Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37690

Multi-agent Reinforcement Learning for Coordinated Emergency Building Sweeps

Lechen Ding[1,*,†], Yunsheng Zhang[1,†], Jiamei Du[2,†], Yuanhang Sun[1,†]

[1]*Beijing Huiwen Middle School, Beijing, China*
[2]*The Affiliated High School of SCNU, Guangzhou, China*
**Corresponding Author. Email: 3103360784@qq.com*
†These authors contributed equally to this work and should be considered as co-first author

Abstract. We develop a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework for coordinating emergency responders during building sweeps. The problem is formulated as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) that captures building topology, occupant vulnerability, dynamic hazards, and partial observability. In the six-room office baseline, MADDPG reaches a 134-second sweep time after 5,000 episodes, within 6.3% of the 126-second theoretical optimum and outperforming PPO and DQN. Extensions to a two-story school and a U-shaped warehouse achieve 120 seconds with five responders and 105 seconds with four responders, respectively, with diminishing returns beyond these staffing levels. Under communication failure, coverage remains about 98% while sweep time rises 15–20%; reduced staffing triggers workload redistribution, and dynamic hazards cause threatened areas to be prioritized. Occupancy sensing, mesh networks, and environmental sensing are projected to reduce sweep time by 35–50%. Sensitivity analysis confirms robust performance across realistic parameter ranges.

Keywords: Emergency Evacuation, Multi-Agent Reinforcement Learning, MADDPG, Building Sweep Optimization

1. Introduction

Building fires remain a major life-safety hazard in the United States, causing thousands of deaths and injuries each year [1, 2]. Modern materials and synthetic furnishings can accelerate fire development, leaving only two to three minutes before conditions become untenable [3, 4]. This short window makes rapid evacuation and reliable verification of cleared spaces essential.

OSHA evacuation guidance requires designated personnel to check spaces before leaving so that occupants are fully accounted for [5]. Sweep efficiency therefore directly affects the reliability of evacuation completion.

Optimal sweeping is complicated by four interacting factors: building topology, occupant characteristics, time-varying hazards, and incomplete information about teammates. These factors make static centralized optimization poorly suited to conditions in which routes and communications can change during the operation.

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

DOI: 10.54254/2755-2721/2026.BA37690

We therefore model emergency sweeping as a Dec-POMDP and use MADDPG with centralized training and decentralized execution [6-8]. Buildings are represented as graphs in which rooms are nodes and connections are edges [9, 10]. The framework is evaluated on a single-story office, a two-story elementary school, and a heterogeneous warehouse, and is later extended to communication failure, moving occupants, dynamic hazards, and reduced responder availability.

2. Assumptions and justifications

Assumption 1: The building's room connectivity, corridors, stairs, and exits remain fixed during the sweep; pre-incident floor plans support graph-based representation.

Assumption 2: Responders move at constant speeds and room inspection times are fixed by room type, consistent with standardized training procedures.

Assumption 3: A space is cleared only after physical entry or direct line-of-sight verification, following conservative life-safety practice.

Assumption 4: The baseline assumes radio contact and shared position and clearance updates; communication failures are examined later.

Assumption 5: Designated egress routes remain passable and conditions remain tenable throughout the sweep, consistent with protected egress requirements.

3. Solution of requirement one

3.1. Core challenge

The evacuation sweep is a dynamic multi-objective optimization problem under uncertainty. With multiple responders, the central difficulty is coordinating division of labor, limiting redundant coverage, and adapting as conditions deteriorate.

3.2. Why traditional methods fail

Classical static optimization and fixed heuristics do not naturally handle decentralized information and changing hazards. Multi-agent reinforcement learning instead learns coordinated policies from interaction while allowing each responder to act from local observations.

3.3. Formal problem structure

Let $A = \{a1 ,...,an\}$ denote $n$responders. The building state at time $t$is:

$$s(t) = (p(t), r(t), h(t)) \tag{1}$$

where p(t) gives responder positions, r(t) records room-clearance status, and h(t) represents hazard intensity.

Each responder can move to adjacent nodes or inspect a room. The joint policy $\pi = (\pi 1 ,...,\pi n)$ is chosen to minimize completion time T subject to full coverage and safety constraints.

This is a Dec-POMDP: transitions depend on joint actions and emergency dynamics, observations are local, and rewards balance speed, coverage, and safety.

DOI: 10.54254/2755-2721/2026.BA37690

3.4. The baseline scenario

The baseline is a single-story office with six rooms along a central hallway, two opposite exits, and two firefighters.

3.5. Graph-based environmental representation

The building is represented as a weighted undirected graph that preserves essential spatial relationships:

$$G = (V, E, w) \tag{2}$$

V contains six room nodes, one hallway node, and two exit nodes. Treating the hallway as a transit node reduces the state space without losing the building's essential topology.

Edges encode room-hallway and hallway-exit adjacency, while edge weights represent traversal costs:

$$w\left(v_i, v_{\text{hall}}\right) = \frac{d_{\text{room-to-hall}}}{v_{\text{walk}}} + \tau_{\text{door}} \tag{3}$$

With a room-to-hall distance of about 3 m, walking speed of 1.5 m/s, and a 2-second door cost, each room-hall transition is about 4 seconds. Each office requires 30 seconds for inspection.

3.6. Multi-agent reinforcement learning formulation

We formulate the sweep as a cooperative multi-agent system in which each firefighter acts as an autonomous learning agent. The global state at time t is:

$$s_t = (p_t, c_t) \tag{4}$$

where pt gives the two agent positions and ct records the six room-clearance states.

Each agent operates under partial observability, receiving local observations:

$$o_t^j = (p_t^j, \mathbf{c}_t^{\text{local}}, p_t^{-j}) \tag{5}$$

Each observation contains the agent's position, locally visible clearance status, and any communicated teammate position, allowing the same policy structure to be used when communication degrades.

The action space for agent *j* is:

$$\mathscr{A}^j = \{\text{move}(v) : v \in N(p_t^j)\} \cup \{\text{inspect}\} \tag{6}$$

where N(p) denotes adjacent nodes. Movement consumes edge time and inspection is available only at room nodes. The reward uses a time penalty, clearance bonus, redundancy penalty, and completion bonus:

$$r_t = -\alpha \cdot \Delta t + \beta \cdot \mathrm{I}\left(\text{new cleared}\right) - \gamma \cdot \mathrm{I}\left(\text{redundant}\right) + \zeta \cdot \mathrm{I}\left(\text{all cleared}\right) \tag{7}$$

DOI: 10.54254/2755-2721/2026.BA37690

With $\alpha = 0.1$, $\beta = 10.0$, $\gamma = 5.0$, and $\zeta = 100.0$, the reward balances speed, coverage, and coordination efficiency.

3.7. Multi-agent deep deterministic policy gradient

MADDPG is used to learn coordinated policies through centralized training with decentralized execution, reducing multi-agent non-stationarity.

For agent *j*, the actor network maps observations to actions:

$$\pi^j(a_t^j \mid o_t^j) = \mu_{\theta_j}^j(o_t^j) \tag{8}$$

The actor uses a three-layer feedforward network with 128 ReLU units per layer and a softmax output. The centralized critic estimates action values from the global state and joint actions:

$$Q_{\phi_j}^j\left(s_t, a_t^1, a_t^2\right) \tag{9}$$

The centralized critic improves credit assignment because it observes the full state and joint actions during training.

Training alternates between experience collection and network updates. The critic minimizes temporal-difference error, while the actor maximizes expected return.

$$\mathcal{L}(\phi_j) = \mathbb{E}[(Q_{\phi_j}^j(s_t, a_t^1, a_t^2) - y_t^j)^2] \tag{10}$$

Target networks generate the bootstrapped value estimate used in the critic update.

$$\nabla_{\theta_j} J(\theta_j) = \mathbb{E}[\nabla_{\theta_j} \mu_{\theta_j}^j(o_t^j) \cdot \nabla_{a^j} Q_{\phi_j}^j(s_t, a^1, a^2)|_{a^j = \mu_{\theta_j}^j(o_t^j)}] \tag{11}$$

We use Adam with actor and critic learning rates of 10^{-3} and 10^{-4}, replay buffer size 50,000, discount factor $\gamma = 0.99$, soft update rate $\tau = 0.01$, and exploration noise decaying from 0.5 to 0.05 over 5,000 episodes.

Together, the graph representation, Dec-POMDP formulation, and MADDPG architecture provide a baseline that can be extended to more complex buildings and degraded information conditions.

3.8. Results and analysis

After 5,000 episodes, the two agents consistently divide the six offices between opposite exits, demonstrating implicit spatial coordination without explicit room assignment (Figure 1).

DOI: 10.54254/2755-2721/2026.BA37690

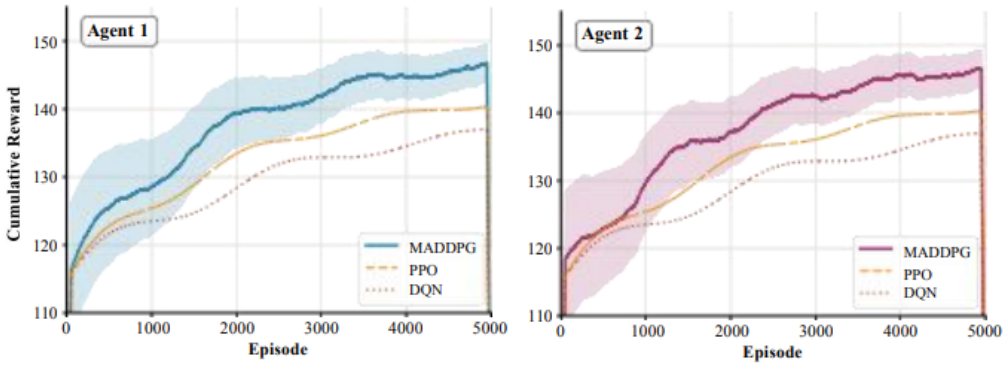

Figure 1. Graph representation and learned sweep strategy for the baseline office building. (a) Building topology as weighted graph with traversal times on edges. (b) Converged coordination strategy showing symmetric spatial partitioning between agents

MADDPG achieves higher cumulative rewards and more stable convergence than PPO and DQN because its centralized critic conditions value estimates on joint actions, reducing the non-stationarity faced by single-agent methods in cooperative settings.

Sweep time falls from more than 275 seconds during early exploration to about 134 seconds after 5,000 episodes, close to the 126-second theoretical lower bound. The remaining gap is 6.3%.

4. Solution of requirement three

The baseline confirms that MADDPG can coordinate two responders in a symmetric single-floor office. We next test the same framework on more complex layouts and coordination constraints.

4.1. Scenario design and complexity dimensions

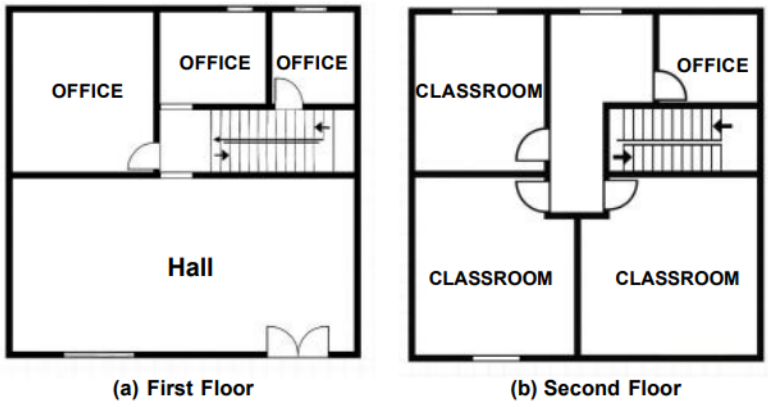

Figure 2. Two-story elementary school floor plans. (a) First floor: three offices, central hallway, one exit, and stairwell. (b) Second floor: three classrooms (high-priority), one office, and stairwell connection with no direct exits

Scenario 1: Two-Story Elementary School (Figure 2). The first floor contains three offices, one exit, and a stairwell; the second floor contains three high-priority classrooms and one office with no direct exit. The shared stairwell creates a vertical bottleneck that requires coordinated floor sequencing.

DOI: 10.54254/2755-2721/2026.BA37690

Scenario 2: Retail store with a Single-Story (Figure 3). The U-shaped warehouse has three 90-second storage zones, three 20-second service sections, a central corridor, and two exits. Uniform occupant priority but highly unequal inspection times tests spatial partitioning under irregular topology.

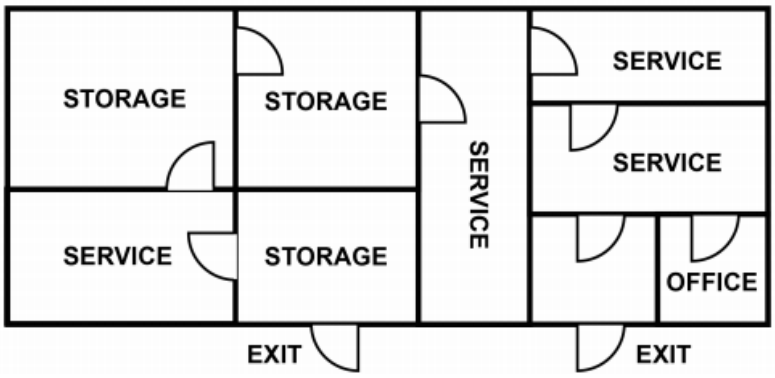

Figure 3. Single-story retail warehouse with U-shaped layout featuring three storage zones (extended inspection), three service sections (rapid inspection), central corridor, and two symmetric exits

4.2. Model adaptations for complex scenarios

The baseline graph and MADDPG model are extended only where needed for multi-floor structure, priority weighting, and larger responder teams.

Multi-Folder Graph Structure. For multi-story buildings, we construct a layered graph:

$$G_{\text{school}} = \left(V^{(1)} \cup V^{(2)} \cup V^{\text{stair}}, E^{\text{horiz}} \cup E^{\text{vert}}, w\right) \tag{12}$$

V(1) and V(2) represent the two floors and Vstair links them. Horizontal edges connect rooms within a floor, while vertical edges connect stairwell nodes across floors.

$$w\left(v_{\text{stair}}^{(1)}, v_{\text{stair}}^{(2)}\right) = \alpha_{\text{stair}} \cdot w\left(v_{\text{stair}}^{(2)}, v_{\text{stair}}^{(1)}\right) \tag{13}$$

An upward stair penalty αstair = 1.8 encourages upper-floor positioning when feasible. The warehouse remains single-floor but uses 90-second storage inspections and 20-second service inspections.

Extended State Representation. Multi-floor buildings expand the state to track floor information:

$$s_t = (p_t, f_t, c_t, h_t) \tag{14}$$

ft records each agent's floor, extending the baseline position state.

The clearance vector spans all rooms and the hazard field varies by floor. Agent observations therefore include floor information:

$$o_t^j = (p_t^j, f_t^j, c_t^{\text{local}}, \{(p_t^k, f_t^k)\}_{k \neq j}) \tag{15}$$

This supports both vertical and horizontal coordination while preserving the same move-and-inspect action structure.

DOI: 10.54254/2755-2721/2026.BA37690

Priority-Weighted Rewards. For the school scenario, classrooms receive priority weight $\rho i = 2.0$ while offices have $\rho i = 1.0$. We modify the baseline clearance reward:

$$r_t^{\text{clear}} = \beta \cdot \rho_i \cdot \mathrm{I}\,(\text{room } i \text{ newly cleared}) \tag{16}$$

The school gives classrooms priority weight ρi = 2.0 and offices ρi = 1.0; the warehouse keeps uniform priority because inspection-time differences already shape behavior.

Scaled Multi-Agent Coordination. For the school, MADDPG extends the same actor-critic structure to 4–5 agents.

Each actor uses local observations, while each centralized critic conditions on the global state and all agents' actions.

The critic input grows with team size while the actor structure remains unchanged. Curriculum learning begins with two agents and progressively adds agents to learn spatial partitioning before multi-floor coordination.

4.3. Results and analysis

The warehouse converges to 105 seconds by about episode 2500, while the school approaches 120 seconds after about 3500 episodes because stairwell coordination requires more exploration (Figure 4).

The difference reflects coordination complexity: warehouse zones can be swept largely in parallel, whereas the school requires floor sequencing, bottleneck management, and vertical load balancing.

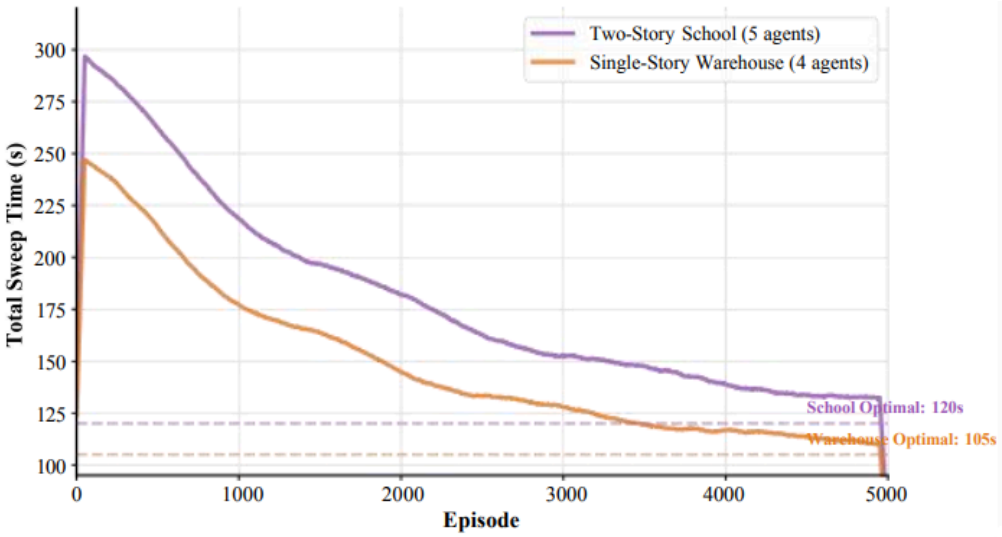

Figure 4. Training convergence comparison. Warehouse converges faster through horizontal partitioning; school demands extended exploration for multi-floor strategies

DOI: 10.54254/2755-2721/2026.BA37690

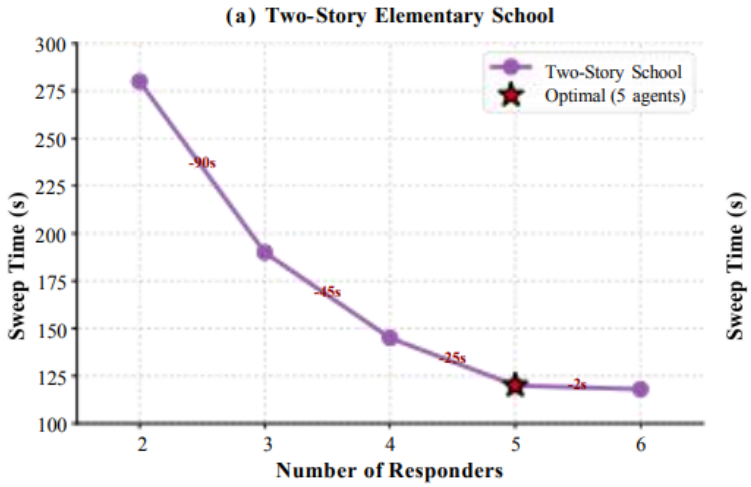

Figure 5. Responder scaling analysis. (a) School optimal at 5 agents. (b) Warehouse optimal at 4 agents. Marginal gains negligible beyond these points

5. Solution of requirement four

The previous scenarios assume reliable communication, static hazards, stationary occupants, and full staffing. We now modify observations, states, and rewards to test the same MADDPG framework under degraded operational conditions.

5.1. Communication failures

We model probabilistic message delivery between agents:

$$P_{\text{comm}}\left(d_{jk}, h\right) = \begin{cases} 1 & \text{if } d_{jk} < r_0 \text{ and } h < h_0 \\ \exp\left(-\lambda d_{jk} - \mu h\right) & \text{otherwise} \end{cases} \tag{17}$$

where djk is inter-agent distance, h is average hazard intensity along the line of sight, and r0 and h0 define reliable communication thresholds. Under failure, teammate information may become stale or unavailable:

$$o_t^j = \left(p_t^j, f_t^j, c_t^j, \tilde{p}_t^{j,\text{local}}\right) \tag{18}$$

During training, messages are randomly dropped so agents learn spatial coordination that does not depend on continuous updates.

Because execution is decentralized, each actor still acts from local observations.

Complete communication blackout increases sweep time by about 15–20%

while maintaining coverage, indicating robust implicit coordination.

5.2. Dynamic hazard propagation

We model fire spread as discrete-time diffusion over the building graph:

DOI: 10.54254/2755-2721/2026.BA37690

$$h_t^{(i)} = \text{clip}\left(h_{t-1}^{(i)} + \kappa \sum_{j \in \mathcal{N}(i)} \frac{h_{t-1}^{(j)}}{|\mathcal{N}(i)|} - \delta h_{t-1}^{(i)}, 0, 1\right) \tag{19}$$

The spread rate κ and suppression term δ update hazard intensity; rooms above θblock become impassable, changing the traversable graph. The reward therefore includes a hazard-exposure penalty.

$$r_t = r_t^{\text{baseline}} - \eta \sum_{j=1}^{n} h_t^{\left(p_t^j\right)} \tag{20}$$

Randomized ignition locations and spread rates train agents to reroute and clear threatened areas before access is lost.

5.3. Moving occupants and environmental non-stationarity

We model occupants as stochastic agents *O*= *{o*1 *,...,om}*with simple behavioral policies:

$$\pi_{\text{occupant}}\left(a \mid p_o\right) = \begin{cases} 0.7 & a = \text{move toward nearest exit} \\ 0.2 & a = \text{stay} \\ 0.1 & a = \text{reverse direction} \end{cases} \tag{21}$$

Moving occupants make room status non-stationary, so clearance is time-stamped:

$$c_t = \{(i, \tau_i) : \text{room } i \text{ last cleared at time } \tau_i\} \tag{22}$$

Rooms whose clearance time exceeds the staleness threshold must be checked again, and occupant positions are added to the global state.

$$s_t = \left(\mathbf{p}_t^{\text{agents}}, \mathbf{p}_t^{\text{occupants}}, \mathbf{f}_t, \mathbf{c}_t, \mathbf{h}_t\right) \tag{23}$$

The episode ends only when all rooms have been cleared recently and all occupants have exited.

Training produces verification patterns in which some responders cover exits while others sweep rooms, reducing re-entry into cleared areas.

5.4. Sub-optimal responder availability

Policies trained with n responders are also tested with fewer than n deployed agents.

The decentralized structure adapts because absent teammates simply leave more rooms uncovered.

In the school, using three instead of five responders increases sweep time by 40–45% but preserves complete coverage through workload redistribution; extra responders add little overhead because redundant inspection is penalized (Figure 5).

5.5. Technology integration

Building technologies can improve observability and communication without replacing the learned coordination policy.

DOI: 10.54254/2755-2721/2026.BA37690

Occupancy Sensors: Direct occupancy data can reduce room verification from 30–90 seconds to about 5–10 seconds and is projected to cut sweep time by 35–50%.

Mesh Networks: Greater reliable communication range improves teammate-position accuracy, especially around stairwells.

Environmental Sensors: Direct hazard measurements support earlier routing decisions, particularly in warehouse storage zones.

AR Wayfinding: Real-time maps can show teammate positions and clearance status, moving observations closer to global awareness.

6. Sensitivity analysis

Sensitivity analysis examines environmental parameters, coordination mechanisms, and training hyperparameters to test robustness across realistic operating ranges.

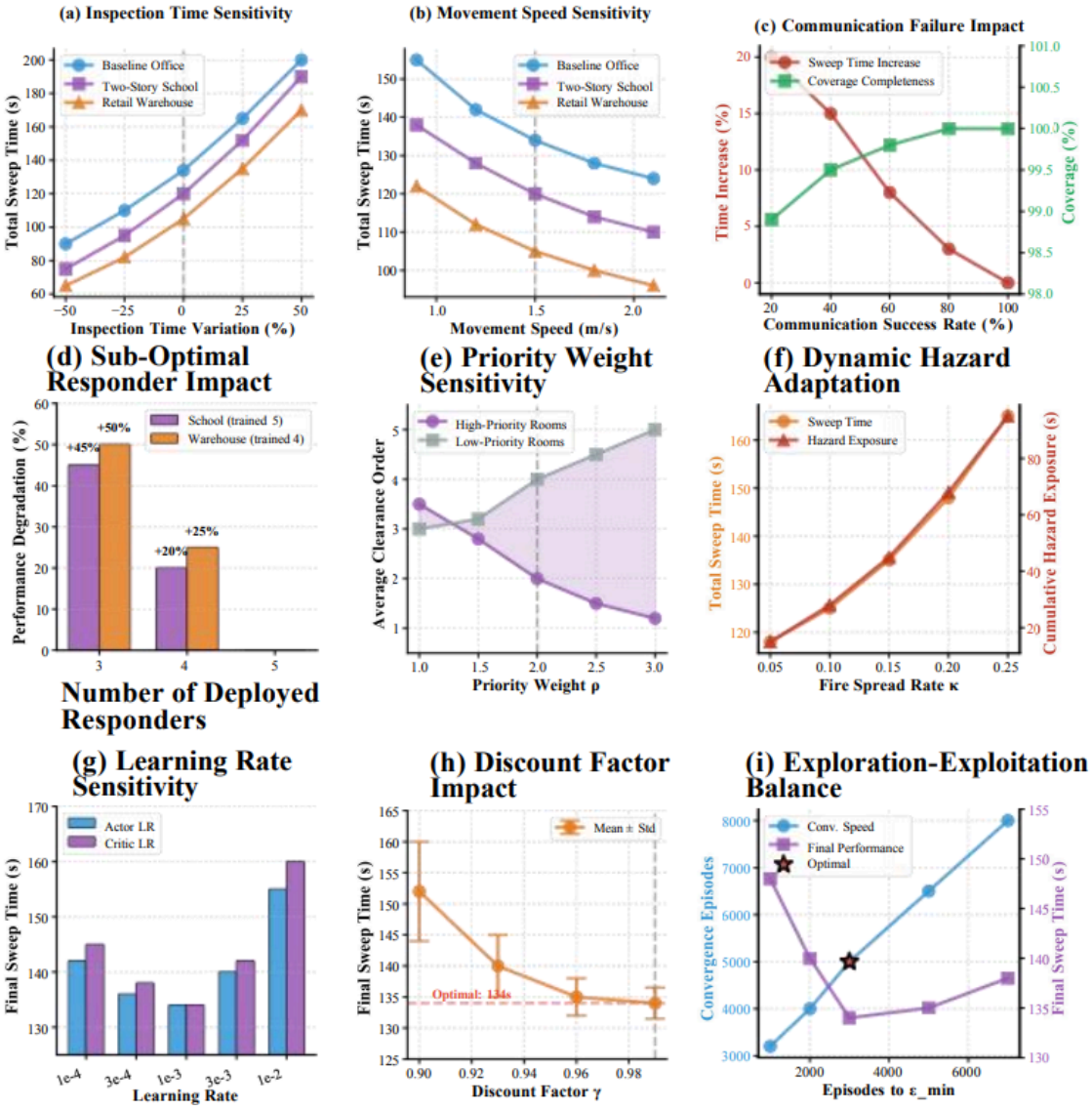

Figure 6. Comprehensive sensitivity analysis across environmental parameters (top row), co-ordination mechanisms (middle row), and training hyperparameters (bottom row)

DOI: 10.54254/2755-2721/2026.BA37690

Environmental variations cause gradual rather than catastrophic degradation (Figure 6). Inspection time has the strongest effect on total sweep time, while faster movement yields smaller gains because inspection dominates duration. Even at 20% communication success, coverage remains above 98.9% and sweep-time overhead stays near 20%.

Coordination also adapts to staffing, priority, and hazard changes. Using three rather than five school responders increases sweep time by about 45% without losing coverage, higher priority weights move vulnerable rooms earlier in the sequence, and faster fire spread increases both sweep time and hazard exposure.

Training is most stable near the selected hyperparameters: learning rates around 10^{-3}, discount factor $\gamma = 0.99$, and exploration decay reaching its minimum near episode 3000. Deviations slow convergence or worsen final sweep time.

Overall, the framework remains stable across the tested parameter ranges and degrades gracefully when communication or responder availability is reduced.

7. Conclusion

MADDPG learns effective sweep coordination across different building layouts and operational constraints, reaching within 6–10% of theoretical optima while remaining robust to communication loss, dynamic hazards, and reduced staffing. Operationally, responders should use available exits, pre-position teams on upper floors where appropriate, prioritize vulnerable spaces, and size teams to architectural bottlenecks rather than floor area alone.

Future work should address topology changes caused by structural damage and validate the policies through live drills with fire departments. Despite these limits, the results show that multi-agent reinforcement learning can support rapid, reliable coordination in emergency sweeps.

References

[1] National Fire Sprinkler Association. (2024, April). Investigating structure fire trends: NFSA provides critical data. *National Fire Sprinkler Association*. https://nfsa.org/2024/04/30/structure-fire-trends-in-the-us/

[2] U.S. Fire Administration. (2023). Residential fire estimate summaries (2014-2023). *U.S. Fire Administration*. https://www.usfa.fema.gov/statistics/residential-fires/

[3] Madrzykowski, D., & Kerber, S. (2008). Fire fighting tactics under wind driven conditions: Laboratory experiments. National Institute of Standards and Technology, Technical Note NIST TN 1455-1.

[4] Underwriters Laboratories. (2012). Analysis of changing residential fire dynamics and its implications on firefighter operational timeframes. UL Fire Safety Research Institute, Tech. Rep.

[5] Occupational Safety and Health Administration. (2024). Evacuation plans and procedures e Tool. *Occupational Safety and Health Administration*. https://www.osha.gov/etools/evacuation-plans-procedures

[6] Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (Volume 30)* (pp. 6379–6390).

[7] Iqbal, S., & Sha, F. (2019). Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML* (pp. 2961–2970).

[8] Malysheva, A., Sung, T. T. K., Sohn, C.-B., Kudenko, D., & Shpilman, A. (2018). Deep multi-agent reinforcement learning with relevance graphs. *arXiv preprint arXiv:1811.12557*. Presented at NeurIPS 2018 Deep Reinforcement Learning Workshop.

[9] Jiang, J., & Lu, Z. (2018). Graph convolutional reinforcement learning for multi-agent cooperation. *arXiv preprint arXiv:1810.09202*. Cited widely in multi-agent coordination research.

[10] Nayak, S., Choi, K., Ding, W., Dolan, S., Gopalakrishnan, K., & Balakrishnan, H. (2023). Scalable multi-agent reinforcement learning through intelligent information aggregation. In *Proceedings of the 40th International Conference on Machine Learning, ICML* (pp. 25752–25777).