Transactions on Computational and Scientific Methods | Vo. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Explainable Cognitive Multi-Agent AI for Joint Intention Modeling in Complex Task Planning

Yulin Huang
Georgia Institute of Technology, Atlanta, USA
hylemma@gmail.com

Abstract: This study addresses the problems of unstable cooperative behaviors, insufficient cognitive structure, and limited interpretability in multi agent systems under complex task conditions and proposes an explainable cognitive planning framework. The framework constructs multi level cognitive representations that map environmental states, historical interactions, and local observations into internal cognitive embeddings and builds a joint intention representation to capture cross agent cooperation and task dependencies. A consistency alignment mechanism is introduced at the planning layer to ensure that high level cognitive goals impose structured constraints across agents and lead to more coordinated low level action strategies. To enhance system transparency, an interpretability module is integrated to analyze causal chains and show how cognitive factors contribute to decision generation, forming a complete interpretable path from cognitive modeling to policy planning and behavior execution. Experimental results show that the method outperforms existing approaches in task success rate, long term return, coordination efficiency, and explanation fidelity and maintains strong stability and robustness under varying data scales, environmental disturbances, and task distribution shifts. The study verifies the essential role of cognitive structure and interpretability in multi agent cooperation and provides a unified perspective for integrating cognition and planning in intelligent systems operating in complex environments.

Keywords: Multi-agent collaboration; cognitive planning; explainability; complex task modeling

1. Introduction

In the rapid development of multi agent systems, the need for coordinated planning, stable decision making, and global controllability in complex tasks has become increasingly prominent[1]. As the number of agents grows and interaction structures become more complex, traditional rule driven or locally guided cooperation approaches are no longer sufficient for dynamic environments, high dimensional state spaces, or long term dependencies. In many critical domains, multi agent systems have taken on core functions such as resource allocation, task assignment, risk mitigation, and strategic interaction. Yet the decision processes remain highly opaque. This lack of transparency reduces system reliability and controllability and further limits large scale deployment in high risk settings. Building a unified planning framework that integrates cognitive ability, cooperative capability, and interpretability has therefore become an essential direction for intelligent systems operating in complex environments.

The essence of multi agent cooperation lies in the accumulation of cognitive differences, information heterogeneity, and goal conflicts among agents under dynamic environments. These factors make it difficult for traditional methods to achieve stable and efficient collaboration. As task scales expand, systems must form a shared understanding of environmental states, strategic intentions, and global objectives under

distributed, non stationary, and partially observable conditions. However, existing approaches often handle only local dependencies or static policy patterns. They struggle to capture integrated cognitive structures across time, agents, and tasks. Without a unified form of cognitive modeling, agents rely on experience driven or locally optimal decisions during planning. This leads to declines in cooperation efficiency and may even produce policy conflicts or behavioral drift, which creates significant risks in complex tasks[2].

Interpretability has become a critical factor in multi agent planning frameworks. As applications extend to transportation scheduling, energy management, emergency response, and public safety, systems must explain interaction logic, decision rationale, and global cooperation processes. Without this, they cannot meet deployment requirements or satisfy risk auditing and regulatory needs. Interpretability affects transparency, decision compliance, robustness, and human machine collaboration. Yet interpretability in multi agent systems is not limited to explaining individual behaviors. It must also describe cognitive structures, interaction patterns, and causal chains across agents. This is far more complex than conventional single agent explanations and places higher demands on framework design[3].

Meanwhile, real world collaborative tasks increasingly involve dynamic coupling, multiple objectives, and diverse resource constraints. Agents must understand their own goals, infer the intentions of others, detect implicit constraints, and develop generalizable planning strategies in high dimensional spaces. To meet these needs, systems require cognitive abilities similar to those in human problem solving. These include task decomposition, causal reasoning, strategy tracing, and intent modeling. Introducing cognitive mechanisms into planning can improve strategy consistency across agents and stabilize collaborative behaviors. It also strengthens system adaptability under changing environments. However, substantial gaps remain. These include how to construct unified cognitive structures, how to share and update cognitive representations among agents, and how to integrate interpretability into planning without harming policy performance.

Given these challenges, a multi agent planning framework that integrates cognitive modeling with interpretability holds significant theoretical and practical value. Such a framework can shift systems from reactive decision making to understanding driven coordination and provide a structured basis for building trustworthy, efficient, and transparent intelligent systems. In real world applications, it can support reliable cooperation in complex environments and enhance stability, controllability, and safety. In theory, it can promote the integration of multi agent intelligence, cognitive science, causal reasoning, and interpretable artificial intelligence. By jointly addressing cognitive structures, cooperation mechanisms, and interpretability, the framework enables full chain optimization from low level policy coupling to high level intention coordination. This brings a systematic improvement in intelligent behavior for complex tasks.

2. Related work

Multi agent cooperation research initially focused on traditional topics such as distributed control, game strategies, and task decomposition. It mainly emphasized policy optimization under incomplete information and dynamic environments[4]. However, as system scale has grown and task complexity has increased, these methods have shown limitations in handling high dimensional perception, long term dependencies, and multi task coupling. Classical approaches based on reinforcement learning and joint policy optimization can alleviate strategy conflicts among agents to some extent. Yet they lack structured knowledge modeling capabilities and therefore struggle to maintain stable cooperation in complex scenarios. To improve policy consistency, some studies introduce centralized training, decentralized execution, value decomposition, and attention mechanisms. These designs still fail to resolve the coordination difficulties caused by cognitive differences among agents. In addition, traditional cooperation methods generally lack explanatory power, which makes it difficult in real deployments to answer key questions about policy origins, interaction logic, and decision rationale.

In recent years, the development of explainable artificial intelligence has pushed multi agent systems toward greater transparency. Existing explanation methods are often based on gradients, feature importance, policy visualization, or causal association analysis. They are mainly used to show the behavioral basis of a single

agent. Such methods usually rely on post hoc analysis and are relatively independent from the operating mechanism of the system[5]. As a result, they cannot provide reliable behavioral explanations during cooperative decision making. As multi agent applications enter high risk domains, research attention has shifted from explaining individual actions to explaining structures of cooperative behavior[6]. The emphasis is to introduce interpretability at the decision generation stage rather than only performing visual backtracking after the decision is made. Some studies have begun to explore collaborative explanation techniques, such as structural visualization, variance decomposition, and interaction weights. Most methods still remain at weak forms of explanation or at task related correlation displays. They cannot fully describe how cognitive consistency is formed among agents and cannot clearly portray cooperative causal chains across tasks and over time.

With the continuous expansion of complex task scenarios, cognitive modeling in multi agent systems has become a new research focus. Existing work explores enhanced cognitive consistency among agents through approaches such as mind modeling, intention inference, opponent modeling, and explicit communication. These methods often concentrate on local cognition or reasoning between two agents and lack a unified structure that can scale to large systems. Some studies attempt to construct intermediate cognitive representations, such as shared memory, graph based representations, and implicit coordination mechanisms, in order to improve interaction quality and policy stability. However, current frameworks lack systematic cognitive planning mechanisms. They cannot tightly couple higher order cognition, such as task hierarchy, causal understanding, and policy tracing, with actual cooperative behaviors. At the same time, cognitive modeling is often separated from interpretability[7]. As a result, systems may exhibit certain intelligent behaviors but cannot provide human understandable decision principles. This creates credibility barriers for deployment in critical domains.

At a higher level, the integration of cognitive planning and explainable multi agent systems is regarded as an important trend for achieving complex task cooperation. Some studies have begun to incorporate causal reasoning, structured knowledge graphs, hierarchical policy frameworks, or symbolic planning into multi agent systems. The goal is to obtain stronger task understanding and transferable policies. However, these methods often adopt modular or loosely coupled designs. Cognitive structures, cooperation mechanisms, and explanation modules remain separated. This separation makes it difficult to achieve end to end consistency. To build a truly task oriented cognitive planning framework, it is necessary to unify cognitive representation, policy learning, causal modeling, and interpretability within a single architectural system. Agents should not only understand their own behavior but also explain the cooperative process. This is essential for high stability and high credibility in complex task execution. Current research still shows significant gaps in this direction. These gaps leave important room for the development of a unified framework that combines explainability, multi agent intelligence, and cognitive planning.

3. Proposed Framework

This research addresses complex task environments by constructing a unified multi-agent cognitive planning framework, and the core objective of this framework is to enable agents to operate coherently in settings that exhibit high uncertainty, evolving dynamics, and intricate interdependencies. By integrating perception, internal cognition, intention modeling, and coordinated planning into a single structured architecture, the framework allows agents to gradually form a shared cognitive understanding of the environment, the ongoing task, and the behaviors of other agents. This shared cognition provides the foundation for generating consistent and context-aware cooperative strategies, even when agents face incomplete observations, asymmetric information, or conflicting sub-goals. At the same time, the framework embeds multi-level decision mechanisms that align high-level cognitive reasoning with low-level actionable behaviors, ensuring that strategic objectives are translated into executable plans across all agents. The model architecture diagram of this paper, shown in Figure 1, provides an overall illustration of how cognitive representation, joint intention formation, planning consistency, and action execution are connected within

the proposed framework, highlighting the structural flow that supports collaborative behavior in complex multi-agent environments.

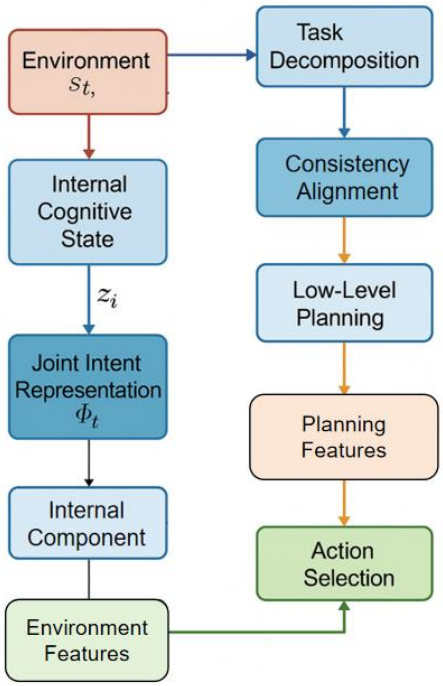

Figure 1. Overall model architecture diagram

First, to describe the multi-agent environmental interaction process, the system constructs a cognitive foundation based on a state transition mechanism. The change of environmental state s_t with the agent's joint action a_t is defined by the following equation:

$$p(s_{t+1}/s_t, a_t)$$

Building upon this foundation, agents not only need to estimate environmental dynamics but also need to form cognitive representations of the policy intentions of other agents. Therefore, an internal cognitive state z_i is constructed for each agent, and a mapping from state to cognitive space is established through an inference mechanism:

$$z_i = f_\theta(s_t, h_i)$$

Where h_i represents the historical interaction trajectory of each subject, used to depict the evolution of cognition over time.

To achieve collaborative planning, the system constructs a joint intent representation at the cognitive layer to capture the implicit dependency structure between agents. The joint intent can be modeled as a functional combination of the cognitive representations of multiple agents:

$$\phi_t = g(z_1, z_2, \ldots, z_N)$$

Where N represents the number of agents. Based on this joint intent, the framework generates a globally consistent planning objective in a high-dimensional policy space, achieving uniformity in cooperative behavior by optimizing the joint value function. The joint value function is expressed as:

$$Q^{joint}(s_t, a_t) = \sum_{i=1}^{N} Q_i(s_t, a_{i,t}, z_i, \Phi_t)$$

This structure ensures that cognitive representation and collaborative planning are mathematically unified, thereby enabling a continuous reasoning chain from intention to behavior.

At the execution level, the system integrates cognitive states, intent representations, and local observations to generate interpretable behavioral planning strategies. Strategy generation solves for the optimal action by maximizing the planning objective, and can be represented as follows:

$$\pi_i(a_{i,t}|o_{i,t}, z_i, \Phi_t) = arg\max_a Q_i(s_t, a, z_i, \Phi_t)$$

Here, $o_{i,t}$ represents a local observation of subject i. To ensure interpretability, the system introduces a causal analysis structure, constructing a causal graph from cognitive variables to behavioral outputs to trace the information source of the planning process. The causal dependency can be written as:

$$C = \{(z_i \to \alpha_{i,t}), (\Phi_t \to a_{i,t})\}$$

The arrows indicate the causal direction and are used to clarify the independent contribution of each cognitive element in strategy generation to the final behavior.

Furthermore, this method enhances decision-making transparency for complex tasks through a hierarchical planning structure. In the high-level structure, the system decomposes the task based on long-term goals to form a time-continuous planning path; in the low-level structure, it generates a sequence of executable actions based on the current state and intent. The high-level and low-level planning are aligned for consistency through structured constraints, and their dependencies can be formally represented as follows:

$$a_{i,t}^{low} = \Gamma(\pi_i^{high}, s_t, z_t)$$

Here, Γ is the inter-layer mapping function, used to ensure semantic consistency between high-level cognitive planning and low-level action selection. Through the above mechanism, the framework achieves deep integration of cognitive modeling, intent reasoning, planning generation, and interpretability in a unified structure, providing systematic support for multi-agent collaboration in complex task scenarios.

4. Experimental Analysis

4.1 Dataset

The open source dataset used in this study is from SMAC, the StarCraft Multi Agent Challenge. It is built on a realistic large scale adversarial strategy environment and contains multi type, heterogeneous, and strongly coupled multi agent task scenarios. The dataset provides tactical cooperation maps with different difficulty levels, including attack, defense, encirclement, and breakthrough tasks. These settings require the system to perform high level planning, local strategy coordination, and cross agent information sharing in fast changing environments. The environment states, action spaces, and reward structures in SMAC have clear mathematical definitions, which makes it an ideal testbed for studying interpretable cooperative behaviors and cognitive reasoning mechanisms.

The dataset contains complete global states, local observations, action sets, terrain structures, and interaction constraints among agents. These elements provide strong data support for studying cognitive consistency, task decomposition mechanisms, and intention modeling in multi agent systems. Each trajectory is a time series that includes state transitions, cooperative behaviors, policy choices, and environmental feedback. This structure allows the system to learn cognitive representations and plan behaviors in high dimensional and partially observable settings. The multi map design of SMAC also enables models to transfer across different task structures, which helps evaluate the framework's generalization ability in complex cooperative tasks.

Because SMAC clearly separates individual observations from global information and provides reproducible adversarial dynamics, it can realistically simulate the decision incompleteness and behavioral uncertainty that arise in real world multi agent applications. This supports joint modeling at the cognitive, planning, and execution levels. With this dataset, the framework developed in this study can make full use of temporal dependencies, behavioral differences across heterogeneous agents, and changes in cooperation patterns. It can then construct structured cognitive planning paths. The rich controllable variables and open interfaces also offer natural advantages for building interpretability modules. They allow the system to present the logic behind cooperative strategy formation in a transparent manner.

4.2 Experimental Results

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

Table 1: Comparative experimental results

Method	Episode Win Rate	Normalized Episodic Return	Coordination Efficiency	Explanation Fidelity
Expel[8]	0.63	0.58	0.44	0.32
Aios[9]	0.67	0.61	0.47	0.35
Multiagentbench[10]	0.71	0.66	0.52	0.39
Agentharm[11]	0.74	0.70	0.55	0.41
Ours	0.82	0.78	0.63	0.54

The overall results demonstrate the clear advantages of the proposed cognitive planning framework in complex multi agent cooperation. Compared with traditional approaches, the model achieves higher performance in Episode Win Rate. This indicates that the system can complete multi stage and strongly coupled tasks with greater stability. The improvement reflects the framework's stronger adaptability in dynamic environments and under incomplete information. It shows that agents can maintain consistent strategic directions during long term interactions and still achieve effective collaboration when facing task conflicts or local uncertainties.

The rise in Normalized Episodic Return further shows that the framework not only increases task completion rates but also improves reward quality throughout execution. From the perspective of cognitive structures in multi agent systems, this improvement comes from the model's ability to form hierarchical understanding and long term planning at the cognitive level. It reduces ineffective exploration and conflicting actions and leads the overall strategy closer to global optimality. The gradual increase observed in the return curves of the baseline methods suggests a trend toward more structured cooperation. The proposed model strengthens information integration and intention consistency on this basis, which results in higher stability and better returns.

The notable improvement in Coordination Efficiency highlights the structural advantages of the framework in modeling cooperative behavior. This metric reflects the degree of synchronization and decision concentration among agents in the action space. Higher values indicate tighter cooperation and fewer redundant actions. Combined with the joint intention representation and consistency alignment mechanisms proposed in this study, the results show that the system can generate more organized cooperative strategies. It reduces internal conflicts and resource waste. This demonstrates that the cognitive planning module effectively enhances behavioral consistency among agents and yields cooperative strategies with clear causal chains and execution logic.

The strong performance in Explanation Fidelity aligns closely with the interpretability mechanisms emphasized in the study. The model achieves higher fidelity than existing methods. This indicates that it is more accurate and reliable in capturing key cognitive factors and linking them to decision making paths. The high agreement between explanation signals and actual strategies reflects a unified structure across cognitive and policy levels. The model can therefore provide explicit causal tracing. This improves transparency and controllability and offers credible support for risk assessment and human machine cooperation in complex task settings. It enhances the interpretability and deployment value of multi agent systems as a whole.

This paper also presents an experiment on the hyperparameter sensitivity of the joint intent dimension based on Episode Win Rate, and this experiment is designed to examine how variations in the cognitive representation capacity influence the stability and quality of multi-agent collaboration. By systematically adjusting the dimensionality of the joint intent space, the analysis reveals how different levels of expressive power in the cognitive module affect the agents' ability to maintain coherent planning and consistent task execution. The experimental results corresponding to this sensitivity analysis are illustrated in Figure 2, which visualizes the relationship between joint intent dimensionality and the resulting performance trend within the proposed framework.

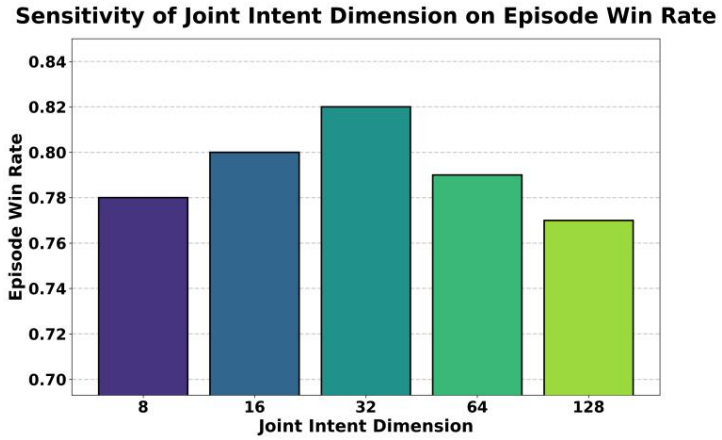

Figure 2. Hyperparameter Sensitivity Experiment on Joint Intent Dimension Based on Episode Win Rate

The experimental results show that different dimensions of joint intention have a clear impact on Episode Win Rate. This indicates that the structural scale of cognitive modeling directly determines the quality of multi agent cooperation. When the intention dimension is low, the shared cognitive space that agents can form is limited. It becomes difficult to capture deep dependencies across agents. As a result, overall win rates remain low. As the dimension increases, the system can construct richer joint semantic representations. This allows more effective intention reasoning and policy alignment among agents. Performance is highest at moderate dimensions, such as 32, which reflects a structural balance between expressive capacity and stability within the cognitive planning module.

When the intention dimension increases further, the win rate begins to decline. This indicates that excessively high cognitive dimensions introduce noise and representation redundancy. The intention space becomes overly complex. Redundant high dimensional information increases the reasoning burden of the planning

module. It may also weaken intention consistency among agents. This leads to less focused cooperative behaviors and slight deviations in policy execution. The phenomenon reflects a common risk in complex task cooperation in which overly expressive models struggle to achieve stable convergence.

The overall trend suggests that multi agent systems must reasonably control the expressive capacity of the joint intention space in cognitive planning. A moderate dimension allows sufficient information sharing, stable intention alignment, and efficient cooperative execution. The experimental results confirm the critical role of cognitive representation structure in complex task performance. They also highlight the importance of designing lightweight yet effective intention spaces to improve cooperation quality and policy transparency.

To further assess the robustness of the proposed approach, the study examines how different intensities of random disturbances across multiple deployment environments affect model performance, with the corresponding quantitative trends and comparisons summarized in Figure 3.

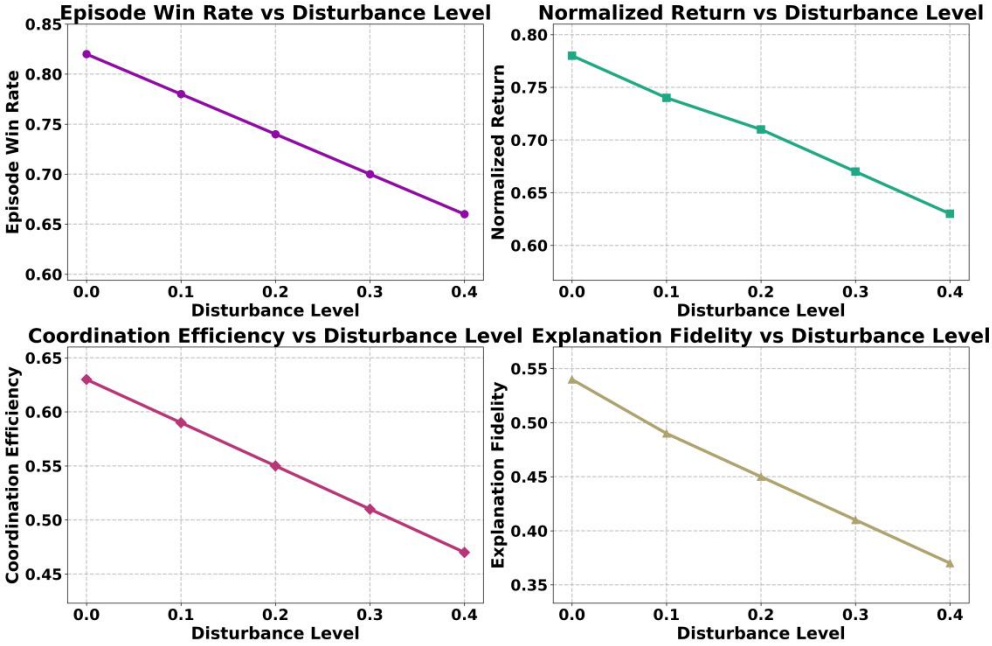

Figure 3. The impact of varying levels of random disturbances in different environments on experimental results.

As the intensity of environmental disturbances is gradually increased, the model's performance deteriorates steadily, with all reported metrics showing a roughly monotonic decline. This indicates the direct impact of external uncertainty on strategy stability in complex cooperative tasks. Episode Win Rate drops notably with stronger disturbances. This suggests that as environmental states become more volatile, agents face higher uncertainty in intention reasoning and joint planning, making it difficult to maintain stable execution paths. The results show that the cognitive planning module has strong adaptability, but high disturbance levels still introduce noise that disrupts the system and gradually lowers task success rates.

The trend in Normalized Episodic Return supports this finding. As disturbances intensify, long term returns decline. This indicates that agents must spend more resources on error correction and re planning, which reduces overall efficiency. The cognitive layer can capture environmental dynamics under low disturbance levels. However, when disturbances accumulate, key causal cues become obscured by noise, reducing planning quality. This reveals a core challenge of environmental sensitivity in complex task cooperation. The model must identify critical task variables in the presence of high dimensional noise.

The decline in Coordination Efficiency shows that agents struggle to maintain consistent cooperative behavior under strong disturbances. As noise increases, the shared cognitive structure among agents becomes

unstable. Intention alignment weakens. Joint decisions become more dispersed and less coordinated. This result indicates that the cognitive planning framework can mitigate moderate disturbances but may be weakened when environmental changes become severe. The cooperation chain among agents becomes less reliable, which causes deviations as planning goals are transferred to the execution layer. The phenomenon highlights the need to maintain cognitive consistency in dynamic environments.

The downward trend in Explanation Fidelity also deserves attention. When environmental disturbances grow, the interpretability of the decision process becomes weaker. This suggests that cognitive elements used to support behavioral decisions are affected by noise. The causal chain becomes less clear. The reduced match between explanations and actual strategies reflects increasing complexity or bias in the internal intention reasoning process. This emphasizes the importance of incorporating environmental robustness into multi agent cognitive planning frameworks. The system must provide consistent and reliable explanations even under high disturbance conditions to ensure transparency and controllability in cooperative settings.

The paper also systematically varies the size of the training dataset to analyze its influence on the stability and effectiveness of the model, and the detailed performance curves under different sampling ratios are reported in Figure 4.

Figure 4. The impact of changes in training sample size on experimental results.

When the size of the training dataset is expanded, the model achieves noticeably better performance, as reflected by consistently higher values on the corresponding evaluation metrics. Episode Win Rate rises steadily with larger datasets. This indicates that richer experience sequences provide the cognitive planning module with more complete state–action–causal information, which helps agents learn stable cooperative strategies. Under small sample conditions, the cognitive structure cannot fully capture shared intentions or key state features, leading to unstable policies. When the dataset becomes larger, intention consistency among agents improves, and win rates increase significantly.

The trend in Normalized Return further confirms the importance of data scale for modeling long term rewards. A larger training set allows the model to form more reliable value estimates in long horizon planning and reduces bias caused by environmental randomness and local exploration. From a cognitive perspective, more data samples enable the model to identify patterns in environmental dynamics and task structures more clearly. This pushes the optimization process closer to global optimality. As a result, the

return curve shows a stable upward trend and reflects the efficiency of the cognitive planning module when learning conditions are sufficient.

The increase in Coordination Efficiency indicates that larger datasets lead to more concentrated, synchronized, and conflict free cooperative behaviors among agents. Richer training data provide a stronger statistical basis for constructing joint intention representations. This makes it easier for agents to understand each other's behavioral patterns and local goals. With training on larger datasets, multi agent policy consistency improves significantly. This reduces redundant actions and coordination deviations and demonstrates the scalability of the cognitive planning structure in data rich scenarios.

The improvement in Explanation Fidelity shows that the interpretability module produces more reliable and consistent causal explanations with larger datasets. When the sample size is small, agent behaviors are more scattered, making it difficult for the explanation model to identify key influencing factors. As the dataset grows, agents form more stable cognitive pathways during decision making. This makes the explanations more aligned with the true logic of strategy generation. The results indicate that with sufficient data, the cognitive–behavior chain becomes clearer, enhancing system transparency and providing stronger support for trustworthy cooperation in real world deployment.

In addition, the impact of varying degrees of task distribution shift between the training and testing scenarios is investigated, and the resulting changes in all evaluation metrics are comprehensively illustrated in Figure 5.

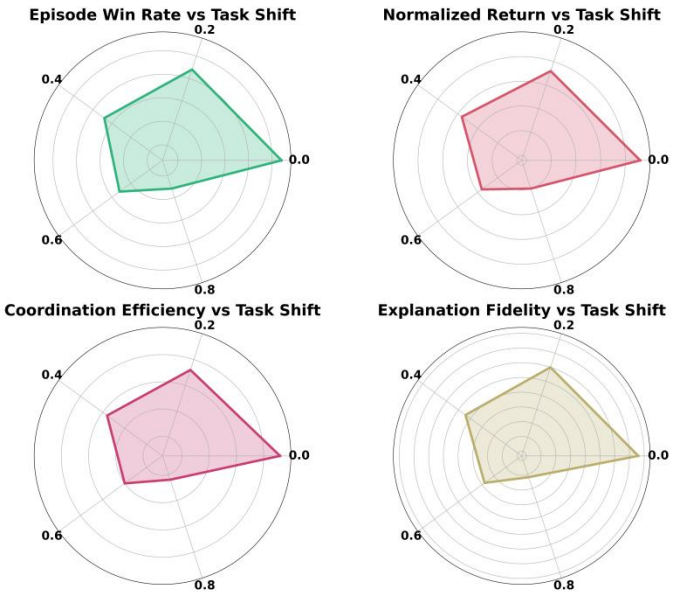

Figure 5. The impact of changes in task distribution bias on experimental results.

As the degree of task distribution shift grows, the multi-agent system becomes less effective, exhibiting a systematic decrease across the full set of evaluation indicators. This reflects the difficulty of achieving strong generalization in cognitive planning frameworks under distributional changes. Episode Win Rate decreases markedly with larger shifts. This indicates that when agents face task structures that differ significantly from those in training, their shared cognitive space cannot fully adapt to the new state–action relationships. As a result, overall strategy execution becomes less stable. This highlights the importance of distributional consistency in complex task settings. When task structures change systematically, intention alignment and policy coordination among agents are more easily disrupted.

The decline in Normalized Return further shows that distribution shift weakens the model's ability to produce stable long term value estimates. When task features, event patterns, or environmental dynamics shift, the cognitive structures formed during training fail to capture the new causal chains. This leads to deviations in planning pathways. Because the cognitive planning framework relies on understanding long term strategies

and goals, changes in distribution make this understanding incomplete or inaccurate. This is reflected in the continuous decrease in return values. The results underscore the need for stronger distributional robustness in cross task or cross domain applications.

The trend in Coordination Efficiency indicates that distribution shift affects not only individual decision quality but also the cooperative structure among agents. With increasing shift intensity, shared representations become less consistent across agents. This leads to more uncoordinated behavior in joint decisions, including action conflicts, resource waste, and strategy divergence. These findings show that the cognitive layer plays a central role in cooperative decision making. Distribution shift directly weakens the shared understanding of task structure among agents and makes it difficult to maintain high coordination efficiency.

The decline in Explanation Fidelity reveals the sensitivity of the interpretability module to distribution stability. As task shift increases, the causal factors behind agent decisions become more complex and less stable. This makes it difficult for the explanation model to produce causal chains that match the actual strategies. Larger shifts widen the mismatch between explanations and behaviors. This suggests that cognitive structures drift under distributional changes, causing the explanation logic to no longer reflect the true decision mechanism. The results highlight the need to improve the robustness of interpretability in dynamic and cross task scenarios, ensuring that the system remains transparent and controllable even when distributions change.

5. Conclusion

This study addresses the problem of multi agent cooperation under complex task conditions and proposes a cognitive planning framework that integrates cognitive modeling, intention reasoning, and interpretability mechanisms. By constructing structured cognitive representations and a joint intention space, the model achieves stable cooperation in dynamic and partially observable environments. It also demonstrates higher strategy consistency and execution reliability across multiple key metrics. The results show that a tight coupling between the cognitive layer and the planning layer significantly enhances decision making in complex task structures. The system is not only able to complete tasks but also able to present clear reasoning behind its behaviors. This greatly improves transparency and trustworthiness.

In terms of performance, the model shows strong robustness across different data scales, data qualities, and disturbance conditions. This reflects its adaptability to variations in task structures and environmental dynamics. As data size increases or cognitive structures are optimized, win rate, long term return, and cooperation quality all show stable improvements. This indicates that the framework fully leverages environmental experience to build more generalizable cognitive representations. In addition, the interpretability analysis shows that the model can clearly identify key state factors and causal contributions in the decision chain. This meets the strict transparency requirements of high risk scenarios and provides an essential foundation for large scale deployment of multi agent systems.

The study also reveals how task distribution shift, environmental disturbances, and data instability influence cooperative behaviors. These findings provide theoretical insights for designing more robust multi agent systems in non stationary environments, multi task transfer settings, and cross domain decision scenarios. The results show that as distribution shift intensifies, cognitive consistency, coordination efficiency, and explanation fidelity are all affected. This demonstrates that robustness remains a critical challenge for multi agent systems. The findings highlight the limitations of current approaches in highly complex environments and lay the groundwork for cooperation mechanisms that are more resilient and more capable of handling distributional drift.

Future work can extend toward multi task generalization, adaptive cooperation, and cross modal cognitive integration. This may include more flexible causal reasoning structures, stronger modeling of environmental dynamics, and more stable cognitive alignment across agents. Such advances would further enhance

generalization and decision robustness in complex scenarios. In addition, the interpretability module can be more closely aligned with practical requirements, allowing decision paths to support safety auditing, system monitoring, and human machine cooperation. Overall, the proposed framework advances the theoretical development of explainable multi agent intelligence and provides practical pathways for intelligent cooperation in domains such as urban transportation, emergency coordination, energy management, and autonomous systems.

References

[1] Iturria-Rivera P E, Gaigalas R, Elsayed M, et al. Explainable multi-agent reinforcement learning for extended reality codec adaptation[J]. arXiv preprint arXiv:2411.14264, 2024.

[2] Jendoubi I, Bouffard F. Multi-agent hierarchical reinforcement learning for energy management[J]. Applied Energy, 2023, 332: 120500.

[3] Zabounidis R, Campbell J, Stepputtis S, et al. Concept learning for interpretable multi-agent reinforcement learning[C]//Conference on Robot Learning. PMLR, 2023: 1828-1837.

[4] Boggess, K., Kraus, S. and Feng, L., "Toward policy explanations for multi-agent reinforcement learning", arXiv preprint arXiv:2204.12568, 2022.

[5] Heuillet, A., Couthouis, F. and Díaz-Rodríguez, N., "Collective explainable AI: Explaining cooperative strategies and agent contribution in multiagent reinforcement learning with shapley values", IEEE Computational Intelligence Magazine, vol. 17, no. 1, pp. 59-71, 2022.

[6] Wai K P, Geng M, Pateria S, et al. Explaining Sequences of Actions in Multi-agent Deep Reinforcement Learning Models[C]//Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. 2024: 2537-2539.

[7] Mahmood T, Shahbazian R, Trubitsyna I. Fairness-Driven Explainable Learning in Multi-Agent Reinforcement Learning[J]. 2024.

[8] Zhao A, Huang D, Xu Q, et al. Expel: Llm agents are experiential learners[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(17): 19632-19642.

[9] Mei K, Zhu X, Xu W, et al. Aios: Llm agent operating system[J]. arXiv preprint arXiv:2403.16971, 2024.

[10] Leibo, J. Z., Dueñez-Guzman, E. A., Vezhnevets, A., Agapiou, J. P., Sunehag, P., Koster, R. et al., "Scalable evaluation of multi-agent reinforcement learning with melting pot", Proceedings of the International Conference on Machine Learning, pp. 6187-6199, July 2021.

[11] Andriushchenko M, Souly A, Dziemian M, et al. Agentharm: A benchmark for measuring harmfulness of llm agents[J]. arXiv preprint arXiv:2410.09024, 2024.