The Efficacy of Dense Local Incentives vs. Sparse Global Penalties in Multi-Agent Wildfire Suppression

Abdul Rahman Hussain Siddique (ah244)
Anirudh Ramesh (aramesh8)
Yashwanth Pulimi (ypulimi)

November 23, 2025

Abstract

In Multi-Agent Reinforcement Learning (MARL) for disaster response, designing effective reward structures is a critical challenge. We investigated this by training a squad of 3 helicopters in the `FirecastRL` environment using multiple strategies: *Cooperative* (Global Penalty), *Greedy* (Local Reward), *Curriculum Learning*, and advanced architectures including *CNN Vision* and *Distance Shaping* ("Heat Seeker"). Contrary to the standard hypothesis, pure Cooperative agents failed to contain the fire due to the credit assignment problem. The Greedy approach significantly outperformed them, reducing burnt area by $\approx 14\%$. Further experiments with CNNs and distance-based shaping improved agent tracking behavior but revealed a fundamental limitation in the discrete action space, leading to "wall-hugging" behavior. Our results suggest that in high-dimensional spatial environments, dense local feedback is superior to sparse global objectives, but solving the navigation problem requires continuous control.

1 Introduction

Wildfire suppression is a race against time. While single-agent RL has shown promise, it struggles to cover large geographical areas. This project explores scaling to multi-agent systems ($N = 3$). The core challenge in MARL is reward shaping: how do we encourage agents to work together? We tested the hypothesis that agents sharing a global penalty for fire spread would learn superior containment strategies compared to agents acting purely on self-interest.

2 Methodology

2.1 Environment Setup

We utilized the `FirecastRL` environment modified to support multiple agents controlled by a centralized PPO policy (Stable-Baselines3). The state space includes a 160×240 grid of fire ignition times and agent coordinates. The action space is a `MultiDiscrete` vector.

2.2 Experimental Conditions

We conducted four distinct experiments to isolate the drivers of agent performance:

1. **Cooperative Agent:** Received +10 for extinguishing, -2 for wasting water, and a global penalty of $-0.1 \times$ (Total Burning Cells).

2. **Greedy Agent:** Received +10 for extinguishing and -2 for wasting water. **Zero penalty** for fire spread. Agents only maximize individual hits.

3. **Curriculum Agent:** Started with Greedy rewards to learn basic mechanics, linearly transitioning to Cooperative rewards over time.

4. **V3 (CNN Vision) & V4 (Heat Seeker):** Addressed the "blindness" of flat-vector agents by reshaping inputs to $(1, H, W)$ for a Convolutional Neural Network (CNN) and adding distance-based reward shaping to lure agents toward the fire.

3 Results

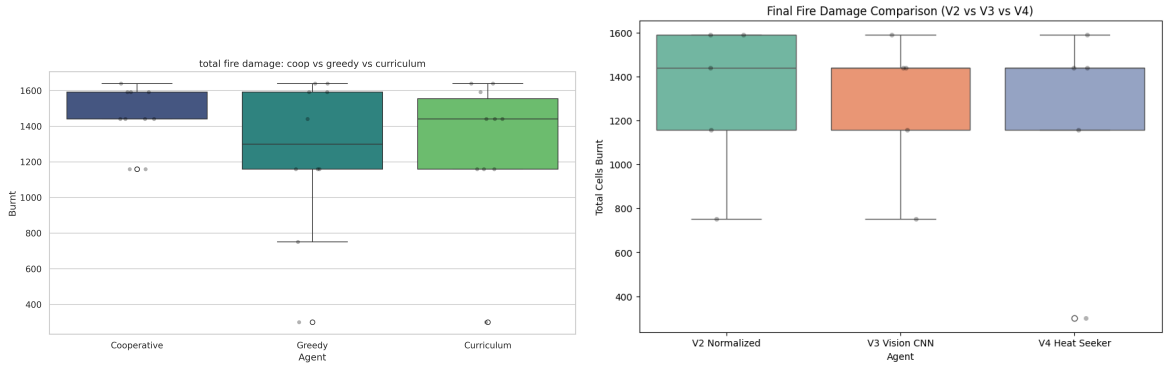
We evaluated all models over 10 deterministic episodes. The Greedy strategy surprisingly yielded the best raw performance, while advanced methods highlighted behavioral limitations.

Strategy	**Avg Cells Burnt** (↓)	**Reliability (Std Dev)**	**Avg Steps**
Cooperative	1449.8	172.01	728.1
Greedy	**1243.4**	440.78	432.1
Curriculum	1297.3	399.12	519.2
V3 (CNN Vision)	1408.6	396.56	641.1
V4 (Heat Seeker)	1323.2	417.72	582.2

Table 1: Comparative performance statistics. Greedy agents saved the most forest, but Curriculum offered a middle-ground. Vision-based agents (V3/V4) struggled to outperform the simple Greedy vector policy.

3.1 Analysis of Damage (Final Impact)

As shown in Figure 1, the Cooperative agent converged to a sub-optimal solution (high damage), whereas the Greedy agent achieved a lower median damage. The V4 (Heat Seeker) agent showed improved reliability over V3 but could not surpass the aggressive Greedy baseline.



(a) Vector Agents Comparison (b) Vision Agents Comparison (V2-V4)

Figure 1: **Total Fire Damage.** The Greedy strategy (Orange) consistently resulted in fewer burnt cells compared to the Cooperative strategy. Vision-based agents (V4) reduced variance but struggled with navigation.

3.2 Suppression Dynamics (Speed)

Figure 2 analyzes the active fire size over time. The Greedy and Curriculum agents show a steeper downward slope in active fire cells, indicating a faster reaction time compared to the

hesitant Cooperative agents.

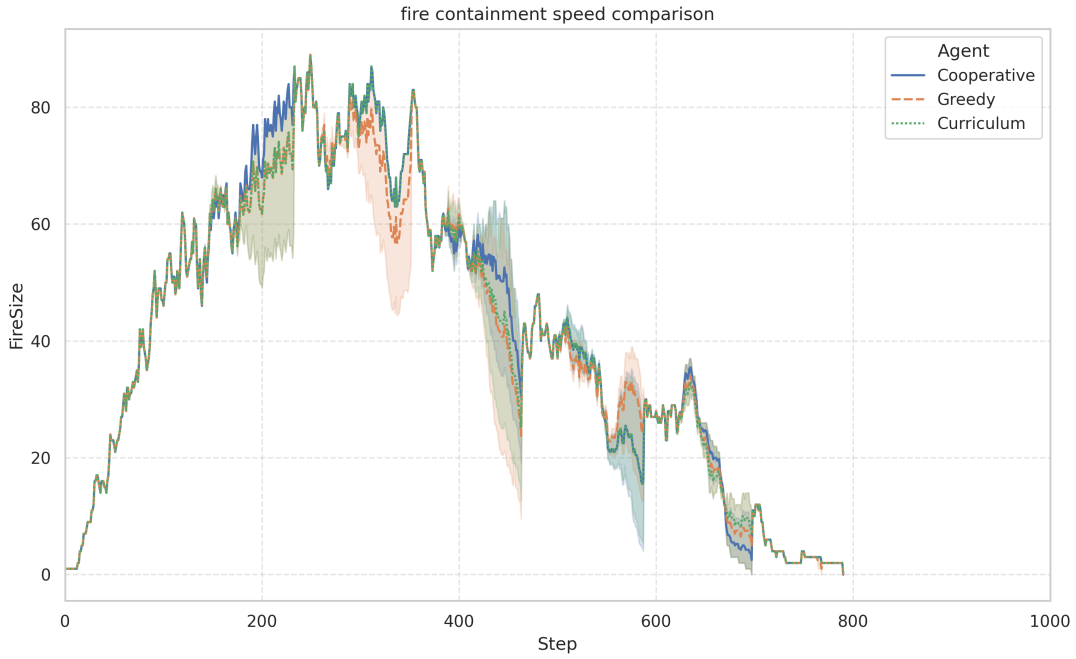
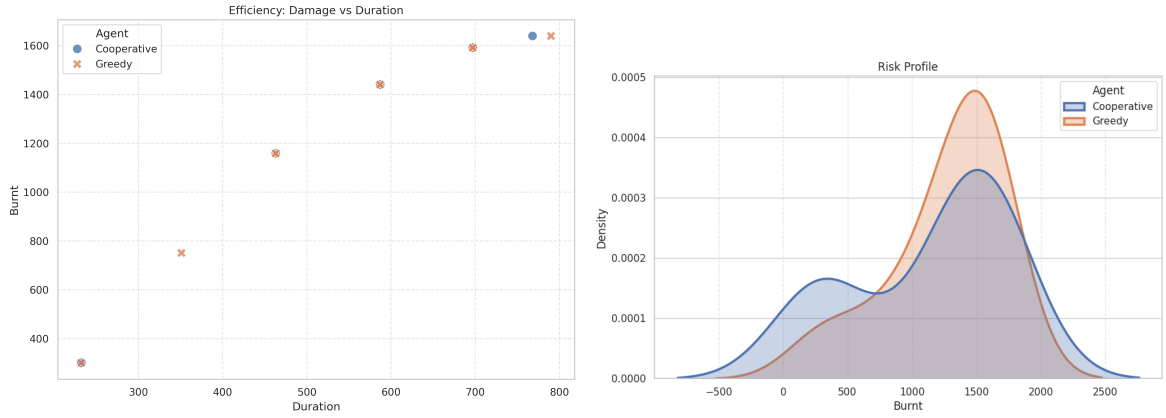


Figure 2: **Containment Trajectory.** Greedy and Curriculum agents extinguish the fire significantly faster.

3.3 Efficiency and Risk

We examined the trade-off between episode duration and damage (Figure 3a) and the probability distribution of outcomes (Figure 3b).



(a) Efficiency Scatter Plot

(b) Risk Density Profile

Figure 3: Efficiency and Risk analysis showing Greedy agents clustering in the low-damage/fast-duration region.

4 Discussion

The failure of the Cooperative reward can be attributed to the **Signal-to-Noise Ratio**. The global penalty created too much noise for the agents to credit specific actions. The Greedy agent received a clean, dense signal ("I dropped water → I got points"), creating a direct cause-and-effect loop that was easier to learn.

4.1 The "Blindness" and "Wall-Hugging" Phenomenon

In Experiments V3 and V4, we attempted to fix the agent's spatial awareness using CNNs and Distance Shaping.

- **Vision (V3):** Giving the agent "eyes" (CNN) was insufficient. Without guidance, the agent could not distinguish the fire from background noise in the massive 160×240 grid.
- **Heat Seeker (V4):** Adding distance rewards successfully lured agents toward the fire initially (as seen in render analysis). However, agents still exhibited "wall-hugging" behavior later in the episode.

This suggests the limitation lies in the **Discrete Action Space** combined with the environment's physics (inertia). Once an agent hits a boundary, the discrete actions make it difficult to turn around smoothly, leading to a "stuck" state that no reward function can easily fix.

5 Conclusion

This study demonstrates that in MARL applications for wildfire suppression, **dense local incentives (Greedy)** are superior to **sparse global penalties (Cooperative)**. While it is intuitive to design rewards around the "ultimate goal" (saving the forest), the difficulty of credit assignment hinders learning. Furthermore, we identified that solving the navigation problem in large grids requires more than just visual perception; future work should explore **Continuous Control** to overcome the mobility limitations observed in the discrete setting.