INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY TRENDS

Vol. 06 No. 01 (2026)

e. ISSN: 2409-5222

https://ijictt.ictt.info

Curriculum Reinforcement Learning for Progressive Coordination in Complex Multi-Step Decision Tasks

Pierre Dupont1, Marie Lefèvre2, Antoine Martin3*

Department of Computer Science, École Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France

Corresponding author: *a.martin@polytechnique.edu*

Abstract : *Training collaborative agents directly on complex tasks often leads to unstable learning and poor convergence. This study proposes a curriculum reinforcement learning strategy in which task difficulty is gradually increased based on coordination complexity and dependency depth. A dynamic curriculum scheduler is designed using reward variance and policy entropy as progression signals, while PPO is used for policy optimization. The framework is evaluated on 12,000 tasks with increasing levels of complexity, from simple coordination to multi-stage planning with constraints. Compared with direct training, the curriculum-based approach improves final task success rate by 18.7% and accelerates convergence by 26.9%. In addition, failure cases related to coordination breakdown are reduced by 31.2%. The results indicate that structured training progression significantly enhances learning stability and performance in collaborative decision environments.*

Keywords: *Curriculum learning; Reinforcement learning; PPO; Multi-agent coordination; Task complexity*

1. INTRODUCTION

Complex multi-step decision tasks are widely encountered in real-world applications such as autonomous systems, distributed control, traffic management, and resource allocation. In these scenarios, agents must make sequential decisions over long horizons while coping with delayed rewards, partial observability, and strong interdependence among actions. The quality of the final outcome depends not only on the correctness of individual decisions, but also on whether effective coordination can be maintained throughout the entire decision process. Reinforcement learning, particularly multi-agent reinforcement learning (MARL), has been extensively studied for these problems because it enables agents to learn adaptive policies through repeated interaction with dynamic environments [1,2]. Recent work on structured modeling and representation learning for complex inference processes has further highlighted the importance of capturing shared relational structure and intermediate dependencies when solving long sequential tasks [3]. This perspective is highly relevant to multi-step collaborative decision problems, in which coordination quality is often determined by how well agents can represent and exploit evolving task structure across multiple stages. Despite these advances, robust learning in complex multi-step settings remains difficult, especially when coordination requirements become stronger and environmental conditions vary over time [4]. A major difficulty in such tasks lies in the interaction between long-term dependency and multi-agent coordination. In many practical settings, an action taken at an early stage may affect not only later rewards, but also the feasible decisions of other agents in subsequent stages. As a result, local errors can accumulate over time and eventually lead to large coordination failures.

Existing studies have shown that MARL still faces several core challenges, including non-stationarity, sparse and delayed rewards, high variance in policy updates, and difficulty in assigning credit to individual agents under shared outcomes [5,6]. These issues become more severe in long-horizon decision problems, where successful behavior depends on stable coordination over multiple steps rather than on short-term reactive actions alone. Even when a policy performs adequately in simpler environments, its effectiveness may decline sharply once dependency depth and interaction complexity increase. Among existing policy optimization methods, proximal policy optimization (PPO) has become one of the most widely adopted baselines because of its relative simplicity, stable training behavior, and strong empirical performance across a broad range of reinforcement learning tasks [7]. PPO has also been frequently used in multi-agent settings due to its practical implementation advantages and compatibility with actor-critic training. However, direct training with PPO on difficult multi-step coordination tasks often leads to slow convergence and unstable learning dynamics, particularly when agents must discover coordinated behaviors from sparse feedback [8]. In such cases, the optimization process may spend a large number of training iterations exploring poorly structured behaviors before meaningful coordination emerges. This not only reduces sample efficiency but also increases the risk of training collapse or premature convergence to suboptimal interaction patterns. To alleviate these difficulties, curriculum reinforcement learning has attracted growing attention. Rather than training agents directly on the full target problem, curriculum learning arranges training tasks from relatively simple settings to more difficult ones, allowing policies to gradually acquire useful knowledge and coordination skills. This idea is motivated by the observation that the order in which tasks are presented can substantially influence the speed and stability of learning. In single-agent reinforcement learning, curriculum design has been explored from several perspectives, including task distribution shaping, self-paced learning, goal ordering, and skill discovery [9,10]. Studies on long-horizon tasks have also shown that task-structured curricula can improve learning stability, enhance sample efficiency, and guide policies toward more effective exploration patterns [11]. These findings suggest that training progression is not merely a procedural choice, but an important factor that shapes representation learning, behavioral refinement, and final policy quality. Curriculum-based ideas have also been extended to multi-agent settings to improve coordination under increasing difficulty. Existing work has examined reverse curricula, adaptive partner selection, and task scheduling mechanisms designed to expose agents to progressively more challenging coordination demands [12]. More recent studies have applied curriculum learning to spatial coordination, robotic control, and cooperative planning, showing that gradual exposure to harder tasks can improve the consistency of learned interaction strategies [13]. Related work in multi-task and sequential adaptation further indicates that task diversity and ordered transfer can improve robustness, although these aspects are still insufficiently addressed in current MARL frameworks [14]. Even so, much of the current evidence is obtained from relatively controlled environments, and it remains unclear whether these gains can be sustained in more complex multi-stage decision problems with stronger action dependencies and deeper coordination chains. Several limitations therefore remain in the current literature. Many curriculum methods rely on manually specified training schedules or hand-crafted task sequences, which are often closely tied to a particular environment and difficult to generalize. Their effectiveness may depend heavily on prior knowledge of task difficulty, yet such knowledge is often unavailable or unreliable in real applications. In addition, empirical evaluations are frequently restricted to small-scale benchmarks with simplified interaction patterns, making it difficult to judge whether the proposed methods remain effective in larger and more realistic multi-step decision settings. Another issue is that many existing studies emphasize either task difficulty or agent heterogeneity, while paying less attention to their combined influence. In practice, these two sources of complexity often increase together. As coordination depth grows, agents must respond not only to harder tasks but also to more complicated inter-agent dependencies, which can substantially increase the difficulty of stable policy learning. Related efforts in offline

reinforcement learning and feedback-based optimization have improved training stability in some settings, yet they do not fully resolve the challenges of online adaptation in complex cooperative environments [15,16]. Coordination failures remain common as tasks become more demanding, especially when success depends on maintaining coherent behavior over long chains of decisions [17]. Another weakness lies in the way curricula are advanced during training. Many approaches still depend on fixed progression rules, such as predefined stages, manually selected switching points, or static performance thresholds. These rules do not always reflect the actual learning state of the agents. Training may therefore move to harder tasks before stable coordination has been formed, or remain too long in easier settings after the policy has already saturated, which reduces overall learning efficiency. Recent studies suggest that adaptive curricula driven by learning signals can provide a better reflection of policy development and task readiness [18]. However, this idea remains underexplored in multi-agent multi-step decision tasks, particularly when coordination quality depends on long dependency chains and progressively increasing interaction complexity. These limitations point to the need for a curriculum learning framework that can adapt task progression to the evolving learning state of the agents while explicitly accounting for the multi-step and cooperative nature of the target problem. The present study is motivated by this need. It focuses on complex multi-step decision tasks in which coordination must be sustained over multiple stages and where both coordination complexity and dependency depth can increase during training. In such settings, a useful training strategy should do more than simply arrange tasks from easy to hard. It should also identify whether the current policy has formed sufficiently stable coordination patterns and whether the agents are ready to handle a more challenging level of interaction. To this end, this study proposes a curriculum reinforcement learning framework for progressive coordination in complex multi-step decision tasks. The core idea is to organize training according to increasing coordination complexity and dependency depth, so that agents can gradually develop stable cooperative behavior before confronting harder task structures. A dynamic curriculum scheduler is introduced to adjust task difficulty during training by using reward variance and policy entropy as indicators of learning progress and behavioral stability. PPO is adopted as the optimization backbone because of its favorable empirical performance and practical stability in reinforcement learning and multi-agent training. Within this framework, curriculum progression is not treated as a fixed external schedule, but as an adaptive process guided by the current state of policy learning. The significance of this study lies in its attempt to connect curriculum design more closely with the actual coordination dynamics of multi-step decision making. Rather than relying on manually defined transitions, the proposed method seeks to align task progression with observable learning signals, thereby reducing mismatch between training stage and policy readiness. This design is expected to improve learning stability, increase task success rate, and reduce coordination failures across tasks with different complexity levels. More broadly, the study aims to provide a more effective training strategy for complex collaborative decision systems in which long-horizon dependency, online policy adaptation, and coordination robustness must be addressed simultaneously. By doing so, it contributes to the development of MARL methods that are better suited to realistic multi-step applications and offers a more practical perspective on how progressive training can support stable coordination in increasingly challenging environments.

2.Materials and Methods

2.1. Task Samples and Environment Description

This study used a simulated multi-agent decision environment with progressive coordination demands. A total of 12,000 tasks were generated. These tasks covered several levels of difficulty, ranging from simple coordination with weak dependency to multi-step planning with constraints and delayed rewards. Each task involved 3–6 agents. Agents made decisions under partial observability and only had access to local information. Stochastic transitions were introduced to represent uncertainty in the environment. Task settings were changed across episodes to reduce the risk of overfitting to fixed interaction patterns. The full task set was divided into training, validation, and test subsets at a ratio of 70%, 15%, and 15%, respectively.

2.2. Experimental Design and Baseline Setting

Two training settings were examined: direct training and curriculum-based training. In the direct training setting, agents were trained on tasks with mixed difficulty from the start. In the curriculum-based setting, training began with easier tasks and then moved to harder ones in stages. Task progression was controlled by a dynamic scheduler based on learning status. To keep the comparison fair, both settings used the same policy network, optimizer, training budget, and hyperparameter values. Each experiment was repeated five times with different random seeds. The final results were reported as the average across runs. This design made it possible to assess the effect of curriculum learning under the same experimental conditions.

2.3. Measurement Methods and Quality Control

Performance was assessed using three indicators: task success rate, convergence speed, and coordination failure rate. Task success rate was defined as the proportion of tasks completed within a fixed decision horizon. Convergence speed was measured by the number of training episodes needed to reach a stable performance level. Coordination failure was recorded when agents could not complete the shared objective because their actions were not properly aligned. To reduce the influence of random variation, all results were averaged over repeated runs, and standard deviations were also calculated. The same hardware and software environment was used in all experiments. Hyperparameters were kept unchanged across different settings. Validation-based early stopping was applied to limit overfitting and to improve result reliability.

2.4. Data Processing and Model Formulation

Training data were collected as state-action-reward trajectories generated during agent interaction. Input observations were normalized before training to improve numerical stability. Reward values were scaled to a fixed interval to avoid unstable updates caused by large variation in return magnitude. Policy learning was carried out with PPO. Its clipped objective function is written as

$$L^{PPO}(\theta)=E_t\left[\min\left(r_t(\theta)\widehat{A}_t,\ \mathrm{clip}(r_t(\theta),1-\epsilon,1+\epsilon)\widehat{A}_t\right)\right]$$

where $r_t(\theta)$ is the ratio between the new policy and the old policy, and $\widehat{A}_t$ is the estimated advantage at step t.

To guide curriculum progression, reward variance and policy entropy were used as monitoring signals. Reward variance was calculated as

$$\sigma_r^2=\frac{1}{N}\sum_{i=1}^{N}(r_i-\bar{r})^2$$

where r_i is the reward obtained in the i-th sample and $\bar{r}$ is the mean reward. Policy entropy was defined as

$$H(\pi)=-\sum_{a}\pi(a|s)\log\pi(a|s)$$

where $\pi(a|s)$ denotes the probability of selecting action a under state s. A higher entropy value indicates broader exploration, while a lower value suggests that the policy has become more concentrated. Curriculum stages were updated only when both indicators met the preset conditions.

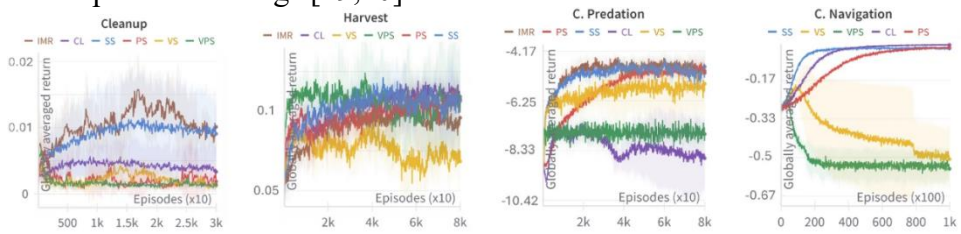
2.5. Training Procedure and Implementation Details

Training was performed for 1.5 million interaction steps. Under the curriculum setting, agents first learned on the easiest task group and then moved step by step to more difficult tasks. The scheduler checked reward variance and policy entropy every 5,000 steps and then decided whether the current task level should be advanced. The learning rate was set to 3×10^{-4}, and the discount factor was 0.99. Each batch contained trajectories collected from parallel environments to improve sampling efficiency. The model was implemented in PyTorch and trained on a GPU platform. After training, the learned policy was tested on unseen tasks to evaluate its generalization ability under different coordination demands.

3.Results and Discussion

3.1. Overall Performance under Progressive Task Difficulty

The curriculum-based training strategy produced better results than direct training across all task levels. When agents were trained on mixed-difficulty tasks from the beginning, the learning process was unstable and performance on tasks with deep dependency remained limited. In contrast, the curriculum setting led to a more stable increase in return and fewer sudden drops during training. As shown in Fig. 1, performance improved steadily as task difficulty increased step by step. On the test set, the success rate increased by 18.7%, and coordination failure was reduced by 31.2%. These results suggest that gradual exposure to harder tasks helps agents form stable coordination before facing more complex constraints. Similar trends have been reported in prior work, where structured training improved learning outcomes in cooperative settings [19,20].



(a) Cleanup (b) Harvest (c) C.Predation (d) C.Navigation

Figure 1 Learning curves for curriculum training and direct training across tasks of increasing difficulty.

3.2. Convergence Behavior and Learning Stability

Clear differences were observed in convergence speed and training stability. Under direct training, the reward curve showed large fluctuations, especially in the middle stage. The policy required more exploration before reaching a stable level. In the curriculum setting, convergence was reached earlier, and the training curve was smoother. The number of episodes required to reach stable performance decreased by 26.9%. This reduction suggests that early training on simpler tasks lowers the difficulty of policy updates. In multi-agent systems, unstable updates often arise from changing interactions between agents. The staged training process reduces this effect by limiting early coordination complexity. The result is consistent with previous findings that stable training schedules improve policy quality in cooperative learning [21,22].

3.3. Effect of Dynamic Scheduling on Coordination Quality

The dynamic scheduler had a clear effect on coordination quality. Reward variance was used to measure stability, while policy entropy reflected the level of exploration. When both measures were considered together, task progression was better aligned with the current learning stage. This avoided two common problems. First, agents were not moved to harder tasks too early. Second, training did not remain too long on simple tasks. As shown in Fig. 2, coordination failures caused by inconsistent actions became less frequent after agents passed intermediate stages. Compared with fixed curricula, the adaptive schedule responded more directly to training signals. This makes it less dependent on manually chosen thresholds or predefined stages. Similar issues with fixed schedules have been noted in earlier MARL studies, where limited flexibility reduced performance in complex environments [23,24].

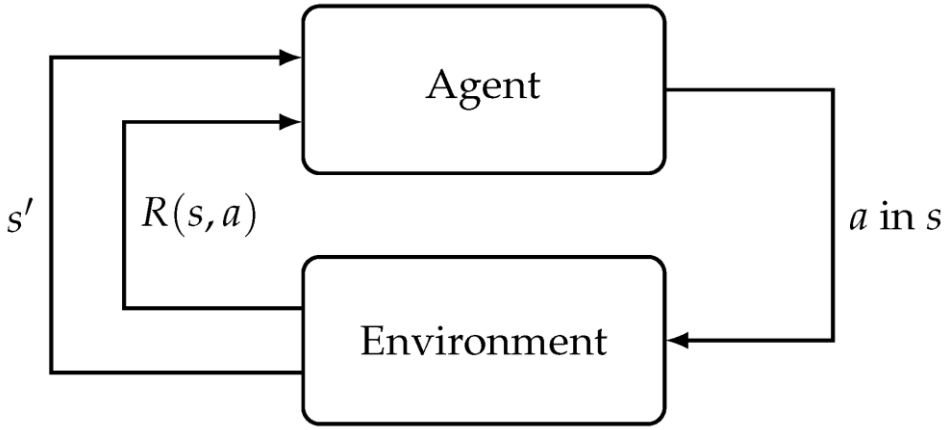


Figure 2. Coordination failure rates under different training settings across task levels.

3.4. Comparison with Existing Studies and Practical Implications

Compared with earlier studies, the present framework places more emphasis on task dependency and coordination depth during training. Many existing methods focus on algorithm design or specific applications, while the structure of the training process receives less attention. In addition, some studies rely on small benchmark sets with limited task variation. This makes it difficult to evaluate performance in more complex settings. In this study, the task set includes a wide range of coordination patterns and dependency structures. The results show that training order plays an important role in policy learning. However, several limitations remain. The environment is still simulated, and the number of agents is moderate. The scheduler also depends on predefined indicators. Future work may consider larger systems and learned scheduling strategies. Even so, the results support the use of structured training to improve robustness in collaborative decision tasks.

4.Conclusion

This study examined a curriculum-based training scheme for multi-agent decision tasks with increasing coordination demands. Tasks were arranged according to coordination difficulty and dependency depth. This setting led to more stable training and fewer failures caused by inconsistent actions. Compared with direct training, the proposed approach reached stable performance earlier and achieved higher success rates across task levels. The use of reward variance and policy entropy allowed task progression to follow the current learning state, which reduced the risk of moving too early or too late between stages. These results show that training order affects policy learning and should be treated as a key part of the learning process. The method can be applied to scenarios such as autonomous coordination, scheduling, and distributed control, where tasks involve multiple steps and strong dependencies. However, several limitations remain. The experiments were conducted in a simulated environment, and the number of agents was limited. The scheduler also depends on predefined signals. Future work may consider larger systems, more diverse benchmarks, and data-driven scheduling methods to improve adaptability and robustness.

References

Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review, 55(2), 895-943.

Xu, D., Liu, H., Qiu, D., & Ma, Q. (2026). Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models.

Shaukat, S. U., Khan, S., & Parkinson, S. (2025). A Review on Multi-step Attack Detection. IEEE Access.

Liu, S., Liu, X., & Feng, H. (2025, November). Research on AI-Driven Visual Design and Immersive Interactive Experiences Based on Multimodal Cognition and User. In

Proceedings of the 2025 International Conference on Digital Society and Intelligent Computing (pp. 734-740).
Abadi, A. S., & Soh, L. K. (2025). Challenges in Credit Assignment for Multi-Agent Reinforcement Learning in Open Agent Systems. arXiv preprint arXiv:2510.27659.
Gao, G., Gao, R., Gao, R., & Zhou, H. (2026). Engineering Analysis and Quantitative Research on the Platform-Based Evolution of Enterprise Communication Systems.
Del Rio, A., Jimenez, D., & Serrano, J. (2024). Comparative analysis of A3C and PPO algorithms in reinforcement learning: A survey on general environments. IEEe Access, 12, 146795-146806.
Xu, D., Chen, H., & Gui, H. (2026). Unified Online Estimation Method for SOC, SOH, and Power Capacity Considering Safety Boundary Consistency in Battery Management Systems.
Gupta, K., Mukherjee, D., & Najjaran, H. (2022). Extending the capabilities of reinforcement learning through curriculum: A review of methods and applications. SN Computer Science, 3(1), 28.
Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.
Lagomarsino, M., Merlo, E., Pupa, A., Birr, T., Krebs, F., Secchi, C., ... & Ajoudani, A. (2025). Intuitive Programming, Adaptive Task Planning, and Dynamic Role Allocation in Human–Robot Collaboration. Annual Review of Control, Robotics, and Autonomous Systems, 9.
Zhang, Y., Gu, W., & Wang, J. (2025). Research on First Article Inspection (FAI)-Driven Quality Assurance Methods for Wind Turbine Installation and Operation & Maintenance and Their Effect on Reliability Improvement. Available at SSRN 6094206.
Thang, T. B., & Binh, H. T. T. (2024). Multitask Augmented Random Search in deep reinforcement learning. Applied Soft Computing, 160, 111605.
Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.
Massaoudi, M. S., Abu-Rub, H., & Ghrayeb, A. (2023). Navigating the landscape of deep reinforcement learning for power system stability control: A review. IEEE access, 11, 134298-134317.
Zhang, Y., & Wang, J. (2026). Design and Implementation of a Computer-Aided Full Lifecycle Quality Management System for Wind Farms in Upgrades, Renovations, and Subcontractor Supervision.
Thaariq, Z. Z. A., & Karakuş, M. (2025). Designing a Methodological Framework for Assessing Teacher Readiness in Adaptive Curriculum Implementation within Indonesian Context. Journal of Curriculum and Learning Technologies (JCLT-Krisno), 1(2), 54-69.
Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). Quantitative Study on the Construction and Application Effectiveness of Graffiti Wall Painting Teaching Models in Public Space Contexts.
Damayanti, I., Baqy, I., & Rokhmat, J. (2025). A Systematic Review of Research Trends in Cooperative Learning Models in Science Learning. Contextual Natural Science Education Journal, 3(2), 1-14.
Wang, Y., Yin, X., Chen, J., & Wang, Y. (2026). Evidence-Based Study on Low-Burden Digital Phenotyping for Precision Screening of Oral Anti-Obesity Drug Efficacy.
Nungu, L., Mukama, E., & Nsabayezu, E. (2023). Online collaborative learning and cognitive presence in mathematics and science education. Case study of university of Rwanda, college of education. Education and Information Technologies, 28(9), 10865-10884.
Xu, D., Gui, H., & Chen, H. (2026). Research on Layered Control and Fault Recovery Mechanisms for Fast Charging Safety Diagnosis of High Voltage Battery Systems Under Charging Network Interoperability Conditions.

Kaven, L., Huke, P., Göppert, A., & Schmitt, R. H. (2024). Multi agent reinforcement learning for online layout planning and scheduling in flexible assembly systems. Journal of Intelligent Manufacturing, 35(8), 3917-3936.