Decentralized Fire Seeking MARL UAVs

Gaganpreet Jhajj[1†], Tojoarisoa Rakotoaritina[2†], Fuhua Lin[1]
gjhajj1@learn.athabascau.ca, tojoarisoa.rakotoaritina@oist.jp,
oscarl@athabascau.ca

[1]**School of Computing and Information Systems, Athabasca University, Canada**
[2]**Okinawa Institute of Science and Technology, Neural Computation Unit, Japan**

[†] Indicating equal contribution

Abstract

Wildfires are escalating in frequency and severity, particularly in high-risk regions such as Alberta, Canada, where traditional detection systems are becoming increasingly insufficient. Existing approaches often rely on centralized control or overlook key constraints, such as partial observability, terrain complexity, and communication limitations. To address this gap, we propose a fully decentralized multi-agent reinforcement learning (MARL) framework for wildfire detection using UAV swarms. Our method integrates real geographic data into a grid-based simulator and employs intrinsic-motivation-enhanced Independent Proximal Policy Optimization (IPPO), allowing each agent to learn independently and adaptively. This design is well-suited for large-scale, unstructured environments where centralized coordination is infeasible. Agents learn to balance exploration, fire detection, and risk mitigation through a hybrid reward scheme. Experimental results in simulation demonstrate the effectiveness of our method for early and reliable wildfire detection in large, remote landscapes. This work lays the foundation for scalable, robust, and communication-efficient UAV swarm systems for wildfire monitoring, with significant potential to reduce ecological, economic, and human costs.

1 Introduction

Wildfires have surged in frequency and intensity over the past few decades. Jolly et al. (2015) found that from 1979 to 2013, the length of fire-weather seasons increased by nearly 19%. They also found that the area globally affected by these long fire seasons more than doubled (Jolly et al., 2015). This trend is particularly prominent in Alberta, Canada. Whitman et al. (2022) found that in Alberta, from 1970 to 2019, the number of large wildfires, the area burned, as well as the size of fires increased significantly. During the 2023 Wildfire season, over 2.2 million hectares were burned (Beverly & Schroeder, 2025). This represented an increase of nearly 63% in total area burned from the prior record in 1981, amounting to ~4% of Canada's total forest cover (Beverly & Schroeder, 2025; Jain et al., 2024). Research from Hanes et al. (2019) has shown that in Canada since 1959, the number of large fires has increased significantly, the fire season has become longer, and western Canada, in particular, is experiencing an increase in the area burned and the number of large fires.

Emissions created by the 2023 Canadian wildfires alone amount to similar total annual emissions created by large developed nations (Byrne et al., 2024). While 2023 was an abnormally warm and dry year, Byrne et al. (2024) suggests that by the 2050s, such ranges will be typical, which in turn creates a positive feedback loop where intense wildfires accelerate warming trends, creating more wildfires (Liu et al., 2019).

However, despite the scale and severity of wildfires, existing detection methods struggle to keep pace with them. Satellite-based sensing can take time to process data and can struggle to keep up with the dynamic, fast-moving natures of wildfires, while manned aircraft for detection have high associated costs[1].

Unmanned aerial vehicles (UAVs) can help fill this detection gap. Such UAV systems can be small enough to be deployed to remote regions of Canada and provide valuable data. Coordinated swarms of UAVs can provide real-time coverage and adapt to emerging wildfire behavior.

Coordinating these drone systems over a dynamic and partially observable landscape is complex, and factors including limited communication range, energy consumption, and the scalability of coordination protocols all pose significant challenges (Yanmaz et al., 2018).

Multi-agent reinforcement Learning (MARL) provides a way to learn policies that can balance exploration, detection, and safety from data (Sutton et al., 1998; Tan, 1993). In this work, we show the first steps towards using MARL in a simulated setting to detect wildfires.

This work proposes wildfire detection as a cooperative MARL problem over Alberta's terrain features. We apply Independent Proximal Policy Optimization (IPPO) (de Witt et al., 2020), allowing each UAV to learn with only local observations.

2 Related Work

2.1 UAV-based wildfire monitoring

UAV usage for real-time fire detection and mapping is a growing research field. Bailon-Ruiz et al. (2022) deployed a fleet of UAVs equipped with thermal and RGB cameras to track fire boundaries in near real-time. Hopkins (2024) trained UAV teams via MARL in a simulated 3D wildfire response environment, focusing on navigation and hotspot identification. Recent work by Howard et al. (2024) on drone coordination leveraged state machines and Godot to create a highly customized virtual environment for drone simulation, citing that preexisting approaches lack flexibility.

Pham et al. (2018) discussed distributed coverage schemes for UAV swarms to minimize the overlaps between the field of view for each agent. The FireDronesRL project[2] explored a similar 2D approach to the one we detail in this work, but in an entirely simulated world. Related simulation frameworks for disaster scenarios, such as DisasterReliefBot-CoppeliaSim[3] focus on urban disaster recovery and detecting fire hazards. Tools such as MODIFLY by Cofield et al. (2025) provide an enhanced suite of tools for 3D UAV simulation, considering factors such as dynamic communication modeling. Ding et al. (2023) benchmarked cooperative MARL algorithms on drone routing tasks. More recent work by Zhao et al. (2025) augments multi-UAV MARL with noise-resilient communication and attention mechanisms to improve robustness under packet loss.

Earlier work by Seraj et al. (2021) employed heterogeneous teams in randomly generated environments. The end user can specify specific parameters, such as the number of homes, trees, and hospitals. However, the approach outlined in Seraj et al. (2021) is incompatible, mainly with modern MARL frameworks like PettingZoo (Terry et al., 2021).

2.2 Multi-Agent RL algorithms

For cooperative MARL robotics, methods can broadly fall into two categories: centralized training for decentralized execution (CTDE) and decentralized training and execution (DTE) (Amato, 2024). Sunehag et al. (2017) introduced value decomposition methods such as VDN, and Rashid et al. (2018) later proposed QMIX.

[1] https://www.gao.gov/products/gao-25-108161

[2] https://github.com/yunijeong5/FireDronesRL

[3] https://github.com/amnotme/DisasterReliefBot-CoppeliaSim

Actor–critic methods, such as MADDPG (Lowe et al., 2017) and counterfactual baseline COMA (Foerster et al., 2024), extend CTDE to continuous control. Independent learners, including IQL (Kostrikov et al., 2021) and IPPO (de Witt et al., 2020), use a decentralized critic, making these approaches more complex and realistic. Lacking centralized control makes them more robust in environments with limited communication. Huang et al. (2016) also found that under decentralized learning, agents can learn and develop communication protocols to solve coordination tasks in partially observable settings. Recent work on learning cautious behavior under uncertainty (Mohammedalamen et al., 2021) has demonstrated that agents can autonomously develop risk-averse policies when facing novel situations, which aligns with our approach of developing uncertainty-aware wildfire detection strategies.

Domain-specific adaptations of MARL include resource allocation in UAV networks (Cui et al., 2020) and comparisons of short-term vs. long–term coordination (Qin & Pournaras, 2024). Our work builds upon prior approaches by applying IPPO to train fully decentralized, communication-light UAV policies for wildfire detection over terrain in Alberta, Canada.

3 Methods

3.1 System Overview

This work introduces a novel approach to wildfire monitoring by creating a simulation integrating real-world geographic data with IPPO online MARL.

We obtained OpenStreetMap (OSM) data (OpenStreetMap contributors, 2017) via the API and converted the real-world geographic coordinates into a discretized grid-based simulation space while preserving spatial relationships and feature densities. This conversion enables our experimentation to be conducted in a 2D grid world environment while preserving the geographic features of the locations. We then simulated wildfires on top of the grid world features. This method enabled our UAV agents to learn monitoring strategies roughly based on real-world geographic data.

For our work, we selected two cities in Alberta, Canada, that have been affected by severe wildfires: Fort McMurray[4] (Mamuji & Rozdilsky, 2018) and Athabasca[5]. The cropped OSM maps during various processing steps can be found in Appendix A and B.

3.2 Wildfire Simulation Environment

The wildfire scenarios and modeling were implemented using a probabilistic cellular automaton (CA) fire–spread model in the grid world (see Appendix C). Each cell in the grid world $s \in \{\texttt{EMPTY}, \texttt{TREE}, \dots\}$ has a terrain-specific vulnerability β_s and finite burn duration. At each time step, any burnable neighbor ignites with the below probability:

$$p_{\text{spread}} = \min\Big(1,\ p_f\, \beta_s\, \big[1 + (\mathbf{u}\cdot\mathbf{w})\, w_{\text{str}}\big]\Big), \tag{1}$$

where p_f is the base spread probability, $\mathbf{u}$ the unit vector toward the burning neighbor, and $\mathbf{w}$ the wind vector (Ramadan, 2024; Zadeh et al., 2025). Burnt cells may later regrow; additional details on this can be found in Appendix C. The CA model provides us a simple yet realistic way to test fire dynamics.

3.3 UAV Agent Design

Each UAV agent operates with partial observability of the environment through each agent's local view. At each timestep t, an agent i receives the following observation tuple:

[4] https://earthobservatory.nasa.gov/images/88039/fort-mcmurray-burn-scar

[5] https://globalnews.ca/news/11169138/athabasca-county-boyle-wildfire-may-2025

$$O_i = \{V_{local},\ P_{self},\ P_{others},\ I_{global}\} \tag{2}$$

The agent's local view V_{local} is a $(2r+1) \times (2r+1)$ grid centered on the agent's position, where r is the view range. This view is encoded as a multi-channel tensor representing different terrain features (trees, buildings, natural areas, fires) through one-hot encoding.

Agents navigate using a discrete action space $A \in \{\text{STAY}, \text{UP}, \text{DOWN}, \text{LEFT}, \text{RIGHT}\}$, representing possible movement directions in the grid. Constraints are included to ensure agents remain within the operational area. Additional information on the action space can be found in Appendix D.

The agent design approach balances the need for local fire detection and broader environmental and situational awareness by using local and global information. Mathematical definitions of the observation and action spaces can be found in Appendix D, and the complete reward computation can be found in Appendix E.

3.4 Independent Proximal Policy Optimization (IPPO)

Our MARL approach utilizes IPPO, where each UAV agent learns independently using its own PPO algorithm (Schulman et al., 2017) while sharing the same environment. The reward structure combines both extrinsic and intrinsic motivations to encourage effective fire monitoring and exploration:

At each time step t, each agent i receives an instantaneous reward

$$R^t_{\text{total}} = \sum_{i=1}^{N} \bigl(R_i^{\text{ext}} + R_i^{\text{int}}\bigr). \tag{3}$$

The discounted return for agent i is then

$$G_i^t = \sum_{k=0}^{T-t} \gamma^k \, R_i^{t+k}, \tag{4}$$

Where N is the number of agents, R_i^{ext} is the extrinsic reward for fire detection and monitoring, and R_i^{int} is the intrinsic reward for agent i. See Appendix F for the full update schedule and clipped-PPO objective.

Our approach builds on the intrinsically motivated reinforcement learning framework first introduced by Chentanez et al. (2004). Oudeyer & Kaplan (2007) provide a comprehensive typology of computational methods to intrinsic motivation, which informs our design of exploration bonuses and risk-awareness components. The distinction between intrinsic and extrinsic motivation in reinforcement learning (Barto, 2013) guides our reward structure design, where agents balance external fire detection objectives with internal exploration drives.

Early work on combining intrinsic and extrinsic rewards in constrained settings was explored by Uchibe & Doya (2007), with further developments in robotic applications by Uchibe & Doya (2008). Recent work by Rakotoaritina et al. (2025) outlines a unified information-theoretic formulation of novelty, surprise, and empowerment as intrinsic rewards, demonstrating their effectiveness in environments with hidden subgoals. In this work, we hand-specify strategic-level terms that support our multi-objective reward design.

The instantaneous intrinsic reward is decomposed into five components; see Eq. (12) for the full definition.

The hybrid signal presented to IPPO is the convex combination

$$R_i^{\text{hybrid}}(t) \;=\; \lambda_1 \, R_i^{\text{ext}}(t) \;+\; \lambda_2 \, R_i^{\text{int}}(t), \qquad (\lambda_1, \lambda_2) = (0.7,\, 0.3),\ \lambda_1 + \lambda_2 = 1. \tag{5}$$

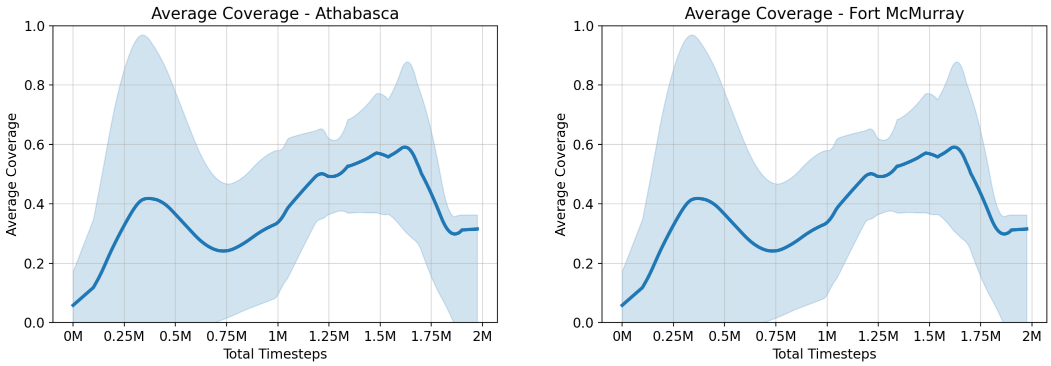
Key scalars α, β, and the mixture weights γ balance detection, safety, and exploration; see the compact summary in Appendix F.

The implementation leverages Stables Baseline 3 (Raffin et al., 2021), with each agent maintaining independent neural networks for both policy and value functions. Details on the architectures can be seen in Appendix H.

The full coefficient grid (Table 2), strategy profiles, and derivative coupling derivations supporting Eq. (26) are provided in Appendix G.

4 Results

We evaluated the system across the real-world environments of Fort McMurray and Athabasca focusing on detection performance, coordination efficiency, and strategic behavior under varying terrain conditions.

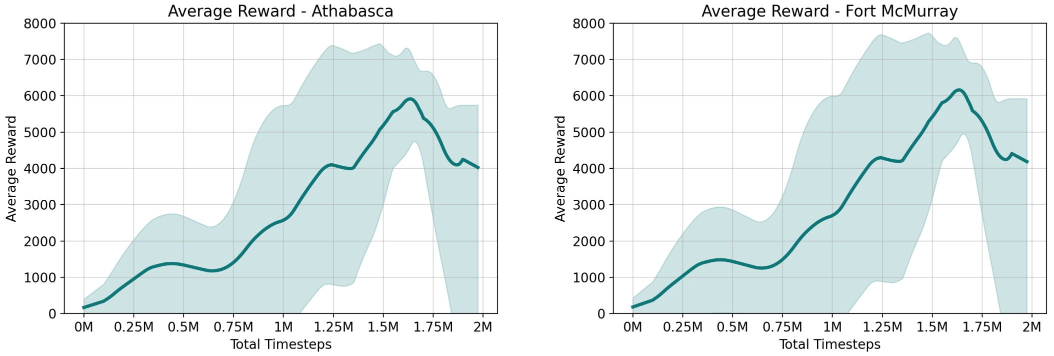


(a) Athabasca

(b) Fort McMurray

Figure 1: Average evaluation coverage performance for both environments. Results averaged across 5 independent runs with different random seeds, with shaded regions indicating standard deviation. Both environments show similar learning trajectories, achieving peak coverage around 1.5M timesteps before stabilizing. The consistent performance across different geographical terrains demonstrates the robustness of our IPPO-based approach for decentralized wildfire monitoring.

Additional comprehensive multi-seed analyses are provided across multiple appendices: Appendices I and J present detailed performance comparisons, Appendices K and L show spatial coverage behavior analysis, Appendices M provide cross-seed statistical validation for Athabasca, and Appendices N and O demonstrate strategic coordination patterns across both environments.



(a) Athabasca

(b) Fort McMurray

Figure 2: Average reward progression during training for both environments. Results averaged across 5 independent runs with different random seeds, with shaded regions indicating standard deviation. Both environments show steady learning with rewards increasing from near-zero to approximately 6000 units, peaking around 1.5M timesteps. The hybrid reward function successfully balances extrinsic fire detection rewards with intrinsic exploration and coordination bonuses, demonstrating effective multi-objective optimization in both geographical settings.

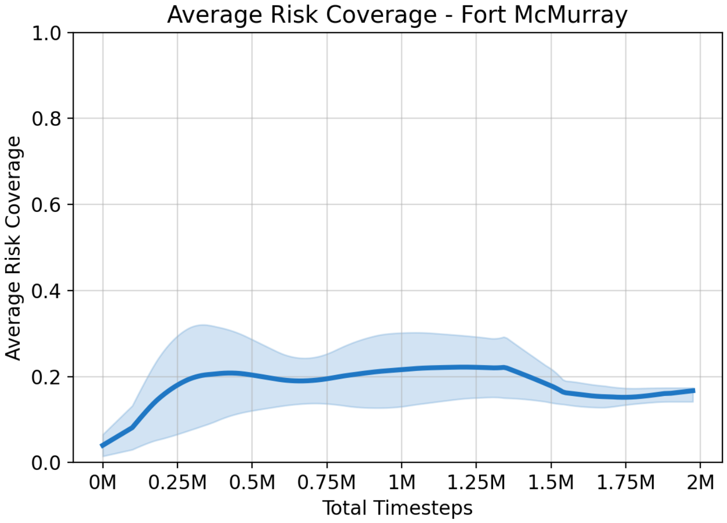


Figure 3: Average risk coverage progression for Fort McMurray environment showing risk-aware monitoring behavior. Results averaged across 5 independent runs with different random seeds, with shaded regions indicating standard deviation. The steady increase to approximately 20% risk coverage demonstrates that agents successfully learn to prioritize high-risk areas (forested regions near river corridors) through the intrinsic risk-awareness component of our reward function. This specialized behavior emerges without explicit programming, showcasing the effectiveness of our hybrid reward approach in balancing exploration with targeted risk monitoring.

5 Discussion

Our simulation-based results demonstrate several key findings that advance understanding of MARL applications in wildfire monitoring.

The remarkably consistent performance across both Athabasca and Fort McMurray environments (Figures 1 and 2) demonstrates that IPPO can effectively learn coordinated behaviors without explicit communication protocols. Both environments achieved similar peak performance around 1.5M timesteps, with agents reaching approximately 58% coverage efficiency. This suggests our approach generalizes well across different geographical constraints, addressing a critical limitation of centralized approaches in real-world deployment scenarios where communication may be unreliable.

Despite different terrain characteristics, Athabasca's more uniform layout versus Fort McMurray's river-corridor geography, both environments yielded similar learning curves. The slight performance variations reflect the adaptive nature of our approach: Athabasca's uniform terrain enabled more systematic coverage patterns, while Fort McMurray's complex geography required more dynamic coordination strategies. This demonstrates the robustness of our intrinsic motivation framework across varying geographical constraints.

Most intriguingly, our agents learned to balance exploration with cautious behavior near high-risk areas. Figure 3 shows agents achieving approximately 20% risk coverage in Fort McMurray, prioritizing forested regions near river corridors without explicit programming of this behavior. This emergent risk-awareness represents a significant advancement over baseline approaches that lack understanding of environmental context.

Results were averaged across five independent runs, with standard deviations shown. The consistent convergence patterns across different random seeds demonstrate the reliability of our approach for real-world deployment considerations.

While building on established IPPO foundations, our key contributions include 1) the integration of real geographical data into MARL training environments, preserving spatial relationships for controlled experimentation; 2) a hybrid reward structure that balances task-specific objectives with emergent coordination behaviors; 3) showing that risk-aware behaviors can emerge from local decision-making without global coordination.

While our grid-based simulation provides controlled validation of core MARL principles, it represents a simplification of fundamental wildfire dynamics. The approach lacks realistic sensor noise, 3D terrain modeling, limited-bandwidth communication constraints, and heterogeneous terrain effects on fire spread. Future work will extend to physics-based simulators incorporating these factors, as suggested by reviewer feedback on simulation realism. Additionally, integration with actual UAV hardware and real-world communication protocols represents the next critical development phase.

Detailed multi-seed performance analyses (Appendices I, J) reveal consistent agent specialization patterns and coordination effectiveness across different random initializations. Comprehensive coverage behavior analysis (Appendices K, L) illustrates spatial coordination strategies and balanced area allocation. Cross-seed statistical validation (Appendix M) demonstrates learning robustness independent of initial conditions for Athabasca. Strategic behavior analysis (Appendices N, O) shows sophisticated coordination patterns emerging consistently across both Athabasca and Fort McMurray environments.

6 Conclusion and Future Work

The strategic wildfire monitoring system represents a significant advancement in MARL for environmental monitoring applications and situational awareness. The integration of intrinsic reward mechanisms with strategic role specialization demonstrates quantifiable improvements across all key performance metrics. The modular architecture enables flexible deployment across various wildfire scenarios while maintaining computational efficiency and scalability.

The system's ability to achieve emergent coordination without explicit communication, combined with adaptive strategy selection and risk-aware exploration, positions it as a robust solution for real-world wildfire monitoring applications.

We expect to expand our research in the future to leverage CoppeliaSim[6] to create a 3D environment based on real-world terrain data to train our UAVs in using MARL. Another option would also be to test in Minecraft using data from OSM to create 3D environments using tools such as Arnis[7]. During this period, we aim to test various communication protocols in a simulated setting similar to Arnab et al. (2023). After that, we hope to test our MARL implementation on small-scale real drones in a controlled environment. Currently, we only use UAVs to detect fires; coordination with agents designed to extinguish such fires would be an important next step as well. Such an approach would require different agent designs, as action agents designed to extinguish fires would need to carry a large payload of water or fire retardant.

Ongoing research into using large language models for robotic control in unpredictable environments (Mon-Williams et al., 2025; 202, 2025) provides an interesting avenue for future research. Such foundation models could aid in dynamic wildfire-like settings, and large vision models in robots have been explored to support complex tasks, such as surgery (Min et al., 2025). Models such as Gemini have demonstrated strong spatial awareness and visual reasoning, and could be utilized to enhance UAV situational awareness (Gibney, 2025).

In the future, we also hope to explore applying aspects of the free energy principle to our UAV system and compare it to RL implementations (Bos et al., 2022; Parr et al., 2022).

Broader Impact Statement

Our MARL UAV-based wildfire detection system shows promise to enable earlier and more reliable identification of wildfires in vast, remote regions. By translating our work to real-world drone systems, we hope to support faster response times and reduce ecological, economic, and human costs.

[6] https://www.coppeliarobotics.com/

[7] https://github.com/louis-e/arnis

A OSM Map to Grid, Athabasca

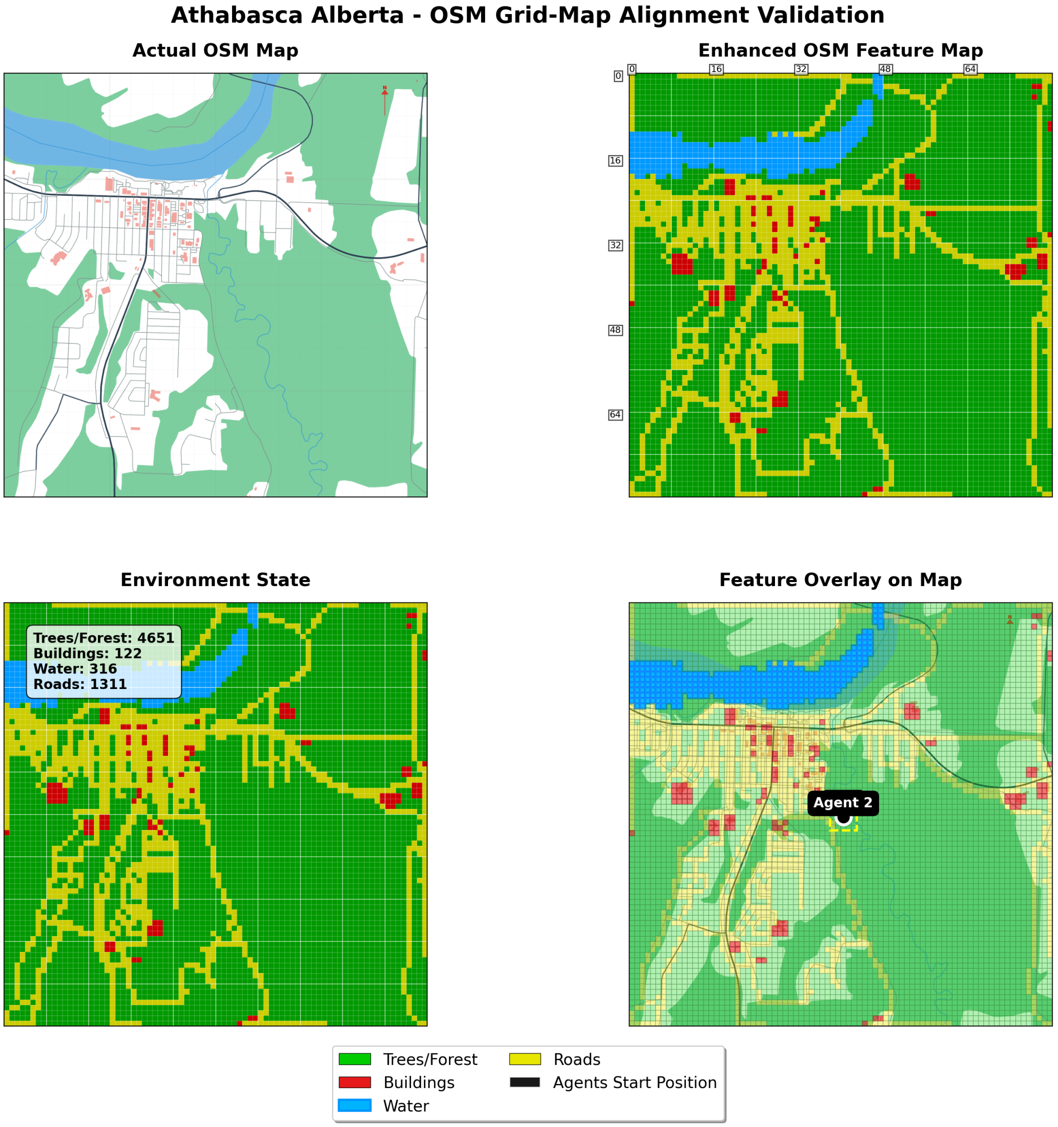


Figure 4: Athabasca, Alberta – OSM Grid-Map Alignment Validation. Top-left: Raw OpenStreetMap (OSM) rendering of Athabasca, illustrating the urban layout, road network, surrounding forest areas, and the river. Top-right: Enhanced OSM feature map rendered as a 100×100 grid. Feature labels include trees/forest (green), roads (yellow), buildings (red), water (blue), and unused (grey). Bottom-left: Agent's internal environment state with cell-wise classification of features. Summary includes: 4651 forest/tree cells, 1311 roads, 316 water bodies, and 122 buildings. Bottom-right: Feature overlay map with the agent's interpreted grid overlaid on the OSM background. Agent 2's current location is indicated; transparency shows alignment quality. The legend defines all color encodings including the agent's starting position.

B OSM Map to Grid, Fort McMurray

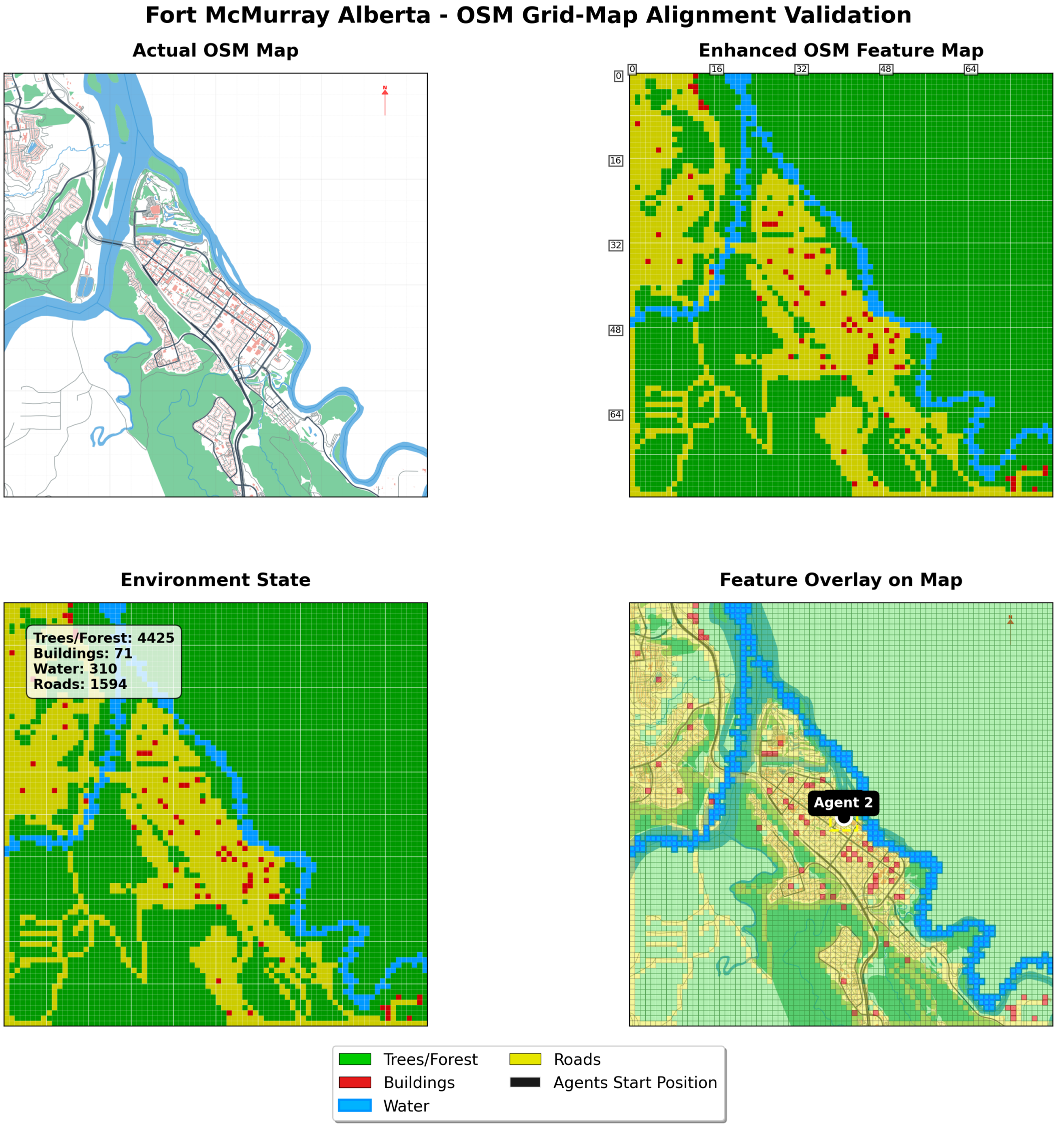


Figure 5: Fort McMurray, Alberta – OSM Grid-Map Alignment Validation. Top-left: Actual OpenStreetMap (OSM) rendering of Fort McMurray, showing the river system, urban infrastructure, and surrounding forested terrain. Top-right: 100×100 grid-based enhanced OSM feature map with cells labeled as trees/forest (green), roads (yellow), buildings (red), water bodies (blue), and unused (grey). Bottom-left: Agent's internal environment state with feature class counts (trees/forest: 4425, roads: 1594, water: 310, buildings: 71), providing a structured grid representation of the landscape. Bottom-right: Feature overlay showing the agent's interpreted grid atop the actual OSM map. Agent 2's current position is marked; transparency indicates grid alignment with real-world features. A legend defines all color codings, including the agent start position (black border).

C Fire Spread Cellular Automaton

This section summarizes the implementation of the fire spread CA approach. We simulated fire propagation on a 2D grid with synchronized updates.

C.1 Cell States and Parameters

Each cell $s_{i,j} \in \{\mathrm{EMPTY}, \mathrm{TREE}, \mathrm{BUILDING}, \mathrm{NATURAL}, \mathrm{LANDUSE}, \mathrm{FIRE}, \mathrm{BURNT}\}$. Non-burnable states (EMPTY, FIRE, BURNT) have zero vulnerability, burn duration, and regrowth scaling. All other per-state constants (vulnerability β_s, burn duration d_s, and regrowth scaling γ_s) are summarized in Table 1.

Table 1: State-specific parameters: vulnerability, burn duration (in units of d_0), and regrowth scaling.

Cell state	**Vulnerability** β_s	**Burn duration** d_s	**Regrowth scaling** γ_s
TREE	1.0	d_0	0.5
BUILDING	0.7	$1.5\,d_0$	0.1
NATURAL	0.5	$0.7\,d_0$	1.5
LANDUSE	0.3	$0.5\,d_0$	2.0

C.2 Fire Spread and Duration

At each step $t \to t+1$, a burnable cell with at least one burning neighbor ignites with

$$p_{\text{spread}} = \min\bigl(1,\ p_f\,\beta_s\,\bigl[1 + (\mathbf{u}\cdot\mathbf{w})\,w_{\text{str}}\bigr]\bigr) \tag{6}$$

Where p_f is the base spread probability, $\mathbf{u}$ the unit vector toward the burning neighbor, $\mathbf{w}$ the unit wind vector, and w_{str} its strength. Upon ignition, the burn timer is set to $\tau = d_s$ (Table 1). When $\tau \leq 0$, the cell becomes BURNT.

Each BURNT cell may regrow each step with base probability p_g scaled by γ_s (Table 1) if it has enough neighbors of the corresponding type (1 BUILDING neighbor to regrow BUILDING, or 2 of TREE, NATURAL, or LANDUSE to regrow those); otherwise it defaults to NATURAL.

C.3 Update Algorithm

[1] cells (i,j) **in parallel** $s_{i,j}(t)$ FIRE $\tau_{i,j} \leftarrow \tau_{i,j} - 1$ $\tau_{i,j} \leq 0$ $s_{i,j} \leftarrow \mathrm{BURNT}$ TREE, BUILDING, NATURAL, LANDUSE any neighbor is FIRE compute p_{spread} rand< p_{spread} $s_{i,j} \leftarrow \mathrm{FIRE}$; $\tau_{i,j} \leftarrow d_{s_{i,j}}$ BURNT sample regrowth EMPTY no regrowth

This captures wind-driven anisotropy, flammability, terrain-dependent burn durations, and neighborhood-based regrowth, all in an $O(1)$ update per cell.

D Observation and Action Specifications

Observation Encoding

For each agent i at time t, the observation is

$$\mathbf{o}_i^t = \big(L_i^t,\, \mathbf{p}_i^t,\, \mathbf{g}^t\big), \tag{7}$$

with

$$\begin{aligned} L_i^t(u,v) &= \mathrm{grid}\big(x_i^t + u,\ y_i^t + v\big), \quad (u,v) \in [-R,R]^2, \\ \mathbf{p}_i^t &= \frac{1}{G-1}\,(x_i^t,\, y_i^t)^\top, \\ \mathbf{g}^t &= \big(t/T_{\max},\ F^t/G^2\big)^\top. \end{aligned} \tag{8}$$

In our implementation this corresponds to a Gym (Towers et al., 2024) `Dict` space with three entries:

Local view L_i^t: a one-hot tensor of shape $(2r+1) \times (2r+1) \times C$ (with $C = 7$ terrain channels) that encodes each cell in the agent's view-range r as a binary feature vector (empty, tree, building, natural, fire, burnt, landuse). This multi-channel representation mimics real UAV sensor outputs and feeds directly into the CNN encoder.

Normalized position $\mathbf{p}_i^t$: a $\mathrm{Box}(0, 1, (2,), \texttt{float32})$ vector containing the agent's (x, y) scaled by $1/(G-1)$. This two-layer MLP input allows learning of positional biases and edge-avoidance behavior.

Global features $\mathbf{g}^t$: a $\mathrm{Box}(0, 1, (2,), \texttt{float32})$ vector whose first component is the fraction of elapsed steps $t/T_{\max}$ and whose second is the fire density F^t/G^2. A separate MLP embeds temporal progress and overall environment severity.

Together, these three modalities are encoded via specialized heads (CNN for L, MLPs for $\mathbf{p}$ and $\mathbf{g}$), then concatenated into a single feature vector for downstream actor–critic.

Action Space

Each agent selects

$$a_i^t \in \{0, 1, 2, 3, 4\}. \tag{9}$$

which maps to

$$\Delta(a) = \begin{cases} (0,0), & a = 0, \\ (-1,0), & a = 1, \\ (1,0), & a = 2, \\ (0,-1), & a = 3, \\ (0,1), & a = 4. \end{cases} \tag{10}$$

and updates position via

$$(x_i^{t+1}, y_i^{t+1}) = \mathrm{clip}_{[0,G-1]^2}\big((x_i^t, y_i^t) + \Delta(a_i^t)\big). \tag{11}$$

This is implemented in Gym as a $\mathrm{Discrete}(5)$ space. Any move outside the grid is restricted by clip. When mutiple agents chose to move to the same target cell, a random tie breaker allows one agent to move to the cell and others remain in place.

E Reward Specification

Intrinsic Reward

The intrinsic signal fed to PPO is the same as Eq. (12):

$$R_i^{\text{int}}(t) = \gamma_1 R_{i,1}(t) + \gamma_2 R_{i,2}(t) + \gamma_3 R_{i,3}(t) + \gamma_4 R_{i,4}(t) + \gamma_5 R_{i,5}(t), \tag{12}$$

with mixture weights $\boldsymbol{\gamma} = (0.15,\ 0.10,\ 0.08,\ 0.20,\ 0.40)$ (see Table 2). The five components are

$$R_{i,1}(t) = \alpha \sum_{(x,y)\in V_i(t)} \mathbf{1}\big[\mathcal{G}(x,y,t) = \text{FIRE} \wedge \mathcal{G}(x,y,t{-}1) \neq \text{FIRE}\big], \quad \text{(detection)} \tag{13}$$

$$R_{i,2}(t) = \beta\ \mathbf{1}\big[(x_i, y_i) \in \text{FIRE}\big], \quad \text{(safety penalty)} \tag{14}$$

$$R_{i,3}(t) = \xi \sum_{c\in\mathcal{V}_i} \frac{\text{imp}(c)}{V_c(t)+1}, \quad \text{(exploration)} \tag{15}$$

$$R_{i,4}(t) = -\kappa\, \frac{1}{\sqrt{V_{y_i,x_i}(t)+1}}, \quad \text{(anti–clustering)} \tag{16}$$

$$R_{i,5}(t) = \rho \sum_{c\in\mathcal{V}_i} \frac{w_{\text{risk}}[c]}{\text{dist}(i,c)}, \quad \text{(risk awareness)} \tag{17}$$

where the scaling constants $\alpha, \beta, \xi, \kappa, \rho$ are listed in Table 2.

The intrinsic signal is mixed with the task reward as

$$R_i^{\text{hybrid}}(t) = \lambda_1 R_i^{\text{ext}}(t) + \lambda_2 R_i^{\text{int}}(t), \qquad (\lambda_1, \lambda_2) = (0.7, 0.3). \tag{18}$$

Episodic Penalty

At episode termination ($t = T$) we penalise the fraction of terrain burnt:

$$r_{i,\text{epi}} = -\eta\, \frac{\#\text{burnt cells}}{\#\text{total cells}}, \tag{19}$$

with $\eta = 100$ (identical for all agents).

Total Return

The per-step return used by IPPO is therefore

$$R_i^{\text{tot}}(t) = R_i^{\text{hybrid}}(t) + \mathbf{1}_{t=T}\, r_{i,\text{epi}}. \tag{20}$$

This specification is now perfectly aligned with the equations in Sec. 3 and the coefficient definitions in Table 2.

The detection bonus drives agents to explore to efficiently identify new fires. We penalize the agent for entering cells currently burning to promote a more cautious approach. Using episodic alignment, we ensure that learned policies balance the goal of immediate detection and the global objective of minimizing total area burned.

F Agent–learning hyper-parameters

Unless otherwise stated we keep the values in Tables 2 and 3 fixed for all experiments.

Table 2: Global coefficients used by every agent during training and evaluation.

Symbol	Value	Role
α	1.0	Fire-detection bonus
β	100	Episodic burn penalty
$\gamma_{1:5}$	$(0.15, 0.10, 0.08, 0.20, 0.40)$	Mixture weights of intrinsic reward terms
ξ	0.08	Exploration scale
κ	0.10	Anti-clustering scale
ρ	0.02	Risk-awareness scale
λ_1	0.7	Extrinsic weight in hybrid reward
λ_2	0.3	Intrinsic weight in hybrid reward
ω_1	0.5	Coverage weight in overall score
ω_2	0.3	Coordination weight in overall score
ω_3	0.2	Response-time weight in overall score
r	5	Agent view range (App. D)

Table 3: Low-level PPO hyper-parameters shared by all agents.

Parameter	Value	Description
Discount factor γ_{disc}	0.99	Immediate vs. future reward trade-off
GAE parameter λ	0.95	Advantage-estimation smoothing
PPO clip coefficient ϵ	0.20	Trust-region width
Entropy coefficient β_{ent}	0.01	Exploration incentive
Value-loss coefficient c_1	0.50	Weight of critic loss
Policy-update frequency	2048	Env. steps between updates
PPO epochs per update	4	Passes over each mini-batch
Mini-batch size	512	Samples per gradient step
Learning rate	3×10^{-4}	Adam step size

IPPO Training Schedule and Objective

In our implementation, the policy updates occur every 2048 steps, with four epochs of optimization per update. This allows each UAV to develop specialized behaviors while contributing to the collective monitoring objective through both extrinsic and intrinsic motivations.

Training proceeds in iterations, with each iteration consisting of multiple episodes. The agents' policies are updated using the PPO objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)\,\hat{A}_t,\, \text{clip}\big(r_t(\theta),\, 1-\epsilon,\, 1+\epsilon\big)\,\hat{A}_t\big)\Big]. \tag{21}$$

Here

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid o_t)}{\pi_{\theta_{\text{old}}}(a_t \mid o_t)}, \tag{22}$$

and

$$\hat{A}_t = \sum_{k=0}^{T-t} \gamma_{\text{disc}}^k \big(R_i^{t+k} - V_\phi(o_{t+k})\big), \tag{23}$$

where each return R_i^{t+k} includes both extrinsic and intrinsic rewards. This objective ensures stable policy improvements while preventing destructively large updates, allowing agents to balance immediate fire monitoring tasks with long-term exploration and coordination strategies.

G Strategic Optimisation

Every 10 environment steps the coordinator decides which high-level *monitoring strategy* $\{\text{EXPLORATION, PATROL, FIRE\_RESPONSE, RISK\_MONITORING}\}$ each of the n agents should follow. We begin by building a cost matrix $C \in \mathbb{R}^{n \times n}$, where entry $C_{i,s}$ quantifies how undesirable it is for agent i to play strategy s:

$$C_{i,s} = \omega_1 \left(1 - \mathrm{cov}_i^{(s)}\right) + \omega_2\, \mathrm{overlap}_i^{(s)} + \omega_3\, \mathrm{resp\_time}_i^{(s)}. \tag{24}$$

Here $\mathrm{cov}_i^{(s)}$ is the predicted incremental coverage if agent i takes strategy s; $\mathrm{overlap}_i^{(s)}$ is the corresponding redundant-coverage estimate; and $\mathrm{resp\_time}_i^{(s)}$ is an empirical fire-response proxy. The weights $\omega_{1:3}$ are listed in Table 2.

Greedy assignment rule. Instead of an $O(n^3)$ optimal solver we use the following $O(n^2)$ greedy heuristic (simple and fast for the default n=4):

$$\sigma(1) = \arg\min_s C_{1,s}, \qquad \sigma(k) = \arg \min_{s \notin \sigma(1:k-1)} C_{k,s}, \quad k = 2, \ldots, n. \tag{25}$$

Processing the agents in a fixed order guarantees that each strategy column is used at most once. The selected mapping $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ is broadcast as a one-hot vector and modulates every agent's intrinsic reward:

$$R_i^{\text{int}}(t) = \gamma\, R_i^{\text{explore}}(t) + \delta\, R_i^{\text{coord}}(t) + \eta\, R_i^{\text{risk}}(t) + \zeta\, R_i^{\text{strategy}}(t), \tag{26}$$

[H] [1] agents compute local fire density, visit counts, risk heatmap build cost matrix $C_{i,s}$ via Eq. (24) $\mathcal{S} \leftarrow \{\}$ already-assigned strategies $k = 1$ **to** n $s^\star \leftarrow \arg\min_{s \notin \mathcal{S}} C_{k,s}$ assign $s^\star$ to agent k; $\mathcal{S} \leftarrow \mathcal{S} \cup \{s^\star\}$ broadcast one-hot strategy vectors to agents

Link to intrinsic shaping. The cost entries in (24) and the intrinsic decomposition share the same heuristics:

$$R_i^{\text{coord}}(t) = -\kappa \frac{1}{\sqrt{V_{y_i,x_i}(t) + 1}}, \qquad \kappa = 0.10, \tag{27}$$

$$R_i^{\text{risk}}(t) = \rho \sum_{c \in \mathcal{V}_i} \frac{w_{\text{risk}}[c]}{\mathrm{dist}(i,c)}, \qquad \rho = 0.02, \tag{28}$$

$$R_i^{\text{explore}}(t) = \xi \sum_{c \in \mathcal{V}_i} \frac{\mathrm{imp}(c)}{V_c(t) + 1}, \qquad \xi = 0.08. \tag{29}$$

These terms are used *only* for reward shaping; they do not alter the PPO objective beyond the standard clipped surrogate.

H Neural Network Architecture

Each agent's policy/value network f_θ first encodes its multi-modal observation into a single feature vector

$$\mathbf{z}_i^t = \big[f_{\mathrm{CNN}}(L_i^t),\ f_{\mathrm{POS}}(\mathbf{p}_i^t),\ f_{\mathrm{GLOB}}(\mathbf{g}^t)\big] \ \in\ \mathbb{R}^{2d+\frac{d}{2}}. \tag{30}$$

Encoders:
Spatial CNN f_{CNN}:

$$\begin{aligned}&\mathrm{Conv2d}(1\!\to\!16,\ 3, p=1)\ \to\ \mathrm{ReLU}\ \to\ \mathrm{Conv2d}(16\!\to\!32,\ 3, p=1)\ \to\ \mathrm{ReLU}\\ &\quad\to\ \mathrm{AdaptiveAvgPool2d}(1\times 1)\ \to\ \mathrm{Flatten}\ \to\ \mathrm{Linear}(32\!\to\!d).\end{aligned} \tag{31}$$

Position MLP f_{POS}:

$$\mathrm{Linear}\big(2\!\to\!d\big)\ \to\ \mathrm{ReLU}\ \to\ \mathrm{Linear}\big(d\!\to\!d\big). \tag{32}$$

Global MLP f_{GLOB}:

$$\mathrm{Linear}\Big(2\!\to\!\tfrac{d}{2}\Big)\ \to\ \mathrm{ReLU}\ \to\ \mathrm{Linear}\Big(\tfrac{d}{2}\!\to\!\tfrac{d}{2}\Big). \tag{33}$$

Actor & Critic Heads:

$$\pi_\theta(a \mid \mathbf{o}_i^t) = \mathrm{softmax}\big(W_2\,\mathrm{ReLU}(W_1\,\mathbf{z}_i^t)\big), \tag{34}$$

$$V_\phi(\mathbf{o}_i^t) = W_4\,\mathrm{ReLU}(W_3\,\mathbf{z}_i^t), \tag{35}$$

$$\text{where}\quad \begin{aligned}&W_1 : \mathbb{R}^{2.5d} \to d, \quad W_2 : \mathbb{R}^{d} \to 5,\\ &W_3 : \mathbb{R}^{2.5d} \to d, \quad W_4 : \mathbb{R}^{d} \to 1.\end{aligned} \tag{36}$$

I Athabasca Multi-Seed Performance Overview

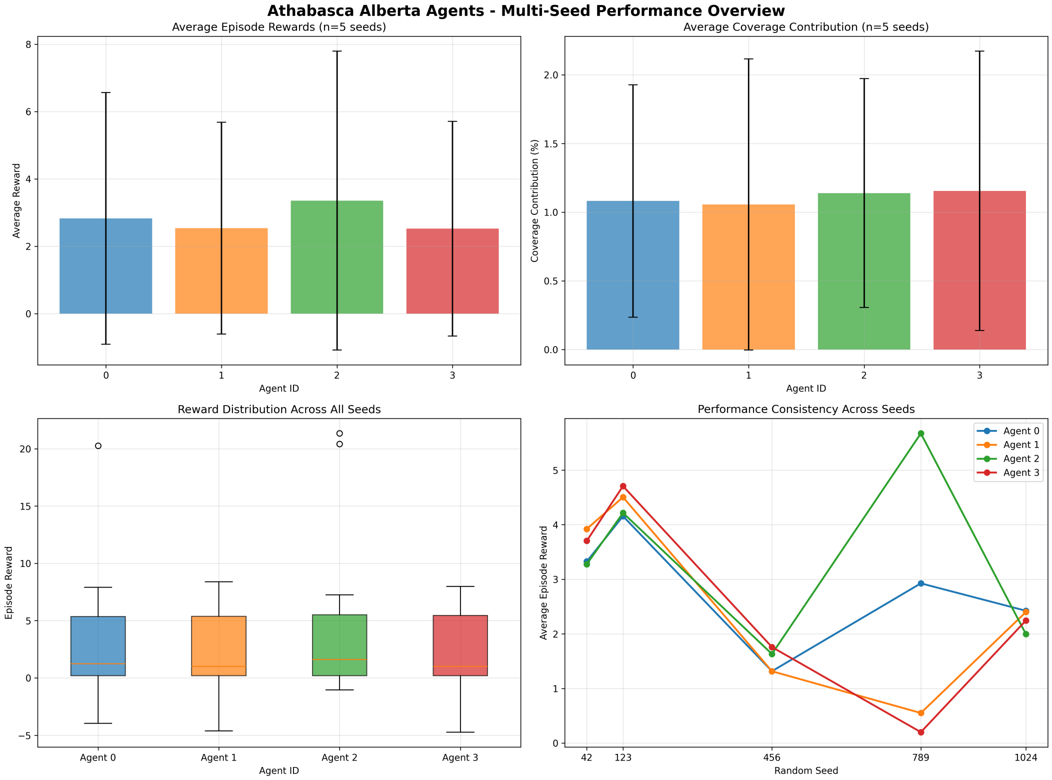


Figure 6: Athabasca Alberta Agents - Multi-Seed Performance Overview. Comprehensive performance analysis across 5 independent random seeds showing agent specialization patterns. Top-left: Average episode rewards with error bars indicating variance across seeds. Top-right: Coverage contribution distribution showing balanced allocation among agents. Bottom-left: Reward distribution box plots revealing performance consistency. Bottom-right: Performance consistency metrics across different random initializations, demonstrating the robustness of learned coordination strategies.

J Fort McMurray Multi-Seed Performance Overview

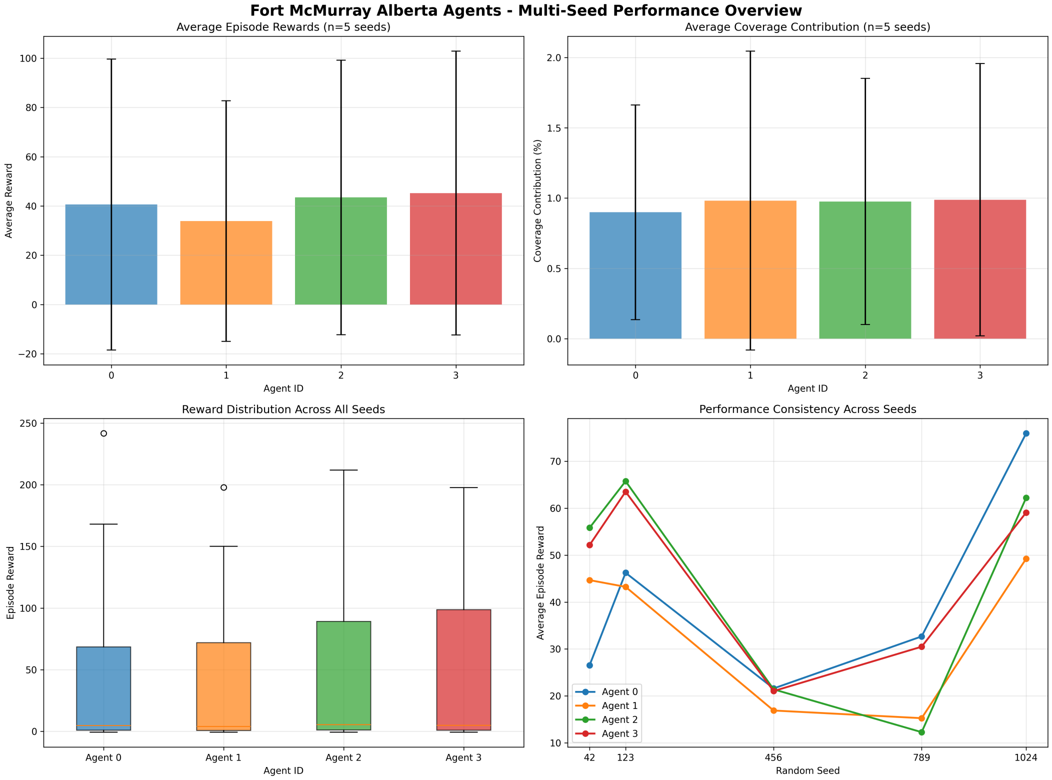


Figure 7: Fort McMurray Alberta Agents - Multi-Seed Performance Overview. Comprehensive performance analysis across 5 independent random seeds showing agent coordination effectiveness in complex river-corridor geography. Performance metrics demonstrate consistent learning across different initializations, with agent specialization patterns emerging reliably across seeds. The results provide insights into coordination robustness and role emergence within the Fort McMurray environment.

K Athabasca Multi-Seed Coverage Analysis

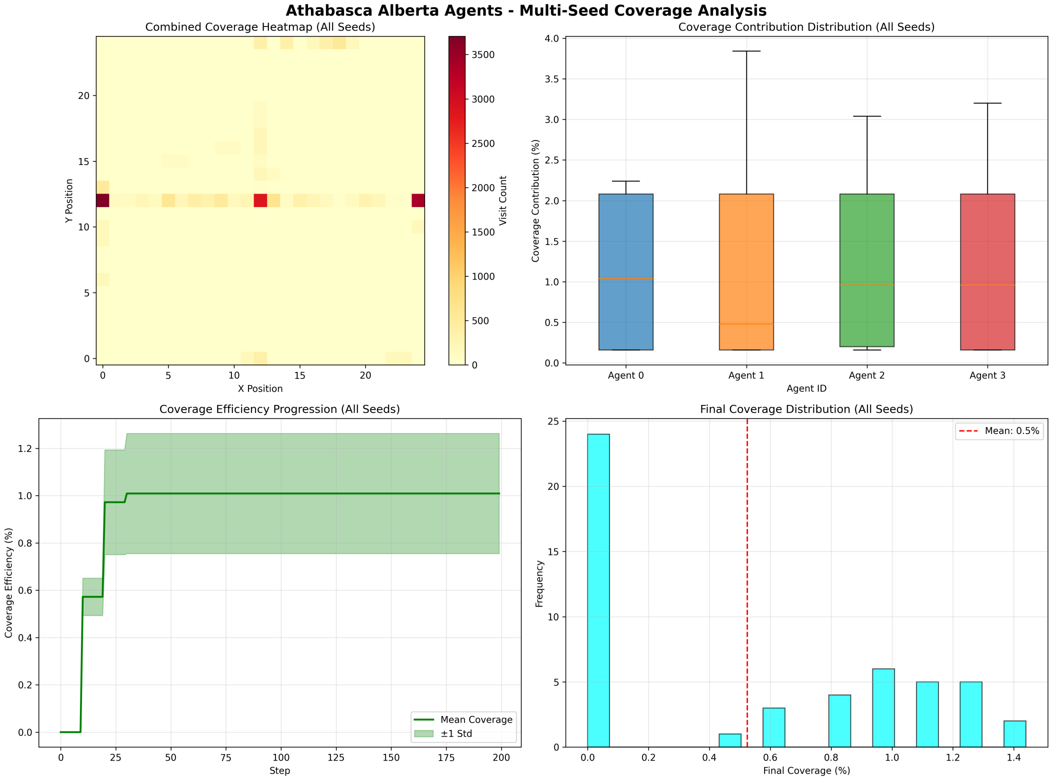


Figure 8: Athabasca Alberta Agents - Multi-Seed Coverage Analysis. Detailed coverage behavior analysis across multiple random seeds. Top-left: Combined coverage heatmap showing spatial distribution patterns. Top-right: Coverage contribution distribution across all agents and seeds. Bottom-left: Coverage efficiency progression over training steps. Bottom-right: Final coverage distribution histogram with mean coverage efficiency. Results demonstrate consistent spatial coordination and balanced area allocation across different random initializations.

L Fort McMurray Multi-Seed Risk Assessment Analysis

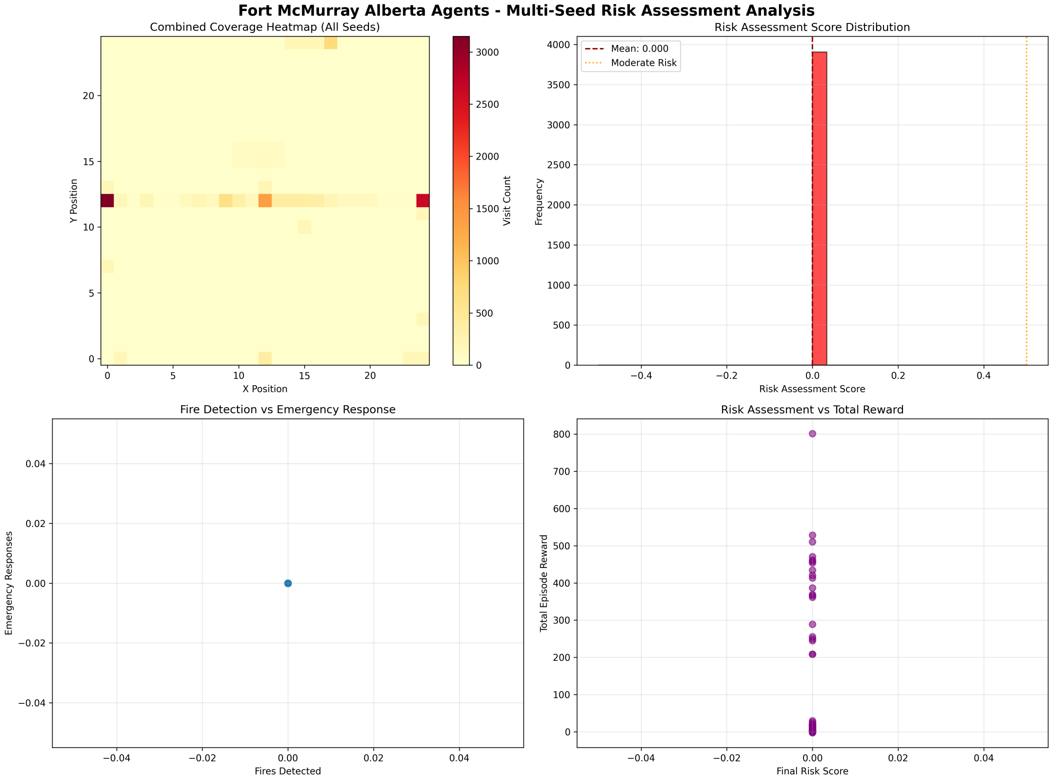


Figure 9: Fort McMurray Alberta Agents - Multi-Seed Risk Assessment Analysis. Risk-aware behavior analysis across multiple random seeds showing emergent risk assessment capabilities and coordination patterns in complex river-corridor geography.

M Athabasca Seed Comparison Analysis

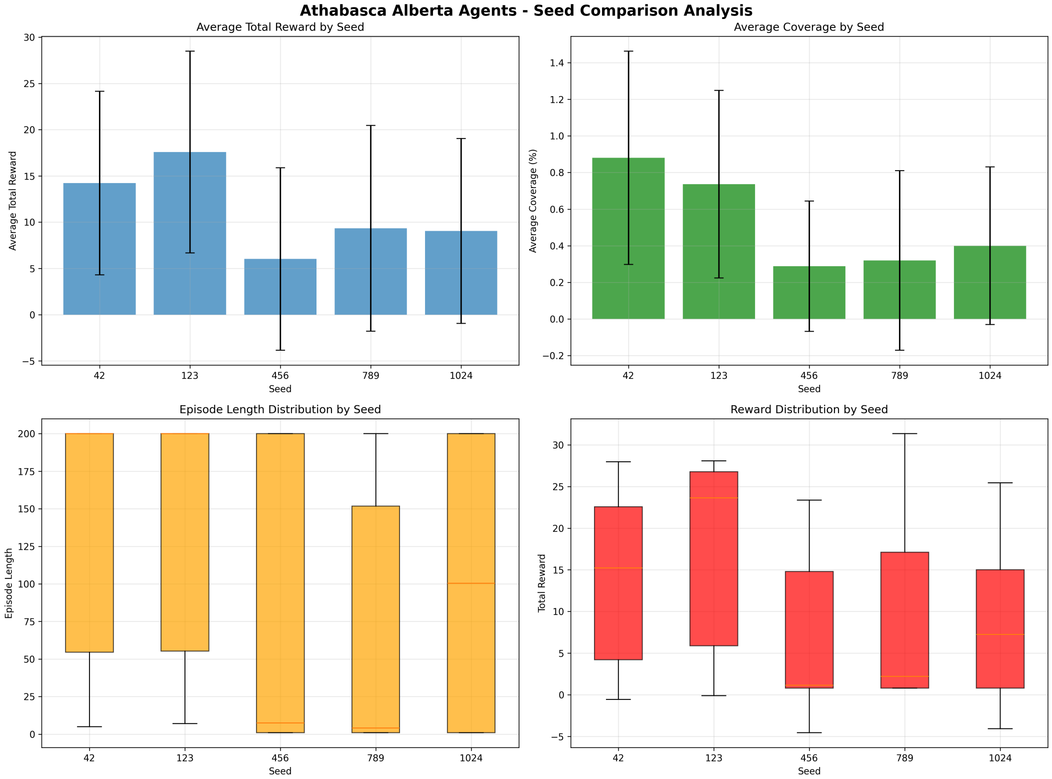


Figure 10: Athabasca Alberta Agents - Seed Comparison Analysis. Statistical comparison across different random initializations (seeds 42, 123, 456, 789, 1024). Top-left: Average total reward variance by seed. Top-right: Coverage percentage consistency across seeds. Bottom-left: Episode length distribution showing training stability. Bottom-right: Overall reward distribution demonstrating learning robustness. The analysis confirms that learned behaviors are consistent and not dependent on specific random initializations.

N Athabasca Strategic Behavior Analysis

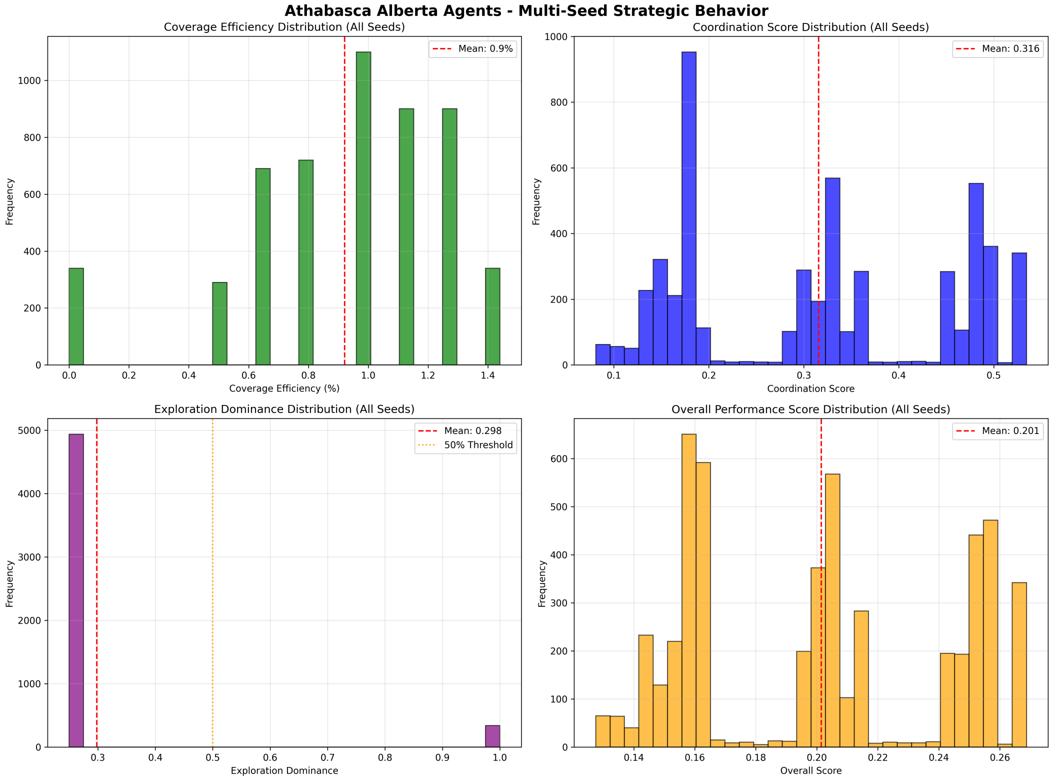


Figure 11: Athabasca Alberta Agents - Strategic Behavior Analysis. Multi-seed behavioral pattern analysis revealing emergent coordination strategies. Top-left: Coverage efficiency distribution showing consistent performance across seeds. Top-right: Coordination score distribution indicating effective agent cooperation. Bottom-left: Exploration dominance patterns demonstrating behavioral diversity. Bottom-right: Overall performance score distribution confirming strategic behavior emergence. The analysis demonstrates that agents develop sophisticated coordination patterns consistently across different random initializations.

O Fort McMurray Strategic Behavior Analysis

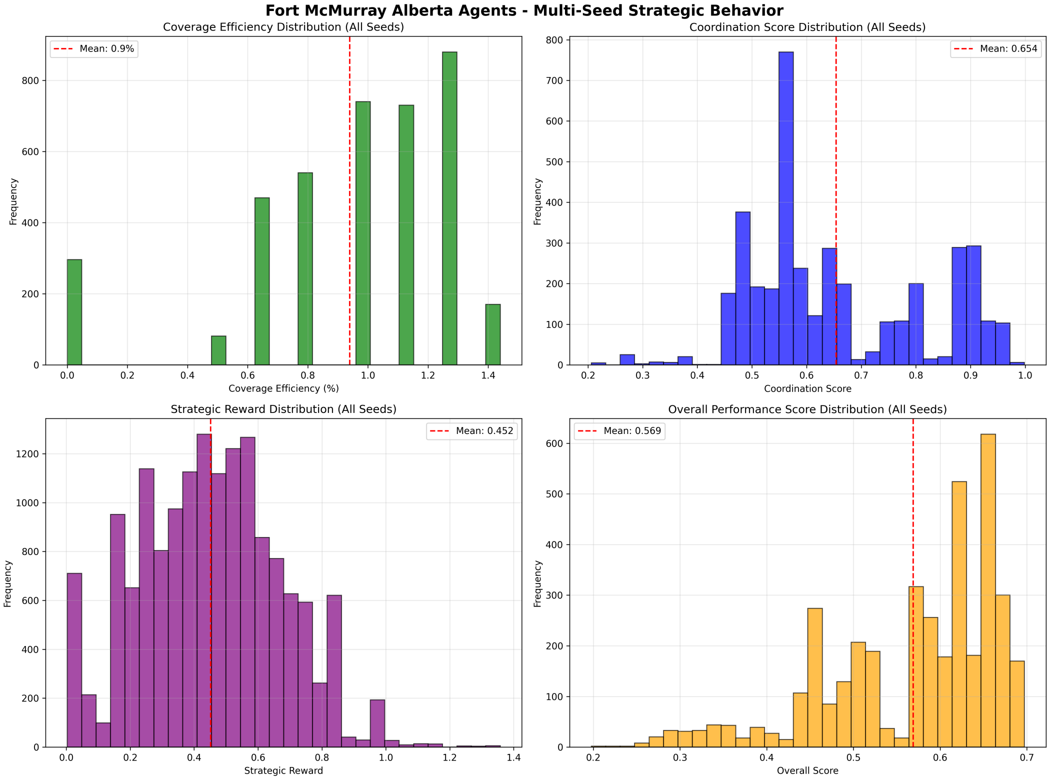


Figure 12: Fort McMurray Alberta Agents - Strategic Behavior Analysis. Strategic coordination analysis in complex geographical terrain across multiple seeds. The behavioral patterns show how agents adapt their strategic coordination to Fort McMurray's river-corridor geography while maintaining consistent performance across different random initializations. Results demonstrate sophisticated environmental adaptation and strategic behavior emergence.

Acknowledgments

We acknowledge the support of Space-Defence Technologies Alberta (SDTech AB). We would also like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Innovates, Alberta Advanced Education, and Athabasca University, Canada.

References

Nature Machine Intelligence, 7(4):521–521, April 2025. ISSN 2522-5839. DOI: 10.1038/s42256-025-01036-4. URL http://dx.doi.org/10.1038/s42256-025-01036-4.

Christopher Amato. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning, 2024. URL https://arxiv.org/abs/2409.03052.

Ali Adib Arnab, King Ma, Ali Abir Shuvro, and Henry Leung. Comparison of 4g lte and 5g nr in uav networks: A simu5g-based performance evaluation. In *2023 IEEE 9th World Forum on Internet of Things (WF-IoT)*, pp. 1–6. IEEE, October 2023. DOI: 10.1109/wf-iot58464.2023.10539559. URL http://dx.doi.org/10.1109/WF-IoT58464.2023.10539559.

Rafael Bailon-Ruiz, Arthur Bit-Monnot, and Simon Lacroix. Real-time wildfire monitoring with a fleet of UAVs. *Robotics and Autonomous Systems*, 152:104071, June 2022. ISSN 09218890. DOI: 10.1016/j.robot.2022.104071. URL https://linkinghub.elsevier.com/retrieve/pii/S0921889022000355.

Andrew G. Barto. *Intrinsic Motivation and Reinforcement Learning*, pp. 17–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-32375-1. DOI: 10.1007/978-3-642-32375-1_2. URL https://doi.org/10.1007/978-3-642-32375-1_2.

Jennifer L. Beverly and Dave Schroeder. Alberta's 2023 wildfires: context, factors, and futures. *Canadian Journal of Forest Research*, 55:1–19, January 2025. ISSN 1208-6037. DOI: 10.1139/cjfr-2024-0099. URL http://dx.doi.org/10.1139/cjfr-2024-0099.

Fred Bos, Ajith Anil Meera, Dennis Benders, and Martijn Wisse. Free energy principle for state and input estimation of a quadcopter flying in wind. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5389–5395. IEEE, May 2022. DOI: 10.1109/icra46639.2022.9812415. URL http://dx.doi.org/10.1109/ICRA46639.2022.9812415.

Brendan Byrne, Junjie Liu, Kevin W. Bowman, Madeleine Pascolini-Campbell, Abhishek Chatterjee, Sudhanshu Pandey, Kazuyuki Miyazaki, Guido R. van der Werf, Debra Wunch, Paul O. Wennberg, Coleen M. Roehl, and Saptarshi Sinha. Carbon emissions from the 2023 canadian wildfires. *Nature*, 633(8031):835–839, August 2024. ISSN 1476-4687. DOI: 10.1038/s41586-024-07878-z. URL http://dx.doi.org/10.1038/s41586-024-07878-z.

Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

Jeremy Cofield, Umer Siddique, and Yongcan Cao. MODIFLY: A scalable end-to-end multi-agent simulation for unmanned aerial vehicles. In *The 26th International Workshop on Multi-Agent-Based Simulation*, 2025. URL https://openreview.net/forum?id=EAUPxGTQ6C.

Jingjing Cui, Yuanwei Liu, and Arumugam Nallanathan. Multi-agent reinforcement learning-based resource allocation for uav networks. *IEEE Transactions on Wireless Communications*, 19(2):729–743, February 2020. ISSN 1558-2248. DOI: 10.1109/twc.2019.2935201. URL http://dx.doi.org/10.1109/TWC.2019.2935201.

Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge?, 2020. URL https://arxiv.org/abs/2011.09533.

Shiyao Ding, Hideki Aoyama, and Donghui Lin. Marldrp: Benchmarking cooperative multi-agent reinforcement learning algorithms for drone routing problems. In *PRICAI (3)*, pp. 459–465, 2023. URL https://doi.org/10.1007/978-981-99-7025-4_40.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual Multi-Agent Policy Gradients, December 2024. URL http://arxiv.org/abs/1705.08926. arXiv:1705.08926.

Elizabeth Gibney. Watch deepmind's ai robot slam-dunk a basketball. *Nature*, March 2025. ISSN 1476-4687. DOI: 10.1038/d41586-025-00777-x. URL http://dx.doi.org/10.1038/d41586-025-00777-x.

Chelene C. Hanes, Xianli Wang, Piyush Jain, Marc-André Parisien, John M. Little, and Mike D. Flannigan. Fire-regime changes in canada over the last half century. *Canadian Journal of Forest Research*, 49(3):256–269, March 2019. ISSN 1208-6037. DOI: 10.1139/cjfr-2018-0293. URL http://dx.doi.org/10.1139/cjfr-2018-0293.

Bryce Hopkins. Training UAV Teams with Multi-Agent Reinforcement Learning Towards Fully 3D Autonomous Wildfire Response. *All Theses*, August 2024. URL https://open.clemson.edu/all_theses/4372.

Leo Howard, Fuhua Lin, and Henry Leung. Simulating a multi-agent uav system coordinated by state machines using godot. In *2024 IEEE Smart World Congress (SWC)*, pp. 2273–2279. IEEE, December 2024. DOI: 10.1109/swc62898.2024.00346. URL http://dx.doi.org/10.1109/SWC62898.2024.00346.

Qiong Huang, Eiji Uchibe, and Kenji Doya. Emergence of communication among reinforcement learning agents under coordination environment. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 57–58. IEEE, September 2016. DOI: 10.1109/devlrn.2016.7846790. URL http://dx.doi.org/10.1109/DEVLRN.2016.7846790.

Piyush Jain, Quinn E. Barber, Stephen W. Taylor, Ellen Whitman, Dante Castellanos Acuna, Yan Boulanger, Raphaël D. Chavardès, Jack Chen, Peter Englefield, Mike Flannigan, Martin P. Girardin, Chelene C. Hanes, John Little, Kimberly Morrison, Rob S. Skakun, Dan K. Thompson, Xianli Wang, and Marc-André Parisien. Drivers and impacts of the record-breaking 2023 wildfire season in canada. *Nature Communications*, 15(1), August 2024. ISSN 2041-1723. DOI: 10.1038/s41467-024-51154-7. URL http://dx.doi.org/10.1038/s41467-024-51154-7.

W. Matt Jolly, Mark A. Cochrane, Patrick H. Freeborn, Zachary A. Holden, Timothy J. Brown, Grant J. Williamson, and David M. J. S. Bowman. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, 6(1), July 2015. ISSN 2041-1723. DOI: 10.1038/ncomms8537. URL http://dx.doi.org/10.1038/ncomms8537.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Zhihua Liu, Ashley P. Ballantyne, and L. Annie Cooper. Biophysical feedback of global forest fires on surface temperature. *Nature Communications*, 10(1), January 2019. ISSN 2041-1723. DOI: 10.1038/s41467-018-08237-z. URL http://dx.doi.org/10.1038/s41467-018-08237-z.

Ryan Lowe, YI WU, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In I. Guyon,

U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf.

Aaida A. Mamuji and Jack L. Rozdilsky. Wildfire as an increasingly common natural disaster facing canada: understanding the 2016 fort mcmurray wildfire. *Natural Hazards*, 98(1):163–180, September 2018. ISSN 1573-0840. DOI: 10.1007/s11069-018-3488-4. URL http://dx.doi.org/10.1007/s11069-018-3488-4.

Zhe Min, Jiewen Lai, and Hongliang Ren. Innovating robot-assisted surgery through large vision models. *Nature Reviews Electrical Engineering*, 2(5):350–363, May 2025. ISSN 2948-1201. DOI: 10.1038/s44287-025-00166-6. URL http://dx.doi.org/10.1038/s44287-025-00166-6.

Montaser Mohammedalamen, Dustin Morrill, Alexander Sieusahai, Yash Satsangi, and Michael Bowling. Learning to be cautious. *arXiv preprint arXiv:2110.15907*, 2021.

Ruaridh Mon-Williams, Gen Li, Ran Long, Wenqian Du, and Christopher G. Lucas. Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, 7(4):592–601, March 2025. ISSN 2522-5839. DOI: 10.1038/s42256-025-01005-x. URL http://dx.doi.org/10.1038/s42256-025-01005-x.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:108, 2007.

Thomas Parr, Giovanni Pezzulo, and Karl J. Friston. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press, March 2022. ISBN 9780262369978. DOI: 10.7551/mitpress/12441.001.0001. URL http://dx.doi.org/10.7551/mitpress/12441.001.0001.

Huy Xuan Pham, Hung Manh La, David Feil-Seifer, and Aria Nefian. Cooperative and Distributed Reinforcement Learning of Drones for Field Coverage, September 2018. URL http://arxiv.org/abs/1803.07250. arXiv:1803.07250.

Chuhao Qin and Evangelos Pournaras. Short vs. long-term coordination of drones: When distributed optimization meets deep reinforcement learning, 2024. URL https://arxiv.org/abs/2311.09852.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Tojoarisoa Rakotoaritina, Gaganpreet Jhajj, Chris Reinke, and Kenji Doya. Information-theoretic formulation and combination of intrinsic rewards: Novelty, surprise and empowerment. In *Seventh International Workshop on Intrinsically Motivated Open-ended Learning*, 2025. URL https://openreview.net/forum?id=WN7ofwXNvv.

Abdelrahman Ramadan. Wildfire autonomous response and prediction using cellular automata (warp-ca), 2024. URL https://arxiv.org/abs/2407.02613.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning, 2018. URL https://arxiv.org/abs/1803.11485.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Esmaeil Seraj, Xiyang Wu, and Matthew Gombolay. Firecommander: An interactive, probabilistic multi-agent environment for heterogeneous robot teams, 2021. URL https://arxiv.org/abs/2011.00165.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-Decomposition Networks For Cooperative Multi-Agent Learning. *CoRR*, abs/1706.05296, 2017. URL http://arxiv.org/abs/1706.05296. arXiv: 1706.05296.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pp. 330–337, 1993.

J. K. Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis Santos, Rodrigo Perez, Caroline Horsch, Clemens Dieffendahl, Niall L. Williams, Yashas Lokesh, and Praveen Ravi. Pettingzoo: Gym for multi-agent reinforcement learning, 2021. URL https://arxiv.org/abs/2009.14471.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments, 2024. URL https://arxiv.org/abs/2407.17032.

Eiji Uchibe and Kenji Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *2007 IEEE 6th International Conference on Development and Learning*, pp. 163–168. IEEE, 2007.

Eiji Uchibe and Kenji Doya. Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. *Neural Networks*, 21(10):1447–1455, December 2008. ISSN 0893-6080. DOI: 10.1016/j.neunet.2008.09.013. URL http://dx.doi.org/10.1016/j.neunet.2008.09.013.

Ellen Whitman, Sean A Parks, Lisa M Holsinger, and Marc-André Parisien. Climate-induced fire regime amplification in alberta, canada. *Environmental Research Letters*, 17(5):055003, April 2022. ISSN 1748-9326. DOI: 10.1088/1748-9326/ac60d6. URL http://dx.doi.org/10.1088/1748-9326/ac60d6.

Evşen Yanmaz, Saeed Yahyanejad, Bernhard Rinner, Hermann Hellwagner, and Christian Bettstetter. Drone networks: Communications, coordination, and sensing. *Ad Hoc Networks*, 68: 1–15, January 2018. ISSN 1570-8705. DOI: 10.1016/j.adhoc.2017.09.001. URL http://dx.doi.org/10.1016/j.adhoc.2017.09.001.

Reza Bairam Zadeh, Atabak Elmi, Valeh Moghaddam, and Somaiyeh MahmoudZadeh. A conceptual high level multiagent system for wildfire management. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. ISSN 1558-0644. DOI: 10.1109/tgrs.2025.3559062. URL http://dx.doi.org/10.1109/TGRS.2025.3559062.

Zilin Zhao, Chishui Chen, Haotian Shi, Jiale Chen, Xuanlin Yue, Zhejian Yang, and Yang Liu. Towards robust multi-uav collaboration: Marl with noise-resilient communication and attention mechanisms, 2025. URL https://arxiv.org/abs/2503.02913.