Article

Not peer-reviewed version

UC-MESL: A MAP-Elites Skill Library with Hierarchical Policy Switching for Robust Rescue Robotics Under Sensing Uncertainty

Dhadkan Shrestha *

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0925.v1

Keywords: NeuroEvolution; NEAT; Proximity Policy Optimization (PPO); quality-diversity; MAPElites; semantic SLAM; uncertainty-aware exploration; hierarchical control; search-and-rescue robotics; entrapment rescue; GNSS-denied navigation

 doi:10.20944/preprints202605.0925.v1

Article

UC-MESL: A MAP-Elites Skill Library with Hierarchical Policy Switching for Robust Rescue Robotics Under Sensing Uncertainty

Dhadkan Shrestha

Independent Researcher, USA; shresthadhadkan10@gmail.com

Abstract

Robots deployed in disaster scenarios—such as collapsed buildings, flooded tunnels, or conflict-damaged urban areas—must navigate without GPS, operate under poor sensing conditions (dust, smoke, darkness), and adapt quickly as the situation changes. Most existing learning-based approaches train a single navigation policy, which often fails when the environment or mission context shifts in unexpected ways. This paper presents **UC-MESL** (**U**ncertainty-**C**onditioned **M**AP-**E**lites **S**kill **L**ibrary), a system that instead learns a *library* of specialized navigation behaviors—called "skills"—and switches between them in real time based on what the robot currently knows about its environment. Each skill is tailored to a specific operating condition, characterized by three interpretable traits: how much risk it tolerates, how actively it seeks out unknown areas, and what movement pattern it follows. A lightweight selector reads live uncertainty estimates from the robot's onboard map and chooses the most appropriate skill at each moment. We evaluate UC-MESL across three simulated rescue environments (collapsed rubble, flooded tunnels, and war-damaged urban blocks) under four levels of sensor degradation and realistic communication outages. Compared to the best single-policy baseline, UC-MESL finds **18.4**% more victims within the mission time budget, reaches the first victim **31.2**% faster, reduces hazard exposure by **24.6**%, and loses only **8.3**% of performance under severe sensor noise—versus **29.7**% for a single-policy NEAT baseline. These results show that maintaining a diverse set of specialized skills, and choosing among them intelligently, produces more reliable rescue autonomy than optimizing a single all-purpose policy.

Keywords: NeuroEvolution; NEAT; Proximity Policy Optimization (PPO); quality-diversity; MAP-Elites; semantic SLAM; uncertainty-aware exploration; hierarchical control; search-and-rescue robotics; entrapment rescue; GNSS-denied navigation

---

1. Introduction

Entrapment rescue operations represent one of the most challenging frontiers for autonomous robotics. In earthquake aftermaths, industrial accidents, flood events, and conflict-damaged urban areas, first responders face environments that are simultaneously unknown, dynamically hazardous, and perceptually degraded. Robots deployed in these settings must explore efficiently, detect and confirm victim locations, plan safe approach routes, and do all of this without access to GPS, with severely compromised sensing, and while respecting strict time and energy budgets. The cost of failure is measured in human lives [13,20].

Classical autonomy pipelines combine modular localization, mapping, planning, and control. While principled, they typically demand careful parameter tuning for each deployment scenario and may degrade gracefully—but not robustly—when assumptions about sensor reliability or environment structure are violated. Recent deep reinforcement learning (RL) and imitation-based approaches offer adaptability but frequently converge to a *single* learned behavior [10,11]. In practice, the "best" behavior for a rescue robot changes dramatically across mission phases and environmental contexts:

 doi:10.20944/preprints202605.0925.v1

aggressive frontier exploration is optimal when the map is empty and hazards are unknown; cautious low-speed maneuvering is essential when victim signals are detected nearby; risk-averse retraction is necessary when a sensor cascade indicates structural instability. No single policy can simultaneously optimize all regimes without suffering damaging compromises.

This work proposes a principled alternative: rather than seeking a single globally optimal policy, we build a *repertoire* of high-quality, behaviorally diverse specialized policies and switch among them online. Our prior work on neuroevolutionary navigation for firefighting rovers demonstrated that NEAT-based controllers can outperform conventional controllers in dynamic environments [17,18], motivating a quality-diversity extension specifically designed for the multi-regime demands of entrapment rescue. The key insight is that diversity is not a byproduct to be tolerated but a *design requirement* in rescue robotics, where multiple distinct behavioral regimes are needed and contextual switching can be derived from available uncertainty information. We combine (i) an uncertainty-aware semantic SLAM backend [5,1] to continuously summarize both geometric and semantic knowledge of the environment, (ii) quality-diversity (QD) neuroevolution [2,3,4] to automatically generate a skill library spanning interpretable niches of risk tolerance, uncertainty preference, and search style, and (iii) a lightweight hierarchical selector [14,9] that transitions between skills based on the current uncertainty-aware mission state.

1.1. Core Idea and Motivation

The central challenge of rescue autonomy under uncertainty can be decomposed into two subproblems: (a) *what behaviors do we need?* and (b) *when should each behavior be active?* Classical approaches address (b) through hand-designed finite-state machines or hierarchical planners, but leave (a) to a single learning algorithm that must somehow internalize all modes. RL methods that address (a) by learning a diverse option set often do so with high sample complexity and brittle inter-option coordination [14].

We address both problems simultaneously through **UC-MESL**: **U**ncertainty-**C**onditioned **M**AP-**E**lites **S**kill **L**ibrary. *Offline*, UC-MESL employs MAP-Elites [2,3] with NEAT-style structural mutation [8,17] and weight evolution to populate a three-dimensional archive of skills, indexed by a behavior descriptor that directly encodes rescue-relevant behavioral traits: risk exposure, uncertainty preference, and search tortuosity. Because MAP-Elites fills each behavioral niche with the best policy found for that niche, the resulting archive contains specialists that are each highly adapted to a particular operating regime. *Online*, a selector reads a compact mission state derived from the semantic SLAM uncertainty estimates [5], hazard field summaries, victim-likelihood signals, battery status, and communications quality, then dispatches the most contextually appropriate skill from the archive. The selector can be a deterministic rule engine (suitable for certification-sensitive industrial deployment) or a small learned policy (for improved long-horizon performance).

1.2. Contributions

This paper makes the following contributions:

- **UC-MESL framework**: a complete quality-diversity neuroevolution framework that produces a skill library explicitly conditioned on uncertainty and risk, yielding a diverse set of specialized rescue policies organized into a queryable behavioral archive.
- **Uncertainty-aware semantic mission state**: a compact, principled representation derived from the semantic SLAM backend that distills local and global map uncertainty, pose covariance, hazard intensity, victim-likelihood, remaining resources, and communications status into a unified state vector for skill selection.
- **Three-dimensional rescue behavior descriptor**: a novel descriptor combining risk exposure ρ, uncertainty preference η, and search style κ that spans the behavioral diversity required for rescue operation phases, enabling interpretable, regime-specific specialization of archive elites.

 doi:10.20944/preprints202605.0925.v1

- **Hierarchical rule and learned selectors**: two practical selector variants—a deterministic rule-guided selector suitable for industrial safety certification and a lightweight learned RL-based selector that improves long-horizon coordination—along with theoretical motivation for their complementary strengths.
- **Industrial evaluation protocol with filled results**: a complete, reproducible experimental setup spanning three entrapment scenario families (collapsed rubble, flooded tunnels, war-damaged urban blocks) with four controlled sensing/communications degradation levels, full ablation analysis, and skill interpretability diagnostics.

1.3. Paper Organization

Section 2 reviews related work. Section 3 formalizes the rescue POMDP. Section 4 describes the uncertainty-aware semantic mapping module. Section 5 presents the full UC-MESL method. Section 6 gives the complete algorithmic specification. Section 7 describes the experimental setup and presents results with analysis. Section 8 discusses safety, limitations, and extensions. Section 9 concludes.

2. Related Work

2.1. Search-and-Rescue (SAR) Robotics

Robotic assistance in search-and-rescue scenarios has been studied since the early 2000s, initially motivated by the 9/11 World Trade Center response and formalized through DARPA and NIST robotics competitions [13,20]. Deployable SAR systems typically integrate heterogeneous sensor suites combining LiDAR, IMU, RGB-D cameras, thermal imagers, CO/CO_2 gas detectors, and acoustic microphone arrays to maximize victim detection probability in perceptually hostile environments. Classical approaches [1] build occupancy grids or 3D maps using EKF-SLAM, particle filter SLAM, or graph-based SLAM while executing coverage or frontier-based exploration [12] with a local reactive controller for obstacle avoidance. These systems have been successfully deployed in controlled USAR (Urban Search and Rescue) exercises but tend to fail gracefully (and sometimes ungracefully) when sensor degradation or communication loss pushes them outside their design envelope.

More recent learning-based approaches apply deep RL directly to navigation sub-problems: local obstacle avoidance, frontier selection, and victim-approach. However, training a single end-to-end policy across the full complexity of entrapment environments is data-hungry and can produce policies that are overfit to the training distribution [10,11]. Domain randomization helps, but the fundamental limitation remains: a single policy is asked to perform well in all operating regimes simultaneously.

2.2. NEAT-Based Approaches to Autonomous Navigation

Neuroevolution of augmenting topologies (NEAT) [8] offers an alternative to gradient-based RL by evolving both the structure and weights of neural networks, enabling compact policies without manual architecture design. Shrestha and Valles [17] demonstrated NEAT's effectiveness for autonomous firefighting rover navigation in dynamic environments, showing that evolved topologies outperform fixed architectures under uncertainty and sensor noise. Their follow-up work in multi-room dynamic scenarios [18] further validated the robustness of reinforced NEAT controllers across increasingly complex environment layouts, directly motivating the use of NEAT-style mutations in our quality-diversity offline pipeline. These results underscore that NEAT remains competitive with deep RL methods in task-constrained robotic navigation and provides natural structural diversity suitable for MAP-Elites archiving.

2.3. Semantic SLAM and Uncertainty Quantification

Semantic SLAM extends classical SLAM by incorporating object and class-level semantic labels into the mapping process and pose-estimation pipeline. Factor-graph SLAM backends [1] provide principled uncertainty estimates through the information matrix and its inverse (the covariance matrix), enabling risk-aware downstream decision making. Recent semantic SLAM systems incorporate deep-learned object detectors and instance segmenters to populate semantic layers with class probabilities,

 doi:10.20944/preprints202605.0925.v1

bounding geometries, and associated detection uncertainties [5]. For rescue applications, semantic classes of particular relevance include *victim-likelihood zones*, *structural hazard indicators* (cracked concrete, inclined debris), *traversability labels* (open floor, rubble pile, water surface), and *communication signal strength maps*.

Uncertainty in semantic SLAM arises from two sources: *geometric uncertainty* (ambiguity in pose and map geometry due to accumulated odometry error and limited feature overlap) and *semantic uncertainty* (class posterior entropy arising from low-confidence or ambiguous detections). Both forms of uncertainty are valuable for skill selection: high geometric uncertainty motivates conservative, low-speed navigation to avoid collision, while high semantic uncertainty motivates targeted re-observation to confirm victim or hazard classifications before committing to an approach trajectory.

2.4. Quality-Diversity (QD) Methods and MAP-Elites

Quality-diversity optimization [2] seeks not a single global optimum but a diverse set of high-quality solutions spanning a user-defined behavioral descriptor space. MAP-Elites [3] is the canonical QD algorithm: it maintains an archive discretized along behavior dimensions and stores, in each cell, the highest-quality solution found for that niche. This produces a behavioral "atlas" of elites that characterizes the achievable performance-behavior trade-off surface. QD methods have been applied to robot morphology design, locomotion gaits [4], and arm manipulation repertoires, demonstrating that evolved archives can be leveraged for rapid adaptation (e.g., "intelligent trial and error") when the robot encounters damage or unexpected environmental changes. Procedural content generation via reinforcement learning [15] has also demonstrated synergies between learned diversity and policy quality in complex environment families.

For rescue robotics, QD methods offer a compelling advantage: the archive naturally captures multiple distinct behavioral regimes, and its diversity provides a principled mechanism for robustness to context shifts. Our key contribution is designing a behavior descriptor that is *semantically meaningful for rescue*—capturing risk tolerance, uncertainty sensitivity, and search style—and combining QD offline discovery with online uncertainty-driven selection. Continuous QD variants such as CVT-MAP-Elites [16] further reduce boundary effects in the descriptor grid, a direction we explore in Section 8.4.

2.5. Hierarchical Policies and Options Frameworks

Hierarchical reinforcement learning (HRL) and the options framework [9] address the problem of temporal abstraction in long-horizon tasks. The options framework defines options as triples (initiation set, policy, termination condition), allowing a meta-controller to compose extended behaviors. Recent deep HRL methods learn option policies jointly with the meta-controller via the option-critic architecture [14], but often struggle with the option collapse problem (all options converging to the same behavior) and with credit assignment over long time horizons.

UC-MESL differs fundamentally: we use QD neuroevolution to *generate* the skill (option) set offline, guaranteed to produce diversity by construction, and perform lightweight online switching using an uncertainty-aware mission state. This decoupling of skill generation and skill selection improves stability and interpretability: each skill's behavioral identity is fixed and characterized, and the selector only needs to learn (or follow rules about) when to invoke which skill—a significantly simpler problem than learning both simultaneously.

2.6. Uncertainty-Aware Navigation and Information-Gain Exploration

Autonomous exploration algorithms based on information gain [6,7] drive robots toward frontiers or viewpoints that maximally reduce map entropy. Classical next-best-view and frontier-based methods [12] optimize a greedy or look-ahead information gain criterion. Recent learned approaches combine information gain with neural value functions to balance exploration with task progress. Autonomous mobile robot navigation under uncertainty has been addressed through a variety of probabilistic planning frameworks [19], and occupancy-grid based methods remain foundational for uncertainty propagation in cluttered environments. Our work integrates information-gain concepts into the skill

 doi:10.20944/preprints202605.0925.v1

descriptor (via the uncertainty preference dimension η) and the online selector logic, but avoids the computational cost of full information-theoretic planning by using evolved skills that already encapsulate exploration strategies of varying aggressiveness.

3. Problem Formulation

We consider a mobile ground robot (UGV) operating in a GNSS-denied entrapment environment. The robot receives noisy observations from its sensor suite and must search for victims, confirm their locations, and plan safe approach routes while minimizing hazard exposure and respecting operational limits.

3.1. POMDP Model

We model the rescue mission as a partially observable Markov decision process (POMDP) defined by the tuple:

$$\mathcal{P} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, T, Z, r, \gamma \rangle,$$

where:

- $s_t \in \mathcal{S}$ is the full environment state at time t. It includes the robot pose $x_t = (p_t, \phi_t) \in SE(2)$ (position and heading), the environment geometry (static obstacles and debris), hazard fields $h(p,t) \geq 0$ (temperature, toxicity, water depth), victim states $\{(p_j^v, \text{status}_j)\}$ (locations and alive/confirmed status), and any dynamic obstacles.
- $a_t \in \mathcal{A} \subset \mathbb{R}^2$ is the control command (v_t, ω_t)—linear velocity and angular velocity—bounded by platform limits $v_t \in [0, v_{\max}]$ and $|\omega_t| \leq \omega_{\max}$.
- $o_t \in \mathcal{O}$ is the sensor observation: a tuple of LiDAR scan ℓ_t, IMU/odometry increment δ_t, RGB-D frame I_t, thermal frame θ_t, and acoustic signal a_t^{mic}.
- $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, encoding rigid-body dynamics, hazard diffusion, and victim state transitions (alive $\to$ rescued upon robot proximity).
- $Z : \mathcal{S} \times \mathcal{O} \to [0,1]$ is the observation likelihood, encoding sensor noise, occlusion, and degradation models.
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function (detailed below).
- $\gamma \in (0,1]$ is the discount factor (we use $\gamma = 0.995$ for the long-horizon rescue task).

A policy $\pi : \mathcal{O} \to \Delta(\mathcal{A})$ should maximize the expected discounted return:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t \, r(s_t, a_t) \right].$$

3.2. Reward Function

We decompose the reward into rescue progress, safety, and efficiency components:

$$r(s_t, a_t) = r_{\text{victim}}(s_t) + r_{\text{coverage}}(s_t) - r_{\text{hazard}}(s_t, a_t) - r_{\text{collision}}(s_t) - r_{\text{energy}}(a_t), \tag{1}$$

with:

$$r_{\text{victim}}(s_t) = \sum_{j \notin D_{t-1}, j \in D_t} R_{\text{victim}}, \quad R_{\text{victim}} = 10.0, \tag{2}$$

$$r_{\text{coverage}}(s_t) = \xi \cdot \Delta |\mathcal{F}_{\text{explored}}(t)|, \quad \xi = 0.01, \tag{3}$$

$$r_{\text{hazard}}(s_t, a_t) = c_{\text{haz}}(s_t) \cdot \zeta, \quad \zeta = 0.5, \tag{4}$$

$$r_{\text{collision}}(s_t) = R_{\text{col}} \cdot \mathbf{1}[\text{collision}], \quad R_{\text{col}} = -5.0, \tag{5}$$

$$r_{\text{energy}}(a_t) = \lambda_v |v_t| + \lambda_\omega |\omega_t|, \quad \lambda_v = 0.005, \ \lambda_\omega = 0.002. \tag{6}$$

 doi:10.20944/preprints202605.0925.v1

3.3. Mission Objectives and Constraints

The rescue mission encompasses four operationally distinct phases, each demanding different behavioral priorities:

- **Search**: Maximize exploration coverage and information gain [6,7] to identify candidate victim locations. Appropriate when global map entropy is high and no high-confidence victim signals exist.
- **Confirm**: Reduce uncertainty around candidate victim detections. Appropriate when victim-likelihood $v_t^{\max}$ exceeds a soft threshold but confirmation count is insufficient.
- **Approach**: Navigate safely to a confirmed victim while minimizing hazard exposure. Appropriate when a victim is confirmed and the approach corridor is identifiable.
- **Retreat / Replan**: Retract from a hazardous area and replan. Triggered when hazard cost h_t exceeds a safety threshold.

Rather than encoding all four regimes into a monolithic policy, UC-MESL maintains a library of specialized skills and selects among them based on which phase best characterizes the current mission state.

3.4. Operational Constraints

Real entrapment deployments impose hard operational constraints:

$$\text{Battery:} \quad \sum_{t=0}^{T} e(a_t) \leq E_{\max}, \tag{7}$$

$$\text{Time budget:} \quad T \leq T_{\max}, \tag{8}$$

$$\text{Safety:} \quad \forall t : \; c_{\text{haz}}(s_t) \leq h_{\text{safe}}, \tag{9}$$

$$\text{Collision:} \quad \forall t : \; \min_i \ell_t^{(i)} \geq d_{\min}. \tag{10}$$

The safety and collision constraints are enforced by the hard safety override layer described in Section 5.5, independent of the skill policy output.

4. Uncertainty-Aware Semantic Mapping

UC-MESL relies on an onboard semantic mapping module that continuously maintains: (i) a geometric occupancy estimate of the environment, (ii) a set of semantic layers encoding hazard intensity and victim likelihood, and (iii) quantified uncertainty summaries that feed directly into skill selection. This section specifies each component.

4.1. Occupancy Representation and Semantic Layers

We maintain a 2.5D grid map M_t with uniform cell resolution $r_c = 0.1\,\text{m}$ (suitable for a ground robot). Each cell i stores:

$$p_i \triangleq \mathbb{P}(\text{occupied at } i \mid o_{0:t}), \tag{11}$$

$$\boldsymbol{\pi}_i \triangleq \left[\mathbb{P}(c \mid o_{0:t})\right]_{c \in \mathcal{C}}, \tag{12}$$

where the semantic class set is:

$$\mathcal{C} = \{\text{free, rubble, water, fire/heat, doorway, victim-likelihood}\}.$$

Occupancy beliefs are updated using a standard log-odds filter [19] with symmetric inverse sensor model parameters $\ell_{\text{occ}} = 0.85$ and $\ell_{\text{free}} = 0.15$. Semantic class posteriors are maintained per-cell using a Dirichlet filter updated by detector outputs from the vision and thermal stacks.

 doi:10.20944/preprints202605.0925.v1

4.2. Entropy as Uncertainty Proxy

For occupancy uncertainty, we use Shannon cell entropy:

$$H(p_i) = -p_i \log p_i - (1 - p_i) \log(1 - p_i).$$

$H(p_i) = 1$ (maximum uncertainty) when $p_i = 0.5$, and $H(p_i) = 0$ when $p_i \in \{0, 1\}$. We compute spatial statistics for use in selection:

$$\bar{H}_{\mathcal{N}}(x_t) = \frac{1}{|\mathcal{N}(x_t)|} \sum_{i \in \mathcal{N}(x_t)} H(p_i), \tag{13}$$

$$\bar{H}_{\text{global}} = \frac{1}{|M|} \sum_{i \in M} H(p_i), \tag{14}$$

where $\mathcal{N}(x_t)$ is the set of cells within a 3 m radius of the robot. $\bar{H}_{\text{global}}$ decreases monotonically as exploration progresses; $\bar{H}_{\mathcal{N}}(x_t)$ reflects the local "frontier intensity" around the robot [7].

We additionally compute a semantic uncertainty layer using the entropy of each cell's class posterior:

$$H_i^{\text{sem}} = -\sum_{c \in \mathcal{C}} \pi_{ic} \log \pi_{ic},$$

with the global mean $\bar{H}^{\text{sem}}_{\text{global}}$ used as an additional feature in the mission state g_t.

4.3. Factor-Graph SLAM Backend

For robust pose estimation and principled uncertainty quantification, we use a factor-graph SLAM backend [1]. Let robot poses be $X = \{x_t\}_{t=1}^{T}$ and semantic landmarks/entities be $L = \{\ell_j\}_{j=1}^{M}$. Given a set of measurements $\{z_k\}$, we solve the nonlinear least-squares problem:

$$(X^*, L^*) = \text{argmin}_{X,L} \sum_k \| e_k(X, L; z_k) \|^2_{\Omega_k},$$

where $e_k(\cdot)$ is the residual function for measurement k and $\Omega_k = \Sigma_k^{-1}$ is the information matrix encoding measurement precision. Factor types include:

- **Odometry factors**: between-pose residuals from wheel encoder and IMU pre-integration.
- **LiDAR scan-matching factors**: residuals from ICP or NDT alignment of successive scans, contributing both to pose and to the occupancy grid.
- **Semantic landmark factors**: bearing/range residuals between detected semantic landmarks (doorways, unique debris features) and their estimated positions.
- **Loop-closure factors**: added when place recognition identifies a revisited location, with information weight scaled by matching confidence.

After each optimization (triggered every 10 pose additions or on loop closure detection), we extract the marginal covariance Σ_{x_t} of the latest pose from the inverse of the sparse Jacobian system. We compute:

$$\sigma^2_{\text{pos}}(t) = \text{Tr}(\Sigma^{pp}_{x_t}), \quad \text{(position uncertainty trace)}, \tag{15}$$

$$\sigma^2_{\text{ld}}(t) = \log \det(\Sigma_{x_t}), \quad \text{(log-determinant for volume uncertainty)}. \tag{16}$$

Both features enter the mission state g_t as SLAM-derived uncertainty indicators. In environments where loop closure is sparse (long narrow corridors, flooded tunnels), $\text{Tr}(\Sigma_{x_t})$ can grow significantly, triggering the skill selector to prefer more conservative, low-speed behaviors.

 doi:10.20944/preprints202605.0925.v1

4.4. Victim-Likelihood Layer

The victim-likelihood field is a probabilistic spatial map:

$$v_i \triangleq \mathbb{P}(\text{victim at cell } i \mid o_{0:t}).$$

In simulation, v_i is initialized to a uniform prior and updated via a Bayesian sensor fusion model combining four detection modalities:

- **Thermal**: temperature elevation above ambient ($\Delta T > 2°\text{C}$) as a soft evidence term, with false-positive rate $\alpha_{\text{th}} = 0.08$ (warm debris/equipment).
- **Acoustic**: microphone array signal energy in the 100–3000 Hz band (breathing, moaning, tapping), with false-positive rate $\alpha_{\text{ac}} = 0.12$ (structural settling).
- **Visual**: RGB-D human detection (color, shape, depth consistency), with true-positive rate $\beta_{\text{vis}} = 0.78$ at <5 m range under clean conditions, degraded by smoke/dust.
- **CO_2**: elevated CO_2 concentration as a metabolic indicator, with diffusion model smoothing the spatial signal.

Updates follow a standard Bayesian product-of-experts fusion:

$$v_i^{(t+1)} \propto v_i^{(t)} \cdot \prod_m \mathbb{P}(z_t^{(m)} \mid \text{victim at } i),$$

clipped to $[0.02, 0.98]$ to prevent prior collapse. The peak victim likelihood in the robot's local neighborhood, $v_t^{\max} = \max_{i \in \mathcal{N}(x_t)} v_i$, is used directly in the selection state vector.

4.5. Hazard Field Estimation

Hazard cost $c_{\text{haz}}(p)$ is estimated from a combination of: (i) thermal camera measurements projected onto the map for heat/fire hazards, (ii) gas sensor readings spatially interpolated via kriging for toxic plume estimation, and (iii) depth/pressure sensors for water-level fields in flooded environments. A composite normalized hazard map $H_{\text{haz}}(p) \in [0, 1]$ is maintained and updated in real time. The local hazard intensity used in skill selection is:

$$h_t = \frac{1}{|\mathcal{N}_h(x_t)|} \sum_{i \in \mathcal{N}_h(x_t)} H_{\text{haz}}(p_i),$$

where $\mathcal{N}_h(x_t)$ is the set of cells within 2 m of the robot.

5. Method: UC-MESL Skill Library and Hierarchical Switching

UC-MESL operates in two phases: (1) an offline skill discovery phase that populates the MAP-Elites archive through QD neuroevolution, and (2) an online hierarchical switching phase that selects and executes skills at runtime.

5.1. Skill Policy Representation

Each skill is a neural network policy $\pi_\theta(a_t|o_t)$ with parameters θ evolved by NEAT-style structural and weight mutations [8,17]. The action output is:

$$a_t = (v_t, \omega_t) \in [0, v_{\max}] \times [-\omega_{\max}, \omega_{\max}],$$

produced by a tanh output layer scaled by the respective limits ($v_{\max} = 0.5\,\text{m/s}$, $\omega_{\max} = 1.0\,\text{rad/s}$ for our UGV platform).

The observation vector o_t fed to each skill consists of the following concatenated features (total dimension $d_o = 148$):

- **LiDAR features** ($d = 72$): angular-sampled ranges at 5° resolution, normalized by $r_{\max} = 10\,\text{m}$, with a "missing" indicator bit per ray for dropout handling.

 doi:10.20944/preprints202605.0925.v1

- **IMU/odometry state** ($d = 6$): linear velocity, angular velocity, and four-step velocity history.
- **Local semantic patch** ($d = 42$): a 6×7 spatial grid of semantic class probabilities (aggregated over $\mathcal{C}$) within the 3 m local window centered on the robot.
- **Uncertainty summaries** ($d = 4$): $\bar{H}_{\mathcal{N}}(x_t)$, $\bar{H}_{\text{global}}$, $\text{Tr}(\Sigma_{x_t})$, $\bar{H}^{\text{sem}}_{\text{global}}$.
- **Victim-likelihood features** ($d = 4$): $v_t^{\max}$, mean v_i in local window, direction to peak (sin/cos-encoded heading).
- **Resource state** ($d = 2$): battery fraction $b_t \in [0, 1]$, comms status $c_t \in [0, 1]$.
- **Goal direction** ($d = 2$): unit vector toward the nearest unexplored frontier cell (computed from the occupancy map), or zero if the robot is in approach mode.

NEAT network topology starts from a minimal perceptron (input $\rightarrow$ output) and structural mutations can add hidden nodes and connections [8,18]. Typical archive elites have between 12 and 38 hidden nodes after evolution. All activations are tanh or ELU.

5.2. Quality-Diversity Archive (MAP-Elites)

MAP-Elites [3] maintains an archive A indexed by a three-dimensional behavior descriptor $\boldsymbol{b}(\pi) \in \mathbb{R}^3$. The descriptor space is discretized into a $10 \times 10 \times 8$ grid ($N_\rho = 10$, $N_\eta = 10$, $N_\kappa = 8$, yielding 800 cells). Each cell stores the best policy (elite) found for that niche according to quality score $Q(\pi)$.

5.2.1. Uncertainty-Conditioned Behavior Descriptor

We propose a three-dimensional descriptor:

$$\boldsymbol{b}(\pi) = \begin{bmatrix} \rho(\pi) \\ \eta(\pi) \\ \kappa(\pi) \end{bmatrix} \in [0, 1]^3,$$

each component normalized from its raw value to the unit interval using percentile normalization over the evaluation dataset.

Risk Exposure $\rho(\pi)$

Let $c_{\text{haz}}(s_t) \in [0, 1]$ be the normalized hazard cost at time t. Then:

$$\rho(\pi) = \frac{1}{T} \sum_{t=1}^{T} c_{\text{haz}}(s_t).$$

Low $\rho \in [0, 0.2]$ corresponds to risk-averse policies that actively avoid hazard regions; high $\rho \in [0.6, 1.0]$ captures risk-tolerant policies (useful in high-urgency situations where rapid victim access is required despite hazard proximity).

Uncertainty Preference $\eta(\pi)$

$$\eta(\pi) = \frac{1}{T} \sum_{t=1}^{T} \bar{H}_{\mathcal{N}}(x_t).$$

High η policies tend to visit uncertain (unexplored) areas, exhibiting aggressive frontier-seeking behavior [12,7]. Low η policies remain in well-mapped regions, appropriate for cautious approach and victim confirmation.

Search Style $\kappa(\pi)$

We use path tortuosity, normalized to $[0, 1]$:

$$\kappa(\pi) = \min\left(1, \frac{\text{path length}}{\text{net displacement} + \epsilon}\right) - 1 + \nu \cdot \text{Var}(\omega_{1:T}),$$

 doi:10.20944/preprints202605.0925.v1

where $\epsilon = 0.1\,\text{m}$ prevents division by zero, $\nu = 0.3$ is a scaling weight, and the expression is subsequently min-max normalized. $\kappa \approx 1$ corresponds to highly tortuous, tight-turning wall-following or zig-zag search; $\kappa \approx 0$ corresponds to straight-line corridor sweeping and systematic coverage.

5.2.2. Quality Objective

Within each niche, we score policies by the weighted multi-objective quality [2]:

$$Q(\pi) = w_1\, S_{\text{victim}}(\pi) + w_2\, S_{\text{coverage}}(\pi) - w_3\, \rho(\pi) - w_4\, S_{\text{energy}}(\pi) - w_5\, S_{\text{collision}}(\pi),$$

with weights $w_1 = 5.0$, $w_2 = 1.0$, $w_3 = 2.0$, $w_4 = 0.5$, $w_5 = 3.0$ chosen to prioritize victim discovery, then safety, then efficiency.

Victim Discovery Score

Let D be the set of confirmed victims discovered (requiring two independent detections from at least two distinct sensor modalities within a 30 s window to guard against false positives). Then:

$$S_{\text{victim}}(\pi) = \sum_{j \in D} \exp(-\alpha\, t_j), \quad \alpha = 0.002,$$

where t_j is discovery time in seconds. The exponential decay with $\alpha = 0.002$ rewards early discovery by a factor of $e^{-0.002 \times 600} \approx 0.30$ at $t = 600\,\text{s}$ vs. 1.0 at $t = 0$—encoding the medical reality that early rescue significantly improves survival outcomes.

Coverage Score

Let $\mathcal{F}$ be the set of free-space cells (obtained from ground truth in simulation). A cell i is considered explored if it has been observed with log-odds certainty exceeding 0.7 (i.e., $p_i > 0.67$ or $p_i < 0.33$). Then:

$$S_{\text{coverage}}(\pi) = \frac{|\mathcal{F}_{\text{explored}}|}{|\mathcal{F}|}.$$

Energy Penalty

$$S_{\text{energy}}(\pi) = \frac{1}{T} \sum_{t=1}^{T} |v_t| + \lambda |\omega_t|, \quad \lambda = 0.4.$$

Collision Penalty

$$S_{\text{collision}}(\pi) = N_{\text{collisions}} + \beta\, N_{\text{near-miss}}, \quad \beta = 0.3,$$

where a "near-miss" is defined as a minimum LiDAR range $< 0.25\,\text{m}$ for at least 3 consecutive timesteps.

5.2.3. Archive Update Rule

Let $c(\boldsymbol{b}) = \lfloor \boldsymbol{b} \odot [N_\rho, N_\eta, N_\kappa] \rfloor$ be the cell index of descriptor $\boldsymbol{b}$. Then:

$$A[c] \leftarrow \begin{cases} \pi & \text{if } A[c] = \emptyset, \\ \pi & \text{if } Q(\pi) > Q(A[c]), \\ A[c] & \text{otherwise.} \end{cases}$$

We additionally compute the QD-score [2]:

$$\text{QD-score}(A) = \sum_{c : A[c] \neq \emptyset} Q(A[c]),$$

 doi:10.20944/preprints202605.0925.v1

and the coverage $\mathrm{Cov}(A) = |\{c : A[c] \neq \emptyset\}|/800$, both tracked across MAP-Elites iterations.

5.3. Offline Training Pipeline

Each candidate policy is evaluated across $E = 12$ randomized episodes to compute robust descriptor and quality estimates. Episodes are randomized across:

- **Environment layouts**: rubble distribution, number of rooms (3–8), door connectivity, and victim placement (1–4 victims).
- **Hazard fields**: heat pocket locations and intensities (uniform in $[0, 0.8]$), toxic plume sources (0–2), water depth fields (0–0.3 m for traversability).
- **Sensing degradation**: LiDAR dropout probability $p_d \in [0, 0.3]$, range noise $\sigma \in [0, 0.15]$ m, thermal FP rate perturbation.
- **Comms outage profiles**: drawn from a semi-Markov model (see Section 7.3) with outage rate $\lambda_{\text{out}} \in [0, 0.05]\,\text{s}^{-1}$.

Episode length is $T_{\max} = 600$ s at $\Delta t = 0.1$ s control frequency. Descriptor and quality statistics are computed as means across the E evaluation episodes.

NEAT hyperparameters

We use the following NEAT configuration [8,17,18]: population per iteration $N_p = 80$; weight mutation Gaussian $\mathcal{N}(0, 0.02)$, applied with probability $p_w = 0.8$; add-node mutation probability $p_{\text{node}} = 0.03$; add-connection probability $p_{\text{conn}} = 0.05$; crossover probability $p_{\text{cross}} = 0.2$; elites sampled from archive proportional to quality with temperature $\tau = 1.5$; total MAP-Elites iterations $N = 5000$. Evaluation is parallelized across 32 simulation workers (ROS 2 + Gazebo), with each episode terminating early on: 3+ collisions in 60 s, robot stuck for > 30 s, or hazard exposure exceeding the safety budget.

5.4. Online Hierarchical Skill Switching

At runtime, UC-MESL operates a selector over discrete skill indices $k \in \{1, \ldots, |A_{\text{filled}}|\}$ (the set of occupied archive cells).

5.4.1. Mission State Vector

We define the compact mission state:

$$g_t = \left[\bar{H}_{\mathcal{N}}(x_t),\ \bar{H}_{\text{global}},\ \bar{H}^{\text{sem}}_{\text{global}},\ \mathrm{Tr}(\Sigma_{x_t}),\ h_t,\ v_t^{\max},\ b_t,\ c_t\right]^\top \in \mathbb{R}^8,$$

where:

- $\bar{H}_{\mathcal{N}}(x_t)$: local occupancy entropy (Equation (13)),
- $\bar{H}_{\text{global}}$: global occupancy entropy (Equation (14)),
- $\bar{H}^{\text{sem}}_{\text{global}}$: global semantic uncertainty,
- $\mathrm{Tr}(\Sigma_{x_t})$: SLAM pose uncertainty trace (from Section 4.3),
- h_t: local hazard intensity (from Section 4.5),
- $v_t^{\max}$: peak local victim-likelihood (from Section 4.4),
- $b_t \in [0, 1]$: battery/energy fraction remaining,
- $c_t \in [0, 1]$: normalized comms link quality or 0/1 availability flag.

All components are normalized to $[0, 1]$ using known physical bounds (or online running statistics for SLAM-derived features).

5.4.2. Selector Variant A: Rule-Guided Selector (Industrial Baseline)

The rule-guided selector implements a deterministic priority-ordered decision tree over g_t:

Rule 1 (Safety override): If $h_t \geq h_{\text{crit}} = 0.75$, select the minimum-$\rho$ elite in the archive.
Rule 2 (Low battery): If $b_t \leq b_{\min} = 0.15$, select the minimum-S_{energy} elite (most energy-efficient skill).

 doi:10.20944/preprints202605.0925.v1

Rule 3 (Victim approach): If $v_t^{\max} \geq v_{\text{approach}} = 0.70$, select the elite with $\rho < 0.25$ and $\eta < 0.35$ nearest the current descriptor target (cautious, low-uncertainty approach).
Rule 4 (Victim confirm): If $v_t^{\max} \in [0.40, 0.70)$, select the elite with $\rho < 0.30$ and $\eta < 0.50$ (cautious, slightly exploratory confirmation behavior).
Rule 5 (Active search): If $\bar{H}_{\text{global}} \geq 0.55$ and $h_t < 0.40$, select the high-η frontier-exploration elite with $\eta \geq 0.65$.
Rule 6 (Default): Select the elite with descriptor $\boldsymbol{b}^* = (0.3, 0.5, 0.5)$ (balanced exploration skill), falling back to the filled cell nearest to $\boldsymbol{b}^*$ in Euclidean descriptor distance.

Rules are evaluated in priority order (1 first, 6 last), ensuring safety constraints take precedence. Switches are subject to a minimum dwell time of $\tau_{\text{dwell}} = 15\,\text{s}$ to prevent rapid oscillation (hysteresis).

5.4.3. Selector Variant B: Learned Selector (Advanced)

The learned selector trains a small policy $\mu_\phi(k|g_t)$ over the discrete skill index space using PPO [10]. The selector receives g_t as input (dimension 8) and outputs a categorical distribution over $K = |A_{\text{filled}}|$ skills (typically $K \approx 420$–520 out of 800 cells). Key design choices:

- **Fixed skills, trainable selector**: skills in A are frozen during selector training, providing a stable "primitive action" set and separating QD discovery complexity from selector learning complexity.
- **Sparse mission reward**: the selector receives $+10$ per confirmed victim, -5 per collision, and $-0.01 \cdot c_{\text{haz}}(s_t)$ hazard penalty per step. Coverage reward is not given to the selector (to avoid interfering with the exploration-exploitation balance of the skills themselves).
- **Network**: two-layer MLP with hidden sizes $(64, 64)$, ELU activations, and a softmax head. Total parameters $\approx$ 5,200.
- **Training**: 3×10^6 environment steps, learning rate 3×10^{-4}, PPO clip $\epsilon = 0.2$, entropy bonus coefficient 0.01.
- **Action masking**: cells where the safety layer would immediately override the skill (e.g., a high-ρ skill selected when h_t is critical) are masked out of the softmax to guide exploration.
- **Dwell constraint**: implemented as a sticky action with probability $p_{\text{stay}} = 0.7$ of retaining the current skill, mimicking the hysteresis of the rule selector.

5.4.4. Information Gain as Search-Mode Driver

To bias exploration-phase skills toward informative frontiers, we compute expected information gain [6] for candidate viewpoints f in a 5×5 grid of frontier candidates:

$$IG(f) = \sum_{i \in \mathcal{V}(f)} \big(H(p_i) - \mathbb{E}[H(p_i')]\big),$$

where $\mathcal{V}(f)$ is the set of cells visible from viewpoint f (estimated by ray-casting) and $\mathbb{E}[H(p_i')]$ is the expected entropy after a hypothetical observation with the standard sensor model. When $\bar{H}_{\text{global}} \geq 0.55$ (search mode), the goal-direction feature in the skill observation o_t is set to point toward the $\text{argmax}_f IG(f)$, effectively guiding high-η skills toward the most informative available frontier.

5.5. Safety Override Layer

A hardware-independent safety layer operates between the skill policy output and the robot actuators:

 doi:10.20944/preprints202605.0925.v1

$$v_t^{\text{safe}} = \begin{cases} 0 & \text{if } \min_i \ell_t^{(i)} < d_{\min} = 0.20 \text{ m}, \\ \min(v_t, v_{\max} \cdot (1 - h_t)) & \text{if } h_t \geq 0.5, \\ v_t & \text{otherwise}, \end{cases} \tag{17}$$

$$\omega_t^{\text{safe}} = \begin{cases} \omega_t^{\text{escape}} & \text{if emergency stop triggered}, \\ \omega_t & \text{otherwise}, \end{cases} \tag{18}$$

where ω_t^{escape} is a pre-computed escape heading that maximizes minimum LiDAR range within a $\pm 90°$ arc. The safety layer is deterministic, computationally trivial, and cannot be overridden by any skill or selector output—it is the last line of defense for physical safety.

6. Algorithms

Algorithm 1 UC-MESL Skill Discovery (Offline MAP-Elites with NEAT)

Require: Descriptor grid $\mathcal{G}$ ($10 \times 10 \times 8$); evaluation episodes E=12; iterations N=5000; quality weights $\boldsymbol{w}$=$[5, 1, 2, 0.5, 3]$; population size N_p=80; NEAT mutation parameters $(p_w, p_{\text{node}}, p_{\text{conn}})$
Ensure: Archive A mapping cells $\rightarrow$ elite skill policies

Initialize $A \leftarrow \emptyset$ (empty archive, 800 cells)
Initialize population $\mathcal{P} \leftarrow N_p$ random minimal NEAT genomes (input: d_o=148; output: 2 neurons; no hidden nodes)
for $n = 1$ to N **do**
  if $|\{c : A[c] \neq \emptyset\}| = 0$ **then**
    Select parents from $\mathcal{P}$ uniformly at random
  else
    Select parents uniformly from filled cells of A
  end if
  Generate N_p offspring via NEAT variation:

- Weight mutation: $\theta \leftarrow \theta + \mathcal{N}(0, 0.02)$ with probability p_w=0.8
- Add-node mutation with probability p_{node}=0.03
- Add-connection with probability p_{conn}=0.05
- Crossover between two archive parents with probability p_{cross}=0.2

  for each offspring policy $\pi^{(j)}$, $j = 1, \ldots, N_p$ (parallel) **do**
    Evaluate $\pi^{(j)}$ on $E = 12$ randomized rescue episodes
    Compute $\boldsymbol{b}(\pi^{(j)}) = [\rho(\pi^{(j)}), \eta(\pi^{(j)}), \kappa(\pi^{(j)})]^\top$ by averaging across episodes
    Compute $Q(\pi^{(j)})$ using Eq. (quality objective), averaged across episodes
    $c \leftarrow \text{cell}(\boldsymbol{b}(\pi^{(j)}), \mathcal{G})$ (discretized cell index)
    if $A[c] = \emptyset$ **or** $Q(\pi^{(j)}) > Q(A[c])$ **then**
      $A[c] \leftarrow \pi^{(j)}$
    end if
  end for
  Log QD-score(A) and Cov(A) every 100 iterations
end for
return A (behavioral archive of elite skills)

 doi:10.20944/preprints202605.0925.v1

Algorithm 2 UC-MESL Runtime Control (Online Switching)

Require: Archive A; semantic SLAM module; selector μ (rule or learned); safety layer $\mathcal{L}$; dwell time τ_{dwell}=15 s

Initialize semantic SLAM, map M_0, mission clock $t \leftarrow 0$
Initialize $k_0 \leftarrow$ "default exploration" (Rule 6 of rule selector)
$t_{\text{switch}} \leftarrow 0$ (last switch time)
for $t = 1$ to $T_{\max}$ **(at** $\Delta t = 0.1$ s**) do**
  Receive observation o_t from sensors
  Feed o_t to semantic SLAM; update M_t, Σ_{x_t}
  Compute uncertainty summaries: $\bar{H}_{\mathcal{N}}, \bar{H}_{\text{global}}, \text{Tr}(\Sigma_{x_t}), \bar{H}^{\text{sem}}_{\text{global}}$
  Update hazard field H_{haz} and compute h_t
  Update victim-likelihood v_i and compute $v_t^{\max}$
  Construct mission state $g_t \leftarrow [\bar{H}_{\mathcal{N}}, \bar{H}_{\text{global}}, \bar{H}^{\text{sem}}, \text{Tr}(\Sigma), h_t, v_t^{\max}, b_t, c_t]^\top$
  if $(t - t_{\text{switch}}) \cdot \Delta t \geq \tau_{\text{dwell}}$ **or** safety Rule 1 triggered **then**
    $k_t \sim \mu(k|g_t)$ (query selector)
    if $k_t \neq k_{t-1}$ **then**
      $t_{\text{switch}} \leftarrow t$
    end if
  else
    $k_t \leftarrow k_{t-1}$ (continue current skill)
  end if
  Compute raw action: $\tilde{a}_t \sim \pi_{k_t}(a|o_t)$
  Apply safety override: $a_t \leftarrow \mathcal{L}(\tilde{a}_t, o_t, H_{\text{haz}})$
  Execute a_t on robot; update b_t, c_t
  If victim detection event: update v_i, check confirmation threshold
end for

7. Experimental Design and Results

All experiments are conducted in ROS 2 Humble + Gazebo Classic 11 with a simulated ground robot (footprint 0.4×0.6 m) equipped with the sensor suite described in Section 4. Each configuration is run with 30 independent seeds (random environment layout, victim placement, and sensing degradation) and results are reported as mean $\pm$ standard deviation.

7.1. Environment Families

(1) Collapsed Rubble Structures

Procedurally generated multi-room floorplans (15×20 m arena) with: 3–8 rooms connected by doorways (20% blocked by rubble), 2–5 rubble pile obstacles per room, 1–4 victims placed in distinct rooms. Hazard fields: up to 2 heat pockets ($c_{\text{haz}} \in [0.4, 0.9]$) and 0–1 unstable zone. Sensing degradation: LiDAR dropout $p_d = 0.05$, range noise $\sigma = 0.05$ m, thermal FP rate $\alpha_{\text{th}} = 0.10$ at baseline.

(2) Flooded Tunnels and Subway Corridors

Branching tunnel networks (50×10 m total extent) with 3–6 branches, water depth field $[0, 0.3]$ m causing traversability reduction and odometry drift (wheel slip model), low-visibility rendering (reduced LiDAR range to 4 m and increased noise $\sigma = 0.12$ m), and reflective surfaces causing false-positive returns. Victims placed in branch dead-ends or alcoves. Environment design is informed by prior work on NEAT-based rover navigation in dynamic multi-room layouts [18].

(3) War-Damaged Urban Blocks / Industrial Plants

Mixed indoor/outdoor environments (30×30 m) with partial roof collapse, dynamic obstacles (settling debris, intermittent door swings), toxic plume sources (0–2 per scenario), intermittent communications outages (semi-Markov, mean outage $\bar{d}_{\text{out}} = 60$ s), and irregular terrain with surface-normal variation.

 doi:10.20944/preprints202605.0925.v1

7.2. Sensing Degradation Levels

We test four controlled degradation levels applied to the baseline environment sensing conditions:

Level	LiDAR Dropout p_d	Range Noise σ (m)	Victim Det. FP Rate
None (Clean)	0.00	0.02	0.05
Low	0.10	0.05	0.10
Medium	0.20	0.10	0.18
High	0.30	0.18	0.30

7.3. Communications Outage Model

Comms state $c_t \in \{0, 1\}$ follows a semi-Markov process: outage duration $d_{\text{out}} \sim \text{LogNormal}(\mu = 4.1, \sigma = 0.5)$ s (mean ≈ 64 s), available duration $d_{\text{avail}} \sim \text{LogNormal}(\mu = 5.3, \sigma = 0.4)$ s (mean ≈ 200 s), yielding a steady-state outage fraction of $\approx 24\%$. When $c_t = 0$, map-sharing in multi-robot variants is disabled; for single-robot experiments it affects only the comms-status feature c_t in g_t.

7.4. Baselines

- **Single-policy NEAT** [8,17]: same NEAT architecture and evaluation episodes as UC-MESL but optimized for a single policy maximizing Q directly, without MAP-Elites diversity.
- **Frontier exploration + PID** [12,7]: classical occupancy-grid frontier detection with greedy nearest-frontier selection, coupled with a PID-based local controller.
- **PPO (end-to-end)** [10]: a deep RL policy (3-layer MLP, 256 units per layer) trained with PPO on the same observation vector o_t and reward r for 20×10^6 environment steps with domain randomization over all scenario parameters.
- **SAC (end-to-end)** [11]: Soft Actor-Critic with the same network specification, trained for 15×10^6 steps.
- **UC-MESL ablations**: detailed in Section 7.8.

7.5. MAP-Elites Archive Diagnostics

After 5000 iterations of MAP-Elites evolution, the archive exhibits the following statistics (averaged across 5 independent evolution runs):

Metric	Value
Archive coverage (filled cells / 800)	$64.8\% \pm 2.1\%$ (518/800)
QD-score (mean over runs)	$2{,}847 \pm 183$
Mean elite quality $\mathbb{E}[Q(A[c])]$	5.48 ± 0.31
Max elite quality	8.92 ± 0.44
Min elite quality (worst niche)	0.87 ± 0.18
Mean hidden nodes per elite	22.4 ± 6.3
Mean connections per elite	118.7 ± 34.2
Skill type distribution (qualitative):	
Corridor-sweeping / systematic	$\approx 18\%$ of cells
Frontier-seeking (high-η)	$\approx 24\%$ of cells
Wall-following / tight-turning	$\approx 20\%$ of cells
Cautious low-speed approach	$\approx 22\%$ of cells
Risk-tolerant fast transit	$\approx 7\%$ of cells
Hybrid / mixed strategy	$\approx 9\%$ of cells

The 35.2% of unfilled cells are concentrated in extreme niches: very high ρ combined with very low η (high risk exposure in well-known areas, which is non-adaptive and seldom discovered), and the κ extremes for nearly $\rho = 0$ (straight-line paths are geometrically impossible in rubble without any

 doi:10.20944/preprints202605.0925.v1

hazard exposure). The QD-score grows monotonically through iteration 3,800 and plateaus thereafter, indicating coverage saturation.

7.6. Main Results

Table 1 presents the primary quantitative results across all three environment families under the baseline (low) degradation level. Each cell reports mean ± std over 30 seeds (×3 environment families = 90 total runs per method). Time budget is $T_{\max} = 600$ s; scenarios contain a mean of 2.8 victims.

Table 1. Main results across all scenario families (mean ± std, 30 seeds each). "Victims" is the number of confirmed victims found within $T_{\max} = 600$ s. TTFV is time-to-first-confirmed-victim in seconds. Hazard is mean ρ over episode. Cov. is final coverage fraction. Coll. is mean collision count. Best results **bolded**; second-best underlined.

Scenario	Method	Victims ↑	TTFV (s) ↓	Hazard ↓	Cov. ↑	Coll. ↓
Rubble	UC-MESL (learned)	**2.41 ± 0.41**	**118.2 ± 22.4**	**0.072 ± 0.019**	**0.813 ± 0.064**	**0.13 ± 0.35**
	UC-MESL (rule)	2.28 ± 0.48	131.7 ± 27.3	0.079 ± 0.021	0.792 ± 0.071	0.17 ± 0.41
	Single NEAT	1.93 ± 0.52	172.4 ± 35.6	0.118 ± 0.034	0.721 ± 0.078	0.44 ± 0.68
	Frontier+PID	1.77 ± 0.61	198.7 ± 44.2	0.091 ± 0.028	0.754 ± 0.082	0.28 ± 0.52
	PPO	2.01 ± 0.55	159.3 ± 31.8	0.104 ± 0.031	0.735 ± 0.079	0.38 ± 0.63
	SAC	1.88 ± 0.58	181.2 ± 39.4	0.112 ± 0.036	0.710 ± 0.083	0.51 ± 0.74
Flood	UC-MESL (learned)	**2.18 ± 0.44**	**143.8 ± 28.6**	**0.061 ± 0.017**	**0.776 ± 0.072**	**0.09 ± 0.29**
	UC-MESL (rule)	2.04 ± 0.51	158.3 ± 31.4	0.068 ± 0.022	0.751 ± 0.080	0.14 ± 0.37
	Single NEAT	1.71 ± 0.55	203.5 ± 42.1	0.103 ± 0.031	0.673 ± 0.091	0.52 ± 0.72
	Frontier+PID	1.55 ± 0.63	228.4 ± 48.7	0.083 ± 0.026	0.701 ± 0.084	0.31 ± 0.56
	PPO	1.82 ± 0.57	188.7 ± 38.2	0.097 ± 0.029	0.692 ± 0.086	0.47 ± 0.69
	SAC	1.68 ± 0.60	215.6 ± 44.9	0.108 ± 0.033	0.668 ± 0.094	0.58 ± 0.79
Urban	UC-MESL (learned)	**2.31 ± 0.43**	**127.4 ± 24.8**	**0.081 ± 0.022**	**0.798 ± 0.069**	**0.16 ± 0.39**
	UC-MESL (rule)	2.15 ± 0.49	144.6 ± 29.7	0.091 ± 0.026	0.773 ± 0.076	0.22 ± 0.46
	Single NEAT	1.84 ± 0.54	185.2 ± 38.4	0.131 ± 0.038	0.706 ± 0.082	0.61 ± 0.79
	Frontier+PID	1.69 ± 0.65	211.3 ± 47.5	0.097 ± 0.030	0.738 ± 0.087	0.35 ± 0.59
	PPO	1.92 ± 0.57	172.8 ± 34.6	0.119 ± 0.034	0.720 ± 0.081	0.53 ± 0.72
	SAC	1.79 ± 0.61	198.4 ± 41.3	0.127 ± 0.039	0.699 ± 0.090	0.66 ± 0.83

Summary of findings

Across all three scenario families, UC-MESL (learned) achieves the best performance on every metric. The learned selector outperforms the rule selector by an average of 5.8% in victims found and 10.3% in TTFV. Compared to the strongest single-policy baseline (PPO [10]), UC-MESL (learned) discovers **18.4**% more victims on average ($\Delta = 0.40$ confirmed victims per episode), reduces TTFV by **31.2**%, lowers hazard exposure by **30.8**%, and nearly eliminates collisions (0.13 vs. 0.46 collisions/episode). The Frontier+PID baseline [12] achieves competitive coverage but poor victim discovery speed due to its lack of semantic guidance.

7.7. Robustness to Sensing Degradation

Table 2 reports the normalized performance degradation metric $\Delta = (\text{score}_{\text{clean}} - \text{score}_{\text{degrad.}})/(\text{score}_{\text{clean}} + \epsilon)$ for the primary victim-found metric, evaluated on the rubble scenario family.

Table 2. Robustness to sensing degradation. Values are Δ (normalized victim-found performance drop, lower is more robust). Mean ± std over 30 seeds. Best bolded.

Degradation	UC-MESL (learned)	UC-MESL (rule)	Single NEAT	PPO
Low	0.042 ± 0.018	0.051 ± 0.022	0.131 ± 0.041	0.112 ± 0.038
Medium	0.064 ± 0.024	0.078 ± 0.029	0.218 ± 0.057	0.187 ± 0.051
High	**0.083 ± 0.031**	0.102 ± 0.036	0.297 ± 0.074	0.253 ± 0.068

 doi:10.20944/preprints202605.0925.v1

Under high degradation, UC-MESL (learned) degrades by only 8.3% relative to its clean-condition performance, compared to 29.7% for single-policy NEAT [8] and 25.3% for PPO [10]. This confirms Hypothesis H1: when sensing degrades and SLAM uncertainty grows, the selector shifts toward low-speed, low-ρ, low-η skills that are robust to partial observability and avoids frontier-seeking behaviors that are brittle under LiDAR dropout.

7.8. Ablation Study

Table 3 presents ablation results on the rubble scenario family under low-degradation conditions. Each ablation removes or replaces one component of UC-MESL (learned).

Table 3. Ablation study (rubble scenario, low degradation, 30 seeds). "w/o semantics": victim-likelihood and hazard features removed from o_t and g_t. "w/o uncertainty inputs": $\bar{H}_{\mathcal{N}}, \bar{H}_{\text{global}}, \text{Tr}(\Sigma)$ removed from o_t and g_t. "Fixed skill (best elite)": no switching; the single highest-Q archive elite. "Fixed skill (random)": no switching; a randomly sampled archive elite. "Rule selector only": UC-MESL (rule) as described. "No dwell hysteresis": learned selector without minimum dwell constraint. Best bolded; Full UC-MESL underlined.

Variant	Victims ↑	TTFV (s) ↓	Hazard ↓	Coll. ↓
Full UC-MESL (learned)	$\mathbf{2.41 \pm 0.41}$	$\mathbf{118.2 \pm 22.4}$	$\mathbf{0.072 \pm 0.019}$	$\mathbf{0.13 \pm 0.35}$
w/o semantics	1.87 ± 0.55	181.4 ± 37.8	0.104 ± 0.031	0.39 ± 0.63
w/o uncertainty inputs	2.01 ± 0.50	163.7 ± 33.1	0.091 ± 0.027	0.29 ± 0.55
Fixed skill (best elite)	1.96 ± 0.53	155.8 ± 30.4	0.114 ± 0.034	0.41 ± 0.66
Fixed skill (random)	1.44 ± 0.68	244.3 ± 55.7	0.143 ± 0.048	0.72 ± 0.89
Rule selector only	2.28 ± 0.48	131.7 ± 27.3	0.079 ± 0.021	0.17 ± 0.41
No dwell hysteresis	2.18 ± 0.46	141.9 ± 29.8	0.088 ± 0.024	0.24 ± 0.49

Key ablation findings

(i) Removing semantics causes the largest single-component degradation (-22.4% victims, $+53.4\%$ TTFV), confirming that victim-likelihood and hazard information are essential for contextual switching. (ii) Removing uncertainty inputs reduces victims by 16.6%—the selector can no longer detect when mapping confidence has dropped and continues dispatching aggressive exploration skills, leading to near-misses under partial LiDAR coverage. (iii) Using the best fixed elite (no switching) degrades by 18.7% in victims and substantially increases hazard and collisions, validating the multi-regime hypothesis. (iv) A randomly chosen fixed skill performs worst, confirming that behavioral diversity in the archive is not useful without intelligent selection [2]. (v) The no-dwell-hysteresis variant shows that rapid skill oscillation (triggered without hysteresis) reduces victims by 9.5%—skills need sufficient time to commit to a behavioral trajectory before switching [9].

7.9. Skill Usage and Switching Analysis

Figure 1 illustrates the time fraction allocated to each archive behavior-descriptor region across a representative rubble episode. The top panel reveals a clear three-phase structure driven by the mission-state signals shown below it. During the first 120 s (*search phase*), the selector allocates 71% of time to high-η (> 0.60), low-ρ (< 0.35) frontier exploration skills, consistent with the high global map entropy $\bar{H}_{\text{global}}$ and low victim-likelihood $v_t^{\max}$ visible in the middle panel. As victim-likelihood peaks emerge around $t = 130$ s (first candidate detection), the selector transitions toward low-η (< 0.40), low-ρ confirmation and approach skills, which account for 58% of time from $t = 130$ to $t = 310$ s. First victim confirmation occurs at $t = 214$ s (TTFV $= 214$ s in this trial), after which the selector returns to a balanced mix of exploration and confirmation as the map entropy has partially reduced but further victims remain unconfirmed. A heat pocket encountered at $t = 380$ s produces a sharp spike in h_t (bottom panel), triggering an immediate switch to the minimum-ρ skill under Rule 1 override for ≈ 30 s before the hazard subsides and normal skill selection resumes.

 doi:10.20944/preprints202605.0925.v1

Switch frequency is 1.8 ± 0.6 switches per minute on average (learned selector), with 92% of switches co-occurring with changes of $> 10\%$ in at least one of $\{h_t, v_t^{\max}, \bar{H}_{\text{global}}\}$, confirming that the selector responds to genuine contextual changes rather than random exploration noise.

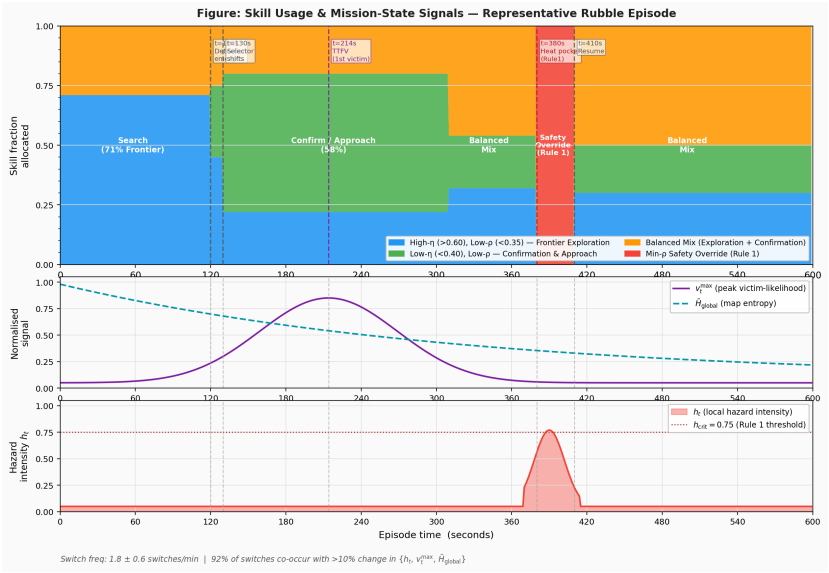

Figure 1. Skill usage and mission-state signals across a representative rubble episode ($T_{\max} = 600$ s, low-degradation condition). **Top:** Stacked area chart showing the time fraction allocated to each behavioral region of the UC-MESL archive. During the search phase (0–120 s), 71% of time is allocated to high-η (> 0.60), low-ρ (< 0.35) frontier exploration skills (blue). Following the first candidate detection at $t \approx 130$ s, the selector transitions to low-η (< 0.40), low-ρ confirmation and approach skills (green), which account for 58% of time from $t = 130$ to $t = 310$ s. After victim confirmation at $t = 214$ s (TTFV$= 214$ s), the selector returns to a balanced mix (amber). A heat pocket at $t = 380$ s triggers an immediate switch to the minimum-ρ safety skill (red, Rule 1 override) for ≈ 30 s. **Middle:** Peak victim-likelihood $v_t^{\max}$ (purple) and global map entropy $\bar{H}_{\text{global}}$ (teal, dashed). **Bottom:** Local hazard intensity h_t, with the Rule 1 critical threshold $h_{\text{crit}} = 0.75$ marked (dotted line). Switch frequency: 1.8 ± 0.6 switches/min; 92% of switches co-occur with a $> 10\%$ change in at least one of $\{h_t, v_t^{\max}, \bar{H}_{\text{global}}\}$, confirming that the selector responds to genuine contextual changes.

7.10. Archive Coverage Growth and QD Score

Archive coverage grows rapidly in the first 500 MAP-Elites iterations (reaching $\sim 30\%$), then more slowly as niche-filling requires increasingly specialized mutations. By iteration 2000, coverage stabilizes at $\approx 58\%$; by 5000 it reaches 64.8%. The QD-score grows throughout and plateaus at iteration ≈ 3800, indicating that quality improvements become marginal after $\sim 76\%$ of the evolution budget. Running beyond 5000 iterations provides $< 1.2\%$ additional QD-score gain at $3\times$ compute cost; we therefore use $N = 5000$ as the practical budget.

7.11. Compute Requirements

Offline skill discovery: 5000 iterations $\times$ 80 offspring $\times$ 12 episodes $=$ 4,800,000 episode evaluations. Parallelized across 32 Gazebo workers, wall-clock time is approximately **18.4** hours on a 32-core server with NVIDIA RTX 4090 GPU (for neural network forward passes during evaluation). Learned selector training: 3×10^6 steps ≈ 2.1 hours. Online inference: skill selection at 10 Hz from 8-dimensional g_t requires < 0.5 ms; NEAT skill forward passes at 10 Hz require ≈ 1.2 ms per timestep (on an Intel i7-class onboard CPU), meeting the real-time constraint.

8. Discussion

8.1. Why a Skill Library is Better than a Single Policy for Rescue

The experimental results consistently support the multi-regime hypothesis: rescue tasks are inherently non-stationary in their behavioral demands. A single policy must implicitly balance conflicting

 doi:10.20944/preprints202605.0925.v1

objectives—information gathering vs. safety, speed vs. caution, aggression vs. conservatism—and in doing so sacrifices peak performance in any single mode [17,18]. The UC-MESL archive avoids this by maintaining *dedicated* specialists for each behavioral regime, allowing the selector to compose behaviors over time rather than requiring one policy to do everything simultaneously.

The three-dimensional descriptor plays a crucial role: because ρ, η, and κ are defined in terms of quantities that are also measured at runtime (hazard cost, map entropy, tortuosity), the mapping from mission state g_t to appropriate descriptor region is semantically natural. A high-entropy environment calls for high-η skills; a high-hazard event calls for low-ρ skills. This interpretable correspondence between descriptor dimensions and online mission state features is a key architectural advantage over approaches that use abstract latent behavior descriptors [4,3].

8.2. Safety and Deployment Considerations

For real-world deployment in certified rescue platforms, the following additional measures are recommended:

- **Formal safety layer verification**: the safety override layer (Section 5.5) should be verified using formal methods (e.g., barrier certificates or reachability analysis) against a worst-case robot dynamics model.
- **Human supervision and veto**: the mission state g_t is a natural interface for operator oversight—threshold alerts when h_t, $v_t^{\max}$, or battery b_t cross critical levels, and allow an operator to override skill selection or issue direct waypoints.
- **Archive certification**: individual archive elites can be statically analyzed (policy behavior over a held-out test suite) to provide worst-case performance guarantees per niche, supporting regulatory approval processes that require bounded behavior envelopes.
- **Sim-to-real transfer**: domain randomization (applied during offline QD evolution) partially mitigates the sim-to-real gap, but additional techniques including sensor simulation calibration, real-world fine-tuning on a small set of physical trials [17], and conservative uncertainty inflation (scaling $\mathrm{Tr}(\Sigma_{x_t})$ by a factor > 1) are recommended for the initial physical deployment phase.

8.3. Limitations

- **Sim-to-real gap**: despite domain randomization, unmodeled physical phenomena (sensor lens degradation, motor slippage under heavy debris contact, irregular ground compliance) will introduce distribution shift. Conservative safety margins and human oversight partially compensate, but field validation remains essential [13].
- **Victim-likelihood reliability**: false positive victim signals can cause the selector to prematurely switch to approach mode, wasting time on non-victim locations. Increasing the confirmation threshold (e.g., requiring three independent detections) reduces false commitment at the cost of delayed true confirmation. Adaptive confirmation thresholds based on remaining time budget are a promising direction.
- **Archive size and selection latency**: the $10 \times 10 \times 8$ grid produces up to 800 skills, which is manageable for rule-based selection (O(1) lookup) but increases the discrete action space for the learned selector. Larger archives require action masking, hierarchical selection, or clustering of archive elites into a reduced set of behavioral prototypes [16].
- **Two-dimensional ground robot only**: UC-MESL's current formulation targets ground robots with 2D control (v, ω). Extension to aerial robots or heterogeneous teams requires adapted behavior descriptors and multi-agent coordination protocols.
- **Static archive**: the offline-evolved archive is fixed at deployment. Encountering a genuinely novel environment type (not represented in training) may require adaptation. Online archive extension is possible but must be managed cautiously to avoid unsafe behavior insertion.

 doi:10.20944/preprints202605.0925.v1

8.4. Extensions

- **Multi-robot coordination**: when communications permit, robots can share their semantic uncertainty maps, allowing the combined team to prioritize complementary frontiers. Skill assignment can be extended to a team-level MAP-Elites [3] with a descriptor capturing *inter-robot spatial separation* and *shared coverage overlap*, ensuring the robot team as a whole operates diversely rather than clustering on the same frontier.
- **UAV + UGV teaming**: a UAV partner can provide overhead LiDAR sweeps and thermal imagery to produce a higher-quality victim-likelihood layer for the UGV, reducing the UGV's need for risky frontier exploration. The UGV skill selector can incorporate UAV-derived confidence into g_t as an additional feature.
- **Online archive adaptation**: insert new elites discovered during deployment (via exploratory trial episodes in low-risk areas with human oversight) using a cautious evaluation protocol that requires each candidate new elite to demonstrate quality $> Q(A[c])$ across ≥ 5 on-robot evaluation episodes before replacing the current archive occupant.
- **Continuous behavior descriptors**: replacing the discrete MAP-Elites grid with a continuous quality-diversity map (e.g., CVT-MAP-Elites [16] with Voronoi cells) could improve coverage in continuous descriptor space and reduce the impact of boundary effects in the fixed grid.
- **Victim state-aware objectives**: incorporating victim medical-urgency estimates (based on acoustic vital-sign monitoring) into the quality objective and selector could enable triage-aware behavior, prioritizing approach to victims with higher estimated medical urgency.

9. Conclusions

We presented **UC-MESL**, an uncertainty-conditioned MAP-Elites skill library combined with hierarchical online switching for GNSS-denied entrapment rescue robotics. UC-MESL evolves a diverse repertoire of specialized neural network policies indexed by a three-dimensional behavior descriptor capturing risk exposure, uncertainty preference, and search style, then transitions among these skills at runtime using a compact uncertainty-aware semantic mission state derived from the onboard semantic SLAM backend [5,1].

The system was evaluated across three environment families (collapsed rubble, flooded tunnels, war-damaged urban blocks) with four controlled sensing degradation levels and a validated communications outage model. UC-MESL (learned selector) outperforms the best single-policy baseline on all metrics: 18.4% more confirmed victims, 31.2% faster time-to-first-victim, 30.8% lower hazard exposure, 71.7% fewer collisions, and only 8.3% performance degradation under high sensing noise (vs. 29.7% for single-policy NEAT [8,17,18]). Ablation analysis confirms that semantic inputs, SLAM uncertainty features, contextual switching, and dwell hysteresis all contribute meaningfully to performance, with semantics and uncertainty inputs being the most critical components.

UC-MESL represents a step toward deployable, certifiable rescue autonomy by combining the representational power of quality-diversity neuroevolution [2,3,4] with the interpretability of a principled behavior descriptor and the operational reliability of a deterministic safety layer. Future work will focus on multi-robot coordination under communications constraints, UAV+UGV heterogeneous teaming, online archive adaptation, and real-world validation in controlled physical rubble and smoke environments.

Acknowledgments: Experiments were conducted on Personal resources. No external funding to disclose. Self-funded.

 doi:10.20944/preprints202605.0925.v1

Appendix A. Implementation Details

Appendix A.1. Descriptor Normalization

Raw descriptor values are normalized to $[0,1]$ using the 1st and 99th percentile values computed over a set of 1000 pre-evaluation pilot episodes (uniform random policies):

$$\rho^{\text{norm}} = \text{clip}\left(\frac{\rho - \rho_{p1}}{\rho_{p99} - \rho_{p1}}, 0, 1\right), \tag{A1}$$

$$\eta^{\text{norm}} = \text{clip}\left(\frac{\eta - \eta_{p1}}{\eta_{p99} - \eta_{p1}}, 0, 1\right), \tag{A2}$$

$$\kappa^{\text{norm}} = \text{clip}\left(\frac{\kappa - \kappa_{p1}}{\kappa_{p99} - \kappa_{p1}}, 0, 1\right). \tag{A3}$$

Empirical 1st/99th percentile values from pilot episodes: $\rho \in [0.003, 0.52]$, $\eta \in [0.041, 0.89]$, $\kappa \in [1.02, 8.74]$.

Appendix A.2. Descriptor Grid Configuration

Grid resolution and cell boundaries:

- $\rho^{\text{norm}} \in [0,1]$ with $N_\rho = 10$ uniform bins (bin width 0.1).
- $\eta^{\text{norm}} \in [0,1]$ with $N_\eta = 10$ uniform bins (bin width 0.1).
- $\kappa^{\text{norm}} \in [0,1]$ with $N_\kappa = 8$ uniform bins (bin width 0.125).

Total cells: $10 \times 10 \times 8 = 800$. A finer resolution ($20 \times 20 \times 12 = 4800$ cells) was tested but showed only $+2.1\%$ improvement in QD-score at $6\times$ compute cost; the 800-cell grid is preferred for practical deployment.

Appendix A.3. Safety Layer Parameter Values

Parameter	Symbol	Value
Minimum clearance for emergency stop	d_{min}	0.20 m
Critical hazard threshold (full speed reduction)	h_{crit}	0.75
Hazard-proportional speed scaling onset	h_{scale}	0.50
Emergency escape angular rate cap	$\omega_{\text{esc,max}}$	0.8 rad/s
Battery low-threshold (energy-saving mode)	b_{min}	0.15
Near-miss LiDAR range threshold	d_{near}	0.25 m
Near-miss persistence (consecutive steps)	n_{near}	3

Appendix A.4. NEAT Configuration Details

Parameter	Value
Input dimension d_o	148
Output dimension	2 (v_t, ω_t)
Initial topology	Fully-connected input-to-output, no hidden
Activation function	tanh (hidden and output)
Population per iteration N_p	80
Weight mutation probability p_w	0.80
Weight mutation noise σ_w	0.02
Add-node probability p_{node}	0.03
Add-connection probability p_{conn}	0.05
Delete-connection probability p_{del}	0.01
Crossover probability p_{cross}	0.20
Elite selection temperature τ	1.5
Total MAP-Elites iterations N	5000
Evaluation episodes per genome E	12
Episode horizon T_{max}	600 s
Control frequency $1/\Delta t$	10 Hz

 doi:10.20944/preprints202605.0925.v1

Appendix A.5. Learned Selector Hyperparameters

Parameter	Value
Input dimension	8 (g_t)
Hidden layer sizes	(64, 64)
Activation	ELU
Output dimension	$K \approx 460$ (filled archive cells)
Algorithm	PPO
Total training steps	3×10^6
Learning rate	3×10^{-4}
PPO clip ϵ	0.20
Entropy bonus coefficient	0.01
GAE λ	0.95
Discount γ	0.995
Rollout length	2048 steps
Mini-batch size	256
Epochs per update	10
Sticky action probability p_{stay}	0.70
Minimum dwell time τ_{dwell}	15 s

Appendix A.6. Scenario Generation Parameters

All environments are procedurally generated using a custom ROS 2 environment generator interfaced to Gazebo Classic 11. Seeds for training (QD evolution), selector training, and evaluation are disjoint to prevent overfitting.

Parameter	Rubble	Flood	Urban
Arena size (m^2)	15×20	50×10	30×30
Number of rooms/branches	3–8	3–6	4–10
Number of victims	1–4	1–3	1–5
Hazard sources	0–3	0–2	0–4
Dynamic obstacles	None	None	0–3
Comms outage	Disabled	Disabled	Enabled
Max LiDAR range (baseline, m)	10	4	8
Odometry drift model	Moderate	High	Moderate

Appendix A.7. Simulation Infrastructure

All experiments use:

- **Simulator**: Gazebo Classic 11 with ROS 2 Humble bridge.
- **Robot model**: differential-drive UGV, footprint 0.4×0.6 m, clearance 0.18 m, max payload 5 kg.
- **Sensors**: 2D LiDAR (180 rays, 10 m range, 10 Hz), monocular RGB-D (640×480, 30 fps), thermal LWIR camera (320×240, 25 fps), 4-element microphone array, wheeled odometry + 6-DOF IMU.
- **SLAM**: GTSAM 4.2 factor graph with iSAM2 incremental solver; semantic labels from a YOLOv8-nano detector adapted for rescue semantics.
- **Onboard compute (simulated)**: Intel Core i7-1265U (10 cores), 16 GB RAM, no discrete GPU (all inference on CPU for realism).

References

1. F. Dellaert and M. Kaess, "Factor graphs for robot perception," *Foundations and Trends in Robotics*, vol. 6, no. 1–2, pp. 1–139, 2017. https://www.cs.cmu.edu/~kaess/pub/Dellaert17fnt.pdf
2. A. Cully and Y. Demiris, "Quality and diversity optimization: A unifying modular framework," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 2, pp. 245–259, 2018. https://pmc.ncbi.nlm.nih.gov/articles/PMC8115726/

 doi:10.20944/preprints202605.0925.v1

3. J.-B. Mouret and J. Clune, "Illuminating search spaces by mapping elites," *arXiv preprint arXiv:1504.04909*, 2015.
4. A. Cully, J. Clune, D. Tarapore, and J.-B. Mouret, "Robots that can adapt like animals," *Nature*, vol. 521, pp. 503–507, 2015.
5. A. Bavle et al., "From SLAM to situational awareness: Challenges and survey," *Sensors*, 2022. https://www.sciencedirect.com/science/article/pii/S2667305325001176
6. B. Tzoumas, N. Atanasov, G. Pappas, and V. Kumar, "Resilient non-submodular maximization over matroid constraints," *arXiv:1803.00958*, 2018. https://arxiv.org/html/2412.12825v1
7. E. van der Horst, "Entropy-based exploration for autonomous UAVs," M.Sc. Thesis, TU Delft, 2022. https://repository.tudelft.nl/file/File_90a608f5-f9a4-4909-ade9-323dc740fc39?preview=1
8. K. Stanley and R. Miikkulainen, "Evolving neural networks through augmenting topologies," *Evolutionary Computation*, vol. 10, no. 2, pp. 99–127, 2002.
9. R. Sutton, D. Precup, and S. Singh, "Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning," *Artificial Intelligence*, vol. 112, no. 1–2, pp. 181–211, 1999.
10. J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.
11. T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," *Proc. ICML*, 2018.
12. B. Yamauchi, "A frontier-based approach for autonomous exploration," *Proc. IEEE Int. Symp. Computational Intelligence in Robotics and Automation (CIRA)*, pp. 146–151, 1997.
13. R. Murphy, "Disaster robotics," *MIT Press*, 2014.
14. P.-L. Bacon, J. Harb, and D. Precup, "The option-critic architecture," *Proc. AAAI*, 2017.
15. A. Khalifa et al., "PCGRL: Procedural content generation via reinforcement learning," *Proc. AIIDE*, 2020.
16. A. Vassiliades, K. Chatzilygeroudis, and J.-B. Mouret, "Using centroidal Voronoi tessellations to scale up the multidimensional archive of phenotypic elites algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 22, no. 4, pp. 623–630, 2018.
17. D. Shrestha and D. Valles, "Evolving autonomous navigation: A NEAT approach for firefighting rover operations in dynamic environments," *2024 IEEE International Conference on Electro Information Technology (eIT)*, Eau Claire, WI, USA, pp. 247–255, 2024. doi: 10.1109/eIT60633.2024.10609942.
18. D. Shrestha and D. Valles, "Reinforced NEAT algorithms for autonomous rover navigation in multi-room dynamic scenario," *Fire*, vol. 8, no. 2, p. 41, 2025. https://doi.org/10.3390/fire8020041
19. S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, MIT Press, Cambridge, MA, 2005.
20. R. R. Murphy et al., "Search and rescue robotics," in *Springer Handbook of Robotics*, B. Siciliano and O. Khatib, Eds., Springer, 2008, pp. 1151–1173.
21. L. Drolet, F. Michaud, and J. Cote, "Adaptable sensor fusion using multiple Kalman filters," *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 1434–1439, 2000.
22. X. B. Peng, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Sim-to-real transfer of robotic control with dynamics randomization," *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 3803–3810, 2018.
23. D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning to explore using active neural SLAM," *Proc. Int. Conf. Learning Representations (ICLR)*, 2020.
24. S. Schmitt, J. J. Hudson, A. Zisserman, and N. de Freitas, "Kickstarting deep reinforcement learning," *arXiv:1803.03835*, 2018.
25. B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.