learning. Adv. neural information processing systems 30 (2017).
9. Florensa, C., Held, D., Wulfmeier, M., Zhang, M. & Abbeel, P. Reverse curriculum generation for reinforcement learning. In Conference on robot learning, 482–495 (PMLR, 2017).
10. Florensa, C., Held, D., Geng, X. & Abbeel, P. Automatic goal generation for reinforcement learning agents. In International conference on machine learning, 1515–1528 (PMLR, 2018).
11. Bloembergen, D., Tuyls, K., Hennes, D. & Kaisers, M. Evolutionary dynamics of multi-agent learning: A survey. J. Artif. Intell. Res. 53, 659–697 (2015).
12. Bu¸soniu, L., Babuška, R. & De Schutter, B. Multi-agent reinforcement learning: An overview. Innov. multi-agent systems applications-1 183–221 (2010).
13. Hernandez-Leal, P., Kartal, B. & Taylor, M. E. A survey and critique of multiagent deep reinforcement learning. Auton. Agents Multi-Agent Syst. 33, 750–797 (2019).
14. Antonio, L. M. & Coello, C. A. C. Coevolutionary multiobjective evolutionary algorithms: Survey of the state-of-the-art. IEEE Transactions on Evol. Comput. 22, 851–865 (2017).
15. Kaelbling, L. P., Littman, M. L. & Moore, A. W. Reinforcement learning: A survey. J. artificial intelligence research 4, 237–285 (1996).
16. Dewey, D. Reinforcement learning and the reward engineering principle. In 2014 AAAI Spring Symposium Series (2014). 9/11
17. Booth, S. et al. The perils of trial-and-error reward design: misdesign through overfitting and invalid task specifications. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 5920–5929 (2023).
18. Laud, A. D. Theory and application of reward shaping in reinforcement learning (University of Illinois at UrbanaChampaign, 2004).
19. Barto, A. G. Intrinsic motivation and reinforcement learning. Intrinsically motivated learning natural artificial systems 17–47 (2013).
20. Bengio, Y., Louradour, J., Collobert, R. & Weston, J. Curriculum learning. In Proceedings of the 26th annual international conference on machine learning, 41–48 (2009).
21. Rohde, D. L. & Plaut, D. C. Language acquisition in the absence of explicit negative evidence: How important is starting small? Cognition 72, 67–109 (1999).
22. Elman, J. L. Learning and development in neural networks: The importance of starting small. Cognition 48, 71–99 (1993).
23. Narvekar, S., Sinapov, J. & Stone, P. Autonomous task sequencing for customized curriculum design in reinforcement learning. In IJCAI, 2536–2542 (2017).
24. Wang, R., Lehman, J., Clune, J. & Stanley, K. O. Paired open-ended trailblazer (poet): Endlessly generating increasingly complex and diverse learning environments and their solutions. arXiv preprint arXiv:1901.01753 (2019).

25. Cobbe, K., Klimov, O., Hesse, C., Kim, T. & Schulman, J. Quantifying generalization in reinforcement learning. In International conference on machine learning, 1282–1289 (PMLR, 2019).
26. Ren, Z., Dong, D., Li, H. & Chen, C. Self-paced prioritized curriculum learning with coverage penalty in deep reinforcement learning. IEEE transactions on neural networks learning systems 29, 2216–2226 (2018).
27. Wu, J. et al. Portal: Automatic curricula generation for multiagent reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 15934–15942 (2024).
28. Wang, K., Zhang, X., Guo, Z., Hu, T. & Ma, H. Csce: Boosting llm reasoning by simultaneous enhancing of casual significance and consistency. arXiv preprint arXiv:2409.17174 (2024).
29. Samvelyan, M. et al. Maestro: Open-ended environment design for multi-agent reinforcement learning. arXiv preprint arXiv:2303.03376 (2023).
30. Parker-Holder, J. et al. Evolving curricula with regret-based environment design. In International Conference on Machine Learning, 17473–17498 (PMLR, 2022).
31. Beyer, H.-G. & Schwefel, H.-P. Evolution strategies–a comprehensive introduction. Nat. computing 1, 3–52 (2002).
32. Miconi, T., Rawal, A., Clune, J. & Stanley, K. O. Backpropamine: training self-modifying neural networks with differentiable neuromodulated plasticity. arXiv preprint arXiv:2002.10585 (2020).
33. Pagliuca, P., Milano, N. & Nolfi, S. Efficacy of modern neuro-evolutionary strategies for continuous control optimization. Front. Robotics AI 7, 98 (2020).
34. Long, Q. et al. Evolutionary population curriculum for scaling multi-agent reinforcement learning. arXiv preprint arXiv:2003.10423 (2020).
35. Yu, C. et al. The surprising effectiveness of ppo in cooperative multi-agent games. Adv. Neural Inf. Process. Syst. 35, 24611–24624 (2022).
36. Song, Y. & Schneider, J. Robust reinforcement learning via genetic curriculum. In 2022 International Conference on Robotics and Automation (ICRA), 5560–5566 (IEEE, 2022).
37. Racaniere, S. et al. Automated curricula through setter-solver interactions. arXiv preprint arXiv:1909.12892 (2019).
38. Cahill, A. Catastrophic forgetting in reinforcement-learning environments. Ph.D. thesis, Citeseer (2011).
39. French, R. M. Catastrophic forgetting in connectionist networks. Trends cognitive sciences 3, 128–135 (1999).
40. Lowe, R. et al. Multi-agent actor-critic for mixed cooperative-competitive environments. Adv. neural information processing systems 30 (2017).
41. Baker, B. et al. Emergent tool use from multi-agent autocurricula. arXiv preprint arXiv:1909.07528 (2019).
42. Vaswani, A. et al. Attention is all you need [j]. Adv. neural information processing systems 30, 261–272 (2017).
43. Chen, J. et al. Variational automatic curriculum learning for sparse-reward cooperative multi-agent problems. Adv. Neural Inf. Process. Syst. 34, 9681–9693 (2021).
44. Qi, X., Zhang, Z., Zheng, H., et al.: MedConv: Convolutions Beat Transformers on Long-Tailed Bone Density Prediction. arXiv preprint arXiv:2502.00631 (2025)
45. Wang, K., Zhang, X., Guo, Z., et al.: CSCE: Boosting LLM Reasoning by Simultaneous Enhancing of Causal Significance and Consistency. arXiv preprint arXiv:2409.17174 (2024)
46. Liu, S., Wang, K.: Comprehensive Review: Advancing Cognitive Computing through Theory of Mind Integration and Deep Learning in Artificial Intelligence. In: Proc. 8th Int. Conf. on Computer Science and Application Engineering, pp. 31–35 (2024)
47. Zhang, X., Wang, K., Hu, T., et al.: Enhancing Autonomous Driving through Dual-Process Learning with Behavior and Reflection Integration. In: ICASSP 2025 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1–5. IEEE, Seoul (2025)
48. Zou, B., Guo, Z., Qin, W., et al.: Synergistic Spotting and Recognition of Micro-Expression via Temporal State Transition. In: ICASSP 2025 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1–5. IEEE, Seoul (2025)
49. Hu, T., Zhang, X., Ma, H., et al.: Autonomous Driving System Based on Dual Process Theory and Deliberate Practice Theory. Manuscript (2025)

50. Zhang, X., Wang, K., Hu, T., et al.: Efficient Knowledge Transfer in Multi-Task Learning through Task-Adaptive Low-Rank Representation. arXiv preprint arXiv:2505.00009 (2025)

51. Wang, K., Ye, C., Zhang, H., et al.: Graph-Driven Multimodal Feature Learning Framework for Apparent Personality Assessment. arXiv preprint arXiv:2504.11515 (2025)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.