Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.34031

CoT-MAORL: Enhancing Offline Multi-Agent Coordination via Chain-of-Thought Alignment

Litian Zheng

School of Artificial Intelligence and Robotics, Xiamen University Malaysia, Selangor, Malaysia
AIT2309093@xmu.edu.my

Abstract. In view of problems such as distributional shifts, non-stationarity and bad credit assignments in offline multi-agent reinforcement learning, this paper proposes CoT-MAORL, an online distributed algorithm for policy-learning based on introducing chain-of-thought reasoning into the space of latent variables. Therefore, the following type of reasoning appears to include both intention and coordination indicators; It has a policy-valued function, while aiming at achieving an overall coherence standard through CoT-based verification mechanisms. Adaptive time steps can be used for adjustment in the following iteration: Dynamics Reasoning. Experiments on the SMAC benchmark show that CoT-MAORL is more effective than prominent offline baselines; it achieved an 86.4 win rate in 2s3z and a 60.8 win rate in terms of good/bad-medium/poor data qualifier combinations for 5m-vs-6m+(-5.2). Abolition experiments provide some support for it has gained about 8.4% due to the reason-based reasoning; The acquisition of recursive Dynamics and Contrastive learning separately contributes more than 5.7%, 4.6%.

Keywords: Offline MARL, Latent Chain-of-Thought, Structured Coordination, Generative Decision Models

1. Introduction

Multi-agent reinforcement learning is becoming more widespread in research on collaborative cooperation among agents under complex environments such as robotics, Traffic Management and Multi-Robot Coordination. Classical CTDE-based methods include QMIX [1] and MADDPG [2], which provide basic solutions for learning decoupled policies via centralised training. However, due to restrictions such as cost and practicability under an online environment for interactivity within a real-world application scenario, offline multi-agent reinforcement learning has emerged instead.

Offline reinforcement learning learns a policy from the given dataset without more interactions. CQL [3] and IQL [4] are typical approaches to address distributional shift problems; they estimate conservative values cautiously, as well as perform policy learning in-sample only. However, when directly applying these methods to the multi-agent scenario, numerous issues arise due to a large number of joint action Spaces and the resulting non-Stationary state.

Recently, some new methods have been proposed to address the aforementioned problems in offline MARL. OMIGA [5] adds implicit global-Local Value Regularization to better capture the Coordination Structure of offline data. InSPO [6] proposes sequential policy optimization within the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.34031

dataset support, reducing out-of-distribution joint actions and improving stability. More recently, diffusion-based approaches such as DoF [7] have been proposed for trajectory generation [8] that use generative diffusion models to model the probability density of joint actions and improve both coordination and data efficiency. Methods show that, in practice, displaying a certain way of distributing collective activity online is problematic. Most existing methods still focus on value regularisation, policy constraints and generative trajectory models; little has been explored about the reason for multi-agent collaboration at this stage yet.

Meanwhile, recent studies in 2025 have also made progress in promoting the development direction of diffusion-based and generative multi-agent learning under structured trajectory representation to achieve scale-out and coordination [7, 8]. Despite this progress, they still operate at the level of action or trajectory generation, without explicitly modeling intermediate decision-making processes such as intent decomposition or coordination reasoning.

Recently, large Language Models (LLMs) have shown excellent ability to perform multi-stage reasoning and planning. Chain-of-thought (CoT) prompt [9] has been proven effective in enhancing reasoning ability by providing clear intermediary stages explicitly. In addition, some recent works on integrating an LLM and several entities for task scheduling and co-operation [10] focus on making these interactions more productive. These methods are more suitable for online or interactive environments; the application to offline multi-agent reinforcement learning is relatively scarce.

To overcome the aforementioned deficiencies, CoT-MAORL is proposed to introduce chain-of-thought reasoning as a structured latent variable in an offline multi-agent reinforcement learning setting. The main idea is to explicitly capture the high-level intentions, sub-goals and coordination information in it, and integrate them into both policies and values simultaneously. Because the previous works are only associated directly with the state-space or trajectories of actions, this paper introduces an agent that can make Decisions through Interpretable Reasoning Structures, enhancing Coordination during Distribution-Shift.

Moreover, a CoT-consistency Objective is presented to make inferences consistent with observed behaviour based on the offline data; thus, the learned reasoning process will not drift from reality too much. This can help stabilise the training and improve generalisation for unknown coordination patterns. Through the integration of structured reasoning and off-line multi-agent reinforcement learning to provide a new way for coordinating failures, credit assignments ambiguity and nonstationarity in multi-agent systems.

Evaluate CoT-MAORL on the benchmark cooperative environments of SMAC, etc. The experimental results show that the proposed approach has consistently outperformed state-of-the-art powerful offline MARL baseline in terms of return, coordinated effects and generalization ability under distribu- tionshifts.

2. Method

2.1. Problem formulation

This paper consider a multi-agent Markov decision process (MMDP) defined by the tuple $M = (I, S, \{A^i\}, T, R, \gamma)$, where $I=\{1,\ldots,N\}$ is the set of agents, S is the global state space, $A^i$is the action space of agent ii, $T:S\times \prod_i A^i \to \Delta(S)$ is the transition function, $R:S\times \prod_i A^i \to R$ is the shared reward function, and $\gamma\in[0,1)$ is the discount factor. This formulation follows the centralized-

DOI: 10.54254/2753-8818/2026.34031

training decentralized-execution (CTDE) paradigm that has become standard in cooperative MARL [1, 2].

In offline MARL, this paper are given a static dataset $D=\{(s_t,a_t,r_t,s_{t+1})\}_{t=1}^{T}$ collected by unknown behavior policies, where $a_t=(a_t^1,\ldots,a_t^N)$ denotes the joint action. No environment interaction is permitted during training, making the problem challenging due to distributional shift and the lack of online exploration. Unlike online MARL where agents can actively coordinate through trial-and-error, offline settings restrict learning to pre-collected data, which exacerbates issues of coordination and credit assignment in multi-agent systems [11, 12].

2.2. Trajectory-level reasoning framework

Unlike prior methods that learn a direct mapping $s\rightarrow a$, this paper introduces a latent reasoning process that decomposes decision-making into two stages:

$$s_t \rightarrow z_t \rightarrow a_t \tag{1}$$

Here, z_t is a structured latent variable that captures high-level intent, subgoals, and coordination state. This paper model z_t as a continuous latent embedding in RdRd, which enables efficient optimization via gradient-based methods and supports contrastive representation learning. This design is inspired by recent advances in latent reasoning, which demonstrate that embedding reasoning within continuous spaces offers richer representational capacity compared to explicit language generation [13].

To reflect the sequence characteristics of reasoning, this paper uses a recurrent reasoning model to update the reasoning status gradually. Since diffusion-based trajectory modelling is computationally intensive and noisy across the entire sequence; however, the per-step recurrence model is simple in calculation and stable under noise. Specifically, the reasoning dynamics are as follows:

$$z_t \leftarrow p_\phi\left(z_t z_{t-1}, s_t, a_{t-1}\right), \quad with\ z_0 = f_i nit\left(s_0\right) \tag{2}$$

where p_ϕ is a probabilistic reasoning model implemented as a GRU-based recurrent network parameterized by ϕ, and f_{init} initializes the reasoning state from the initial observation. This recursive formulation enables the model to maintain and update its understanding of the task dynamically, capturing long-term dependencies that static latent variables cannot express. Unlike static latent variable approaches where z remains fixed within a trajectory, the recurrent reasoning allows the agent to adapt its coordination strategy as the task progresses—a crucial capability for long-horizon collaborative tasks where subgoals change over time [14].

2.3. Global–local reasoning decomposition

Simultaneously support team-level coordination and individual adaptation by dividing the reasoning state into a global component and local component.

DOI: 10.54254/2753-8818/2026.34031

$$z_t = \left(z_t^{global}, z_t^1, \ldots, z_t^N\right) \tag{3}$$

where z_t^{global} represents shared team intent that is common to all agents, and z_t^i represents agent-specific reasoning that captures individual adaptation. The joint distribution factorizes according to:

$$p(z_t|\ldots)=p(z_t^{global}|\ldots)\cdot\prod_i p(z_t^i|z_t^{global},\ldots) \tag{4}$$

Which has a hierarchical structure encoded as follows: Firstly infers the global intention based on the current situation and previous knowledge; Then, for each agent, it conditions this inference with both the overall goal and their own environment. This Design is in a reasonable state of forcing all agents to follow one's plan or enabling independence. Recent studies on offline MARL have demonstrated that modelling multi-modal aspects of joint cooperative policies is essential for success [12], and propose to decompose it into shared-coordination signals and local adaptation components via a global-local approach.

2.4. Diffusion-based policy with reasoning conditioning

Each agent selects its actions based on both the current state and the complete reason state. To capture the multimodal feature of action distribution in offline data sets, this paper uses a diffusion-based policy to generate actions iteratively denoised [15, 16]. The previous work has been applied in reasoning and action at the same time, leading to unstable training and error propagation; thus, this paper limits diffusion to policies while maintaining an ordinary learning module responsible for reasoning.

For every agent ii, its action selection will proceed in reverse-diffusion based on both its own State and reasoning States.

$$p_\theta(a_t^{i,(k-1)}|a_t^{i,(k)},s_t,z_t)=N(a_t^{i,(k-1)};\mu_\theta^i(a_t^{i,(k)},s_t,z_t,k),\sigma_k^2 I) \tag{5}$$

for a_t^i, $a_t^{i,(K)}\sim N(0,I)$ and after K denoising steps obtain the final action $a_t^{i,(0)}$. The joint policy factorizes over agents:

$$p_\theta(a_t|s_t,z_t)=\prod_{i=1}^N p_\theta^i(a_t^i|s_t,z_t) \tag{6}$$

Reflecting the decentralized execution model: Each agent operates according to its own state and a common-reasoning-state-shared state, respectively. Since agents have the same reasoning state, including global intention; thus, no communication channels are needed for them to coordinate their behaviour - this follows the stigmergic coordination rule of agent collaboration via shared environment markers. The Diffusion formulation can capture intricate multi-modal action distribution more effectively; it also helps to better model diverse Strategies for a behaviour Policy in an offline situation.

DOI: 10.54254/2753-8818/2026.34031

2.5. Q-guidance for diffusion policy

In offline RL, value functions serve as stabilizers that prevent the policy from selecting out-of-distribution actions [3]. Rather than treating Q as optional, incorporate it as a guidance mechanism for the diffusion policy. Specifically, this paper trains a Q-function conditioned on the reasoning state: $Q_\psi(s_t, a_t, z_t)$, where ψ are the Q-network parameters. During training, this paper uses Q to guide the diffusion policy via a guidance loss:

$$\mathscr{L}_Q = E\left[\left(Q_\psi\left(s_t, a_t^{(0)}, z_t\right) - y_t\right)^2\right] \tag{7}$$

the $a_t^{(0)}$ is the denoised action from the diffusion policy, and the Bellman target is:

$$y_t = r_t + \gamma \cdot E_{a_{t+1}, z_{t+1}}\left[Q_\psi\left(s_{t+1}, a_{t+1}, z_{t+1}\right)\right] \tag{8}$$

with $a_{t+1} \sim \pi_\theta(\cdot \mid s_{t+1}, z_{t+1})$ and $z_{t+1} \sim p_\phi(\cdot \mid z_t, s_{t+1}, a_t)$. This design enables the value function to propagate credit through both actions and reasoning, addressing a key limitation of conventional offline MARL methods that struggle with credit assignment under sparse rewards [11].

2.6. Contrastive CoT alignment

The most prominent problem with offline-data-based learning of latent reasoning is the lack of ground-truth reasoning labels. Directly predicting reasoning based on the state will result in a hallucination of reason, which deviates from real behaviour. To solve this problem, this paper adopts a contrastive goal to make the inferred reasoning match the observed state-Action pair; This is inspired by recent achievements of contrastive representation learning for reinforcement learning [17] and [18].

Let $h(s_t,a_t)$ be a state-action encoder that produces a normalized embedding. The InfoNCE loss is defined as:

$$\mathscr{L}_CoT = -E\left[log\left(\frac{exp(sim(z_t,h(s_t,a_t)))}{\Sigma_{a'\in A_neg} exp(sim(z_t,h(s_t,a')))}\right)\right] \tag{9}$$

as $sim(\cdot,\cdot)$ denotes cosine similarity, and A_{neg} consists of joint action samples $a'=(a'^1,\ldots,a'^N)$ drawn from other transitions within the mini-batch. Intuitively, this loss pulls the reasoning state z_t close to the embedding of the actually executed joint action a_t , while pushing it away from embeddings of alternative joint actions. This ensures that the learned reasoning is grounded in observable behavior and provides a rich representational signal for training. Unlike regression-based alternatives, this contrastive formulation naturally handles the multimodal nature of joint action distributions in offline datasets [12].

DOI: 10.54254/2753-8818/2026.34031

2.7. Coordination regularization

To further encourage coherent coordination across agents, regularize the alignment between the global intent and each agent's local reasoning. Specifically, this paper maximizes the cosine similarity between z_t^{global} and z_t^i for all agents:

$$\mathcal{L}_coord = -E[(1/N)\Sigma_i(z_t{}^{global} \cdot z_t{}^i)/(\|z_t{}^{global}\|\|z_t{}^i\|)] \tag{10}$$

The regularisation term penalises deviations in team-level and individual reasons to promote shared cognition among multiple agents on this task. In reality, improve the coordination stability and reduce conflictual behaviour. This Design is an improvement of recent progress in multi-agent cooperation that focuses on communication Structure Coordination [19].

2.8. Full training objective

Combine the diffusion policy loss, Q-guidance loss, contrastive-CoT alignment and coordination regularisation to form a unified objective as follows:

$$L=L_{diffusion}^{policy}+\lambda_1 L_Q+\lambda_2 L_{CoT}+\lambda_3 L_{coord} \tag{11}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. The diffusion policy loss is the standard simplified objective [7]:

$$L_{diffusion}^{policy}=E_{k,\epsilon\sim N(0,I),(s,a,z)\sim D}\left[\|\epsilon-\epsilon_\theta(a^{(k)},s,z,k)\|^2\right] \tag{12}$$

as $a^{(k)}= \sqrt{\bar{\alpha}_k} \cdot a + \sqrt{(1-\bar{\alpha}_k)} \cdot \epsilon$. The Q-guidance term stabilizes training by keeping generated actions within the data distribution, while the CoT and coordination terms provide structured inductive biases for learning interpretable and coordinated reasoning.

2.9. Training procedure

The complete training procedure is summarized in Algorithm 1 (Table 1). At each iteration, this paper samples a batch of transitions from the offline dataset and rolls forward the reasoning states using the recurrent dynamics $\bar{\psi}\leftarrow\tau\psi+(1-\tau)\bar{\psi}$. Then sample a noise level k and train the diffusion policy denoising network ϵ_θ to predict the noise added to the actions. The Q-network is updated using the Bellman target, and the contrastive and coordination losses are computed. All parameters are updated jointly via gradient descent. The target Q-network is updated via an exponential moving average. This procedure continues until convergence or after a fixed number of epochs. All networks are optimized jointly using stochastic gradient descent, with gradients backpropagated through both the diffusion denoising process and the recurrent reasoning dynamics—enabling end-to-end learning of the complete reasoning-to-action pipeline.

DOI: 10.54254/2753-8818/2026.34031

Table 1. Algorithm 1

Algorithm 1: CoT-MAORL Training
Input: Offline dataset D, hyperparameters λ_1, λ_2, λ_3
Initialize: Reasoning model p_ϕ, Diffusion policy ϵ_θ, Q-network Q_ψ, target $Q_{\bar{\psi}}$
for each training iteration do
Sample batch $\{(s_t,a_t,r_t,s_{t+1})\}$ from D
// Roll forward reasoning states
$z_0=f_{init}(s_0)$
for t = 1 to T do
$z_t\sim p_\phi(\cdot\|z_{t-1},s_t,a_{t-1})$
end for
// Diffusion policy loss
Sample noise level k, noise $\epsilon\sim N(0,I)$
$a^{(k)}=\sqrt{\bar{\alpha}_k}\cdot a+\sqrt{1-\bar{\alpha}_k}\cdot\epsilon$
$L_{diff}=\|\epsilon-\epsilon_\theta(a^{(k)},s,z,k)\|^2$
// Q-guidance loss
$a^{(0)}=denoise(a^{(K)},\epsilon_\theta)$
$y=r+\gamma\cdot Q_{\bar{\psi}}(s',a',z')$
$L_Q=\left(Q_\psi(s,a^{(0)},z)-y\right)^2$
// Contrastive CoT loss
$L_{CoT}=-\log\left[\frac{\exp(sim(z,h(s,a)))}{\sum_{a'}\exp(sim(z,h(s,a')))}\right]$
// Coordination regularization
$L_{coord}=-\frac{1}{N}\sum_{i=1}^{N}\frac{z^{global}\cdot z^i}{\|z^{global}\|\|z^i\|}$
// Total loss and update
$L=L_{diff}+\lambda_1 L_Q+\lambda_2 L_{CoT}+\lambda_3 L_{coord}$
Update ϕ,θ,ψ using ∇L
Update target: $\bar{\psi}\leftarrow\tau\psi+(1-\tau)\bar{\psi}$
end for
Output: p_ϕ, ϵ_θ

3. Experiment

3.1. Experimental setup

Evaluating CoT-MAORL on the StarCraft II Multi-Agent Challenge (SMAC) [20], a well-used test bed for multiple agents' cooperative performance in a decentralized setting. Because the Environment is complicated, Allied Units need to Plan Their Tactics Efficiently In This Battle And Form A Joint-Operation Strategy Long-Term Together. Considering both SMACv1 and its more difficult version, SMACv2, comprehensively in the evaluation.

In the case of SMACv1, canonical maps such as 2s3z, 3s5z and 5m_vs_6m have increased coordination difficulty and different agents' interaction characteristics. SMACV2 also adds randomised initial positions and different units to increase environmental stochasticity. Across the three races (Protoss, Terran, and Zerg) at different scales to test their reliability as well as generalisation ability.

DOI: 10.54254/2753-8818/2026.34031

In accordance with recent offline MARL protocols, all the experiments in this paper are performed on a set of standardised off-grid multi-agent training data generated by OG-MARL [19, 21]. The dataset spans different levels of behaviour Quality produced by Mappo [22]. After following the usual setting, this study will compare three situations: good experts' demonstration, medium quality of trajectory tracking, and exploratory behaviour. This Design can help systematically evaluate the stability of CoT-MAORL under different data distributions and degrees of suboptimality.

This paper will be compared with several other typical representatives and their theoretical foundations are diverse. CQL [3] and IQL [4] are value-based methods that tackle distributional shifts via conservative value estimation. The sequence modelled by DT [23] and the extension of this work, namely MADT [24], views decision-making as a case of conditional sequential prediction. In addition, this study also adds the strongly offline strongMARL algorithm of OMIGA [5] and recent methods using shared latent representation for coordination, LatentMAS [14]. Through a comparison of both traditional and current state-of-the-art aspects in this regard.

Implementations of the Reasoning module φp are generated using a GRU-based recurrency function with an output feature size of 128 to extract long-term correlation among reasons; The Latent dimension d equals 64. Policy and user.

Three-layer multi-Layer Perceptron (MLP) networks using ReLU activation functions were employed to implement Q-learning. This paper trains all models using the Adam optimizer with a learning rate of4×10^{-4}, batch size 512, and discount factor γ=0.99. The target network updates according to a decay factor of τ = 0.005. Overall objectives are weighted as behaviour cloning: λ_1 = 0.5; Contrastive reasoning Alignment: λ_2 =0.3; Coordination regularisation : λ3 = 0.2. Averaged across 4 random seeds; All results are given as means ± standard deviations. Build experiments based on the OG-MARL architecture to achieve fairness comparison and replication.

3.2. Results

Table 2. SMACv1

Model	IQL	CQL	OMIGA	DT	MADT	LatentMAS	CoT-MAORL (Ours)
2s3z	72.4 ± 3.2	75.6 ± 2.9	78.2 ± 2.5	74.1 ± 3.0	79.5 ± 2.3	83.1 ± 2.1	86.4 ± 1.9
3s5z	58.3 ± 3.8	61.5 ± 3.5	65.8 ± 3.1	59.6 ± 3.6	67.2 ± 2.9	71.5 ± 2.6	75.2 ± 2.4
5m_vs_6m	42.1 ± 3.5	45.3 ± 2.9	49.8 ± 2.7	43.5 ± 3.2	51.2 ± 2.5	55.6 ± 2.3	60.8 ± 2.1

Table 3. SMACv2 - protoss

Model	IQL	CQL	OMIGA	DT	MADT	LatentMAS	CoT-MAORL (Ours)
5_vs_5	58.6 ± 3.5	61.2 ± 3.2	64.5 ± 2.9	59.8 ± 3.4	65.8 ± 2.7	69.2 ± 2.5	73.5 ± 2.3
10_vs_10	53.7 ± 3.6	56.8 ± 3.3	60.2 ± 2.8	54.5 ± 3.5	61.5 ± 2.6	65.3 ± 2.4	69.8 ± 2.2
10_vs_11	49.5 ± 3.8	52.8 ± 3.5	57.2 ± 3.1	50.6 ± 3.7	58.2 ± 2.9	62.4 ± 2.6	67.5 ± 2.4
20_vs_20	48.2 ± 3.9	51.5 ± 3.6	55.8 ± 3.2	49.6 ± 3.8	56.8 ± 3.0	60.5 ± 2.7	65.2 ± 2.5
20_vs_23	41.8 ± 4.1	45.2 ± 3.8	49.6 ± 3.4	42.8 ± 4.0	50.5 ± 3.2	54.8 ± 2.9	59.6 ± 2.7

DOI: 10.54254/2753-8818/2026.34031

Table 4. SMACv3 - terran

Model	IQL	CQL	OMIGA	DT	MADT	LatentMAS	CoT-MAORL (Ours)
5_vs_5	45.2 ± 3.9	48.5 ± 3.6	52.3 ± 3.3	55.5 ± 3.7	62.6 ± 3.1	66.8 ± 2.8	71.4 ± 2.5
10_vs_10	40.5 ± 4.0	43.8 ± 3.7	47.5 ± 3.4	50.6 ± 3.9	57.8 ± 3.2	62.2 ± 2.9	66.6 ± 2.6
10_vs_11	38.2 ± 4.1	39.5 ± 3.8	44.2 ± 3.5	46.2 ± 4.0	54.2 ± 3.3	58.5 ± 3.0	63.5 ± 2.7
20_vs_20	35.8 ± 4.2	39.2 ± 3.9	43.5 ± 3.6	45.8 ± 4.1	53.5 ± 3.4	57.6 ± 3.1	61.2 ± 2.8
20_vs_23	32.5 ± 4.5	35.1 ± 4.2	36.8 ± 3.8	38.5 ± 4.4	46.5 ± 3.6	51.2 ± 3.3	56.2 ± 3.0

Table 5. SMACv2 - zerg

Model	IQL	CQL	OMIGA	DT	MADT	LatentMAS	CoT-MAORL (Ours)
5_vs_5	41.4 ± 3.8	44.1 ± 3.4	47.6 ± 3.1	42.8 ± 3.5	49.2 ± 2.9	53.5 ± 2.6	60.6 ± 2.4
10_vs_10	39.2 ± 3.9	41.8 ± 3.6	45.5 ± 3.3	40.5 ± 3.8	46.8 ± 3.1	51.5 ± 2.8	57.2 ± 2.6
10_vs_11	36.5 ± 4.0	39.2 ± 3.7	43.2 ± 3.4	37.6 ± 3.9	44.5 ± 3.2	49.2 ± 2.9	55.8 ± 2.7
20_vs_20	34.5 ± 4.1	37.2 ± 3.8	41.5 ± 3.5	34.5 ± 4.0	42.8 ± 3.3	47.5 ± 3.0	53.5 ± 2.8
20_vs_23	31.8 ± 4.3	34.5 ± 4.0	38.8 ± 3.7	31.8 ± 4.2	39.8 ± 3.5	44.2 ± 3.2	49.2 ± 3.0

Tables 2, 3, 4 and 5 show the winning-rate results of three experiments in SMAC-v1 and SMAC-v2 with the good-Q dataset. In all 12 cases, CoT-MAORL performs well in all aspects; Compared to these others' baselines, it outperforms significantly. Among them, the improvement over the best-performing baseline latentMAS is as great as 5 per cent in terms of reasoning across multiple parties.

Gains are more prominent in difficult cases, such as 5m_vs_6m and large-scale 20_vs_20 problems. These environments require accurate regulation, and they are not produced by means of focusing on value or following the sequence model. Although the trajectory-based reasoning model does not guarantee that agents' long-term coordinated Strategies will remain consistent and thus have relatively low win rates.

Furthermore, CoT-MAORL still performs well after the transition from SMACv1 to SMACv2 with more random units due to their different combinations. It indicates that the learnt reasoning representations capture more generalised coordination patterns, not only overfitting to particular situations. In terms of advantages in multiple unit environment heterogeneity situations requiring cooperation among various departments, such as the 3-5-Z system and terra-mission.

3.3. Abolition experiment

To gain a deeper understanding of the impact of individual components more comprehensively on this 5m_vs_6m sub-task in detail.

Table 6. Ablation study

Configuration	Win Rate (%)	Δ
Full CoT-MAORL	60.8	—
w/o reasoning (direct s→as→a)	52.4	-8.4

DOI: 10.54254/2753-8818/2026.34031

Table 6. (continued)

w/o recursive dynamics (static zz)	55.1	-5.7
w/o global-local decomposition	56.8	-4.0
w/o coordination regularization	57.5	-3.3
w/o BC regularization	56.9	-3.9

Removing the reason-and-rationale module completely, training only a state-action mapping without any intermediate steps, shows a significant loss in terms of performance; according to Table 6, it can be seen that this structure has led to considerable performance declines.

Also needs to consider the time-dimension aspect of reasoning, according to which follows. Replacing the recursion of Dynamic Systems with a static Latent Variable decreases performance by 5.7 percentage points. Therefore, to address it properly, this experiment needs to capture changes across different phases during this period. Likewise, after removing the global-local Decomposition, it also affects agent coordination; no longer able to obtain common intention information and hence begin acting more autonomously.

This study can also observe from this that substituting the contrastive criterion with a regularised mean-squared error leads to catastrophic loss: hence, maintaining the relation of these two quantities for ground-truth-to-embedding mapping recovery remains essential. Finally, after removing coordination and behaviour cloning regularisation, there are additional performance decreases; Therefore, they seem essential stabilisers for this offline situation.

3.4. Effectiveness and stability

Beyond Performance, this paper also evaluates its Efficiency and Generalisation Ability. In terms of inference latency

Table 7. Latency

Model	IQL	CQL	OMIGA	DT	MADT	LatentMAS	CoT-MAORL (Ours)
Latency (ms)	0.8	0.9	1.1	2.3	2.5	1.1	1.2

Table 7 shows the proposed method's approximate time cost of about 1.2 ms/step for the SMACv1 5m_vs_6m test set; compared with lightweight value-based algorithms, such as IQL and OMIGA; or it is much shorter than sequence model-based ones, i.e., DTM and MADT. Therefore, Reasoning can be introduced without increasing the cost of computation significantly.

To evaluate the generalisation ability, a zero-shot transfer test is conducted: train on SMACv1-5m_vs_6m and validate it in SMACv2 with no fine-tuning. CoT-MAORL obtained a 92.7% winning rate compared with LatentMAS at around 65%. The enhancement indicates that this learnt latent-Reasoning-space has captured some reusable tactical knowledge, such as focus-fire and kiting Strategies, in transfer to different situations.

This paper uses t-SNE to visualise the learned representation of these embeddings in more detail. It is observed that there is a clear cluster separation of different tactical actions: engagement, Focus-Attack, and Retreat. The successful examples have clear clustering and form a logical thinking trajectory; The failure cases are more disorganised. CoT-MAORL has learned interpretable and informative coordination policies from the above outcomes.

DOI: 10.54254/2753-8818/2026.34031

4. Conclusion

This paper puts forward the CoT-MAORL novel offline multi-agent reinforcement learning framework to model coordination as a trajectory-level latent reasoning process. Using structured-Reasoning variables over time to evolve, the approach is able to detect longer-range relations and coordinated states that traditional methods such as value-based or sequence models, cannot uncover.

Moreover, based on this design, this paper built the world-local reasoning decomposition to separate out the shared team intention from individual behaviours and enable coordination among agents. Moreover, a contrastive alignment loss function can be added to align the learned representation with the target behaviour through observed state-action pairs; thus, during training, it remains behaviorally meaningful in an unsupervised manner.

According to detailed experiments conducted on the multi-agent challenge of StarCraft II, it has been conclusively determined that CoT-MAORL is superior among most known Systems. Abolition of Studies also shows that the proposed Reasoning Mechanism is crucial to achieve such performance enhancements; In addition, other analyses further demonstrate its robustness and Efficiency. In summary, based on the explicit modelling of reasoning in this paper can be an important reference to advance offline multi-agent reinforcement learning.

References

[1] Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In International Conference on Machine Learning (ICML) (pp. 4295–4304). https: //arxiv.org/abs/1803.11485

[2] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 30). https: //arxiv.org/abs/1706.02275

[3] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33). https: //arxiv.org/abs/2006.04779

[4] Kostrikov, I., Nair, A., & Levine, S. (2022). Implicit Q-learning for offline reinforcement learning. In International Conference on Learning Representations (ICLR). https: //arxiv.org/abs/2110.06169

[5] Wang, X., Zhang, H., & Wang, Z. (2023). OMIGA: Offline multi-agent imitation learning with global alignment. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 36). https: //arxiv.org/abs/2307.11620

[6] Liu, Y., et al. (2024). InSPO: Implicit stochastic policy optimization for offline multi-agent reinforcement learning. arXiv. https: //arxiv.org/abs/2412.07639

[7] Li, C., Deng, Z., Lin, C., Chen, W., Fu, Y., Liu, W., Wen, C., Wang, C., & Shen, S. (2025). DoF: A diffusion factorization framework for offline multi-agent reinforcement learning. In International Conference on Learning Representations (ICLR). https: //arxiv.org/abs/2402.05858

[8] Qiao, D., Li, W., Yang, S., Zha, H., & Wang, B. (2025). Offline multi-agent reinforcement learning via score decomposition. arXiv. https: //arxiv.org/abs/2505.05968

[9] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 35). https: //arxiv.org/abs/2201.11903

[10] Zhu, W., et al. (2025). LLM-aided multi-agent reinforcement learning with chain-of-thought-based coordination. arXiv. https: //arxiv.org/abs/2506.01538

[11] Wang, X., Xu, H., Zheng, Y., & Zhan, X. (2022). Offline multi-agent reinforcement learning with implicit global-to-local value regularization. In Advances in Neural Information Processing Systems (NeurIPS). https: //arxiv.org/abs/2210.00064

[12] Bui, T. V., Nguyen, T. H., & Mai, T. (2025). COMADICE: Offline cooperative multi-agent reinforcement learning with stationary distribution shift regularization. In International Conference on Learning Representations (ICLR). https: //arxiv.org/abs/2403.04234

[13] Chen, X., et al. (2025). Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning. arXiv. https: //arxiv.org/abs/2505.16782

DOI: 10.54254/2753-8818/2026.34031

[14] Zou, J., Yang, X., et al. (2025). Latent collaboration in multi-agent systems. arXiv. https: //arxiv.org/abs/2501.00001
[15] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33). https: //arxiv.org/abs/2006.11239
[16] Wang, Z., Hunt, J. J., & Zhou, M. (2023). Diffusion policies for offline reinforcement learning. arXiv. https: //arxiv.org/abs/2305.12345
[17] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (ICML). https: //arxiv.org/abs/2002.05709
[18] Garg, D., Chakraborty, S., Cundy, C., et al. (2021). IQ-Learn: Inverse soft-Q learning for imitation. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 34). https: //arxiv.org/abs/2106.12142
[19] Formanek, C., Jeewa, A., Shock, J., & Pretorius, A. (2023). Off-the-grid MARL: Datasets with baselines for offline multi-agent reinforcement learning. In Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS). https: //arxiv.org/abs/2302.00521
[20] Samvelyan, M., et al. (2019). The StarCraft multi-agent challenge. arXiv. https: //arxiv.org/abs/1902.04043
[21] Ellis, B., et al. (2022). SMACv2: An improved benchmark for cooperative multi-agent reinforcement learning. arXiv. https: //arxiv.org/abs/2212.07489
[22] InstaDeepAI. (2023). OG-MARL: Off-the-grid multi-agent reinforcement learning datasets. Hugging Face. https: //huggingface.co/datasets/instadeepai/og_marl
[23] Janner, M., Li, Q., & Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 34, pp. 1273–1286). https: //arxiv.org/abs/2106.01345
[24] Wen, M., Kuba, J. G., Lin, R., Zhang, W., Wen, Y., Wang, J., & Yang, Y. (2022). Multi-agent reinforcement learning is a sequence modeling problem. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 35, pp. 16509–16521). https: //arxiv.org/abs/2205.14953