UAM-MARL: Uncertainty-Aware Modality-Enhanced Multi-Agent Reinforcement Learning with LLM-Guided Graph Policies

Zichen Song*
Sungkyunkwan University
Suwon, Republic of Korea
sls530@skku.edu

Weijia Li
Lanzhou University
Lanzhou, China
forgetfuljre@gmail.com

ABSTRACT

Multi-agent systems (MAS) are increasingly deployed to solve complex embodied tasks, yet coordination efficiency, reward design, and robustness remain persistent challenges. While multi-agent reinforcement learning (MARL) provides a principled framework for cooperation, existing LLM-driven approaches often assume perfect perception and static planning, which is unrealistic in noisy and dynamic environments. In practice, discrepancies between language-based reasoning and uncertain multimodal perception, the *semantic–perception gap*, lead to incorrect subgoal assignments, misaligned rewards, and unstable coordination. To address this limitation, we propose **UAM-MARL**, an uncertainty-aware modality-enhanced MARL framework. UAM-MARL integrates three components: (1) an uncertainty-aware perception module that estimates confidence scores over multimodal inputs and propagates them to the planner, (2) a cross-modal consistency checker that validates the alignment between LLM-generated plans and environment observations, and (3) an uncertainty-weighted reward generator that composes individual and team rewards by scaling reasoning-derived signals with perception confidence. These modules augment the LLM-based planner–critic and homology-guided graph policy, enabling more reliable coordination under noisy observations and dynamic disturbances. Experiments in the AI2-THOR simulator demonstrate that UAM-MARL achieves superior performance compared to centralized-LLM, dialogue-based LLM, and state-of-the-art MARL baselines, yielding higher success rates, shorter completion times, lower token costs, and stronger robustness to perception noise. Ablation studies further confirm the contribution of uncertainty modeling and cross-modal consistency to improving both efficiency and scalability.

KEYWORDS

Multi-Agent, Reinforcement Learning, LLM, Graph Policies

ACM Reference Format:
Zichen Song and Weijia Li. 2026. UAM-MARL: Uncertainty-Aware Modality-Enhanced Multi-Agent Reinforcement Learning with LLM-Guided Graph Policies. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/fp0052FHAS5003

*Corresponding author

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/fp0052FHAS5003

1 INTRODUCTION

Multi-agent systems (MAS) are increasingly deployed for embodied tasks such as household assistance, warehouse logistics [71], search-and-rescue, and environmental monitoring, where teams of agents must coordinate under partial observability, real-time constraints [53], and safety requirements. Despite notable progress, three challenges persist: *coordination efficiency* under non-stationary multi-agent dynamics and tight resource budgets; *reward design* that aligns local behaviors with global objectives over long horizons; and *scalability and robustness* when task graphs contain synchronization, precedence constraints, and cycles that induce deadlocks or redundant communication. These issues compound as the number of agents, the task complexity, and the environmental variability increase [9, 11, 14, 16, 20, 33, 52, 56–58, 63, 64, 70, 72, 83, 89].

Multi-agent reinforcement learning (MARL) provides a principled framework for learning cooperative behavior through interaction, with algorithms that leverage centralized-training and decentralized execution (CTDE) for credit assignment and stability. However, purely learning-based approaches struggle to follow high-level human instructions, to decompose long-horizon tasks into executable subgoals, and to obtain well-shaped, dense rewards across diverse scenarios. Large language models (LLMs) enhance task decomposition and planning, but existing methods often assume reliable perception and static world models. In realistic settings, noisy visual detection, ambiguous multimodal signals, or missing observations create a *semantic–perception gap*, where language-based reasoning diverges from the actual environment. This mismatch propagates errors to subgoal allocation, reward shaping, and dependency graph construction, ultimately degrading coordination and robustness [5, 17, 22, 25, 41, 45, 49, 85, 90–92, 98].

To address these limitations, we present **UAM-MARL**, an uncertainty aware modality-enhanced MARL framework for robust multi-agent collaboration under noisy and dynamic environments. UAM-MARL comprises four tightly coupled components. **(i) LLM-based task decomposition with critic verification:** A planner LLM parses high-level instructions into subtask sequences and an initial dependency graph, while a critic LLM verifies feasibility and correctness to suppress hallucinations. **(ii) Uncertainty-aware perception and cross-modal consistency checking:** Environmental inputs are modeled with probabilistic uncertainty, and a consistency checker aligns LLM-generated plans with multimodal observations, reducing the impact of perception errors. **(iii) Uncertainty-weighted modular reward generation:** Reward functions are generated by LLM reasoning but scaled by input confidence scores, ensuring that unreliable perceptual signals contribute less to shaping. **(iv) Homology-guided dependency construction and coordination:** Dependency graphs are interpreted as

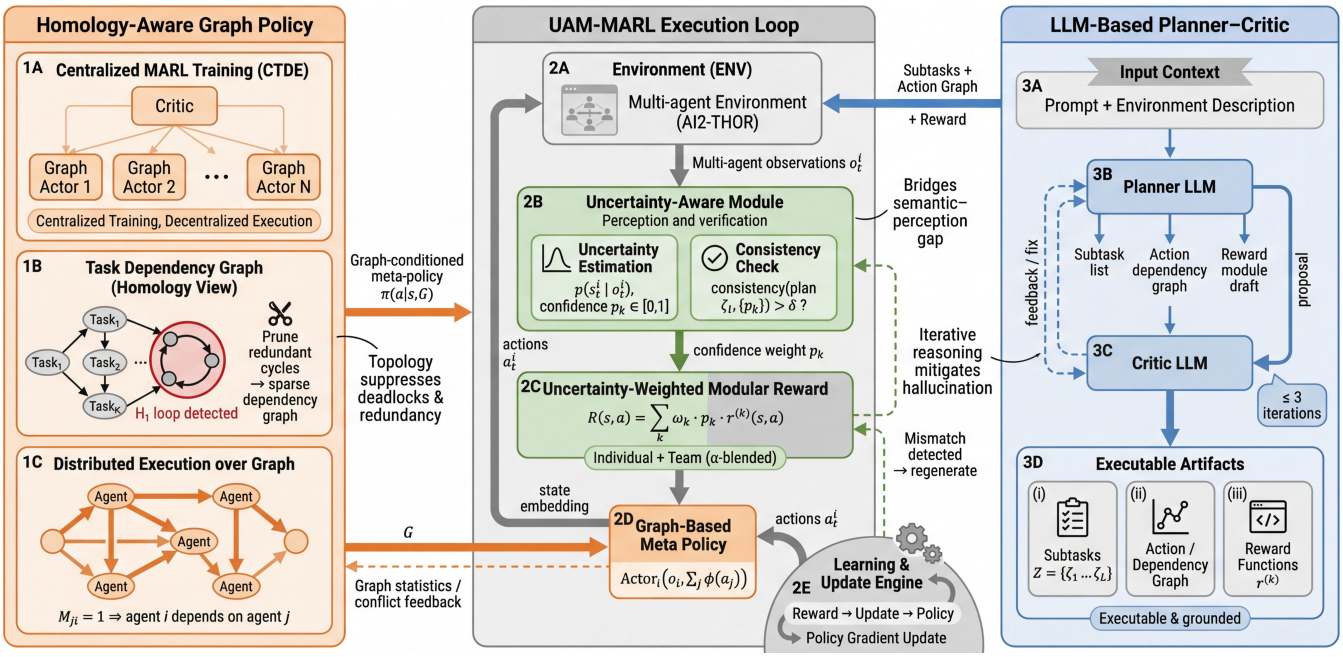

Figure 1: Overview of the proposed UAM-MARL framework. The left panel illustrates centralized training with decentralized execution (CTDE), where a task dependency graph is constructed from a homology perspective to detect and prune redundant cycles, enabling efficient distributed execution over agent graphs. The middle panel shows the UAM-MARL execution loop, integrating an uncertainty-aware module for perception verification, confidence-weighted modular rewards, and a graph-based meta-policy that coordinates individual and team behaviors. The right panel presents an LLM-based planner–critic, which iteratively generates and critiques subtasks, action-dependency graphs, and reward drafts, producing executable and grounded artifacts that bridge semantic planning and multi-agent reinforcement learning.

simplicial complexes; first-order homology detects cycles, and a pruning operator removes redundant loops, yielding a sparse and topology-consistent adjacency matrix for graph-based policy learning. Together, these modules bridge the semantic–perception gap and enable more reliable, efficient, and robust multi-agent coordination. Comprehensive experiments in a 3D embodied simulator demonstrate that UAM-MARL achieves higher success rates, shorter completion times, lower token costs, and stronger robustness to perception noise compared to the dialogue-based LLM systems and the strong MARL baselines.

This work makes the following contributions:

- **A modality-enhanced MARL framework with uncertainty modeling.** We introduce UAM-MARL, which integrates LLM planning, homology-guided coordination, and uncertainty-aware multimodal reasoning to improve cooperation quality, robustness, and scalability.
- **Planner–critic task decomposition.** We design an LLM planner with a critic verifier that eliminates semantic errors and infeasible steps, producing executable subtask sequences and dependency graphs aligned with environmental affordances.
- **Uncertainty-aware and interpretable reward synthesis.** We propose an LLM-based reward generator that emits explicit reasoning steps and composes modular individual and team rewards, scaled by perceptual confidence, enabling consistent shaping across tasks under noisy observations.
- **Cross-modal consistency checking.** We introduce a lightweight consistency checker that validates LLM-generated plans against multimodal perception, reducing plan environment mismatches and mitigating the semantic–perception gap.

2 RELATED WORK

2.1 LLM for Multi-Agent Planning

Large language models (LLMs) have recently been adopted to bridge high-level natural-language goals and low-level action execution in embodied and multi-agent settings. Early efforts leverage LLMs as high-level planners that translate instructions into structured plans, task graphs, or programs (e.g., plan-as-program and code-as-policy paradigms), improving instruction following and long-horizon reasoning. Dialogue-based approaches assign an LLM to each agent and coordinate through inter-agent conversations to negotiate roles, share partial observations, and refine joint strategies; this enhances flexibility but introduces latency, high token consumption, and instability in dynamic environments when frequent re-planning is required. Centralized LLM planners reduce dialogue overhead by producing joint action proposals or synchronized subtask schedules for all agents, yet they often struggle to incorporate individual

agent affordances, local constraints, and asynchronous execution, which are critical under partial observability and the non-stationary dynamics [7, 10, 12, 19, 27, 28, 42, 44, 51, 73, 75, 77, 86, 88, 94, 96].

To improve reliability, planner–checker and the planner-critic paradigms augment LLM planning with verification components that detect semantic inconsistencies, unreachable subgoals, and unsafe steps before deployment. Memory-augmented variants maintain episodic or semantic memories to reduce re-prompting costs and improve temporal coherence, while hierarchical designs separate abstract planning from low-level control policies learned via reinforcement learning (RL). Despite these advances, most current LLM-driven multi-agent systems assume *accurate and noise-free perception*. In realistic settings, mismatches between language-based reasoning and uncertain multimodal observations—the *semantic–perception gap*—cause incorrect subgoal assignments, reward misalignment, and unstable coordination. Our work addresses this gap by introducing uncertainty-aware modules that integrate perception confidence and cross-modal consistency into the LLM-driven planning pipeline.

2.2 Reward Function Learning in MARL

Reward design remains a central challenge in multi-agent reinforcement learning (MARL). Hand-crafted dense rewards are labor-intensive, brittle across tasks, and prone to specification gaming; sparse terminal rewards slow learning and exacerbate credit assignment. Classical approaches employ potential-based reward shaping to preserve optimal policies while accelerating exploration, or exploit centralized training with decentralized execution (CTDE) to learn value functions that aid credit assignment. However, hand-designed shaping terms rarely scale to complex tasks with precedence constraints, safety conditions, and role specialization [6, 13, 15, 18, 26, 29, 43, 46, 50, 54, 76, 78, 87, 97].

Recent work explores automated reward construction. Programmatic reward induction converts language descriptions into executable reward code; IRL and preference-based methods infer rewards from demonstrations or comparisons; and LLM-based approaches generate dense, modular rewards from textual task specifications and environment summaries, sometimes refined by simulation-in-the-loop or human feedback. In single-agent domains, LLM-generated rewards (combined with self-reflection or verifier signals) have improved learning speed and generalization. Extending these ideas to MARL requires decomposing team objectives into individual and joint components, handling inter-agent dependencies, and ensuring shaping consistency across roles. Our method advances this direction by introducing *uncertainty-weighted reward synthesis*, where LLM-derived modular rewards are scaled by perceptual confidence, thereby mitigating the propagation of errors from noisy observations.

2.3 Graph Structure and Topological Methods in RL

Graph representations are natural for modeling multi-agent dependencies, communication, and factorization. Coordination graphs and value factorization methods (e.g., mixing networks) decompose the joint action-value function into local terms with structured interactions, improving scalability under CTDE [52]. Graph neural networks (GNNs) propagate information along communication links or task dependencies to support coordination, role assignment, and implicit scheduling. Recent graph-based coordination strategies build action-dependency graphs to order agent updates, but they typically rely on heuristic or learned edge construction without topological guarantees, leaving cycles and redundant edges that increase variance and communication overhead [2, 8, 31, 32, 37, 40, 61, 66, 68, 69, 72, 79–82, 93].

Topological data analysis (TDA) offers tools to reason about global structure beyond local connectivity. Homology captures holes and cycles via Betti numbers; persistent homology tracks their stability across scales. In RL and control, topology has been used to characterize exploration structure, detect bottlenecks, and analyze policy landscapes. However, explicit homology-based construction of multi-agent dependency graphs remains underexplored. We extend this line of work by combining homology-guided dependency graph pruning with uncertainty-aware reasoning, yielding dependency structures that are both *topology-consistent* and *robust to noisy perceptions* [1, 4, 23, 24, 30, 35, 38, 47, 48, 55, 62, 65, 74, 84].

2.4 Uncertainty and Multimodal Perception in MARL

Handling perceptual uncertainty is a long-standing problem in robotics, control, and reinforcement learning. Bayesian policies, ensemble methods, and distributional RL approaches model epistemic or aleatoric uncertainty to improve robustness under partial observability. In multi-robot systems, sensor fusion and multimodal learning are employed to reduce noise and improve task success, while in embodied AI, vision-language models aim to align natural-language instructions with raw sensory observations. However, most existing MARL frameworks treat perception as accurate input, without explicitly modeling confidence or checking for cross-modal inconsistencies. Our UAM-MARL departs from this assumption by incorporating an uncertainty-aware perception module and a cross-modal consistency checker, which jointly bridge the semantic–perception gap and enable more reliable LLM-guided multi-agent coordination in noisy environments [3, 21, 34, 36, 39, 59, 60, 67, 95].

3 METHOD

3.1 Problem Formulation

We formulate the multi-agent environment as a Multi-Agent Markov Decision Process (MMDP), defined by $\mathcal{M} = (S, \{A_i\}_{i=1}^N, P, R, \gamma)$, where S is the global state space, A_i is the action space for agent i, $P : S \times A \to S$ is the transition probability function, and $\gamma \in (0, 1)$ is the discount factor. The joint action is denoted as $a = (a_1, \ldots, a_N)$, and the joint policy is $\pi = \{\pi_1, \ldots, \pi_N\}$. The objective is to learn a joint policy that maximizes the expected cumulative reward:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

3.2 LLM-Based Task Decomposition and Critic Evaluation

Given a high-level natural language instruction τ, we employ a planner LLM $\mathcal{L}_p$ to decompose it into a sequence of executable

Algorithm 1 UAM-MARL Framework

Require: Environment $\mathcal{E}$, Agents $\mathcal{A}$, Planner $\mathcal{L}_p$, Critic $\mathcal{L}_c$, Reward LLM $\mathcal{L}_r$, Homology module $\mathcal{H}$, Perception model $\mathcal{P}$, Meta-policy π_θ
Ensure: Learned meta-policy π_θ
for each episode **do**
  Generate subgoals with planner–critic; build G_0
  Prune graph $G \leftarrow \mathcal{H}(G_0)$
  Generate modular rewards $\mathcal{R}$ via $\mathcal{L}_r$
  for each timestep t **do**
    for each agent a_i **do**
      Get observation o_t^i; estimate uncertainty $p(s_t^i | o_t^i)$
      Check consistency between g_i and $\{p(s_t^i)\}$; adjust if needed
      Sample $a_t^i \sim \pi_\theta(s_t^i, G)$
      Reward $r_t^i = \sum_k \omega_k p_k r^{(k)}(s_t^i, a_t^i)$
    end for
    Update π_θ with policy gradient
  end for
end for
return π_θ =0

subtasks $\mathcal{Z} = \{\zeta_1, \ldots, \zeta_L\}$ and construct an initial task dependency graph $G_0 = (V, E)$, where $V = \mathcal{Z}$ and E encodes directed dependencies between subtasks. To mitigate hallucinations and semantic errors, we introduce a critic LLM $\mathcal{L}_c$ to validate the plan. The critic evaluates each ζ_l and the graph structure to identify contradictions, infeasible steps, or illogical dependencies. If inconsistencies are detected, feedback is returned to $\mathcal{L}_p$ to revise the plan. This iterative planner–critic loop ensures reliable, grounded subtask graphs.

3.3 Uncertainty-Aware Perception and Cross-Modal Consistency

Unlike prior work that assumes accurate and noise-free perception, our framework explicitly models perceptual uncertainty. For each observation o_t^i from agent a_i, a perception encoder produces a distribution over possible states:

$$p(s_t^i \mid o_t^i) = \text{PerceptionModel}(o_t^i), \tag{2}$$

yielding confidence scores p_k for candidate entities or attributes. These uncertainty estimates are propagated to the planner to guide subgoal generation.

To further reduce mismatches between LLM-generated subgoals and environment reality, we introduce a **cross-modal consistency checker**. Given LLM plan ζ_l and perception outputs $\{p_k\}$ from vision, language, or other modalities, the checker verifies alignment:

$$\mathbb{I}_{\text{cons}}(\zeta_l) = \begin{cases} 1 & \text{if consistency}(\zeta_l, \{p_k\}) > \delta, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where δ is a confidence threshold. If inconsistencies are found, the planner is prompted to regenerate or adjust the subgoal, preventing propagation of faulty instructions.

3.4 Uncertainty-Weighted Modular Reward Function Generator

To provide dense and interpretable feedback that accounts for perception noise, we extend LLM-based reward generation with uncertainty weighting. For each subgoal, the reward generator $\mathcal{L}_r$ produces chain-of-thought (CoT) reasoning that maps to modular reward functions. Each reward module $r^{(k)}(s, a)$ is scaled by the corresponding confidence p_k:

$$R(s, a) = \sum_{k=1}^{K} \omega_k \cdot p_k \cdot r^{(k)}(s, a), \tag{4}$$

where ω_k is a learned or manually tuned weight and $p_k \in [0, 1]$ represents perceptual confidence. We further decompose the global reward into individual and team components:

$$R(s, a) = \alpha R_g(s, a) + (1 - \alpha) \cdot \frac{1}{N} \sum_{i=1}^{N} R_i(s, a_i). \tag{5}$$

This formulation down-weights unreliable perceptual signals, thereby reducing error propagation from noisy observations.

3.5 Homology-Aware Dependency Graph Construction

We interpret the dependency graph G_0 as a simplicial complex $\mathcal{K}$. The 0-simplices are tasks (vertices), and 1-simplices are dependencies (edges). Let C_0, C_1 be the chain groups and $\partial_1 : C_1 \to C_0$ the boundary operator:

$$\partial_1([v_i, v_j]) = v_j - v_i. \tag{6}$$

The first homology group is $H_1(\mathcal{K}; \mathbb{Z}) = Z_1/B_1$, where $Z_1 = \ker(\partial_1)$ and $B_1 = \text{im}(\partial_2)$. Each independent cycle $c_j \in H_1$ is assigned a relevance score:

$$S(c_j) = \frac{1}{|c_j|} \sum_{e \in c_j} f_{\text{rel}}(e). \tag{7}$$

A threshold λ filters low-score cycles, yielding a pruned graph $G = \mathcal{P}_\lambda(G_0)$. The final adjacency matrix $M \in \{0, 1\}^{N \times N}$ encodes agent dependencies, where $M_{ji} = 1$ indicates agent i depends on j.

3.6 Graph-Based Meta Policy Learning

Let $\mathcal{P}_i = \{j \mid M_{ji} = 1\}$ be the parent agents of agent i. The meta-policy is defined as:

$$\pi_i^m(a_i \mid o_i, \{a_j\}_{j \in \mathcal{P}_i}; \mathcal{T}) = \text{Actor}_i\left(o_i, \sum_{j \in \mathcal{P}_i} \phi(a_j)\right). \tag{8}$$

The global objective is:

$$\eta_m = \mathbb{E}_{a \sim \pi(\cdot|s,M)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right], \tag{9}$$

with gradient update:

$$\nabla_{\theta_i} \eta_m = \mathbb{E}_{a_i \sim \pi_i^m}\left[\nabla_{\theta_i} \log \pi_i^m(a_i \mid o_i, \{a_j\}_{j \in \mathcal{P}_i}) Q_i(s, a)\right]. \tag{10}$$

This structure-aware policy promotes efficient, scalable, and robust coordination under uncertain and multimodal observations.

Table 1: Evaluation results on four scenes. We report the average success rate (↑), average time (↓), and normalized token cost (↓) over 100+ runs with 4 agents.

Scene	Metric	Centralized LLM	LLMs Dialog	LGC-MARL (w/o Critic)	LGC-MARL (w/o Reward)	LGC-MARL (w/o Graph)	**UAM-MARL (Ours)**
Scene 1	Success Rate	0.59	0.67	0.77	0.89	0.73	**0.95**
	Avg. Time	99.59 ± 5.32	154.54 ± 9.74	68.61 ± 2.09	73.15 ± 3.39	78.95 ± 3.96	**62.36 ± 1.73**
	Token Cost	6.3	10.6	1.3	1.2	2.2	**0.9**
Scene 2	Success Rate	0.54	0.62	0.71	0.84	0.67	**0.91**
	Avg. Time	106.44 ± 6.07	189.21 ± 11.82	80.94 ± 2.57	83.79 ± 3.46	89.12 ± 4.02	**74.57 ± 2.13**
	Token Cost	7.8	14.2	2.6	2.3	3.0	**1.8**
Scene 3	Success Rate	0.58	0.64	0.73	0.86	0.71	**0.92**
	Avg. Time	105.11 ± 5.83	183.23 ± 10.57	77.72 ± 2.63	80.08 ± 3.58	86.01 ± 3.34	**69.28 ± 2.35**
	Token Cost	7.5	13.6	2.4	2.1	3.4	**1.7**
Scene 4	Success Rate	0.51	0.60	0.71	0.80	0.65	**0.89**
	Avg. Time	115.85 ± 6.56	206.69 ± 11.31	91.34 ± 2.83	89.34 ± 4.24	95.37 ± 3.50	**78.93 ± 2.42**
	Token Cost	8.8	16.3	3.8	3.2	5.1	**2.2**

4 EXPERIMENTS

4.1 Experiment Setup

We conduct comprehensive evaluations of **UAM-MARL** in the **AI2-THOR** simulation platform, a high-fidelity interactive environment widely adopted for embodied AI research. Within this platform, agents must collaboratively complete complex multi-stage tasks such as object search, delivery, and manipulation in cluttered household environments. We consider four diverse and challenging *scenes*, each designed to stress different coordination capabilities including spatial reasoning, temporal ordering, subgoal alignment, and robustness under noisy observations.

Each experimental episode involves a team of $n \in \{4, 5, 6\}$ agents operating simultaneously. To simulate realistic perception errors, we inject controlled *observation noise*, including: (i) visual misclassification (e.g., confusing cup vs. bowl), (ii) object occlusion (e.g., target partially hidden), and (iii) modality dropout (e.g., missing visual or language cues). These perturbations explicitly evaluate the ability of systems to handle the *semantic–perception gap*. Our evaluation involves more than 100 independent runs per configuration to ensure statistical reliability. All models share the same initialization and environment seeds for fairness.

4.2 Metric Definitions and Mathematical Formulation

To rigorously and comprehensively evaluate the performance of our proposed **UAM-MARL** framework, we adopt a set of five complementary metrics. These metrics are designed to quantify multiple aspects of multi-agent coordination, including task completion efficiency, inter-agent cooperation, resilience to environmental uncertainty, and the computational overhead introduced by large language models (LLMs) in the decision-making loop. Formally, we define these metrics as follows.

Success Rate (SR). The success rate captures the proportion of episodes in which the multi-agent system successfully completes the assigned task according to predefined criteria. This metric directly reflects the overall effectiveness and reliability of UAM-MARL in achieving the desired goals:

$$\mathrm{SR} = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[\text{episode } i \text{ is successful}], \tag{11}$$

where M denotes the total number of evaluated episodes and $\mathbb{1}[\cdot]$ is the indicator function, which evaluates to 1 if the episode is successful and 0 otherwise. A higher SR indicates a more reliable multi-agent coordination strategy.

Average Task Time ($\mathcal{T}_{\mathbf{avg}}$). Task efficiency is measured by the average number of environment steps required to complete the task. This metric evaluates how quickly the agents can achieve their objectives and reflects the efficiency of their planning and coordination:

$$\mathcal{T}_{\mathrm{avg}} = \frac{1}{M} \sum_{i=1}^{M} T_i, \tag{12}$$

where T_i represents the number of steps taken in episode i. Minimizing $\mathcal{T}_{\mathrm{avg}}$ is desirable as it indicates both effective coordination and efficient resource utilization.

Token Cost ($C_{\mathbf{token}}$). In LLM-driven multi-agent systems, language model interactions introduce additional computational and communication overhead. To quantify this, we define the token cost as the average number of tokens consumed by all agents per episode:

$$C_{\mathrm{token}} = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N} \sum_{j=1}^{N} \mathrm{Token}_{ij}, \tag{13}$$

where N is the number of agents and Token_{ij} is the total number of tokens used by agent j in episode i. Lower C_{token} indicates a more

Table 2: Ablation study of UAM-MARL. We report success rate (↑), average time (↓), token cost (↓), conflict count (↓), and uncertainty robustness (UR, ↑) over 100+ runs with 4 agents.

Model Variant	Success Rate (↑)	Avg. Time (↓)	Token Cost (↓)	Conflict Count (↓)	UR (↑)
w/o Uncertainty Module	0.83	82.7 ± 3.6	2.1	4.5	0.68
w/o Consistency Check	0.85	81.9 ± 3.3	2.0	4.2	0.71
w/o Weighted Reward	0.86	84.5 ± 3.8	2.2	4.0	0.73
UAM-MARL (Full)	**0.91**	**74.6 ± 2.1**	**1.8**	**3.2**	**0.82**

efficient use of LLM queries, which is crucial for scalability and real-time deployment.

Coordination Conflict Count (C_{conflict}). Multi-agent systems often encounter coordination conflicts such as collisions or redundant task assignments. We define the coordination conflict count as:

$$C_{\text{conflict}} = \frac{1}{M} \sum_{i=1}^{M} (\text{collision}_i + \text{redundancy}_i), \tag{14}$$

where collision_i denotes the number of physical or logical collisions between agents, and redundancy_i counts overlapping task assignments in episode i. Reducing C_{conflict} is critical for ensuring safe and efficient operations in densely populated environments.

Uncertainty Robustness (UR). To assess the resilience of UAM-MARL to imperfect and noisy observations, we introduce a novel metric called *Uncertainty Robustness (UR)*. This metric compares the task success under noisy conditions to that under ideal (clean) conditions:

$$\text{UR} = \frac{SR_{noisy}}{SR_{clean}}, \tag{15}$$

where SR_{clean} and SR_{noisy} represent the success rates measured in clean and noisy observation settings, respectively. A UR value close to 1 indicates that the system maintains high performance despite perceptual uncertainties, highlighting the robustness of the proposed approach. Conversely, lower values suggest sensitivity to noise, which can be a target for further model improvement.

Collectively, these five metrics provide a comprehensive evaluation framework, allowing us to measure not only the task completion performance but also the coordination efficiency, safety, computational overhead, and resilience of UAM-MARL in diverse operational scenarios.

Interpretation. Together, these metrics provide a comprehensive evaluation: 1) Higher SR reflects reliable goal completion; 2) Lower $\mathcal{T}_{\text{avg}}$ implies efficiency; 3) Lower C_{token} reflects economical LLM usage; 4) Lower C_{conflict} indicates improved coordination; 5) Higher UR highlights robustness to noisy and uncertain observations.

4.3 Baselines

We compare our proposed UAM-MARL with a range of competitive and representative baselines [78–85]:

- **Centralized LLM**: A single large language model plans all agent behaviors in a centralized fashion without intra-agent dialog. While capable of global planning, this method lacks flexibility and incurs high token costs.
- **LLMs Dialog**: Each agent interacts with others via language-based dialog to negotiate task roles and coordinate subgoals. Although this improves decentralization, it suffers from dialog ambiguity and lacks noise robustness.
- **TopoMARL (w/o UAM)**: Our framework without uncertainty-aware modules, representing static LLM planning with the homology-guided coordination but no perception uncertainty modeling. This baseline directly measures the contribution of UAM.
- **LGC-MARL (w/o Critic)**: An ablation that omits the critic during task decomposition, testing the importance of semantic verification.
- **LGC-MARL (w/o Reward)**: A important variant replacing the uncertainty-weighted modular reward generator with heuristics, assessing the contribution of CoT-based and confidence-aware reward shaping.
- **LGC-MARL (w/o Graph)**: A model without homology-guided graph pruning, instead using shallow or random connections, evaluating the role of topological priors.
- **UAM-MARL (Ours)**: The complete model integrating the planner–critic task decomposition, uncertainty-aware perception, cross-modal consistency, uncertainty-weighted reward generation, and homology-aware dependency graph construction.

4.4 Quantitative Results

Table 1 reports the evaluation results of all methods across four distinct scenes with four agents. **UAM-MARL** consistently achieves the best performance across all environments, reaching success rates of 0.95, 0.91, 0.92, and 0.89 in Scenes 1 through 4, respectively. These represent substantial improvements over traditional centralized planning and dialog-based methods, whose success rates remain below 0.7 in all cases.

In terms of task efficiency, UAM-MARL requires the fewest steps to complete tasks. For example, in Scene 1 it converges in only 62.36 ± 1.73 steps on average, significantly faster than all ablations and baselines. Token usage is also minimized: UAM-MARL consumes fewer LLM tokens than any alternative (e.g., only 0.9 tokens in Scene 1), indicating superior communication efficiency and inference economy.

These improvements stem from the integration of structured task decomposition, uncertainty-aware perception, cross-modal consistency checking, modular reward shaping, and homology-based dependency pruning. The uncertainty-aware modules in particular enable the system to remain robust under noisy and incomplete

observations, yielding stable improvements across diverse environments. Moreover, the performance gains of UAM-MARL remain consistent across scenes of varying spatial layouts and interaction complexities, suggesting strong generalization beyond a single task configuration. This stability indicates that uncertainty modeling and graph-structured coordination not only improve peak performance but also reduce variance across runs, which is critical for scalable deployment in real-world multi-agent systems.

4.5 Ablation Study

Table 2 presents the ablation study evaluating the contribution of each uncertainty-aware component in UAM-MARL. Several trends emerge: removing the uncertainty module significantly reduces robustness, with success rate dropping to 0.83 and uncertainty robustness (UR) falling to 0.68. Without consistency checking, semantic–perception mismatches increase conflicts to 4.2, degrading success rate to 0.85. Replacing uncertainty-weighted rewards with heuristic shaping slows convergence (84.5 steps) and lowers UR to 0.73. The full UAM-MARL model outperforms all ablations, achieving a success rate of 0.91, the fastest average completion time (74.6 steps), the lowest token cost (1.8), the fewest conflicts (3.2), and the highest robustness to uncertainty (UR = 0.82). These results highlight that each module contributes significantly to overall performance, while the uncertainty-aware design is indispensable for handling noisy observations. Overall, UAM-MARL demonstrates consistent advantages over baselines in terms of task success, efficiency, token economy, and resilience to perception noise.

5 HYPERPARAMETER SETTINGS

We provide a comprehensive description of the hyperparameter configurations used for training and evaluating UAM-MARL to ensure reproducibility, stable convergence, and robust performance across diverse multi-agent scenarios. All hyperparameters were selected based on preliminary ablation studies and prior empirical practices in multi-agent reinforcement learning and LLM-guided planning.

For the environment, each experiment involves $N = 5$ agents interacting in a shared space with dynamic task objectives. The maximum episode length is set to $T_{\max} = 200$ steps, providing sufficient horizon for completing hierarchical tasks while avoiding excessive computational overhead. The discount factor γ is chosen as 0.99 to balance immediate rewards and long-term task completion, and all evaluation metrics, including success rate and average task time, are computed over $M = 100$ episodes to ensure statistical significance and reduce variance in performance measurements. The observation noise level during testing is controlled to evaluate the robustness of the framework under uncertain perception conditions.

The LLM planner is implemented using `GPT-4-mini` with 8 attention layers and a hidden dimension of 1024. The critic LLM is fine-tuned on a curated dataset of task validation examples to effectively identify semantic inconsistencies, infeasible subtasks, or contradictory dependencies in the generated plans. During subgoal generation, the token limit is restricted to 512 tokens per subgoal to maintain computational efficiency, and the iterative planner–critic loop is executed up to three times per high-level instruction to refine the task dependency graph. The cross-modal consistency checker compares LLM-generated subgoals with perception outputs across multiple modalities, employing a confidence threshold $\delta = 0.7$ to determine acceptable alignment. If consistency falls below this threshold, the planner is prompted to regenerate or adjust the subgoal, ensuring high-fidelity task execution.

Perception uncertainty is modeled using a ResNet-18 encoder with a feature output dimension of 256. Softmax probabilities over candidate states provide confidence scores $p_k \in [0, 1]$, which are used to weight the contribution of each subgoal in the reward function. The reward function itself consists of $K = 4$ modular components corresponding to individual task achievement, team coordination, safety adherence, and operational efficiency. Each module is scaled by its corresponding perceptual confidence p_k and a default weight $\omega_k = 1.0$, with the global versus local reward blending factor α set to 0.6.

For homology-aware dependency graph construction, the initial subtask graph G_0 is interpreted as a simplicial complex. The relevance score of each independent cycle is normalized to $[0, 1]$, and cycles with scores below the pruning threshold $\lambda = 0.5$ are discarded, yielding a simplified and task-relevant dependency graph G. Exploration noise is sampled from a Gaussian distribution with standard deviation $\sigma = 0.1$, promoting sufficient exploration while maintaining stability. Training is conducted for 10,000 episodes, with early stopping applied if the success rate does not improve for 500 episodes. All experiments are executed on an NVIDIA A100 GPU with 40GB memory, resulting in an average runtime of approximately 1.2 seconds per episode. These hyperparameters collectively ensure that UAM-MARL achieves high success rates, efficient task completion, and robust coordination under multimodal uncertainty while maintaining reproducibility and computational feasibility.

6 CONCLUSION

In this work, we proposed **UAM-MARL**, an uncertainty-aware multi-agent reinforcement learning framework that addresses the semantic–perception gap in embodied collaboration. By combining LLM-based task decomposition, modular reward generation, and homology-guided graph policies with uncertainty-aware perception, cross-modal consistency checking, and uncertainty-weighted rewards, UAM-MARL bridges the gap between high-level language reasoning and robust execution under noisy observations. Extensive experiments in AI2-THOR demonstrate that UAM-MARL not only surpasses centralized and dialog-based LLM baselines in success rate, efficiency, and token economy, but also achieves strong robustness against perception noise and modality errors through its uncertainty-aware design. These results suggest that integrating symbolic reasoning from LLMs, topological graph structures, and uncertainty modeling is a promising direction for enabling resilient and scalable multi-agent collaboration in complex embodied environments. In future work, we plan to extend UAM-MARL to larger agent populations and real-world robotic systems, exploring its potential in domains such as warehouse logistics, autonomous driving, and disaster response. More broadly, our study highlights the value of unifying language-based reasoning, topological analysis, uncertainty modeling, and reinforcement learning as a principled paradigm for multi-agent intelligence.

REFERENCES

[1] Gürdal Arslan and Serdar Yüksel. 2017. Decentralized Q-learning for stochastic teams and games. *IEEE Trans. Automat. Control* 62, 4 (2017), 1545–1558.

[2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. 31.

[3] J. Bai et al. 2024. A Dynamic Knowledge Graph Approach to Distributed Self-Driving Laboratories. *Nature Communications* 15 (Jan 2024).

[4] B. Balaji, S. Mallya, S. Genc, S. Gupta, L. Dirac, V. Khare, G. Roy, T. Sun, Y. Tao, B. Townsend, E. Calleja, S. Muralidhara, and D. Karuppasamy. 2020. Deepracer: Autonomous racing platform for experimentation with sim2real reinforcement learning. In *IEEE International Conference on Robotics and Automation*. 2746–2754.

[5] Wele Gedara Chaminda Bandara and Vishal M. Patel. 2023. ChangeFormer: A Transformer-Based Siamese Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[6] Wele Gedara Chaminda Bandara and Vishal M. Patel. 2023. ChangeFormer: A Transformer-Based Siamese Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[7] Md A Bashar, Lorenzo Bruzzone, and Francesca Bovolo. 2018. Unsupervised Change Detection in Satellite Images Using Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*. IEEE, 1608–1617.

[8] Evan Brooks, Luke Walls, Robert L. Lewis, and Satinder Singh. 2024. Large Language Models Can Implement Policy Iteration. *Advances in Neural Information Processing Systems* 36 (2024).

[9] H. Chakraa, F. Guérin, E. Leclercq, and D. Lefebvre. 2023. Optimization techniques for multi-robot task allocation problems: Review on the state-of-the-art. *Robotics and Autonomous Systems* (2023), 104492.

[10] Hao Chen, Xiaoyan Li, Wenzhong Shi, and Liangpei Wang. 2021. STANet: A Spatial-Temporal Attention Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 59, 7 (2021), 5769–5783.

[11] Hao Chen, Zilong Wu, Yusheng Duan, Liangpei Wang, and Xiangjian He. 2022. BIT: Bi-Temporal Image Transformer for Change Detection. *ISPRS Journal of Photogrammetry and Remote Sensing* 183 (2022), 91–105.

[12] Kai Chen, Shuo Huang, Wei Wang, and Jian Zhang. 2023. Change Detection with Graph Convolutional Networks for Remote Sensing Images. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 3892–3895.

[13] Kai Chen, Shuo Huang, Wei Wang, and Jian Zhang. 2023. Change Detection with Graph Convolutional Networks for Remote Sensing Images. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 3892–3895.

[14] Y. Chen, J. Arkin, Y. Zhang, N. Roy, and C. Fan. 2024. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems?. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 4311–4317.

[15] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. 2018. Change Detection in VHR Images Using Fully Convolutional Siamese Networks. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 4068–4072.

[16] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. 2018. Change Detection with Fully Convolutional Siamese Networks. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 4063–4067.

[17] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. 2018. FC-Siam-Diff: A Fully Convolutional Siamese Difference Network for Change Detection. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 4068–4072.

[18] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. 2018. FC-Siam-Diff: A Fully Convolutional Siamese Difference Network for Change Detection. In *IEEE International Conference on Image Processing (ICIP)*. IEEE, 4068–4072.

[19] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. 2018. HRSCD: A Benchmark Dataset for High-Resolution Satellite Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 56, 12 (2018), 7123–7139.

[20] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. 2018. Change Detection in Multispectral Remote Sensing Images via Deep Siamese Convolutional Network. *Remote Sensing* 10, 5 (2018), 760.

[21] Y. Ding, X. Zhang, C. Paxton, and S. Zhang. 2019. Robotic Task Oriented Knowledge Graph for Human-Robot Collaboration in Disassembly. *Procedia CIRP* 83 (2019), 105–110.

[22] Shengyu Fang, Kai Li, Jie Shao, and Zhiqiang Li. 2021. SNUNet: A U-Net-Based Siamese Network for Change Detection of Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 14 (2021), 1184–1196.

[23] Jakob Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. 2137–2145.

[24] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI Conference on Artificial Intelligence*.

[25] Peng Gao, Xiaoming Liu, Hao Wang, and Yong Zhang. 2023. A Multiscale Siamese Network With Transformer for Change Detection in Remote Sensing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 5056–5065.

[26] Peng Gao, Xiaoming Liu, Hao Wang, and Yong Zhang. 2023. A Multiscale Siamese Network With Transformer for Change Detection in Remote Sensing Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, 5056–5065.

[27] Ming Gong, Yang Yang, Jing Zhang, Yan Shi, and Li Yu. 2016. Change Detection for SAR Images Using Deep Convolutional Neural Networks. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 5157–5160.

[28] Rui Guo, Hao Chen, and Liangpei Wang. 2023. MANet: Multiscale Attention Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.

[29] Rui Guo, Hao Chen, and Liangpei Wang. 2023. MANet: Multiscale Attention Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.

[30] Jayesh K. Gupta, Maxim Egorov, and Mykel Kochenderfer. 2017. Cooperative multi-agent control using deep reinforcement learning. In *International Conference on Autonomous Agents and Multi-agent Systems*. 66–83.

[31] Shun Hao, Yuan Gu, Han Ma, Jie Hong, Zhi Wang, Dazheng Wang, and Zhi Hu. 2023. Reasoning with Language Model is Planning with World Model. In *NeurIPS 2023 Workshop on Generalization in Planning*.

[32] Edward J. Hu, Peter Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shun Wang, Lei Wang, Wei Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[33] Xin Huang, Hongyang Pan, Hao Chen, Xiangjian He, and Jun Liu. 2022. DMINet: Dual Multiscale Interaction Network for Change Detection of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–15.

[34] D. Höller et al. 2020. HDDL: An Extension to PDDL for Expressing Hierarchical Planning Problems. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. 34. 9883–9891.

[35] Garud N. Iyengar. 2005. Robust dynamic programming. *Mathematics of Operations Research* 30, 2 (2005), 257–280.

[36] Z. Jia, J. Li, X. Qu, and J. Wang. 2025. Enhancing Multi-Agent Systems via Reinforcement Learning with LLM-based Planner and Graph-based Policy. *arXiv preprint arXiv:2503.10049* (2025).

[37] Y.-q. Jiang et al. 2019. Task Planning in Robotics: An Empirical Comparison of PDDL- and ASP-Based Systems. *Frontiers of Information Technology Electronic Engineering* 20 (2019), 363–373.

[38] Erim Kardeş, Fernando Ordóñez, and Randolph W. Hall. 2011. Discounted robust stochastic games and an application to queueing control. *Operations Research* 59, 2 (2011), 365–382.

[39] H. Kasaei and M. Kasaei. 2024. VITAL: Visual Teleoperation to Enhance Robot Learning Through Human-in-the-Loop Corrections. Preprint.

[40] Alexander Levy, George Konidaris, Robert Platt, and Kate Saenko. 2018. Learning Multi-Level Hierarchies with Hindsight. In *International Conference on Learning Representations (ICLR)*.

[41] Hongwei Li, Yu Wang, Shuang Wang, and Xiangrong Zhang. 2023. Spectral-Spatial-Temporal Attention-Based Change Detection Network for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[42] Ming Li, Yu Wang, Chen Sun, and Hao Zhang. 2023. IFNet: Interaction Fusion Network for Change Detection of Remote Sensing Images. *Remote Sensing* 15, 7 (2023), 1760.

[43] Ming Li, Yu Wang, Chen Sun, and Hao Zhang. 2023. IFNet: Interaction Fusion Network for Change Detection of Remote Sensing Images. *Remote Sensing* 15, 7 (2023), 1760.

[44] Qiang Li, Yang Zhao, Bin Wang, and Qian Du. 2023. HSI-ChangeNet: A Deep Learning Framework for Hyperspectral Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[45] Wei Li, Qiang Wu, Yuxin Zhang, Qian Du, and Antonio Plaza. 2023. Spectral–Spatial–Temporal Transformer Network for Hyperspectral Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[46] Wei Li, Qiang Wu, Yuxin Zhang, Qian Du, and Antonio Plaza. 2023. Spectral–Spatial–Temporal Transformer Network for Hyperspectral Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.

[47] Shiau Hong Lim, Huan Xu, and Shie Mannor. 2013. Reinforcement learning in robust Markov decision processes. In *Advances in Neural Information Processing Systems*. 701–709.

[48] Michael L. Littman. 1994. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*. 157–163.

[49] Hongwei Liu, Qian Liu, Bin Wang, and Jian Sun. 2023. MTCDNet: Multitemporal Change Detection Network for Remote Sensing Images. *Remote Sensing* 15, 2 (2023), 389.

[50] Hongwei Liu, Qian Liu, Bin Wang, and Jian Sun. 2023. MTCDNet: Multitemporal Change Detection Network for Remote Sensing Images. *Remote Sensing* 15, 2 (2023), 389.

[51] Jian Liu, Jie Wu, Min Zhang, and Xu Wang. 2023. CRN: Change Residual Network for Remote Sensing Image Change Detection. *Remote Sensing* 15, 10 (2023), 2678.

[52] Ruisheng Liu, Xuefeng Jia, Lei Zhang, Bing Wang, and Yongqiang Zhang. 2021. DTCDSCN: Dual-Task Constrained Deep Siamese Convolutional Network for

Change Detection in Optical Remote Sensing Imagery. *IEEE Transactions on Geoscience and Remote Sensing* 59, 8 (2021), 6555–6567.
[53] W. Liu, K. Leahy, Z. Serlin, and C. Belta. 2023. Robust multi-agent coordination from CATL+ specifications. In *Proceedings of the American Control Conference (ACC)*. IEEE, 3529–3534.
[54] Wei Liu, Qiang Zhang, Qian Du, and Bin Wang. 2023. Multiscale Feature Aggregation Network for Change Detection in Remote Sensing Images. *Remote Sensing* 15, 6 (2023), 1520.
[55] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6379–6390.
[56] Xuewei Luo, Changqing Li, Shunping Song, and Gui-Song Xia. 2022. FCCDN: Feature Constraint Network for VHR Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–13.
[57] Xiangyu Ma, Ming Guo, Shaojun Zhou, Lingfei Zhang, Hao Yu, Zhen Zhang, and Yanning Wang. 2024. DRNCD: Dual Residual Network With CBAM for Remote Sensing Image Change Detection. In *2024 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 5983–5986.
[58] J. G. Martin, F. J. Muros, J. M. Maestre, and E. F. Camacho. 2023. Multi-robot task allocation clustering based on game theory. *Robotics and Autonomous Systems* 161 (2023), 104314.
[59] L. Marzari et al. 2021. Towards Hierarchical Task Decomposition Using Deep Reinforcement Learning for Pick and Place Subtasks. In *Proc. 20th International Conference on Advanced Robotics (ICAR)*. 640–645.
[60] J. Marín et al. 2021. Recipe1M+: A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2021), 187–203.
[61] Raghav Murthy, Sebastian Heinecke, Juan Carlos Niebles, Zhi Liu, Ling Xue, Wei Yao, Yuchen Feng, Zhi Chen, Anil Gokul, Devansh Arpit, et al. 2023. ReX: Rapid Exploration and Exploitation for AI Agents. *arXiv preprint arXiv:2307.08962* (2023).
[62] Arnab Nilim and Laurent El Ghaoui. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* 53, 5 (2005), 780–798.
[63] Marina Papadomanolaki, Maria Vakalopoulou, Sergey Zagoruyko, and Konstantinos Karantzalos. 2019. Siamese Attention Networks for Change Detection in Multispectral Images. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 211–214.
[64] Dezhong Peng, Yi Zhang, Hao Guan, Xingming Zhang, and Qian Wang. 2022. Change Detection in Remote Sensing Images Using Conditional Generative Adversarial Networks. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–15.
[65] Lerrel Pinto, James Davidson, Rahul Sukthankar, and Abhinav Gupta. 2017. Robust adversarial reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. JMLR.org, 2817–2826.
[66] Akshat Prasad, Alexander Koller, Michael Hartmann, Peter Clark, Ashish Sabharwal, Mohit Bansal, and Tushar Khot. 2024. ADaPT: As-Needed Decomposition and Planning with Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 4226–4252.
[67] N. Shinn et al. 2024. Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in Neural Information Processing Systems* 36 (2024).
[68] Nicholas Shinn, Francesco Cassano, Arjun Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in Neural Information Processing Systems* 36 (2024).
[69] Mohit Shridhar, Xiaodan Yuan, M.-A. Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations*.
[70] Zichen Song, Weijia Li, Chao Liu, and Yike Guo. 2022. SnuNet-CD: A Densely Connected Siamese Network for Change Detection of Remote Sensing Images. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2076–2079.
[71] S. S. O. Venkata, R. Parasuraman, and R. Pidaparti. 2023. Kt-bt: A framework for knowledge transfer through behavior trees in multirobot systems. *IEEE Transactions on Robotics* (2023).
[72] Chunlei Wang, Zhanyu Guo, Qiang Li, Jinjun Wang, Jie Gao, and Zhenhua Guo. 2024. WACDN: Weighted Aggregation Change Detection Network for Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), 1–12.
[73] Chao Wang, Dong Xu, Fan Zhou, and Xiaoxiang Zhu. 2023. CDNet: A Robust Change Detection Network for Remote Sensing Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 7455–7464.
[74] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. 2013. Robust Markov decision processes. *Mathematics of Operations Research* 38, 1 (2013), 153–183.
[75] Yue Wu, Jing Liu, Ling Feng, and Shuang Zhang. 2023. Cross-Temporal Attention Network for Change Detection in Remote Sensing Images. *Remote Sensing* 15, 8 (2023), 2112.
[76] Yue Wu, Jing Liu, Ling Feng, and Shuang Zhang. 2023. Cross-Temporal Attention Network for Change Detection in Remote Sensing Images. In *Remote Sensing*, Vol. 15. MDPI, 2112.
[77] Tao Xu, Zhen Zhang, Zhi Liu, and Jian Wang. 2023. Siamese Pyramid Network for Change Detection of Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16 (2023), 4212–4225.
[78] Tao Xu, Zhen Zhang, Zhi Liu, and Jian Wang. 2023. Siamese Pyramid Network for Change Detection of Remote Sensing Images. In *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 16. IEEE, 4212–4225.
[79] Zhiyu Yang, Peng Qi, Shuai Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In *Proc. 2018 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 2369–2380.
[80] Shunyu Yao, Hao Chen, Jian Yang, and Karthik Narasimhan. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. *Advances in Neural Information Processing Systems* 35 (2022), 20744–20757.
[81] Shunyu Yao, Jia Zhao, Dong Yu, Ning Du, Itamar Shafran, Karthik Narasimhan, and Yining Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proc. International Conference on Learning Representations (ICLR)*.
[82] Wei Yao, Sebastian Heinecke, Juan Carlos Niebles, Zhi Liu, Yuchen Feng, Ling Xue, R. Rithesh, Zhi Chen, J. Zhang, Devansh Arpit, et al. 2023. Retroformer: Retrospective Large Language Agents with Policy Gradient Optimization. In *12th International Conference on Learning Representations (ICLR)*.
[83] Tongtong Zhan, Zhen Chen, Hao Xu, Xiuqin Tang, Qiang Wang, Xinyang Li, and Wenzhong Shi. 2023. SwinUnet-CD: A Swin UNet-Based Remote Sensing Change Detection Method. *Remote Sensing* 15, 2 (2023), 319.
[84] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*. 5867–5876.
[85] Lei Zhang, Hao Chen, Liangpei Wang, and Licheng Jiao. 2023. MSTNet: Multiscale Transformer Network for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.
[86] Lei Zhang, Lei Ding, Ruisheng Liu, and Bo Du. 2023. DSIFN: Deep Siamese Interaction Fusion Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.
[87] Lei Zhang, Lei Ding, Ruisheng Liu, and Bo Du. 2023. DSIFN: Deep Siamese Interaction Fusion Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.
[88] Peng Zhang, Yuan Liu, Yu Jin, and Zhiyong Chen. 2023. SiamCRNN: A Siamese Convolutional Recurrent Neural Network for Change Detection of Remote Sensing Images. *Remote Sensing* 15, 5 (2023), 1234.
[89] Yunchao Zhang, Zhiqiang Ma, Weidi Xie, Yiqun Xu, Liang Chen, Gang Sun, Xiao Li, and Hong Tang. 2022. TinyCD: An Efficient Change Detection Network for Remote Sensing Images. *Remote Sensing* 14, 9 (2022), 2182.
[90] Yunchao Zhang, Zhiqiang Ma, Weidi Xie, Yiqun Xu, Gang Sun, Xiao Li, and Hong Tang. 2022. Space-Time-Spectral Attention Neural Network for Change Detection in Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–14.
[91] Yifan Zhang, Chuang Zhang, Haopeng Zhang, Junsong Li, Jie Zhang, and Guojin Wang. 2023. MSPSNet: Multiscale Pyramid Siamese Network for Change Detection of Remote Sensing Images. *Remote Sensing* 15, 3 (2023), 633.
[92] Zizheng Zhang, Bing Wang, Lei Zhang, and Bo Du. 2023. P2V-CD: A Plug-and-Play Framework for Remote Sensing Image Change Detection. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–13.
[93] Aohan Zhao, Da Huang, Qiang Xu, Min Lin, Y.-J. Liu, and Guodong Huang. 2024. EXPEL: LLM Agents Are Experiential Learners. In *Proc. AAAI Conference on Artificial Intelligence*, Vol. 38. 19632–19642.
[94] Wei Zhao, Tao Liu, Ming Zhang, and Jun Li. 2023. MCFNet: Multi-Context Fusion Network for Change Detection in Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–12.
[95] Y. Zhen et al. 2023. Robot Task Planning Based on Large Language Model Representing Knowledge with Directed Graph Structures. Preprint.
[96] Peng Zhou, Zhen Luo, Jian Wu, and Hua Tang. 2023. CGNet: Context-Guided Network for Remote Sensing Image Change Detection. *IEEE Geoscience and Remote Sensing Letters* 20 (2023), 1–5.
[97] Peng Zhou, Zhen Luo, Jian Wu, and Hua Tang. 2023. CGNet: Context-Guided Network for Remote Sensing Image Change Detection. *IEEE Geoscience and Remote Sensing Letters* 20 (2023), 1–5.
[98] Xiang Zhou, Yong Li, Hong Wang, and Zhi Chen. 2023. Change Detection in Remote Sensing Images Using Transformer-Based Models. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 1452–1455.