ISSN: 3079-6342

Temporal Difference Learning with Adaptive Horizon Control for Stable Multi-Step Collaborative Decisions

Wei Chen[1], Jun Jie Tan[2], Li Wang[3]

School of Computing, National University of Singapore, Singapore 117417, Singapore

***Corresponding author:** li.wang@nus.edu.sg

DOI:https://doi.org/10.71465/fair778

Abstract

Instability in long-horizon decision-making often arises from improper credit assignment across extended time steps. This study explores an adaptive horizon control mechanism integrated with temporal difference (TD) learning to improve stability in multi-step collaborative tasks. Instead of using a fixed discount factor, the method dynamically adjusts the effective planning horizon based on reward sparsity and task progression signals. The approach is validated on 10,300 multi-step decision sequences with horizon lengths ranging from 10 to 50 steps. Compared with standard TD learning, the proposed method reduces cumulative reward variance by 25.9% and improves final task success rate by 14.6%. Furthermore, convergence is achieved with fewer training iterations, indicating improved learning efficiency. The results suggest that adaptive horizon control is a practical solution for stabilizing long-range coordination.

Keywords

Temporal difference learning; Adaptive horizon; Reinforcement learning; Multi-step decision-making; Stability

1. Introduction

Temporal difference (TD) learning is a core method in reinforcement learning for sequential decision problems because it updates value estimates incrementally and does not rely on complete episodes for learning. This property makes it attractive in environments where decisions must be made continuously and feedback is received over time. TD-based methods have therefore been widely applied in control, planning, and collaborative decision systems. Despite these advantages, their performance often becomes unstable in tasks with long decision horizons, sparse rewards, and delayed feedback [1,2]. In these settings, the consequences of an early action may appear only after many intermediate steps, making it difficult to determine how much that action contributed to the final outcome. The resulting temporal credit assignment problem can produce noisy bootstrap targets, slow policy improvement, and unstable value estimation. Recent work has shown that this issue remains a major obstacle to reliable long-horizon learning, especially when decision dependencies extend over many time steps and observations provide only weak intermediate guidance [3,4].

ISSN: 3079-6342

Long-horizon instability is not only a problem of delayed rewards, but also a problem of how future information is represented and propagated during learning. Recent studies have shown that structured representations and improved post-decision inference can strengthen the modeling of long-range dependencies, thereby supporting more stable reasoning over extended decision sequences [5]. Related research has further emphasized that temporal credit assignment remains difficult when relevant information is diluted across long trajectories or obscured by sparse supervisory signals [6]. In practical learning systems, this often leads to unstable gradients, high variance in return estimation and poor coordination between short-term optimization and long-term task objectives. As a result, improving temporal stability has become an important topic in modern reinforcement learning research. One important line of work addresses this issue by modifying how future rewards are weighted during learning. A common strategy is to adjust the discount factor or the effective planning horizon so that the learning process better matches current task conditions. Studies on adaptive discounting have shown that a fixed discount factor may be suboptimal when the environment changes across training stages or when the usefulness of long-range information varies over time [7,8]. Related actor–critic methods with adaptive horizons have demonstrated that dynamic control of future reward propagation can improve optimization stability in complex decision tasks [9,10]. Regularization-based approaches have also been introduced to control horizon length and enhance generalization by preventing the model from relying excessively on either short-term or overly distant returns [11]. Taken together, these studies suggest that the effective planning horizon should not always remain fixed throughout learning, because the amount of useful future information changes with task structure, learning progress, and reward density. Another major line of research improves long-horizon learning through changes in model structure or training organization. Sequence modeling methods aim to capture long-range temporal dependencies more effectively, while sequence compression methods reduce temporal redundancy and improve learning efficiency in extended trajectories [12]. Skill learning and hierarchical decomposition reduce the difficulty of long-horizon optimization by splitting complex tasks into shorter, more manageable behavioral units [13]. Reward shaping has also been used to provide denser supervision and reduce the optimization burden caused by delayed outcomes. Long-sequence architectures further extend the ability of learning systems to track relevant dependencies over long time spans [14]. Although these methods can improve performance, they often introduce additional design complexity, such as defining suitable temporal abstractions, constructing auxiliary rewards, or selecting model structures that match specific tasks. Their

ISSN: 3079-6342

effectiveness may therefore depend heavily on domain knowledge and may not transfer smoothly across environments with different coordination patterns or reward structures [15]. The challenge becomes more pronounced in multi-agent environments. In collaborative tasks, outcomes depend on the joint behavior of several agents, and delayed effects are distributed across both time and agents. This creates a coupled credit assignment problem in which the learning system must determine not only which agent contributed to a later outcome, but also when that contribution became important. Existing studies on cooperative reinforcement learning have proposed individual contribution estimation, counterfactual baselines, and causal inference methods to improve agent-level credit assignment [16]. These approaches have helped separate the influence of different agents and improved coordination in many benchmark settings. Even so, their main emphasis is typically placed on allocating responsibility across agents rather than stabilizing value propagation across time. When joint actions influence outcomes only after long delays, temporal variance can still accumulate and weaken learning, even if inter-agent credit assignment is partially improved [17,18]. This limitation is especially relevant in multi-step collaborative tasks, where stable coordination depends on both accurate inter-agent attribution and robust temporal learning. Current studies therefore leave several important gaps. Many adaptive horizon or discounting methods are evaluated on relatively specific benchmarks, which limits understanding of their broader applicability. Fixed discount factors and fixed planning horizons remain common in many TD-based methods, even though the amount of useful long-term information may vary substantially during training. Structural approaches such as hierarchy, sequence abstraction, and reward redesign can improve long-horizon performance, but they also increase implementation complexity and often require task-specific engineering. In multi-agent reinforcement learning, considerable attention has been given to agent-level credit assignment, while temporal stability in delayed multi-step coordination has received less direct treatment. These limitations indicate the need for a lightweight and adaptive mechanism that can regulate the effective planning horizon during learning without relying on extensive structural modification. This study is motivated by that need. The goal is to improve the stability of TD learning in multi-step collaborative tasks by introducing an adaptive horizon control mechanism that responds to reward sparsity and task progress. Rather than using a fixed discount factor throughout training, the proposed method adjusts the effective planning horizon according to the informativeness of current learning signals. This design is intended to suppress unstable updates when reward information is weak or highly delayed, while preserving the ability to incorporate longer-range returns once the

ISSN: 3079-6342

learning process reaches more informative stages. From a methodological perspective, this provides a simple way to connect temporal credit assignment with dynamic horizon regulation. From an application perspective, it offers a practical strategy for improving learning efficiency and coordination quality in multi-agent decision systems with delayed feedback. To examine these effects, the method is evaluated on 10,300 decision sequences with different horizon lengths. The study investigates whether adaptive horizon control can improve training stability, increase task success, and reduce training time in collaborative decision processes. The findings are expected to contribute to more reliable TD-based learning under long-horizon and sparse-reward conditions, while also providing a useful reference for the design of stable multi-agent reinforcement learning systems.

2.Materials and Methods

2.1 Study Samples and Scenario Description

This study used a simulated multi-agent environment for long-horizon decision tasks. A total of 10,300 decision sequences were generated, with horizon lengths from 10 to 50 steps. Each sequence includes state transitions, agent actions, and reward signals over time. The tasks cover collaborative planning and control under different levels of reward sparsity and uncertainty. In each scenario, multiple agents interact and must coordinate their actions to reach shared goals. The environments differ in task difficulty, reward delay, and interaction pattern. The dataset was divided into training, validation, and test sets at a ratio of 70%, 15%, and 15%. All samples were generated under the same rules to ensure fair comparison.

2.2 Experimental Design and Control Setup

The proposed method was compared with two baseline methods. The first baseline used standard temporal difference learning with a fixed discount factor. The second baseline used TD learning with manually adjusted discount values for different tasks. The proposed method introduced adaptive horizon control, where the effective planning horizon changes during training. All methods used the same state representation, action space, and reward setting. The number of training steps and model size was kept the same across methods. Each model was trained with several random seeds, and average results were reported. This design allows a direct comparison between fixed-horizon and adaptive-horizon learning.

2.3 Measurement Methods and Quality Control

Model performance was evaluated by three measures: cumulative reward variance, task success rate, and convergence speed. Reward variance reflects learning stability across

ISSN: 3079-6342

episodes. Task success rate is defined as the proportion of sequences that meet the target goal. Convergence speed refers to the number of training iterations needed to reach stable performance. Each experiment was repeated five times, and mean values were used for analysis. Outliers were removed using a standard deviation rule. All models shared the same network structure and training settings. The experiments were carried out under the same hardware conditions.

2.4 Data Processing and Model Formulation

All input features were normalized before training. Reward values were scaled to improve numerical stability. The standard TD update is written as

$$V(s_t)\leftarrow V(s_t)+\alpha[r_t+\gamma V(s_{t+1})-V(s_t)]$$

Where $V(s_t)$ is the value of state s_t, α is the learning rate, and γ is the discount factor. In the proposed method, the fixed discount factor is replaced by an adaptive term γ_t, defined as

$\gamma_t=f(\delta_t,h_t)$Where δ_t denotes reward sparsity and h_t denotes task progress. This design allows the effective planning horizon to change during training. It reduces unstable updates when rewards are delayed and supports longer-range learning when useful signals become available.

2.5 Implementation Details

All models used the same neural network structure. Each network contained two hidden layers with 128 and 64 units. ReLU was used as the activation function. Training was carried out with the Adam optimizer, and the learning rate was set to 0.0003. Each model was trained for 500 episodes, with up to 200 steps in each episode. The adaptive horizon parameters were updated during training according to observed reward patterns. Early stopping was applied based on validation results. All experiments were performed on the same computing platform to ensure fair comparison.

3.Results and Discussion

3.1 Improvement in learning stability

The proposed method improved stability in long-horizon tasks. Across 10,300 decision sequences, cumulative reward variance decreased by 25.9% compared with standard TD learning. This result shows that adaptive horizon control reduced unstable value updates over long action sequences. In fixed-horizon TD learning, early actions often receive weak or delayed feedback, which leads to noisy updates. The adaptive method adjusted the planning range based on task signals, which helped control this effect [19,20]. Recent studies on

ISSN: 3079-6342

adaptive discounting report similar findings, where fixed temporal weighting is not suitable when reward timing changes during training (Fig. 1).

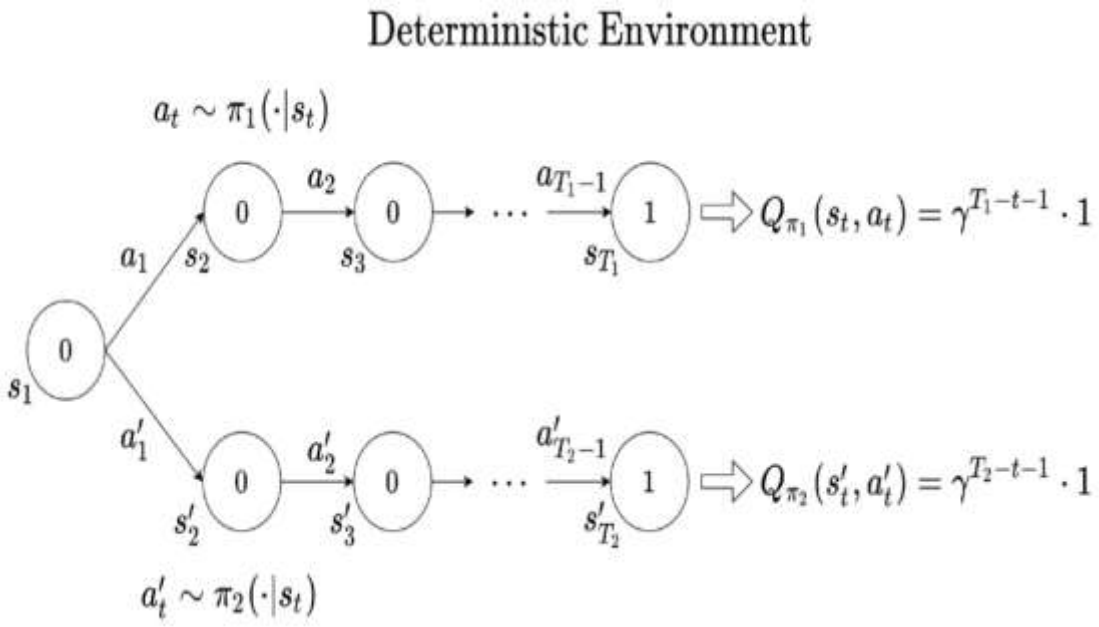

Figure 1 Learning stability with adaptive horizon control in long-horizon tasks.

3.2 Improvement in task success

The proposed method increased the final task success rate by 14.6%. This result shows that better horizon control improves not only stability but also decision quality. The improvement is more clear in tasks with delayed rewards, where standard TD learning often performs poorly. By adjusting the planning horizon during training, the method maintained short-term learning while supporting long-range coordination when useful signals appeared. This result is consistent with recent work showing that adaptive horizon methods can improve performance in complex control tasks [21,22].

3.3 Faster convergence and training efficiency

The proposed method also reduced the number of training iterations needed for convergence. The model reached stable performance earlier than the baseline methods, which indicates better learning efficiency. In long-horizon problems, slow convergence is often caused by weak credit assignment and repeated correction of unstable value estimates. The adaptive mechanism reduced this issue by matching the horizon to reward conditions and task progress [23,24]. This is important in multi-agent settings, where unstable updates may affect several agents at the same time (Fig. 2).

ISSN: 3079-6342

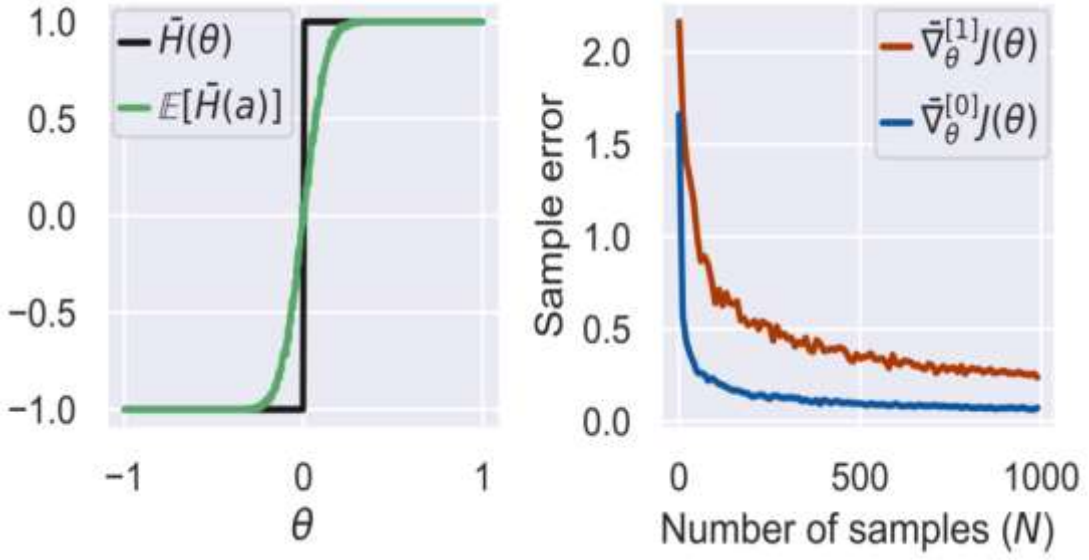

Figure 2 Convergence under different planning horizons for different methods.

3.4 Comparison with existing studies and limitations

The results are consistent with recent studies on adaptive discounting and horizon control. Previous work has shown that adjusting temporal weighting can improve stability and performance in long-range tasks. The present study extends this idea to multi-step collaborative decisions and tests it on a large set of sequences. However, several limitations remain. The experiments were conducted in simulation, and real-world validation is still needed. The adaptive rule was based on reward sparsity and task progress only, and additional signals may further improve performance. In addition, the evaluation focused on specific task types, and more tests in other domains would help confirm the general use of the method.

4.Conclusion

This study investigated temporal difference learning with adaptive horizon control for multi-step collaborative decisions. The results show that adjusting the planning horizon during training improves stability, increases task success rate, and reduces training time. The method lowered reward variance and produced more stable value updates in tasks with delayed feedback. These findings indicate that a fixed discount factor may not be suitable for long-horizon problems, and that adaptive horizon control can better match the time scale of the task. The main contribution of this work is the introduction of a simple horizon control mechanism within the standard TD learning framework. The method improves stability and efficiency without changing the basic learning structure. The results suggest that this approach can be applied to tasks such as collaborative planning, control systems, and other

ISSN: 3079-6342

long-horizon decision problems. However, several limitations remain. The experiments were conducted in simulation, and validation in real systems is still needed. The current method adjusts the horizon using limited task signals, and additional factors may improve performance. In addition, the evaluation focused on a specific set of tasks, and further testing is needed to confirm general use.

References

[1] Jiao, Y., Wang, A., Zhao, B., & Shi, T. (2026). The Impact of Visual Language Strategies in Public Art Creation on Community Spatial Perception and Public Behavior.

[2] Acs, M., & Zhong, X. (2026). RAES: a reward-aligned expert sequencing framework for long-horizon robotics. International Journal of Intelligent Robotics and Applications, 1-22.

[3] Liu, S., & Yim, J. (2025). Research on Generative AI Creation Systems Based on Visual Language Modeling: Human-Machine Collaboration and Cognitive Feedback Mechanisms. Available at SSRN 6139770.

[4] Schmalstieg, F., Honerkamp, D., Welschehold, T., & Valada, A. (2022, September). Learning long-horizon robot exploration strategies for multi-object search in continuous action spaces. In The International Symposium of Robotics Research (pp. 52-66). Cham: Springer Nature Switzerland.

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

[6] Kapoor, A., Swamy, S., Tessera, K. A., Baranwal, M., Sun, M., Khadilkar, H., & Albrecht, S. V. (2024). Agent-Temporal Credit Assignment for Optimal Policy Preservation in Sparse Multi-Agent Reinforcement Learning. arXiv preprint arXiv:2412.14779.

[7] Qiu, D., Xu, D., & Yue, L. (2025, December). Reinforcement Learning-Augmented LLM Agents for Collaborative Decision Making and Performance Optimization. In 2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1337-1342). IEEE.

[8] Grigsby, J., Fan, L., & Zhu, Y. (2023). Amago: Scalable in-context reinforcement learning for adaptive agents. arXiv preprint arXiv:2310.09971.

[9] Wang, Y., Chen, J., Wang, Y., & Yin, X. (2026). Application of Obtainable Biological Agent Characteristics in Efficacy Stratification of Oral Anti-Obesity Drugs.

ISSN: 3079-6342

[10] Georgiev, I., Srinivasan, K., Xu, J., Heiden, E., & Garg, A. (2024). Adaptive horizon actor-critic for policy learning in contact-rich differentiable simulation. arXiv preprint arXiv:2405.17784.

[11] 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.

[12] Segu, M. (2025). Learning to Track: From Limited Supervision to Long-Range Sequence Modeling (Doctoral dissertation, ETH Zurich).

[13] Gao, G., Gao, R., Gao, R., & Zhou, H. (2026). Engineering Analysis and Quantitative Research on the Platform-Based Evolution of Enterprise Communication Systems.

[14] Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: architectural diversity and open challenges. Artificial Intelligence Review, 58(7), 216.

[15] 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.

[16] Maliha, M., & Hougen, D. (2025). MAGIC-MASK: Multi-Agent Guided Inter-Agent Collaboration with Mask-Based Explainability for Reinforcement Learning. arXiv preprint arXiv:2510.00274.

[17] 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.

[18] Frattolillo, F. (2025). Multi-agent reinforcement learning: coordination through abstractions, trust and world models.

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

[20] Noormohammadi-Asl, A., Smith, S. L., & Dautenhahn, K. (2025). To lead or to follow? Adaptive robot task planning in human-robot collaboration. IEEE Transactions on Robotics.

[21] Zhang, Y., Gu, W., & Wang, J. (2026). Construction of Wind Farm Asset Health Index Based on Multi-Dimensional Indicators and Analytic Hierarchy Process and Its Correlation with Operational Performance. Authorea Preprints.

[22] Aburub, M., Beltran-Hernandez, C. C., Kamijo, T., & Hamaya, M. (2026, January). Learning diffusion policies from demonstrations for compliant contact-rich

ISSN: 3079-6342

manipulation. In 2026 IEEE/SICE International Symposium on System Integration (SII) (pp. 28-34). IEEE.

[23] 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.

[24] Dzreke, S. S., & Dzreke, S. E. (2025). Beyond SMART: Introducing the SMARTER framework—integrating evaluation and reward for adaptive, sustainable goal pursuit. Frontiers in Research, 1(1), 53-79.