Scientific Reports

https://doi.org/10.1038/s41598-026-37191-w

Article in Press

End-to-end emergency response protocol for tunnel accidents augmentation with reinforcement learning

Received: 23 July 2025

Accepted: 20 January 2026

Published online: 26 January 2026

Cite this article as: ur Rehman H.M.R., Gul M.J., Younas R. *et al.* End-to-end emergency response protocol for tunnel accidents augmentation with reinforcement learning. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-37191-w

Hafiz Muhammad Raza ur Rehman, M. Junaid Gul, Rabbiya Younas, Muhammad Zeeshan Jhandir, Roberto Marcelo Alvarez, Yini Miro & Imran Ashraf

End-to-end Emergency Response Protocol for Tunnel Accidents Augmentation with Reinforcement Learning

Hafiz Muhammad Raza ur Rehman[1], M. Junaid Gul[1], Rabbiya Younas[1], Muhammad Zeeshan Jhandir[2], Roberto Marcelo Alvarez[3,4,5], Yini Miro[3,6,7], Imran Ashraf[1*]

[1]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan, 38541, Republic of Korea.
[2]Department of Data Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.
[3]Universidad Europea del Atlantico, Isabel Torres 21, Santander, 39011, Spain.
[4]Universidad Internacional Iberoamericana Arecibo, Puerto Rico, 00613, USA.
[5]Universidade Internacional do Cuanza, Cuito, Angola.
[6]Universidad Internacional Iberoamericana, Campeche, 24560, Mexico.
[7]Universidad de La Romana, La Romana, Republica Dominicana.

*Corresponding author(s). E-mail(s): imranashraf@ynu.ac.kr;
Contributing authors: mrazaurrehman@ynu.ac.kr; drmalik@yu.ac.kr; younas.rabbiya@yu.ac.kr; zeeshan.jhandir@iub.edu.pk; roberto.alvarez@uneatlantico.es ; yini.miro@uneatlantico.es;

Abstract

Autonomous unmanned aerial vehicles (UAVs) offer cost-effective and flexible solutions for a wide range of real-world applications, particularly in hazardous and time-critical environments. Their ability to navigate autonomously, communicate rapidly, and avoid collisions makes UAVs well suited for emergency response scenarios. However, real-time path planning in dynamic and unpredictable environments remains a major challenge, especially in confined tunnel infrastructures where accidents may trigger fires, smoke propagation, debris, and rapid environmental changes. In such conditions, conventional preplanned

or model-based navigation approaches often fail due to limited visibility, narrow passages, and the absence of reliable localization signals. To address these challenges, this work proposes an end-to-end emergency response framework for tunnel accidents based on Multi-Agent Reinforcement Learning (MARL). Each UAV operates as an independent learning agent using an Independent Q-Learning paradigm, enabling real-time decision-making under limited computational resources. To mitigate premature convergence and local optima during exploration, Grey Wolf Optimization (GWO) is integrated as a policy-guidance mechanism within the reinforcement learning (RL) framework. A customized reward function is designed to prioritize victim discovery, penalize unsafe behavior, and explicitly discourage redundant exploration among agents. The proposed approach is evaluated using a frontier-based exploration simulator under both single-agent and multi-agent settings with multiple goals. Extensive simulation results demonstrate that the proposed framework achieves faster goal discovery, improved map coverage, and reduced rescue time compared to state-of-the-art GWO-based exploration and random search algorithms. These results highlight the effectiveness of lightweight MARL-based coordination for autonomous UAV-assisted tunnel emergency response.

Keywords: Robotic systems; drones; multi-agents system; path finding; reinforcement learning; tunnel hazards; unmanned aerial vehicles

1 Introduction

Over the past decades, a wide range of natural and man-made disasters, including earthquakes, floods, explosions, and large-scale fires, have caused severe loss of human life and critical infrastructure. Such events frequently lead to collapsed buildings and damaged tunnel systems, trapping victims beneath debris and creating extremely hazardous conditions for emergency response teams. Rapid and effective search-and-rescue (SAR) operations are therefore essential; however, conventional response methods are often constrained by structural instability, toxic environments, and limited accessibility. In this context, UAVs have emerged as a promising technological solution for enhancing SAR operations and improving disaster response efficiency.

Tunnels and underground facilities have been widely developed to support modern transportation and urban infrastructure. Despite their economic and logistical importance, tunnels are fully enclosed environments, which significantly increase risk during emergency situations. In the event of a tunnel fire, trapped individuals often have limited escape routes, resulting in a high likelihood of casualties. For example, a tunnel fire in the Shanxi Yanhou Tunnel in 2014 resulted in 40 fatalities and 12 injuries [1]. Such incidents highlight the critical importance of tunnel fire safety management and efficient emergency response mechanisms.

Tunnel environments pose multiple safety hazards, including fire, smoke propagation, structural collapse, electrical failures, and the presence of hazardous materials. Among these, smoke inhalation remains one of the leading causes of fatalities in tunnel accidents. Fires in confined spaces can rapidly generate dense smoke, severely

reducing visibility and causing asphyxiation. Furthermore, toxic gases released during combustion can induce disorientation and respiratory distress, significantly worsening survival prospects for trapped individuals.

The enclosed nature of tunnels also complicates firefighting and rescue efforts. Limited ventilation accelerates heat accumulation and smoke spread, while narrow passages and structural damage restrict responder mobility. Emergency personnel must often operate under extreme uncertainty, with incomplete situational awareness and rapidly evolving conditions.

Several real-world incidents further illustrate these challenges. In July 2018, a fire occurred in the Tianjin Binhai Tunnel in China due to the ignition of flammable goods transported by a truck, resulting in injuries to tunnel users and firefighters [2]. Similarly, the Sasago Tunnel ceiling collapse in Japan on December 2, 2012, resulted in multiple vehicles being crushed and at least nine fatalities, highlighting the catastrophic consequences of tunnel structural failures and the challenges faced by emergency responders in such confined spaces [3]. These cases emphasize the need for robust safety protocols, continuous infrastructure monitoring, and intelligent emergency response systems capable of operating in hazardous tunnel environments.

Effective tunnel emergency response requires responders to navigate complex, maze-like infrastructures under conditions of low visibility and dynamic obstruction. Traditional path-planning methods often struggle in such scenarios due to smoke, debris, and partial structural collapse. Additionally, satellite-based navigation systems such as GPS are unreliable or unavailable in underground environments, further complicating localization and routing tasks.

Most existing navigation and rescue approaches rely heavily on accurate environmental models or prior knowledge of tunnel layouts [4, 5]. In real-world emergencies, however, such information is frequently incomplete, outdated, or entirely unavailable. Reinforcement learning (RL) offers a viable alternative by enabling autonomous agents to adapt their behavior through interaction with the environment rather than relying on predefined models [6]. RL has been extensively applied to UAV control and robotic navigation tasks [7, 8], including trajectory tracking, path following, and disturbance mitigation. For example, RL-based frameworks have been proposed for UAV motion planning with suspended loads [9], stable trajectory generation [10], adaptive PID control [11], and disturbance compensation in complex airflow conditions [12]. Cooperative UAV path-planning approaches, such as Dubins-based methods [13], have also been explored, although they often struggle with rapid local environmental changes.

Despite these advances, the application of RL to autonomous UAV-based disaster response particularly for mission planning, victim search, and cooperative exploration in confined tunnel environments remains relatively underexplored [14, 15]. Moreover, many existing approaches rely on computationally intensive models, explicit inter-agent communication, or centralized learning architectures, which may limit their applicability in time-critical rescue operations.

Motivated by these challenges, this work focuses on developing a lightweight, adaptive, and cooperative multi-agent learning framework tailored for tunnel emergency response. The proposed approach emphasizes real-time feasibility, efficient exploration, and safety-aware decision-making under partial observability. In the proposed scheme

GWO has been used with RL as policy despite others GWO based RL algorithms where they used GWO for managing exploration and exploitation ratio [16].

The main contributions of this work are summarized as follows:

- Development of a MARL framework for autonomous tunnel emergency response and victim search.
- Adoption of an IQL paradigm to enable real-time decision-making under limited computational resources.
- Integration of frontier-based exploration with graph-based path planning to efficiently navigate partially known environments.
- Design of a reward mechanism that discourages redundant exploration, penalizes unsafe behavior, and prioritizes victim discovery.
- Formulation of observable and hidden state representations to address partial observability in cooperative multi-agent settings.
- Comprehensive simulation-based evaluation demonstrating improved exploration efficiency, rescue time, and collision avoidance compared to baseline methods.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the theoretical foundations and background concepts. Section 4 presents the proposed methodology. Section 5 discusses the experimental setup and performance evaluation. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Literature Review

Deep reinforcement learning (DRL) has drawn significant attention among researchers in Unmanned Aerial Vehicle (UAV) systems, as it addresses the growing need for autonomous aerial vehicles capable of executing complex tasks in dynamic and uncertain environments. Recent literature explores how DRL enhances UAV guidance, navigation, and control (GNC), particularly in unpredictable or GPS-denied scenarios.

For instance, [17] presents an asynchronous deep deterministic policy gradient (ADDPG) method for mapless navigation with mobile robots in challenging environments, demonstrating the applicability of RL to autonomous navigation tasks. In another study, [18] proposes a technique that integrates external memory, enabling neural network models to perform mapping, localization, and navigation decision-making within a unified framework. This configuration allows simultaneous position estimation and map construction alongside continuous control.

For continuous control in autonomous navigation, [19] utilizes sparse LiDAR inputs and relative target locations within a DRL framework, resulting in improved path-planning efficiency and robustness. Moreover, [20] introduces an integrated communication and control architecture based on DDPG for UAV swarm formation management, enabling enhanced control precision and collision avoidance. Despite these advancements, [21] report that DRL models face challenges related to generalization, safety, training stability, and computational overhead, which hinder their deployment in real-world, safety-critical environments.

While general-purpose UAVs demonstrate strong DRL-enabled navigation and control capabilities, deploying them in mission-critical operations such as search and rescue (SAR) introduces additional challenges. SAR missions often involve cluttered, GPS-denied environments and strict time constraints, requiring UAVs to exhibit high levels of autonomy, reliability, and adaptability. Consequently, recent research has focused on UAV systems tailored specifically for SAR applications.

In this context, [22] propose a UAV-based SAR framework that leverages received signal strength (RSS) measurements and a Q-learning-based strategy to detect indoor victims. Their results show that directional antennas improve convergence speed and localization accuracy compared to omnidirectional antennas. Similarly, Donnelly et al. [23] model UAV-based SAR using partially observable Markov decision processes (POMDPs) and deep Q-networks (DQNs), demonstrating improved performance over heuristic methods in complex environments. However, such approaches typically rely on deep learning architectures and centralized training, which may limit real-time applicability.

Maritime UAV-based SAR missions pose further challenges due to large operational areas and rapidly changing conditions. To address this, Wu et al. [24] propose a hybrid genetic algorithm and RL (GA-RL) approach for path planning, embedding Q-learning into evolutionary optimization. Their method achieves improved convergence and solution quality compared to standard optimization techniques. For wilderness SAR, Bhattacharya et al. [25] develop a modular DRL framework for 3D UAV navigation and person detection using curriculum learning, achieving high accuracy in both semi-autonomous and guided navigation tasks.

In multi-agent UAV systems, Wang et al. [26] present a Q-learning-based 3D deployment framework that enables multiple UAVs to dynamically reposition for optimal coverage, outperforming traditional clustering approaches. Nevertheless, many existing multi-agent studies primarily focus on coverage or communication efficiency rather than rescue prioritization or redundant exploration avoidance.

Recent studies have also explored heterogeneous and cooperative multi-agent systems for dynamic monitoring and patrolling tasks. For example, the UAV–UGV cooperative framework presented in [27] investigates coordinated patrolling and energy management in urban environments, demonstrating how task allocation and inter-agent cooperation can improve system endurance and coverage. While effective, such approaches typically rely on explicit coordination strategies and stable communication, which may be difficult to guarantee in confined tunnel environments affected by smoke, fire, or structural damage.

UAV path planning under obstacle-rich environments has also been explored using bio-inspired optimization techniques. Ant Colony Optimization (ACO), inspired by collective ant foraging behavior, has been widely applied to robotic and UAV path planning [28–31]. These methods have been shown to generate collision-free trajectories and optimize routes under various constraints. However, their performance is often sensitive to parameter tuning and may degrade in highly dynamic or partially observable environments.

Classical robotic exploration techniques provide important foundations for autonomous navigation. Frontier-based exploration, introduced by Yamauchi [32],

enables robots to expand their knowledge of unknown environments by targeting boundaries between explored and unexplored regions. Probabilistic mapping and SLAM-based approaches [33] further improve navigation by maintaining belief distributions over the environment. These methods, however, are primarily designed for single-agent settings and lack adaptive coordination mechanisms for cooperative rescue scenarios.

RL-based exploration has traditionally relied on Q-learning and policy gradient methods. Q-learning, introduced by Watkins and Dayan [34], provides a model-free mechanism for learning optimal actions in discrete state-action spaces. Policy gradient methods [35] extend learning to continuous action spaces but generally require higher computational resources. In multi-agent contexts, cooperative exploration strategies have been investigated by Cao et al. [36], demonstrating improved coverage efficiency, though without explicit consideration of real-time rescue constraints or victim prioritization.

In time-critical disaster scenarios, balancing exploration and exploitation becomes particularly important. To address this challenge, recent work has proposed infrastructure-assisted learning frameworks that leverage edge intelligence and communication-aware optimization. For instance, [37] integrates UAV-mounted reconfigurable intelligent surfaces (RIS) and high-altitude platforms (HAPs) to optimize disaster response under strict latency constraints. While such approaches improve exploration efficiency, they require sophisticated communication infrastructure and centralized coordination, limiting their applicability in underground or tunnel-based rescue operations. Scalability in large-scale systems has also been addressed using deep reinforcement learning in communication-centric domains. For example, [38] employs a DRL-based relaying election mechanism to improve energy efficiency in large IoT networks. Although DRL-based solutions offer scalability and performance benefits, they typically require extensive training data, powerful computational resources, and centralized training paradigms, which may not be feasible for real-time emergency response in tunnel environments.

Overall, the literature highlights significant progress in UAV navigation, RL, and multi-agent coordination. However, most existing approaches rely on deep or centralized learning architectures, explicit communication strategies, or computationally intensive optimization methods. Limited attention has been given to lightweight MARL frameworks that operate under real-time constraints, minimize redundant exploration, and ensure safety in confined tunnel environments. These limitations motivate the proposed IQL-based multi-agent framework, which emphasizes computational efficiency, implicit coordination through reward design, and practical applicability for tunnel emergency response.

3 Preliminaries

This section introduces the fundamental concepts and mathematical tools required to understand the proposed multi-agent rescue framework. Specifically, we review graph-based shortest-path planning, artificial potential fields for collision avoidance, and the RL foundations underpinning the IQL paradigm adopted in this work.

3.1 Graph-Based Shortest Path Planning

Graph-based path planning is widely used in robotic navigation to compute collision-free and efficient routes in structured environments. In the context of tunnel rescue, the environment is represented as a weighted graph, where nodes correspond to discrete spatial locations and edges denote traversable connections between them [39].

Let $G = (V, E)$ be a weighted graph, where V is the set of vertices and E is the set of edges. Each edge $(u, v) \in E$ is associated with a non-negative weight $w(u, v)$ representing traversal cost.

Given a source node $s \in V$, Dijkstra's algorithm computes the shortest path distance from s to all other nodes in V by iteratively expanding the closest unvisited node and relaxing adjacent edges. The algorithm is formally defined as:

$$\text{Dijkstra}(G, s) = \{d(v) \mid v \in V\}, \tag{1}$$

where $d(v)$ denotes the minimum cumulative cost from s to node v.

In this work, shortest-path computation is used to guide agents toward frontier cells during exploration, enabling efficient navigation through partially explored tunnel environments.

3.2 Artificial Potential Fields for Collision Avoidance

Artificial Potential Fields (APFs) are a classical motion planning technique used to generate collision-free trajectories by modeling the environment as a combination of attractive and repulsive forces. Goals exert attractive forces, while obstacles and other agents exert repulsive forces[40].

Let $\mathbf{p} = (x, y)$ denote the current position of an agent and $\mathbf{p}_{goal}$ the target position. The total potential field is defined as:

$$U(\mathbf{p}) = U_{att}(\mathbf{p}) + U_{rep}(\mathbf{p}), \tag{2}$$

where the attractive potential is given by:

$$U_{att}(\mathbf{p}) = \frac{1}{2} k_{att} \|\mathbf{p} - \mathbf{p}_{goal}\|^2, \tag{3}$$

and the repulsive potential generated by obstacles is defined as:

$$U_{rep}(\mathbf{p}) = \sum_{i=1}^{N_{obs}} \begin{cases} \frac{1}{2} k_{rep} \left(\frac{1}{\|\mathbf{p} - \mathbf{p}_{obs_i}\|} - \frac{1}{r_{obs}} \right)^2, & \text{if } \|\mathbf{p} - \mathbf{p}_{obs_i}\| < r_{obs}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Here, $\mathbf{p}_{obs_i}$ denotes the position of the i-th obstacle, r_{obs} is the influence radius, and k_{att}, k_{rep} are scaling constants.

In the proposed framework, collision avoidance is not enforced explicitly through force-based control but is incorporated implicitly through reward penalties inspired by APF principles.

3.3 Markov Decision Process Formulation

RL problems are commonly modeled using a Markov Decision Process (MDP), defined by the tuple:

$$\langle S, A, P, R, \gamma \rangle,$$

where:

- S is the set of states,
- A is the set of actions,
- $P(s'|s,a)$ is the state transition probability,
- $R(s,a,s')$ is the reward function,
- $\gamma \in [0,1]$ is the discount factor.

At each time step t, an agent observes state s_t, executes action a_t, receives reward R_{t+1}, and transitions to state s_{t+1}.

3.4 Value Functions and Bellman Equations

The state-value function under policy π is defined as:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid S_0 = s\right]. \tag{5}$$

Similarly, the action-value function is given by:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid S_0 = s, A_0 = a\right]. \tag{6}$$

The optimal action-value function satisfies the Bellman optimality equation:

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right]. \tag{7}$$

3.5 Q-Learning

Q-learning is a model-free RL algorithm that iteratively approximates the optimal action-value function $Q^*(s,a)$ without requiring prior knowledge of transition probabilities[41]. The update rule is defined as:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[R_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\right], \tag{8}$$

where $\alpha \in (0,1]$ is the learning rate.

3.6 IQL in Multi-Agent Systems

In IQL, each agent maintains its own Q-table and learns independently by treating other agents as part of the environment. Although this introduces non-stationarity, IQL remains computationally lightweight and suitable for real-time applications.

In this work, IQL is adopted to ensure scalability and real-time feasibility in tunnel rescue scenarios, avoiding the high computational cost associated with DRL or centralized training paradigms.

3.7 Grey Wolf Optimizer

The GWO is a population-based metaheuristic inspired by the social hierarchy and cooperative hunting behavior of grey wolves. In GWO, candidate solutions are categorized into four hierarchical groups: alpha (α), beta (β), delta (δ), and omega (ω), where α, β, and δ represent the three best solutions guiding the search process, while ω represents the remaining candidates[42, 43].

Let $\mathbf{X}(t)$ denote the position of a search agent at iteration t, and let $\mathbf{X}_\alpha$, $\mathbf{X}_\beta$, and $\mathbf{X}_\delta$ denote the positions of the three best solutions. The position update mechanism in GWO is defined as:

$$\mathbf{D}_k = \left|\mathbf{C}_k \cdot \mathbf{X}_k - \mathbf{X}(t)\right|, \quad k \in \{\alpha, \beta, \delta\} \tag{9}$$

$$\mathbf{X}'_k = \mathbf{X}_k - \mathbf{A}_k \cdot \mathbf{D}_k \tag{10}$$

$$\mathbf{X}(t+1) = \frac{1}{3}\left(\mathbf{X}'_\alpha + \mathbf{X}'_\beta + \mathbf{X}'_\delta\right), \tag{11}$$

where $\mathbf{A}_k = 2a \cdot \mathbf{r}_1 - a$, $\mathbf{C}_k = 2 \cdot \mathbf{r}_2$, and $\mathbf{r}_1, \mathbf{r}_2 \in [0, 1]$ are random vectors. The parameter a decreases linearly from 2 to 0 over iterations, allowing a smooth transition from exploration to exploitation.

GWO has been widely adopted for path planning and optimization tasks due to its simplicity, fast convergence, and low computational overhead. However, standalone GWO methods are prone to premature convergence in complex or dynamic environments.

In this work, GWO is not employed as an independent optimizer. Instead, its exploration behavior is integrated with RL to guide policy exploration and mitigate local optima. This hybridization preserves the lightweight nature of tabular learning while improving search diversity in complex tunnel rescue environments.

4 Material & Methods

Figure 1 illustrates the overall scenario exploitation process of the proposed tunnel emergency response framework.

4.1 Frontier-Based Cooperative Exploration

A frontier-based exploration strategy is employed to enable efficient navigation and mapping of tunnel environments. Initially, the environment is represented as a discretized occupancy grid where all cells are marked as unexplored. As agents traverse the environment, onboard sensors continuously update the grid based on newly observed information.

Frontier cells, defined as the boundary between explored and unexplored regions, are selected as exploration targets. Agents compute collision-free paths toward these frontier cells using the A* algorithm while avoiding static and dynamic obstacles. Exploration continues until all reachable frontiers are exhausted or all victims are successfully located.

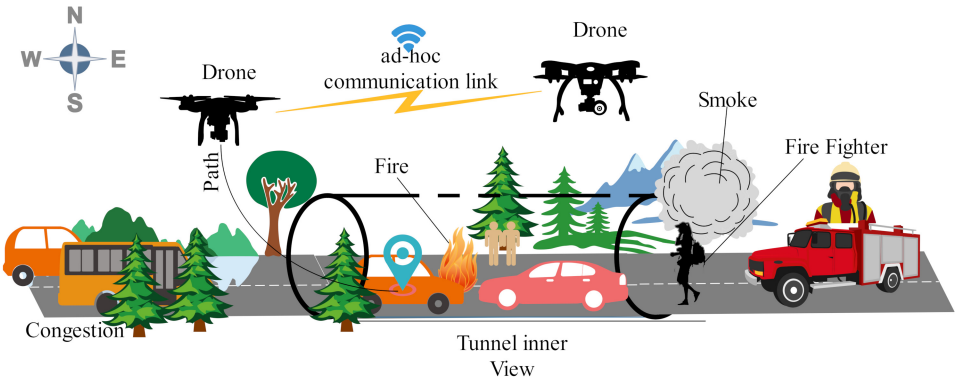


Fig. 1: Scenario exploitation pictorial explanation of the proposed multi-agent rescue system.

4.2 Multi-Agent Rescue System Overview

The proposed rescue system consists of multiple autonomous UAV/robot agents operating cooperatively in a shared tunnel environment. Each agent independently performs navigation, victim detection, obstacle avoidance, and map updating. Cooperation is achieved implicitly through shared environmental feedback rather than explicit inter-agent communication.

The performance of the system is evaluated using the following criteria:

- number of victims successfully rescued,
- coverage of the tunnel environment,
- avoidance of redundant exploration,
- collision-free navigation.

Accordingly, the global objectives are defined as:

$$\text{Maximize: } \omega_1 \rightarrow \text{number of rescued victims,} \tag{12}$$

$$\text{Minimize: } \omega_2 \rightarrow \text{total unexplored area.} \tag{13}$$

The corresponding utility function is formulated as:

$$u(\omega, x_{0:T}) = \arg\max\left(\omega_1 - \omega_2\right). \tag{14}$$

The proposed framework adopts an IQL paradigm, where each agent maintains its own Q-table and updates it independently based on local observations. This design choice is motivated by the strict computational and latency constraints of real-time tunnel rescue operations.

Unlike centralized training or DRL approaches, IQL avoids neural network inference, replay buffers, and extensive training requirements, making it suitable for

deployment in resource-constrained and time-critical environments. Potential non-stationarity associated with IQL is mitigated through reward shaping and structured frontier-based task decomposition[42].

The state space S consists of partially observable states s^1 and partially hidden states s^2. Observable states represent the agent's position:

$$s^1 = [c_x, c_y] \in \mathbb{R}^2, \tag{15}$$

while hidden states correspond to victim locations:

$$s^2 = [v_x, v_y] \in \mathbb{R}^2. \tag{16}$$

The complete state vector of an agent at time t is:

$$s_t = [s_t^1, s_t^2] \in \mathbb{R}^4. \tag{17}$$

Each agent p observes the environment using exteroceptive sensors. The observation space at time t is defined as:

$$o_t^p = [c_t^p, v_t^p, c_t^{(N-p)}] \in \mathbb{R}^{4+2(N-1)}, \tag{18}$$

where $c_t^{(N-p)}$ denotes the relative positions of other agents.

At each time step, an agent can perform one of nine discrete actions corresponding to grid-based movement, as shown in Figure 2. The action space is defined as:

$$a_t^p \in \mathbb{R}^2. \tag{19}$$

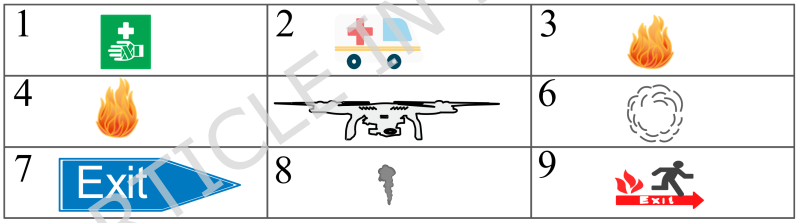


Fig. 2: Grid-based action space for agent movement.

To enable adaptation to dynamic tunnel conditions such as blocked passages, smoke, or fire spread, GWO is integrated into the exploration policy of RL. Rather than acting as a standalone optimizer, GWO biases action selection during exploration to prevent premature convergence.

This sustained exploration mechanism ensures that agents continue adapting their policies online as the environment evolves, allowing exploration to proceed until all victims are rescued and the operation is completed.

A customized reward function is designed to promote cooperative exploration, safety, and efficiency:

$$r_t^p = \begin{cases} +r_v, & \text{if a new victim is discovered,} \\ -r_d, & \text{if battery is depleted or agent fails,} \\ -\lambda, & \text{if an explored cell is revisited,} \\ -\kappa, & \text{if the next cell is occupied by another agent,} \\ -\eta, & \text{otherwise.} \end{cases} \tag{20}$$

The parameters are set as $r_v = 10$, $r_d = 20$, $\lambda = 1$, $\kappa = 3$, and $\eta = 1$.

The duplicate-exploration penalty discourages redundant exploration, while the collision penalty ensures safety during learning. Since agent positions are known in the centralized simulation environment, collisions are detected prior to action execution.

Each agent updates its Q-table using the standard Q-learning update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t^p + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]. \tag{21}$$

Action selection follows an ϵ-greedy strategy augmented with GWO-guided exploration.

Algorithm 1 GWO-Guided IQL for Multi-Agent Rescue

```
Initialize Q-tables Q_p(s, a) for all agents
Initialize learning rate α, discount factor γ
while rescue mission not completed do
  for each agent p do
    Observe state s_t^p
    Select action a_t^p using ε-greedy + GWO guidance
    Execute a_t^p and receive reward r_t^p
    Update Q-table using Q-learning rule
  end for
end while
```


The proposed framework relies on well-established convergence properties of Q-learning and frontier-based exploration. Reward shaping and shared environmental feedback promote stable cooperative behavior and reduce non-stationarity. These properties summarize known results rather than introducing new theoretical guarantees.

4.3 Property 1: Bounded Exploration and Exploitation

Statement: Under standard Q-learning conditions, the proposed IQL framework maintains a bounded balance between exploration and exploitation during the learning process, preventing premature convergence while ensuring policy improvement over time.

Explanation: Each agent updates its action-value function using the classical Q-learning update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t \left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right], \tag{22}$$

where α_t denotes the learning rate and γ is the discount factor.

When the learning rate satisfies the Robbins-Monro conditions,

$$\sum_{t=0}^{\infty} \alpha_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

and when state-action pairs are sufficiently explored through an ϵ-greedy policy, the Q-values are known to converge toward stable estimates in stationary environments.

In the proposed framework, exploration is further regulated through reward shaping and step penalties, which discourage excessive wandering while still allowing agents to explore unvisited regions. This results in a bounded exploration-exploitation trade-off that supports stable learning behavior without introducing additional computational complexity.

4.4 Property 2: Frontier Coverage Behavior

Statement: The integration of frontier-based exploration with RL leads to progressive reduction of unexplored regions while prioritizing victim discovery in partially observable environments.

Explanation: Let $E(t)$ and $U(t)$ denote the sets of explored and unexplored cells at time step t, respectively. Frontier cells are defined as:

$$F(t) = \{c \in U(t) \mid c \in \text{Boundary}(E(t))\}, \tag{23}$$

representing the interface between known and unknown regions.

By selecting actions that guide agents toward frontier cells, the exploration process incrementally expands $E(t)$ while reducing $U(t)$. In the proposed reward formulation, revisiting previously explored cells incurs a penalty, which discourages redundant exploration and promotes efficient coverage of new areas.

As exploration progresses, the number of frontier cells naturally decreases:

$$|F(t)| \to 0 \quad \text{as} \quad t \to \infty,$$

indicating saturation of the reachable environment. Simultaneously, victim discovery events are reinforced through positive rewards, ensuring that exploration remains goal-directed rather than purely spatial. This behavior supports systematic coverage without requiring explicit global coordination.

4.5 Property 3: Cooperative Utility Improvement

Statement: The collective utility of the multi-agent system improves as individual agents learn policies that are shaped by shared environmental feedback and complementary exploration behaviors.

Explanation: Let r_t^p denote the reward received by agent p at time t, and define the cumulative system-level reward as:

$$R_{\text{total}}(t) = \sum_{p=1}^{C} r_t^p, \tag{24}$$

where C is the number of agents.

Although each agent maintains an independent Q-table, coordination emerges implicitly through shared environmental states and reward signals, such as penalties for collisions and redundant exploration. As agents learn to avoid overlapping paths and unsafe actions, the collective reward accumulated over an episode increases.

The expected utility over a finite horizon T can be expressed as:

$$\mathbb{E}[U] = \mathbb{E}\left[\sum_{t=0}^{T} R_{\text{total}}(t)\right], \tag{25}$$

which improves as agents adopt policies that balance individual objectives with system-level efficiency. This property reflects cooperative behavior emerging from decentralized learning rather than guaranteeing global optimality, making it suitable for real-time rescue operations.

5 Performance Evaluation

This section presents an extensive evaluation of the proposed GWO-guided IQL framework in both single-agent and multi-agent tunnel rescue scenarios. The evaluation focuses on exploration efficiency, rescue effectiveness, safety, and execution time under complex and constrained environments. To ensure fair and reliable comparisons, all experiments are conducted under identical environmental settings and averaged over 20 independent simulation runs.

Two representative environments are considered: (i) a maze environment with a single agent and a single rescue goal, and (ii) a road-map maze environment with multiple agents and multiple rescue goals. These environments emulate realistic tunnel accident conditions characterized by limited visibility, narrow pathways, and dynamically distributed victims.

5.1 Evaluation Environments

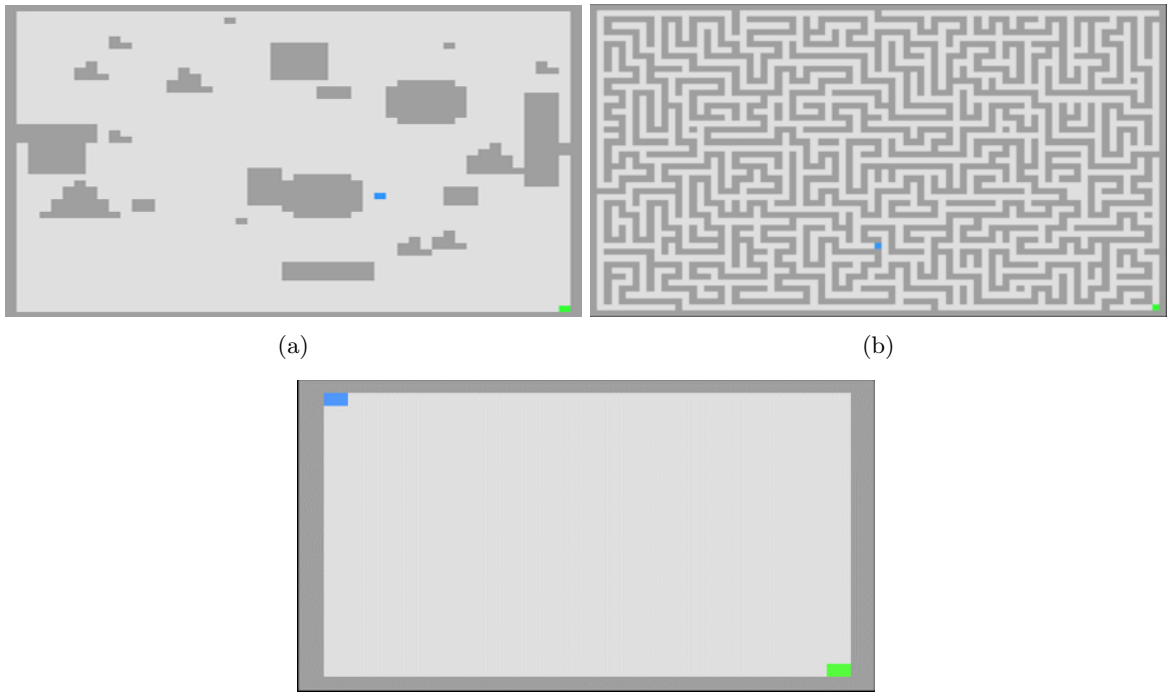
Single-Agent, Single-Goal Maze Environment

To evaluate baseline navigation and exploration capability, three different maze configurations with varying obstacle densities, corridor structures, and goal locations are

employed. These environments are illustrated in Figure 3. In each configuration, the agent is represented by a blue box, while the rescue target (victim) is shown as a green box.

The environments include multiple dead ends and narrow passages, posing challenges for exploration strategies that suffer from premature convergence or inefficient search behavior. The agent is equipped with onboard sensors to perceive nearby obstacles and free space, enabling partial observability similar to real tunnel conditions.

The agent follows a Q-learning-based policy to balance exploration and exploitation. Successful victim discovery yields a positive reward, while collisions with obstacles incur penalties. Additionally, a step penalty of -1 is applied at each time step to discourage unnecessary movement and promote efficient navigation. This reward structure ensures comprehensive environment coverage while prioritizing timely victim rescue and collision avoidance.



(a) (b)

(c)

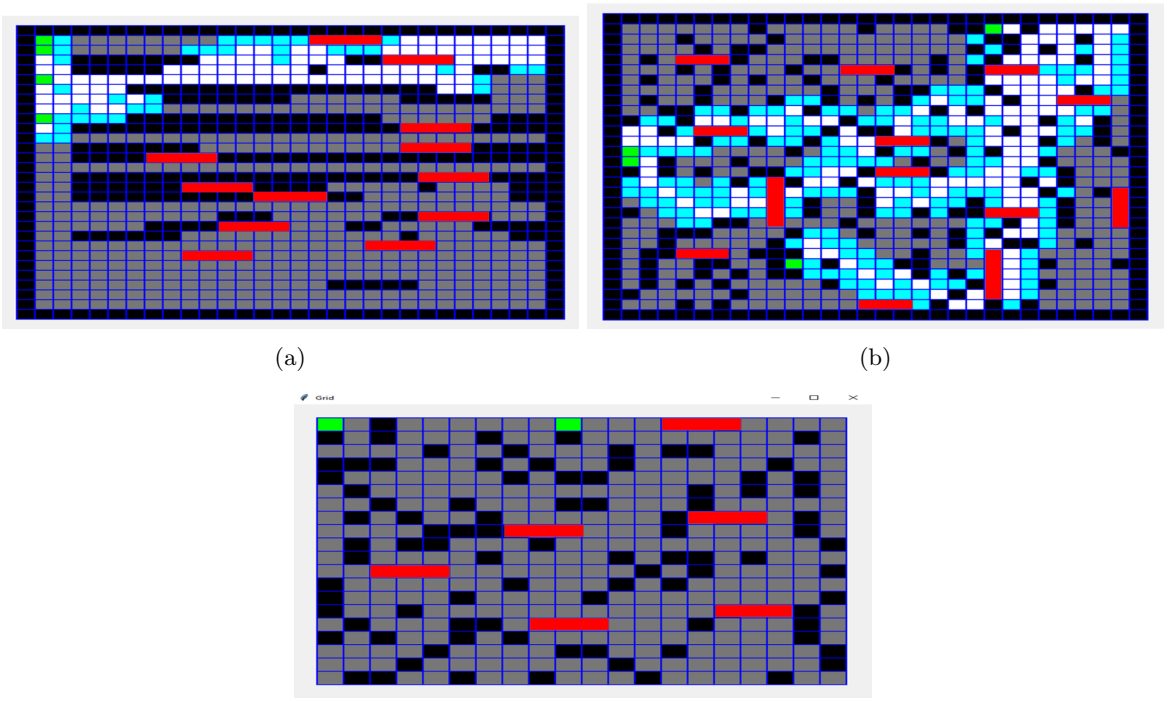
Fig. 3: Single-agent exploration in three maze environments with one rescue target.

Multi-Agent, Multi-Goal Road-Map Maze Environment

To assess cooperative behavior and scalability, three road-map maze environments with varying sizes, obstacle distributions, and victim locations are used. These environments are shown in Figure 4. Victims are depicted in red, obstacles in black, explored regions in white, frontier regions in light blue, and agents in green.

Multiple agents operate simultaneously to locate and rescue all victims. Each agent employs the IQL paradigm with an individual Q-table, while coordination is achieved implicitly through shared environment updates in a centralized simulation framework. This setting reflects practical rescue operations where a command center maintains global situational awareness while individual agents act autonomously.

Positive rewards are assigned for successful victim rescues, while penalties are applied for collisions, redundant exploration, and inefficient movements. A step penalty of -1 further encourages agents to minimize rescue time. This environment provides a comprehensive testbed for evaluating cooperative efficiency, safety during learning, and robustness in complex tunnel rescue scenarios.



(a) (b)

(c)

Fig. 4: Multi-agent exploration in road-map maze environments with multiple rescue targets.

The proposed framework is compared against the following baseline methods:

- Random Search,
- Utility-Based Cooperative Exploration (UCE),
- Cooperative Multi-Agent Exploration (CME),
- GWO-based exploration.

All algorithms are evaluated under identical environmental conditions, agent counts, and termination criteria. Each experiment is repeated 20 times to mitigate randomness and ensure statistical reliability. Performance is measured in terms of explored area, number of iterations to achieve rescue goals, total execution time, and collision avoidance.

5.2 Single-Agent Performance Analysis

Figure 5 illustrates the comparative performance of the proposed framework and baseline methods in single-goal environments. The UCE approach exhibits the lowest exploration efficiency, requiring a significantly higher number of steps to reach the goal. This behavior is primarily attributed to its lack of adaptive exploration mechanisms.

The GWO-based approach demonstrates improved exploration compared to Random Search but still suffers from premature convergence in certain maze configurations. In contrast, the proposed GWO-guided IQL framework consistently achieves faster goal discovery and higher map coverage efficiency across all environments.

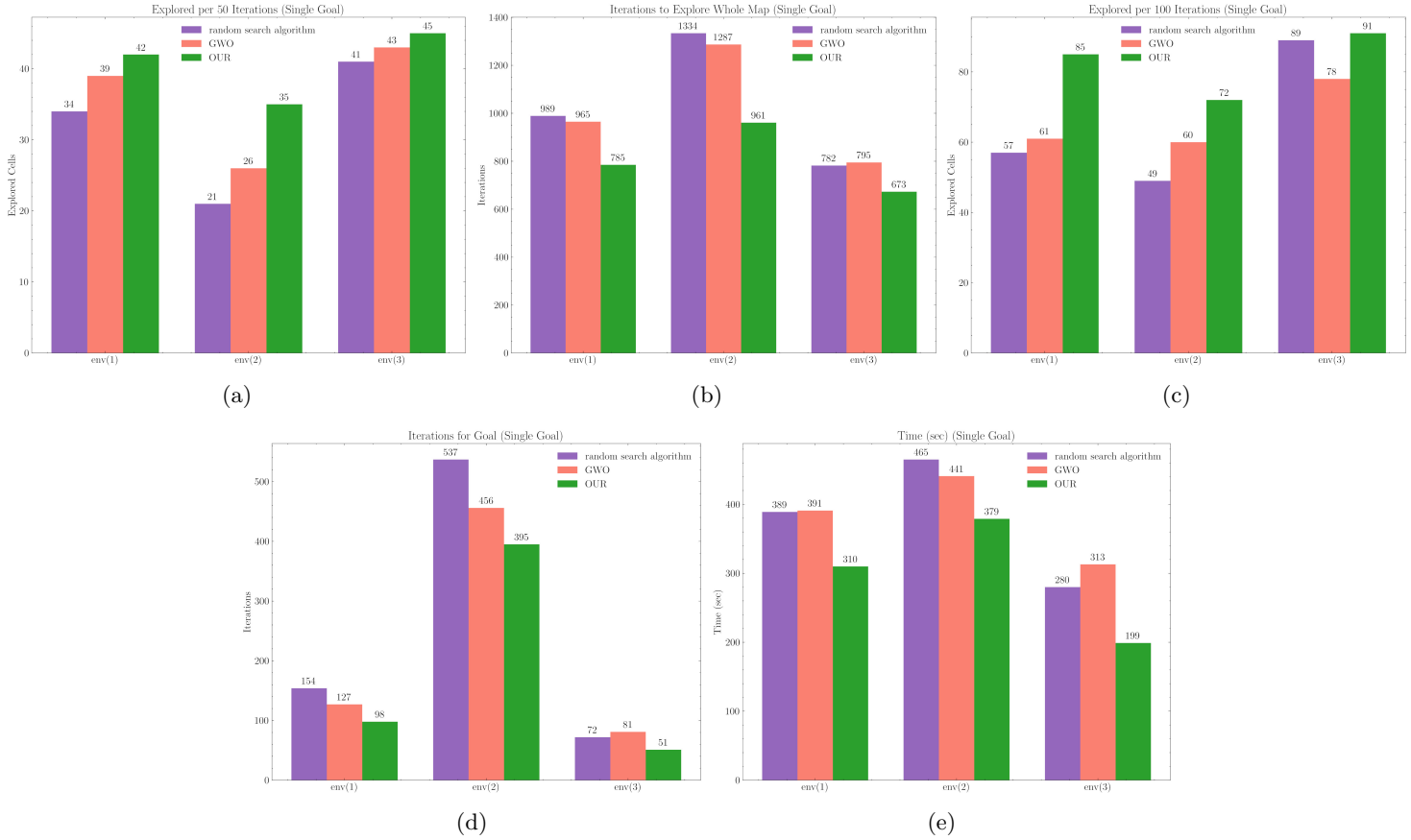


Fig. 5: Performance comparison in single-goal environments.

Quantitative results are summarized in Table 1. The proposed approach achieves the rescue goal using only 51 iterations and 199 seconds, significantly outperforming Random Search (72 iterations, 280 seconds) and GWO (81 iterations, 313 seconds). These gains highlight the effectiveness of reward shaping, step penalties, and GWO-guided exploration in accelerating convergence and improving real-time performance.

Table 1: Single goal evaluation.

Algorithm	**Explored per 50 iteration/step**	**Explored per 100 iteration/step**	**Iteration/step for goal**	**Iteration/step to explore whole map**	**Time (s)**
Random search algorithm (single goal (1))	34	57	154	989	389
GWO (single goal (1))	39	61	127	965	391
OUR (single goal (1))	42	85	98	785	310
Random search algorithm (single goal (2))	21	49	537	1334	465
GWO (single goal (2))	26	60	456	1287	441
OUR (single goal (2))	35	72	395	961	379
Random search algorithm (single goal (3))	41	89	72	782	280
GWO (single goal (3))	43	78	81	795	313
OUR (single goal (3))	45	91	51	673	199

5.3 Multi-Agent Performance Analysis

Figure 6 presents the performance comparison in multi-goal environments. The proposed framework consistently outperforms CME and GWO across all scenarios in terms of both rescue time and exploration efficiency. Notably, the proposed method identifies shorter collective paths that cover all victims, which can be reused for subsequent detailed rescue operations.

Table 2 further demonstrates the scalability of the proposed framework. With four agents, the proposed approach completes the rescue task in 754 seconds using 798 iterations, compared to 876 iterations for Random Search and 950 iterations for GWO. When only two agents are used, the proposed method again achieves superior performance, completing the mission in 780 seconds, compared to 1056 seconds and 1120 seconds for Random Search and GWO, respectively.

These improvements are attributed to implicit coordination through reward shaping, duplicate-exploration penalties, and collision avoidance mechanisms, which collectively enhance cooperative efficiency without introducing communication overhead.

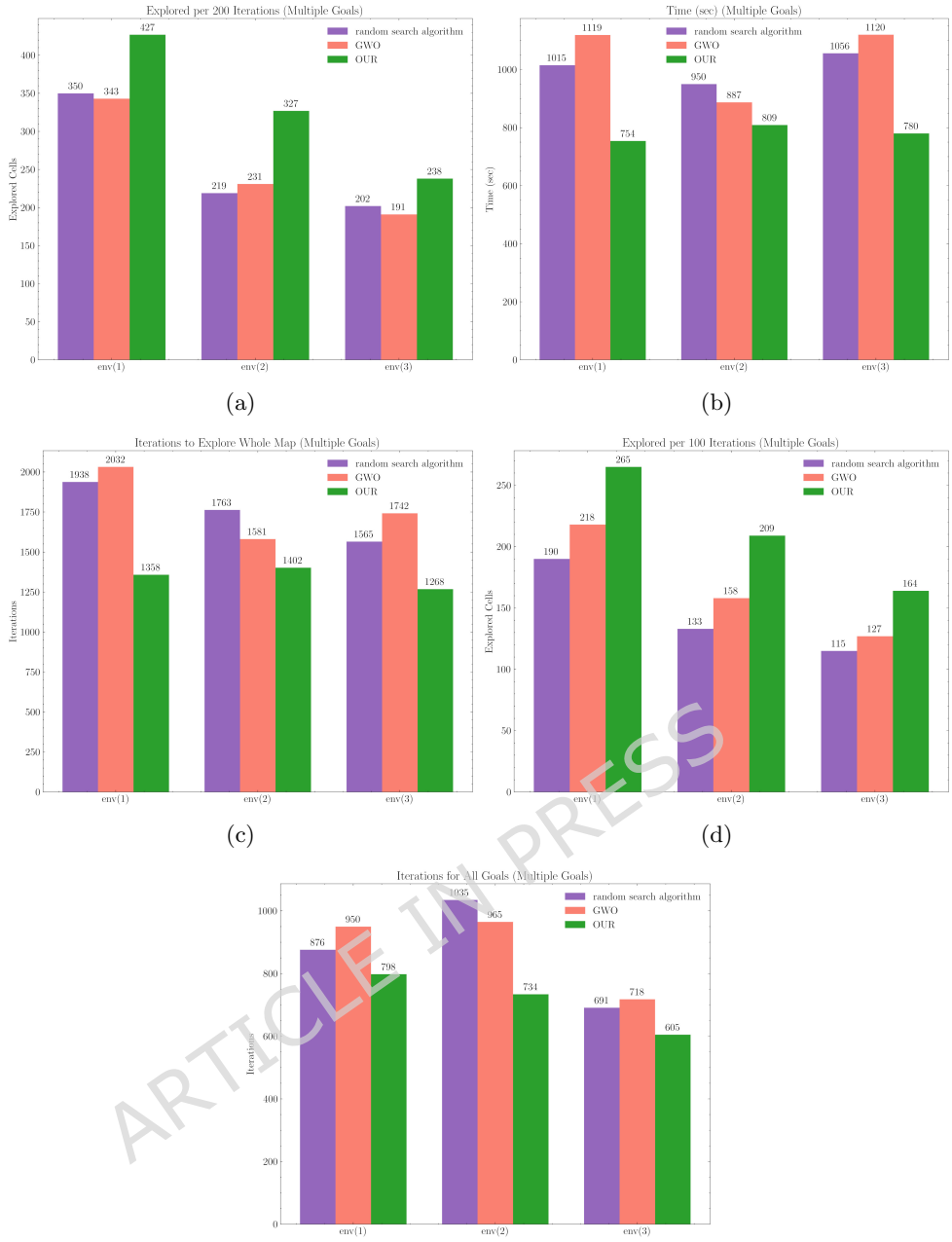


Fig. 6: Performance comparison in multi-goal environments.

Table 2: Multi goal evaluation.

Algorithm	**Explored per 100 iteration/step**	**Explored per 200 iteration/step**	**Iteration/step for all goals**	**Iteration/step to explore whole map**	**Time (s)**	**# of agents**
Random search algorithm (multi goals (1))	190	350	876	1938	1015	4
GWO (multi goals (1))	218	343	950	2032	1119	4
OUR (multi goals (1))	265	427	798	1358	754	4
Random search algorithm (multi goals (2))	133	219	1035	1763	950	4
GWO (multi goals (2))	158	231	965	1581	887	4
OUR (multi goals (2))	209	327	734	1402	809	4
Random search algorithm (multi goals (3))	115	202	691	1565	1056	2
GWO (multi goals (3))	127	191	718	1742	1120	2
OUR (multi goals (3))	164	238	605	1268	780	2

5.4 Discussion and Practical Implications

The experimental results demonstrate that the proposed GWO-guided IQL framework consistently outperforms Random Search, CME, UCE, and standalone GWO-based methods across both single-agent and multi-agent rescue scenarios. These improvements are observed in terms of reduced rescue time, fewer iterations to achieve goals, and higher exploration efficiency, while maintaining collision-free navigation.

A key factor contributing to this performance gain is the integration of GWO into the exploration policy rather than as a standalone optimizer. By guiding exploration without replacing the underlying RL process, the proposed approach avoids premature convergence to suboptimal paths, which is a common limitation of greedy or purely heuristic-based exploration strategies. This sustained exploration capability is particularly important in tunnel environments, where dynamic changes such as blocked passages or newly accessible regions can significantly alter optimal rescue routes during operation.

The use of IQL with tabular representation plays a critical role in enabling real-time feasibility. Unlike DRL or centralized-training paradigms, the proposed framework avoids neural network inference and extensive training overhead, allowing agents to make rapid decisions based on lightweight table lookups. This design choice is well aligned with the operational constraints of emergency response systems, where computational resources, energy availability, and response time are limited.

Another important observation from the results is the effectiveness of reward shaping in achieving implicit coordination among agents. Penalizing duplicate exploration and collisions discourages inefficient or unsafe behaviors without requiring explicit inter-agent communication or task assignment. As demonstrated in multi-agent experiments, this mechanism enables agents to naturally distribute themselves across the environment, reducing redundant coverage and accelerating collective victim discovery.

From a safety perspective, embedding collision avoidance directly into the reward function ensures that unsafe actions are penalized during learning, allowing agents to internalize safety constraints early in the training process. This approach reduces the likelihood of collision-prone policies emerging, which is essential for operation in narrow tunnel environments where maneuvering space is constrained.

In practical deployment scenarios, the proposed framework can be integrated into centralized tunnel monitoring and command systems, where a global situational map is maintained and shared with multiple autonomous agents. The lightweight nature of the learning algorithm makes it suitable for onboard implementation on resource-constrained platforms, while the centralized simulation assumption provides a foundation for future extensions that incorporate communication delays, sensor noise, or decentralized coordination mechanisms.

Despite its advantages, the proposed framework has certain limitations. The current implementation assumes idealized sensing and reliable global map updates, which may not fully reflect real-world tunnel conditions characterized by sensor noise, communication disruptions, or partial observability. Additionally, while the tabular IQL approach is effective for the evaluated environments, scaling to very large or continuous state spaces may require function approximation or hierarchical learning strategies[44].

Overall, the results suggest that the proposed GWO-guided IQL framework offers a practical and effective solution for time-critical tunnel rescue operations. Through balancing exploration efficiency, safety, and computational feasibility, the framework provides a strong foundation for real-world emergency response systems and opens avenues for future research on decentralized coordination, adaptive communication strategies, and integration with physical robot platforms.

6 Conclusion

This paper presented a lightweight multi-agent reinforcement learning (MARL) framework for autonomous UAV-assisted emergency response in tunnel accident scenarios. The proposed approach employs an Independent Q-Learning (IQL) paradigm augmented with frontier-based exploration and policy-level guidance from Grey Wolf Optimization (GWO) to enable efficient, real-time decision-making under partial observability and dynamic environmental conditions. Extensive simulation results across both single-agent and multi-agent environments demonstrate that the proposed framework consistently achieves faster victim discovery, improved map coverage, and reduced overall rescue time when compared with baseline approaches such as random search and standalone GWO-based exploration. In particular, the results show that the proposed reward design effectively discourages redundant exploration, balances exploration and exploitation, and enhances cooperative behavior among agents without requiring explicit inter-agent communication. These characteristics are especially important in confined tunnel environments, where communication may be unreliable and rapid response is critical. From a practical perspective, the proposed method emphasizes computational efficiency and decentralized execution, making it suitable for real-time deployment in emergency scenarios where hardware resources and response time are constrained. By relying on tabular learning and implicit coordination through reward shaping, the framework avoids the heavy training and infrastructure requirements associated with deep or centralized reinforcement learning methods.

Future work will focus on extending the framework to three-dimensional tunnel models, incorporating realistic sensor noise and communication delays, and validating the approach in high-fidelity simulators or real-world testbeds. Additionally, hybrid architectures that combine lightweight IQL with selective deep reinforcement learning components or adaptive communication strategies will be explored to further improve scalability and robustness in large and highly dynamic rescue operations.

Declarations

Funding

This research is funded by the European University of Atlantic.

Conflict of Interest

"The authors declare no conflict of interests."

Ethics approval and Consent to participate

"Not applicable."

Consent for publication

"Not applicable."

Availability of data and materials

The dataset used in this study can be requested from corresponding authors.

Code availability

"Not applicable."

Authors' contributions

HRuR conceptualization, data curation and writing - the original manuscript.
MJG conceptualization, formal analysis and writing - the original manuscript.
RY methodology, and formal analysis and data curation.
MZJ software, methodology and project administration.
RMA investigation, funding acquisition, and visualization.
YM visualization, software, and investigation.
I.A. supervision, validation and writing - review & edit the manuscript.
All authors reviewed the manuscript and approved it.

References

[1] Ma, H., Zhao, J., Huang, H., Wang, Z., Yao, Y.: An experimental investigation into the fire behaviors and smoke characteristics of continuous spill fires in road tunnels. Fire Safety Journal **141**, 104009 (2023) https://doi.org/10.1016/j.firesaf.2023.104009

[2] Chen, Q., Zhao, J.: Case study of the tianjin accident: Application of barrier and systems analysis to understand challenges to industry loss prevention in emerging economies. Process Safety and Environmental Protection **131** (2019) https://doi.org/10.1016/j.psep.2019.08.028

[3] Sasago Tunnel. Wikipedia. https://en.wikipedia.org/wiki/Sasago_Tunnel (2025). https://en.wikipedia.org/wiki/Sasago_Tunnel

[4] La, H.M.: Multi-robot swarm for cooperative scalar field mapping. In: Robotic Systems: Concepts, Methodologies, Tools, and Applications, pp. 208–223. IGI Global, ??? (2020)

[5] La, H.M., Sheng, W., Chen, J.: Cooperative and active sensing in mobile sensor networks for scalar field mapping. IEEE Transactions on Systems, Man, and Cybernetics: Systems **45**(1), 1–12 (2014)

[6] Sutton, R.S., Barto, A.G., *et al.*: Reinforcement Learning: An Introduction vol. 1. MIT press Cambridge, ??? (1998)

[7] La, H.M., Lim, R., Sheng, W.: Multirobot cooperative learning for predator avoidance. IEEE Transactions on Control Systems Technology **23**(1), 52–63 (2014)

[8] La, H.M., Lim, R.S., Sheng, W., Chen, J.: Cooperative flocking and learning in multi-robot systems for predator avoidance. In: 2013 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, pp. 337–342 (2013). IEEE

[9] Faust, A., Palunko, I., Cruz, P., Fierro, R., Tapia, L.: Learning swing-free trajectories for uavs with a suspended load. In: 2013 IEEE International Conference on Robotics and Automation, pp. 4902–4909 (2013). IEEE

[10] Bou-Ammar, H., Voos, H., Ertel, W.: Controller design for quadrotor uavs using reinforcement learning. In: 2010 IEEE International Conference on Control Applications, pp. 2130–2135 (2010). IEEE

[11] Santos, S.R.B., Nascimento, C.L., Givigi, S.N.: Design of attitude and path tracking controllers for quad-rotor robots using reinforcement learning. In: 2012 IEEE Aerospace Conference, pp. 1–16 (2012). IEEE

[12] Waslander, S.L., Hoffmann, G.M., Jang, J.S., Tomlin, C.J.: Multi-agent quadrotor testbed control design: Integral sliding mode vs. reinforcement learning. In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3712–3717 (2005). IEEE

[13] Bellingham, J., Richards, A., How, J.P.: Receding horizon control of autonomous aerial vehicles. In: Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301), vol. 5, pp. 3741–3746 (2002). IEEE

[14] Pham, H.X., La, H.M., Feil-Seifer, D., Van Nguyen, L.: Reinforcement learning for autonomous uav navigation using function approximation. In: 2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), pp. 1–6 (2018). IEEE

[15] Hung, S.-M., Givigi, S.N.: A q-learning approach to flocking with uavs in a stochastic environment. IEEE transactions on cybernetics **47**(1), 186–197 (2016)

[16] Jia, W., Lv, L., Duan, R., Sun, T., Sun, W.: A reinforcement learning-based adaptive grey wolf optimizer for simultaneous arrival in manned/unmanned aerial vehicle dynamic cooperative trajectory planning. Drones **9**(10) (2025) https://doi.org/10.3390/drones9100723

[17] Tai, L., Liu, M.: A robot exploration strategy based on q-learning network.

In: 2016 IEEE International Conference on Real-time Computing and Robotics (RCAR), pp. 57–62 (2016). IEEE

[18] Zhang, J., Tai, L., Liu, M., Boedecker, J., Burgard, W.: Neural slam: Learning to explore with external memory. arXiv preprint arXiv:1706.09520 (2017)

[19] Tai, L., Paolo, G., Liu, M.: Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 31–36 (2017). IEEE

[20] Wei, J., Zhao, Y., Yang, K.: Integrated communication and control for intelligent formation management of uav swarms: A deep reinforcement learning approach. IEEE Wireless Communications Letters (2025)

[21] Azar, A.T., Koubaa, A., Ali Mohamed, N., Ibrahim, H.A., Ibrahim, Z.F., Kazim, M., Ammar, A., Benjdira, B., Khamis, A.M., Hameed, I.A., *et al.*: Drone deep reinforcement learning: A review. Electronics **10**(9), 999 (2021)

[22] Kulkarni, S., Chaphekar, V., Chowdhury, M.M.U., Erden, F., Guvenc, I.: Uav aided search and rescue operation using reinforcement learning. In: 2020 SoutheastCon, vol. 2, pp. 1–8 (2020). IEEE

[23] Zuluaga, J.G.C., Leidig, J.P., Trefftz, C., Wolffe, G.: Deep reinforcement learning for autonomous search and rescue. In: NAECON 2018-IEEE National Aerospace and Electronics Conference, pp. 521–524 (2018). IEEE

[24] Zhan, H., Zhang, Y., Huang, J., Song, Y., Xing, L., Wu, J., Gao, Z.: A reinforcement learning-based evolutionary algorithm for the unmanned aerial vehicles maritime search and rescue path planning problem considering multiple rescue centers. Memetic Computing **16**(3), 373–386 (2024)

[25] Talha, M., Hussein, A., Hossny, M.: Autonomous uav navigation in wilderness search-and-rescue operations using deep reinforcement learning. In: Australasian Joint Conference on Artificial Intelligence, pp. 733–746 (2022). Springer

[26] Liu, X., Liu, Y., Chen, Y.: Reinforcement learning in multiple-uav networks: Deployment and movement design. IEEE Transactions on Vehicular Technology **68**(8), 8036–8049 (2019)

[27] Zhang, Y., Liu, H., Wang, X., Chen, J.: A UAV–UGV cooperative system for patrolling and energy management in urban monitoring. IEEE Transactions on Vehicular Technology **74**(2), 2451–2464 (2025)

[28] Gao, P., Zhou, L., Zhao, X., Shao, B.: Research on ship collision avoidance path planning based on modified potential field ant colony algorithm. Ocean & Coastal Management **235**, 106482 (2023)

[29] De Castro, G.G., Pinto, M.F., Biundini, I.Z., Melo, A.G., Marcato, A.L., Haddad, D.B.: Dynamic path planning based on neural networks for aerial inspection. Journal of Control, Automation and Electrical Systems **34**(1), 85–105 (2023)

[30] Shen, Z., Ding, W., Liu, Y., Yu, H.: Path planning optimization for unmanned sailboat in complex marine environment. Ocean Engineering **269**, 113475 (2023)

[31] Seyyedabbasi, A., Kiani, F., Allahviranloo, T., Fernandez-Gamiz, U., Noeiaghdam, S.: Optimal data transmission and pathfinding for wsn and decentralized iot systems using i-gwo and ex-gwo algorithms. Alexandria Engineering Journal **63**, 339–357 (2023)

[32] Yamauchi, B.: A frontier-based approach for autonomous exploration. In: Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation', pp. 146–151 (1997). IEEE

[33] Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics cambridge. MA: MIT Press [Google Scholar] (2005)

[34] Clifton, J., Laber, E.: Q-learning: Theory and applications. Annual Review of Statistics and Its Application **7**(1), 279–301 (2020)

[35] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)

[36] Cao, Y.U., Fukunaga, A.S., Kahng, A.: Cooperative mobile robotics: Antecedents and directions. Autonomous robots **4**, 7–27 (1997)

[37] Huang, Z., Li, Q., Zhao, Y., Zhang, R.: Optimizing disaster response with UAV-mounted reconfigurable intelligent surfaces and HAP-enabled edge computing in 6G networks. Journal of Network and Computer Applications **221**, 104213 (2025) https://doi.org/10.1016/j.jnca.2025.104213

[38] Alqahtani, S., Nguyen, T., Kim, D.: Energy efficiency relaying election mechanism for 5G internet of things: A deep reinforcement learning technique. In: Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6. IEEE, ??? (2024)

[39] Nawaz, F., Sung, M., Gadginmath, D., D'sa, J., Bae, S., Isele, D., Figueroa, N., Matni, N., Tariq, F.M.: Graph-based Path Planning with Dynamic Obstacle Avoidance for Autonomous Parking (2025). https://arxiv.org/abs/2504.12616

[40] Srivastava, A., Vasudevan, V.R., Harikesh, Nallanthiga, R., Sujit, P.B.: A modified artificial potential field for uav collision avoidance. In: 2023 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 499–506 (2023). https://doi.org/10.1109/ICUAS57906.2023.10156492

[41] Kostrikov, I., Nair, A., Levine, S.: Offline Reinforcement Learning with Implicit Q-Learning (2021). https://arxiv.org/abs/2110.06169

[42] Rehman, H.M.R.U., On, B.-W., Ningombam, D.D., Yi, S., Choi, G.S.: Qsod: Hybrid policy gradient for deep multi-agent reinforcement learning. IEEE Access **9**, 129728–129741 (2021) https://doi.org/10.1109/ACCESS.2021.3113350

[43] Mirjalili, S., Mirjalili, S.M., Lewis, A.: Grey wolf optimizer. Advances in Engineering Software **69**, 46–61 (2014) https://doi.org/10.1016/j.advengsoft.2013.12.007

[44] Younas, R., Raza Ur Rehman, H.M., Lee, I., On, B.-W., Yi, S., Choi, G.S.: Samarl: Novel self-attention-based multi-agent reinforcement learning with stochastic gradient descent. IEEE Access **13**, 35674–35687 (2025) https://doi.org/10.1109/ACCESS.2025.3544961