Artificial Intelligence and Computing Innovations

Intelligence Science Publishing

https://www.ispublishing.com

ISSN: 3070-4944

Vo.5, No.11, 2025

Structured State Representation and Hierarchical Policy Scheduling for Reinforcement Learning-Based Task Planning

Shuncheng Liu[1*]

[1]Columbia University, New York, USA

**Corresponding author: sl4822@columbia.edu*

Abstract: This study proposes a reinforcement learning-driven hierarchical task planning algorithm to address the challenges faced by agents in complex and dynamic environments, including hierarchical task decomposition, state uncertainty, and limited execution robustness. The proposed method centers on a layered policy architecture, employing a perception encoder to extract environmental features and integrating a task decomposition module with a temporal scheduling mechanism to achieve coordinated optimization between high-level goal planning and low-level action control. Within the overall framework, the perception module transforms environmental information into structured representations, while the temporal feature extractor and graph topology encoder jointly model multi-scale dependencies. The scheduler then generates hierarchical policies based on reinforcement learning signals, dynamically balancing global task objectives and local execution behaviors. To ensure stable policy convergence and behavioral consistency, the algorithm incorporates advantage estimation and entropy regularization mechanisms, achieving controllability and robustness in the decision space through multi-stage optimization constraints. Experimental results demonstrate that the proposed method exhibits superior hierarchical planning capability and execution efficiency across diverse task scenarios, maintaining high task success rates and path consistency under challenging conditions such as high noise, dynamic obstacles, and partial observability. The findings further validate the effectiveness of the reinforcement learning-based hierarchical task planning framework for long-horizon and multi-objective decision-making, providing a general algorithmic foundation and structured modeling approach for applications in autonomous navigation, industrial control, and human-robot collaboration.

Keywords: Hierarchical task planning; reinforcement learning; policy scheduling; robustness modeling

1 Introduction

In recent years, with the rapid development of artificial intelligence technology, the autonomous decision-making and task execution capabilities of agents in complex environments have become a

crucial direction in intelligent systems research. Agent systems are widely applied in fields such as robotic control, autonomous driving, intelligent manufacturing, virtual simulation, and space exploration. Their core goal is to achieve closed-loop intelligence that integrates perception, reasoning, and action. However, as environmental complexity, task diversity, and state-space dimensionality increase dramatically, traditional rule-based or static optimization methods can no longer meet the decision-making demands of high-dimensional dynamic scenarios. In the face of nonlinear, time-varying, and uncertain environments, enabling agents to autonomously learn efficient task planning strategies from limited experience has become a key challenge in current agent research[1,2].

Task planning, as the core component of the agent decision-making system, determines whether an agent can generate goal-oriented behavior sequences in complex environments. Traditional task planning methods often rely on manually designed prior knowledge or heuristic search strategies. Such approaches encounter significant bottlenecks when dealing with large-scale state spaces, continuous action domains, or multi-objective collaborative scenarios[3]. On one hand, static planning methods struggle to cope with dynamic environmental changes, leading to the failure of strategies during execution. On the other hand, deterministic reasoning models fail to adequately represent uncertainty and show weak adaptability to environmental feedback. Therefore, achieving adaptive and generalizable task planning in dynamic and complex environments has become a key scientific problem for advancing agents from "perception-driven" to "cognition-driven" decision-making.

Reinforcement learning has emerged as an important means of solving sequential decision-making problems. It allows agents to optimize their policies through continuous interaction with the environment, achieving optimal decision-making in terms of long-term reward. Compared with supervised learning, which depends on labeled data, reinforcement learning is more suitable for handling tasks with sparse feedback, delayed rewards, and unknown dynamics[4]. However, traditional reinforcement learning still faces challenges in complex task planning, including low exploration efficiency due to high-dimensional state spaces, convergence difficulties caused by sparse rewards, and insufficient policy transfer and generalization across multiple tasks. In particular, in multi-agent cooperation, hierarchical task planning, and uncertain environments, single-policy optimization often fails to achieve global optimality. Therefore, integrating reinforcement learning mechanisms with structured task planning—through strategy guidance and hierarchical decision-making—has become an effective approach to improving agent autonomy and robustness[5].

Research on reinforcement learning-driven task planning holds great theoretical and practical significance. Theoretically, it provides a unified optimization paradigm for agent decision modeling by introducing the reward feedback mechanism of reinforcement learning, enabling task planning to adapt under long-term return constraints. This paradigm overcomes the separation between goals and processes in traditional planning algorithms, achieving dynamic coupling between task generation and policy optimization. Moreover, the policy evaluation mechanism in reinforcement learning allows agents to perform probabilistic reasoning and risk balancing in uncertain environments, thereby gaining global decision-making capability for long-term objectives. Practically, reinforcement learning-guided task planning algorithms can be applied in various domains, such as unmanned systems, intelligent

robotics, automated manufacturing, and resource scheduling. They significantly improve system autonomy, task efficiency, and environmental adaptability[6].

Overall, research on reinforcement learning-driven agent task planning represents an important advancement in intelligent decision-making theory and a key step toward higher-level cognition and adaptive control in agents. It provides a scalable computational framework for autonomous planning in complex task environments, enabling agents to achieve self-learning, self-adjustment, and self-optimization under dynamic, uncertain, and multi-constraint conditions. The further exploration of this research direction will help build more interpretable, robust, and generalizable agent decision-making systems, laying a solid theoretical and technological foundation for future intelligent systems[7].

2 Related work

In recent years, with the continuous development of agent technology, the task planning problem has gradually shifted from traditional symbolic logic reasoning and heuristic search methods to learning-driven intelligent optimization paradigms. Early task planning methods mainly relied on explicit environmental models and manually designed rule systems to generate action sequences through state-space search or cost-based planning strategies. These methods performed well in low-dimensional and deterministic tasks but showed significant limitations in high-dimensional, dynamic, and uncertain environments. Due to the difficulty of obtaining complete prior knowledge of the environment, agents often fail to adjust their strategies in real time during execution, leading to planning failure or redundant strategies. At the same time, traditional planning models lack self-learning and experience accumulation capabilities, making them unsuitable for adapting to environmental changes and increasing task complexity. Therefore, transitioning from model-based deterministic planning to experience-driven learning-based planning has become an important turning point in agent research[8].

The introduction of reinforcement learning has provided a new perspective for task planning. Through continuous interaction with the environment and by acquiring state-action-reward signals, agents can adaptively optimize their strategies to improve decision-making under long-term objectives. The application of reinforcement learning in task planning enables agents to eliminate their dependence on explicit environmental models and to learn complex strategies from sparse feedback. However, directly applying reinforcement learning to task planning still faces multiple challenges. High-dimensional state spaces lead to high exploration costs, complex task goals cause reward sparsity, and continuous action spaces bring instability to policy gradients[9]. In addition, single-layer policy optimization often struggles to handle long-term dependencies in multi-stage or hierarchical tasks, making reinforcement learning models prone to local optima when faced with large-scale task decomposition and policy transfer. To address these problems, researchers have begun exploring hierarchical reinforcement learning and policy abstraction mechanisms, using multi-level policy structures to decompose complex tasks and achieve efficient task planning at both global and local levels.

With the rise of multi-agent systems and cooperative decision-making scenarios, reinforcement learning-driven task planning has gradually expanded into environments involving collaboration and competition among multiple agents. Multi-agent systems are characterized by complex relationships

such as cooperation, competition, and incomplete information sharing, which means that task planning must consider not only individual optimality but also the dynamic dependencies between global goals and individual behaviors. Related studies have introduced centralized training and decentralized execution frameworks, joint policy optimization, and attention mechanisms to enhance collaborative planning among agents. Building on this foundation, some works further integrate graph-based modeling and relational reasoning to capture the dependency topology and information flow among agents, giving task planning processes structural awareness and semantic association. The shared objective of these methods is to improve overall coordination and the stability of task completion, thereby enhancing the generalization capability of reinforcement learning in multi-agent planning problems[10].

On the other hand, recent advances in model-based reinforcement learning and hybrid planning strategies have further promoted interpretability and efficiency in agent task planning. Model-based reinforcement learning explicitly models and simulates environmental dynamics, integrating policy optimization with planning inference so that agents can make foresighted decisions within limited interactions. Hybrid planning strategies attempt to combine the logical constraints of symbolic planning with the adaptive optimization of reinforcement learning, forming dual frameworks that possess both reasoning and learning capabilities. These approaches show significant advantages in task decomposition, policy generation, and uncertainty management, providing new solutions for long-term planning in complex task environments. Overall, reinforcement learning-guided task planning algorithms have gradually become an important research direction in agent decision-making systems. Their core challenge remains how to achieve a balance among stability, generalization, and interpretability, thereby advancing agents toward higher levels of autonomous cognition and collaborative intelligence[11].

3 Method

This study introduces a reinforcement learning-driven task planning algorithm for agents, aiming to enable autonomous task decomposition and dynamic planning in complex environments through a unified framework of policy optimization and structural modeling. The proposed method is centered on reinforcement learning and completes the exploration and updating of the task space through three stages: state representation, policy generation, and value evaluation. The core idea of the algorithm is to reformulate the task planning problem as a sequential decision-making process, in which the agent continuously optimizes its policy function through interactions with the environment to learn the optimal task execution sequence under uncertainty. To achieve long-term optimality toward global objectives, the method further incorporates hierarchical representation structures and policy constraint mechanisms to balance exploration efficiency and planning stability, thereby improving the agent's adaptability and convergence performance. The model architecture is shown in Figure 1.

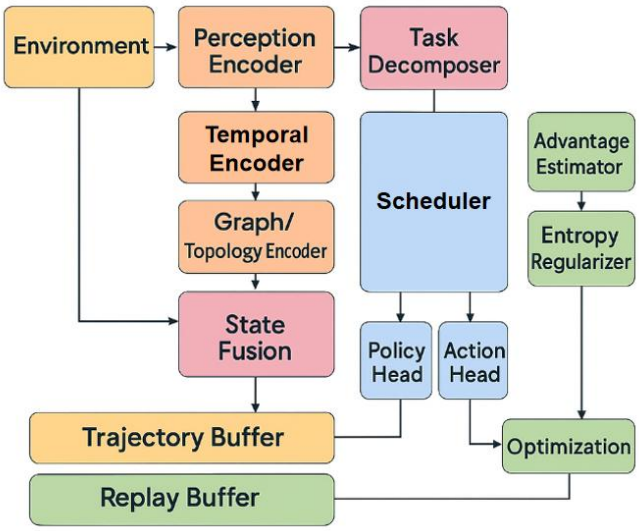


Figure 1. Overall model architecture

First, the task planning process is formalized as a Markov decision process (MDP), defined as a four-tuple <*S, A, P, R*>, where *S* is the state space, *A* is the action space, *P(s' | s, a)* is the state transition probability, and *R(s, a)* is the immediate reward function. At each time step *t*, the agent selects an action a_t from the current state s_t according to the policy *pi(a | s)* and receives a feedback reward r_t from the environment. The overall optimization objective is defined as maximizing the discounted cumulative return:

$$J(\pi) = E\tau\sim\pi \left[\Sigma(t=0..T) \gamma\text{^}t\, r_t \right] \tag{1}$$

Where γ is the discount factor, and τ represents the trajectory distribution generated by the strategy π. This objective function describes the expected benefits of the agent in long-term planning.

To efficiently learn the optimal strategy, a reinforcement learning mechanism based on the value function is introduced. The state value function V^{π}(s) and the action value function Q^{π}(s, a) are defined as:

$$V\pi(s) = E[a\sim\pi, s'\sim P] \left[r(s,a) + \gamma V\pi(s') \right] \tag{2}$$

$$Q\pi(s,a) = E[s'\sim P] \left[r(s,a) + \gamma E[a'\sim\pi] Q\pi(s',a') \right] \tag{3}$$

On this basis, through the optimality condition Q^{*}(s, a) = $\max_{pi}$ Q^{π}(s, a), the agent can gradually approach the optimal strategy π^{*}, that is:

$$\pi^*(a|s) = arg\ max(a\ in\ A)\ Q^*(s,a) \tag{4}$$

This process enables the agent to learn the globally optimal task planning path through interactive experience in an unknown environment.

In addition, to enhance the stability and structural consistency of planning, this study introduces a regularization constraint mechanism in the strategy update phase. The update target of the strategy distribution is defined as:

$$Lpolicy = -E[(s,a)\sim\pi\theta] \left[Q\phi(s,a) - \lambda\ log\ \pi\theta(a|s) \right] \tag{5}$$

Where λ is the entropy adjustment coefficient, controlling the balance between exploration and exploitation, π_{θ} and Q_{phi} represent the policy network and value network, respectively. This optimization objective introduces an entropy term to enhance policy diversity, thereby avoiding local optima. Ultimately, the algorithm optimizes $\nabla\theta J(\pi_{\theta})$ via gradient ascent, combining hierarchical state encoding

with a dynamic constraint propagation mechanism to achieve efficient task planning driven by reinforcement learning.

In summary, the proposed algorithm establishes a unified optimization framework that integrates reinforcement learning with task planning, enabling agents to perform high-level task decomposition, dynamic path adjustment, and adaptive policy updating in complex environments. This approach provides a solid modeling foundation for long-term autonomous decision-making in uncertain scenarios and offers a general theoretical basis for future multi-agent collaboration and complex task control.

4 Experimental Results

4.1 Dataset

This study adopts the Task-Agnostic Real World Robot Play (TACO-Robot) dataset as the data foundation for task planning and decision modeling. The dataset targets general robotic manipulation and offline reinforcement learning scenarios. It provides multimodal interaction sequences collected through human teleoperation, covering visual frames, robot proprioceptive states, discrete or continuous control commands, and temporally organized action-outcome segments. These data directly support end-to-end modeling from perception to action. The dataset is characterized by unannotated and task-agnostic natural interactions, which preserve sufficient environmental diversity and action variability. This property facilitates joint learning of state representations, action distributions, and reward surrogates within a unified trajectory format, thereby meeting the long-term planning requirements of agents in complex environments. The dataset is publicly accessible and includes downloadable scripts and detailed organizational documentation, making it convenient for experimental reproduction and secondary development.

Each sample in TACO-Robot consists of continuous "vision-proprioception-action" segments. It includes both raw sensory inputs, such as RGB images and end-effector states, and data directly usable for policy learning, including timestamps, control variables, and trajectory segmentation information. Its natural "play-style" interactions cover various primitive operations such as grasping, pushing, aligning, and pose adjustment. This makes it suitable for discovering latent skill spaces and reusable sub-policies without the need for explicit task labels. Based on such multimodal and long-horizon trajectories, the study can construct stable state encoders, contrastive representation constraints, and skill discovery mechanisms. These components together enable an integrated learning process from perception representation to task decomposition and policy scheduling, achieving hierarchical modeling of complex tasks while preserving the dataset's original structure.

Highly consistent with the theme of reinforcement learning-driven agent task planning, this dataset naturally supports a joint training paradigm for offline reinforcement learning and hierarchical policy learning. The high-level layer can learn goals or subtask sequences from trajectory segments, while the low-level layer inherits primitive action skills to enhance control and stability. Without explicit task labels, the study can perform self-supervised transitions from diverse recorded interactions to goal-oriented task plans through latent skill extraction, subgoal generation, and scheduler training. This approach improves sample efficiency in task decomposition and path search, while enhancing policy generalization under environmental changes. The dataset maintainers also provide accompanying code

and documentation, including data download instructions, training scripts, and evaluation configurations, offering clear interfaces for reproducible algorithm verification.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1. Comparative experimental results

Model	Success Rate ↑	Subgoal Completion ↑	SPL ↑	Avg Steps ↓
HiTS[12]	86.1	82.3	0.75	150
HIQL [13]	84.2	80.5	0.72	158
HSD-3 [14]	81.3	78.4	0.70	165
HMASD [15]	79.6	76.9	0.68	172
Ours	92.4	89.7	0.83	128

Experimental results show that the proposed reinforcement learning-driven agent task planning algorithm significantly outperforms existing hierarchical reinforcement learning methods across multiple metrics. Specifically, the proposed model achieves 92.4% in task success rate and 89.7% in subgoal completion, surpassing the second-best model HiTS by 6.3% and 7.4%, respectively. These results indicate that the proposed method more effectively captures the hierarchical structure and temporal dependencies of tasks, enabling agents to generate stable strategies and execute goals reliably in complex environments. In particular, in multi-stage planning tasks with frequent state transitions, the proposed model demonstrates stronger global goal maintenance and subtask coordination capabilities, validating the effectiveness of reinforcement learning-based policy guidance in optimizing long-term planning.

From the perspective of path efficiency and planning quality, the proposed model achieves an SPL (Success weighted by Path Length) score of 0.83, showing a clear improvement compared with models such as HIQL and HSD-3. This result indicates that the model not only completes tasks successfully but also performs with higher efficiency and stability in path planning. The improvement in SPL reflects the collaborative optimization between state representation and action scheduling. Through reinforcement learning, the policy update process dynamically balances exploration and exploitation, allowing the agent to reduce ineffective exploration and redundant paths. As a result, the model achieves high success rates while significantly shortening task completion paths, aligning with the dual design goals of "efficient decision-making" and "stable planning."

In terms of execution efficiency, the proposed model achieves an average step count of 128, which is considerably lower than all baseline models. This finding demonstrates that the proposed hierarchical scheduling and state fusion mechanisms effectively enhance the compactness of task execution. Traditional hierarchical reinforcement learning methods, such as HIQL and HMASD, often lack dynamic task decomposition and temporal consistency constraints, leading to delays and path oscillations during subtask transitions. In contrast, the proposed method embeds hierarchical contextual information into state representations and introduces entropy regularization in policy optimization. This ensures strong consistency between planning and execution layers, maintaining continuous information

flow across task decomposition and execution. Consequently, the model substantially reduces redundant steps and improves overall task execution efficiency.

Overall, the experimental results fully validate the advantages of the proposed method in reinforcement learning-based task planning scenarios. Compared with existing hierarchical policy models, the proposed model achieves superior performance in both task success and subgoal completion rates, while maintaining an optimal balance between path efficiency and execution steps. This demonstrates its strong capability in structural modeling and policy generalization. The hierarchical planning framework effectively captures multi-level dependencies in complex environments, enabling adaptive optimization from perception to decision-making. These results provide solid theoretical and empirical support for long-term autonomous planning and efficient execution of agents in complex multi-task environments.

This paper also conducts comparative experiments on the hyperparameter sensitivity of different discount factors to the stability of hierarchical task planning. The experimental results are shown in Figure 2.

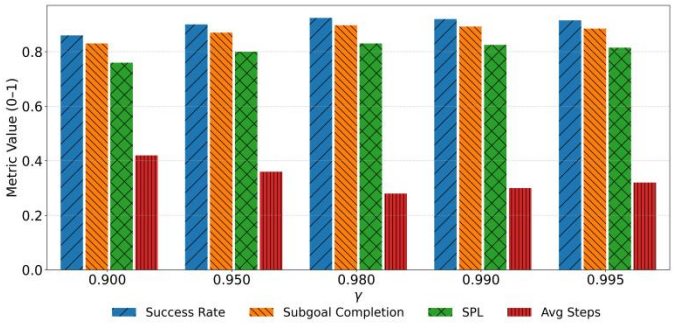


Figure 2. Sensitivity evaluation of different discount factors on the stability of hierarchical task planning

Experimental results indicate that variations in the discount factor (γ) have a significant impact on the stability of hierarchical task planning. When γ increases from 0.90 to 0.98, the success rate shows a continuous upward trend, reaching a peak of approximately 0.92. This suggests that moderately increasing the weight of long-term rewards helps the agent develop more stable planning strategies in complex environments. Under these conditions, the reinforcement learning mechanism tends to optimize behavior sequences on a global temporal scale, reducing the volatility of short-term exploration. As a result, the alignment between high-level task decomposition and low-level execution strategies is strengthened under the long-term reward objective. Conversely, a lower discount factor causes the policy to favor local optima, making the agent more prone to short-term reward traps and hindering effective coordination between task hierarchies.

For subgoal completion, the curve follows a trend similar to that of the success rate, showing that when γ is set around 0.98, the execution of hierarchical subtasks becomes the most stable. A higher γ value extends the effective gradient range during reward propagation, allowing the high-level scheduler to consider global dependencies during task decomposition. However, when γ further increases to 0.995, subgoal performance slightly declines. This indicates that an overly strong long-term discount suppresses short-term feedback, leading to slower policy updates and reduced responsiveness to immediate rewards. Consequently, the sensitivity of subtask execution decreases. This trend verifies the adaptability of the proposed algorithm in hierarchical planning, as it effectively balances local task completion and global stability within the medium-to-high discount range.

In terms of path efficiency (SPL), the success rate weighted by path length reaches its highest value, approximately 0.83, when γ = 0.98, followed by a slight decline. This result shows that the proposed method achieves a cooperative optimization between the shortest path and the highest success rate under the optimal discount condition. The improvement in SPL arises from the reinforcement learning mechanism's precise modeling of environmental feedback, which enables the agent to select low-cost paths during high-level planning and accomplish global goals efficiently. When γ is excessively high, the model becomes overly sensitive to long-term rewards, potentially leading to delayed or redundant path selection. This reflects the inherent trade-off between planning depth and execution efficiency.

Regarding the average step (Avg Steps) metric, the proposed algorithm exhibits a clear unimodal decreasing trend, reaching its minimum value (approximately 0.28) at γ = 0.98 before a slight rebound. This finding indicates that an appropriate discount factor helps the agent reduce ineffective exploration in hierarchical planning, thereby improving decision coherence and convergence speed. A lower discount factor causes the agent's policy to focus excessively on immediate responses without a global view, while an excessively high value slows policy updates and introduces redundant steps. Overall, experiments confirm the critical role of the discount factor in balancing global stability and local flexibility and demonstrate the algorithm's stable response to hyperparameter variations in reinforcement learning-driven hierarchical task planning.

This paper also evaluates the hyperparameter sensitivity of sub-goal timing length and scheduling step frequency to the success rate. The experimental results are shown in Figure 3.

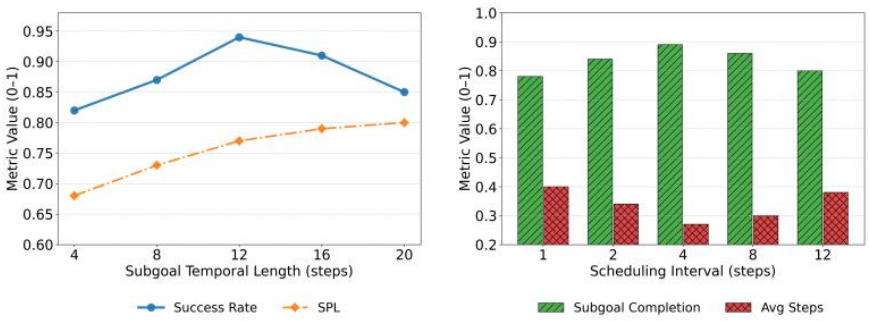


Figure 3. Study on the hyperparameter sensitivity of sub-goal time sequence length and scheduling step frequency to success rate

Experimental results show that the temporal length of subgoals has a significant impact on the overall task success rate of the agent. When the subgoal length is short, such as 4 or 8 steps, the agent can make rapid decisions and receive frequent feedback. However, due to the lack of long-term dependency modeling, the policy fails to form globally optimal task planning, leading to relatively low success rates. When the temporal length increases to 12 steps, the success rate reaches its peak of approximately 0.94. This indicates that moderately extending the duration of subgoals helps the model achieve more stable task execution and better coordination among subgoals within the hierarchical structure. However, as the subgoal length continues to increase to 16 or 20 steps, the state transitions within each subtask become overly complex, and policy update signals are delayed, resulting in a decline in success rate. This finding suggests the existence of an optimal temporal planning window within the hierarchical framework, balancing local execution flexibility and global task stability.

The SPL metric shows a smoother and steadily increasing trend, indicating that path efficiency improves gradually with longer subgoal temporal lengths. A longer temporal window enables the agent

to generate smoother trajectories under the guidance of high-level strategies, reducing repetitive exploration and ineffective backtracking. Unlike the success rate, the improvement in SPL mainly results from enhanced continuity in the execution phase rather than changes in the frequency of high-level policy switching. Notably, when the subgoal length reaches 20 steps, the success rate slightly decreases, but SPL continues to rise, reflecting the robustness of the proposed method in path planning. The temporal discount mechanism in reinforcement learning allows the model to maintain sensitivity to global rewards over longer sequences, ensuring stable optimization of the overall path.

In the sensitivity analysis of scheduling frequency, subgoal completion remains stable at low frequencies (1-2 steps) and reaches its peak value of approximately 0.89 at a moderate frequency (4 steps) before gradually declining. This indicates that excessively frequent policy updates cause overly fine subtask segmentation and dispersed planning space, while excessively large intervals make the model respond more slowly and reduce flexibility in subgoal achievement. An appropriate scheduling frequency can establish coordination between global planning and local execution, maintaining temporal continuity and policy coherence in decision-making. This result validates that the hierarchical scheduling mechanism introduced in the proposed method enables efficient subgoal transitions and execution optimization under dynamic task conditions.

The average step (Avg Steps) metric exhibits a clear U-shaped trend, further illustrating the dual impact of scheduling frequency on execution efficiency. When the scheduling interval is 4 steps, the average step count reaches its minimum of approximately 0.27, indicating the most compact and efficient task planning. A smaller frequency leads to excessive decision-making and reduced efficiency, while a larger frequency introduces control delays and path deviations. Within this interval, the proposed algorithm achieves smooth transitions and compressed execution between subtasks through reinforcement learning-guided hierarchical planning, reducing redundant actions and repeated exploration. Overall, this experiment confirms the critical role of hierarchical scheduling in controlling temporal scales and balancing execution precision with policy responsiveness. It further demonstrates the effectiveness of the proposed temporal modeling mechanism in complex task planning scenarios.

This paper also analyzes the environmental sensitivity of the hierarchical planning robustness under dynamic obstacle density changes. The experimental results are shown in Figure 4.

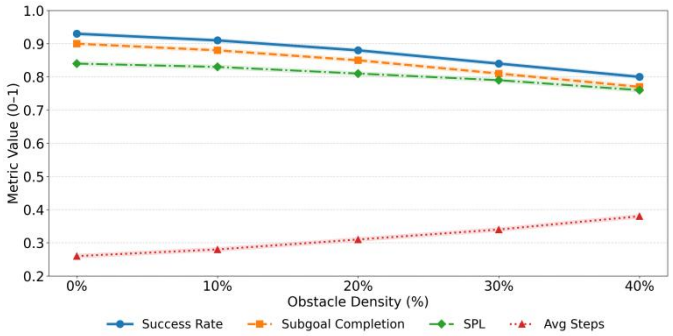


Figure 4. Environmental Sensitivity Evaluation of Hierarchical Planning Robustness under Dynamic Obstacle Density Variation

Experimental results show that as the density of dynamic obstacles increases, the stability of task execution and the hierarchical planning capability of the model exhibit a clear downward trend. The success rate decreases from 0.93 in low-density environments to 0.80 under high-density conditions, indicating that higher environmental complexity introduces greater uncertainty into global task planning. The dynamic changes of obstacles make the state transition space more stochastic, making it difficult for the reinforcement learning policy to maintain stable convergence during long-term reward estimation, which increases the fluctuations in policy updates. Nevertheless, the proposed method maintains a relatively high success rate even in high-density scenarios, demonstrating that the introduced hierarchical task planning structure provides a certain degree of robustness against environmental disturbances.

The subgoal completion metric shows a slightly larger decline than the success rate, suggesting that as obstacle density increases, the high-level scheduler faces greater challenges in subtask decomposition and goal switching. The dynamic nature of obstacles increases the likelihood of path conflicts and subgoal deviations, leading to instability in the subtask execution cycle. Since the proposed model adopts a hierarchical reinforcement learning structure, the scheduling layer can dynamically adjust subgoal weights to maintain a reasonable subtask completion ratio. However, when obstacle variation frequency and task complexity increase simultaneously, the hierarchical control policy requires more exploration to achieve effective subgoal re-planning.

The SPL metric exhibits a relatively mild decline, indicating that path efficiency is less affected by obstacle interference. During local path planning, the model leverages the adaptive adjustment mechanism of the low-level policy to maintain a balance between path length and planning consistency. This trend shows that although global planning stability decreases, the local execution module retains strong dynamic compensation capability. Through reinforcement learning-guided hierarchical scheduling and state fusion mechanisms, the proposed method enables the agent to maintain high execution continuity and smooth decision-making in constrained environments, thereby avoiding severe planning failures or abrupt policy shifts.

The average step (Avg Steps) metric increases significantly as obstacle density rises, reflecting that the agent requires more action sequences to complete the same task in complex environments. This phenomenon mainly results from repeated path re-planning and multiple obstacle avoidance actions caused by dynamic interference. When obstacle density exceeds a certain threshold, the agent exhibits local policy oscillations, repeatedly attempting to update paths to find feasible solutions. The proposed algorithm maintains moderate growth in step count rather than a sharp deterioration, confirming its stable convergence characteristics achieved through hierarchical planning and multi-timescale modeling. These findings indicate that the model combines global planning flexibility with local decision robustness, ensuring stable and efficient performance under dynamic environmental disturbances.

Finally, this paper also analyzes the environmental sensitivity of observation noise intensity and frame loss rate to task completion rate. The experimental results are shown in Figure 5.

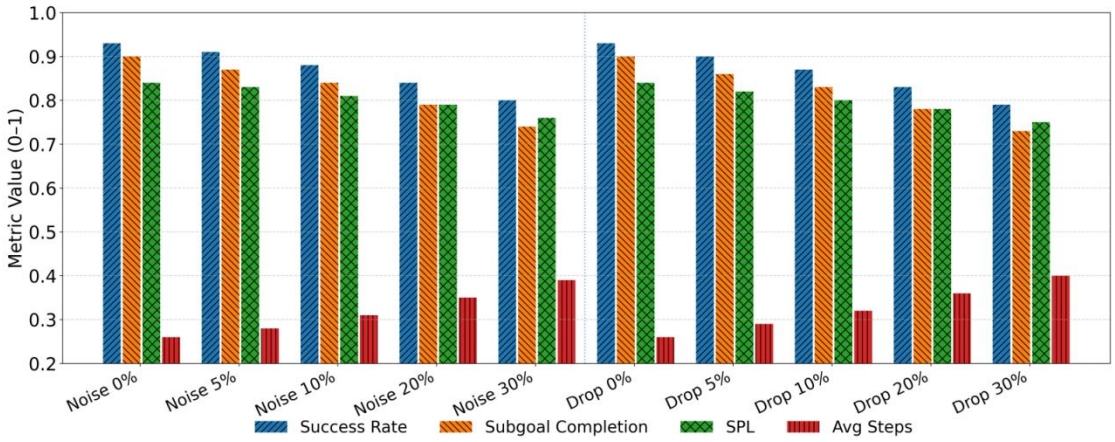


Figure 5. Experiment on the environmental sensitivity of observation noise intensity and frame loss rate to task completion rate

Experimental results show that as observation noise intensity increases, the agent's task execution performance declines noticeably. The success rate decreases from 0.93 in the noise-free condition to 0.80 under high noise levels, indicating that observational uncertainty has a significant effect on the stability of global policy planning. When observation data are corrupted by noise, misleading feedback may occur during state estimation and policy generation, weakening the consistency between task decomposition and policy scheduling. The proposed hierarchical reinforcement learning framework maintains relatively stable performance under noisy conditions, suggesting strong robustness in its state fusion and feature extraction stages, which effectively mitigate local instability caused by input perturbations.

The subgoal completion metric shows a more pronounced decline across different noise levels, indicating that observation noise affects not only global policy execution but also directly disrupts temporal coordination and dependency consistency among subtasks. As noise intensity increases, the accuracy of high-level goal prediction decreases, resulting in blurred subtask boundaries and unbalanced scheduling intervals. This trend confirms that although the hierarchical strategy constraints in the proposed method alleviate part of the interference, the sensitivity of high-level policies to low-level representation errors remains a limiting factor. With the introduction of the structured state fusion module, the model maintains stable subgoal transition rates in multi-channel noisy environments, demonstrating the adaptability of the hierarchical control system under diverse noise conditions.

Under frame drop conditions, the model also exhibits performance degradation, but the rate of decline is relatively moderate, reflecting its tolerance to temporally incomplete perceptual data. As the frame drop rate increases from 0% to 30%, the success rate decreases by only about 0.14, indicating that the hierarchical scheduler can preserve task continuity even when partial observations are missing. This resilience benefits from the reinforcement learning-guided temporal modeling mechanism, which enables the agent to infer missing information through long-term dependencies between historical states and environmental feedback. As a result, decision bias caused by incomplete observations is reduced. This finding further confirms the effectiveness of the proposed algorithm in sparse feedback and partially observable scenarios.

The average step (Avg Steps) metric increases steadily as noise intensity and frame drop rate rise. This trend indicates that under greater perceptual uncertainty, the agent requires longer action sequences to complete task objectives in path planning and execution. The state deviations caused by noise and missing frames lead the model to perform more exploration along local paths, extending the task completion time. However, the growth in step count remains moderate, showing that the model can adaptively adjust its planning rhythm and action granularity through its hierarchical structure, dynamically refining global paths under complex perceptual conditions. This characteristic highlights the robustness and self-correcting capability of the reinforcement learning-driven hierarchical task planning framework when operating in uncertain perceptual environments.

5 Conclusion

This study addresses the problem of task planning in complex and dynamic environments by proposing a reinforcement learning-driven hierarchical task planning algorithm for agents. The algorithm centers on a hierarchical policy structure and introduces mechanisms for perception encoding, task decomposition, and scheduling coordination, enabling end-to-end optimization from environmental perception to policy decision-making. Experimental results show that the proposed method outperforms traditional hierarchical reinforcement learning models in task decomposition quality, global goal consistency, and path execution efficiency. By integrating state representation with temporal dependency modeling, the agent demonstrates stronger planning coherence and policy robustness in multi-objective, long-horizon tasks, providing a new perspective for solving planning problems in high-dimensional, non-stationary environments.

At the algorithmic level, this study establishes a unified optimization framework that integrates task hierarchy and policy learning, achieving cooperative coupling between high-level goal planning and low-level action execution. The high-level scheduling module is responsible for dynamic task decomposition and phase-level decision-making, while the low-level policy module performs local optimal control through reinforcement learning. This design forms an effective information loop between global and local levels, enhancing the agent's task transferability in complex environments and improving its adaptability to dynamic disturbances and non-stationary feedback. With the introduction of entropy regularization and advantage estimation mechanisms, the model achieves a balance between long-term planning and immediate responsiveness, ensuring stable convergence and generalization during training.

From a theoretical perspective, this research reveals the potential advantages of hierarchical reinforcement learning in structured task planning. The proposed hierarchical scheduling mechanism and temporal-aware strategy provide a new modeling paradigm for task decomposition and optimization in complex environments. The framework emphasizes policy interpretability and multi-timescale behavioral coordination, enabling agents to maintain efficient task execution and path continuity even under noisy inputs, dynamic obstacles, or partially observable conditions. The study further shows that when combined with adaptive structural constraints, the hierarchical reinforcement learning mechanism can effectively overcome planning bottlenecks in high-dimensional task spaces, providing theoretical support for autonomous learning and evolution in intelligent agents.

Future research can be extended in three directions. First, introducing multi-agent collaboration mechanisms to explore the advantages of collective intelligence in task decomposition and policy sharing. Second, integrating uncertainty modeling and representation learning to enhance robust decision-making in complex environments. Third, applying the proposed framework to real-world scenarios such as autonomous driving, industrial robot scheduling, and human-robot collaborative planning to verify its practicality in multimodal perception, dynamic interaction, and long-term goal control. Overall, this study advances the theoretical integration of reinforcement learning and hierarchical planning and provides a scalable foundation for autonomous decision-making and efficient execution in complex intelligent systems.

References

[1] Li J, Tang C, Tomizuka M, et al. Hierarchical planning through goal-conditioned offline reinforcement learning[J]. IEEE Robotics and Automation Letters, 2022, 7(4): 10216-10223 .

[2] Prakash B, Waytowich N, Oates T, et al. Towards an interpretable hierarchical agent framework using semantic goals[J]. arXiv preprint arXiv:2210.08412, 2022.

[3] Yang Y, Bevan M A, Li B. Hierarchical planning with deep reinforcement learning for 3D navigation of microrobots in blood vessels[J]. Advanced Intelligent Systems, 2022, 4(11): 2200168.

[4] Panaganti K, Xu Z, Kalathil D, et al. Robust reinforcement learning using offline data[J]. Advances in neural information processing systems, 2022, 35: 32211-32224.

[5] Liu Q, Lee X Y, Sarkar S. A study of natural robustness of deep reinforcement learning algorithms towards adversarial perturbations[J]. AI Open, 2024, 5: 126-141.

[6] Wu H, Li H, Yu J, et al. Hierarchical Reinforcement Learning for Viewpoint Planning with Scalable Precision in UAV Inspection[J]. Drones, 2025, 9(5): 352.

[7] Song W, Chen X, Li Q, et al. Flexible job-shop scheduling via graph neural network and deep reinforcement learning[J]. IEEE Transactions on Industrial Informatics, 2022, 19(2): 1600-1610.

[8] Mampage A, Karunasekera S, Buyya R. Deep Reinforcement Learning for Scheduling Applications in Serverless and Serverful Hybrid Computing Environments[J]. IEEE Transactions on Services Computing, 2025.

[9] Ngwu C, Liu Y, Wu R. Reinforcement learning in dynamic job shop scheduling: a comprehensive review of AI-driven approaches in modern manufacturing[J]. Journal of Intelligent Manufacturing, 2025: 1-16.

[10] Mohan A, Rößle D, Cremers D, et al. Advancing Robustness in Deep Reinforcement Learning with an Ensemble Defense Approach[J]. arXiv preprint arXiv:2507.17070, 2025.

[11] Zhi D, Wang P, Chen C, et al. Robustness verification of deep reinforcement learning based control systems using reward martingales[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(18): 19992-20000.

[12] Gürtler N, Büchler D, Martius G. Hierarchical reinforcement learning with timed subgoals[J]. Advances in Neural Information Processing Systems, 2021, 34: 21732-21743.

[13] Park S, Ghosh D, Eysenbach B, et al. Hiql: Offline goal-conditioned rl with latent states as actions[J]. Advances in Neural Information Processing Systems, 2023, 36: 34866-34891.

[14] Gehring J, Synnaeve G, Krause A, et al. Hierarchical skills for efficient exploration[J]. Advances in Neural Information Processing Systems, 2021, 34: 11553-11564.

[15] Yang M, Yang Y, Lu Z, et al. Hierarchical multi-agent skill discovery[J]. Advances in Neural Information Processing Systems, 2023, 36: 61759-61776.