Neuro-Symbolic Risk Semantics for Safe Multi-Agent Coverage in Hazardous Environments

Nirali Sanghvi
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
Roorkee, India 247667
nirali_s@cs.iitr.ac.in

Rajdeep Niyogi*
Department of Computer Science and Engineering
Indian Institute of Technology Roorkee
Roorkee, India 247667
rajdeep.niyogi@cs.iitr.ac.in

***Abstract*—Ensuring safety in Multi-Agent Reinforcement Learning (MARL) is critical for real-world deployment, but remains brittle under epistemic mismatch, scalar penalties obscure failure mechanisms, and fixed symbolic priors can become miscalibrated under distribution shift. We introduce FMEA-Net, a neuro-symbolic risk semantics module for safe multi-agent coverage that makes engineering-style Failure Mode and Effects Analysis (FMEA) differentiable and adaptive. A neural cause recognizer maps each agent's local observation to hazard causes, which is propagated through learnable cause-to-mode and mode-to-effect tables to obtain interpretable mode/effect beliefs and a severity-weighted risk score. These outputs form a compact safety context that conditions decentralized actors, enabling agents to distinguish why a region is hazardous (e.g., slip vs. collision), rather than merely how much and to learn risk-aware behaviors without collapsing coverage. Across hazardous multi-agent coverage benchmarks, the proposed method improves safe coverage and reduces safety violations and overall risk relative to baselines that rely on unstructured scalar penalties, with consistent gains across unseen maps and risk-model mismatch.**

***Index Terms*—Neuro-symbolic AI, Multi-agent reinforcement learning, Safe coverage**

I. Introduction

Neural policies have become central to Multi-Agent Reinforcement Learning (MARL), but safety remains a major obstacle to real-world deployment. In safety-critical settings, agents must not only achieve high task performance but also avoid rare and severe failures arising from uncertain environments and multi-agent interactions. Existing safe RL and safe MARL methods provide useful tools, including constrained optimization, Lagrangian relaxation, safety games, local shields, and barrier functions [1]–[4]. However, most of these approaches represent safety through scalar costs or fixed constraints, and usually assume that the underlying safety model is correct. This is limiting in practice: very different hazards, such as slippery terrain, obstacle-dense corridors, or localization-poor regions, may produce similar scalar risk values while requiring different responses. Under such mismatch, agents may either become overly conservative or exploit errors in the safety signal rather than learn genuinely safe behavior.

Neuro-symbolic AI offers a promising alternative by combining neural learning with structured symbolic reasoning [5], [6]. Instead of relying only on a single penalty signal, it represents intermediate safety-relevant factors, such as events, constraints, and causal relations, in an interpretable form. This is particularly attractive for safety-critical decision-making because it helps explain why a situation is unsafe, not merely how unsafe it is. Recent work in robotics further suggests that structured symbolic abstractions can improve interpretability and generalization in sequential decision-making [7]. However, existing neuro-symbolic safety methods, including probabilistic logic shielding and differentiable barrier functions [8]–[10], typically assume that the symbolic safety model is fixed and correct. In practice, this assumption may fail as hazard statistics, sensing quality, and interaction patterns change over time. Although runtime enforcement may block unsafe actions, it does not correct mismatch in the underlying risk model.

A related line of work in robotics uses structured causal priors, particularly Failure Mode and Effects Analysis (FMEA), which represents risk through typed chains of the form Cause $\rightarrow$ Mode $\rightarrow$ Effect [11]. This representation provides the interpretability needed in safety-critical systems, but existing FMEA-inspired approaches are generally static and non-differentiable. More broadly, robustness to risk-model mismatch remains underexplored in safe MARL, and common remedies such as domain randomization or worst-case optimization often become overly conservative under distribution shift [12]. These limitations motivate a framework in which the symbolic safety model remains interpretable yet can adapt online when deployment conditions deviate from prior assumptions.

In this work, we propose FMEA-Net, a neuro-symbolic architecture that renders structured risk semantics differentiable and adaptive. We demonstrate that learning a structured, auditable risk model jointly with a neural policy significantly improves robustness when safety priors are imperfect. Our contributions are:

1) **Neuro-symbolic semantics as a learnable safety primitive:** We develop FMEA-Net, a differentiable cause-modeeffect semantics module that parameterizes $P_\theta(M \mid C)$ and $P_\phi(E \mid M)$, is initialized from an expert prior, and supports online calibration from rollout-derived weak supervision with explicit prior anchoring to preserve semantic interpretability.
2) **Architecture-level injection of symbolic knowledge**

*Corresponding author.

into MARL: We present an actor-critic design in which calibrated semantic beliefs are injected as an auxiliary context to decentralized actors within Centralized Training Decentralized Execution (CTDE) training, yielding an explicit interface between symbolic safety reasoning and gradient-based policy optimization.

3) **Mismatch-focused evaluation and diagnostics:** We construct a suite of stress tests that perturb the symbolic prior (e.g., link permutations, sparse corruptions, and severity re-scaling) and evaluate generalization under map and noise shifts. Alongside task and safety metrics, we report semantic calibration indicators (e.g., table distance / anchored drift) to isolate the effect of online semantic correction. baselines.

II. Problem Formulation

We model the multi-agent safe coverage problem as a decentralized partially observable Markov decision process (Dec-POMDP), defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \Omega, O, \mathcal{R}, \gamma \rangle$ on a discrete grid $\mathcal{G} = \{1, \dots, W\} \times \{1, \dots, H\}$.

A. Environment and Dynamics

The system consists of N agents indexed by $\mathcal{N} = \{1, \dots, N\}$. The environment contains a static hazard field $\mathbf{H} : \mathcal{G} \to \{0, \dots, K\}$, where $\mathbf{H}(c) = 0$ denotes a safe cell and $\mathbf{H}(c) = k \in \{1, \dots, K\}$ denotes a hazard of class k.

State The global state at time t is $s_t = (\mathbf{p}_t, \mathbf{H}) \in \mathcal{S}$, where $\mathbf{p}_t = (p_t^1, \dots, p_t^N)$ represents the joint configuration, and $p_t^i \in \mathcal{G}$ is the coordinate position of agent i.

Actions and Transitions At each timestep t, agent i selects an action $a_t^i \in \mathcal{A} = \{\text{up, down, left, right, stay}\}$. Let $\Delta : \mathcal{A} \to \mathbb{Z}^2$ map actions to unit displacements. The state transitions deterministically according to the position update: $p_{t+1}^i = \text{proj}_{\mathcal{G}}\big(p_t^i + \Delta(a_t^i)\big)$ where $\text{proj}_{\mathcal{G}}(\cdot)$ clamps coordinates to the grid boundaries. The joint action is denoted $\mathbf{a}_t = (a_t^1, \dots, a_t^N)$.

Observations Agents operate under partial observability. Each agent i receives a local observation $o_t^i = O(s_t, i) \in \Omega$, which includes: (i) its own position p_t^i, (ii) relative positions of teammates within a communication radius R_{com}, and (iii) a local view of the hazard map $\mathbf{H}$ within a sensing radius R_{obs} centered at p_t^i.

B. Safe Coverage

We define the set of safe cells as $\mathcal{G}_{\text{safe}} = \{c \in \mathcal{G} \mid \mathbf{H}(c) = 0\}$. The history of visited locations by the team up to time t is given by the set $\mathcal{V}_t = \bigcup_{\tau=0}^{t} \bigcup_{i=1}^{N} \{p_\tau^i\}$. The coverage metric $C_{\text{safe}}(t)$ is the fraction of safe terrain explored:

$$C_{\text{safe}}(t) = \frac{|\mathcal{V}_t \cap \mathcal{G}_{\text{safe}}|}{|\mathcal{G}_{\text{safe}}|} \in [0, 1]. \tag{1}$$

C. Violations and Risk

Violations A safety violation occurs when an agent enters a hazardous cell. We define a binary indicator for agent i at time t as $\mathbb{I}_t^i = \mathbb{1}[\mathbf{H}(p_t^i) \neq 0]$. The total number of violations for the team over a horizon T is: $V = \sum_{t=0}^{T-1} \sum_{i=1}^{N} \mathbb{I}_t^i$.

Weighted Risk To differentiate between hazard severities (e.g., minor vs. catastrophic), we assign a scalar risk weight $w_k \geq 0$ to each hazard class k. The instantaneous risk cost for agent i is defined as: $\rho_t^i = w_{\mathbf{H}(p_t^i)} \cdot \mathbb{I}_t^i$. The cumulative risk for the team over an episode is $J_{\text{risk}} = \sum_{t=0}^{T-1} \sum_{i=1}^{N} \rho_t^i$.

D. Learning Objective

The problem is formulated as Centralized Training with Decentralized Execution (CTDE). Agents learn a joint policy $\pi(\mathbf{a}_t \mid \mathbf{o}_t) = \prod_{i=1}^{N} \pi^i(a_t^i \mid o_t^i)$ parameterized by θ. The objective is to maximize the expected discounted return regarding task performance while minimizing cumulative risk:

$$\max_{\theta} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \gamma^t \mathcal{R}_{\text{task}}(s_t, \mathbf{a}_t) \right] \quad \text{s.t.} \quad J_{\text{risk}} \leq \delta, \tag{2}$$

where $\mathcal{R}_{\text{task}}$ is the coverage reward and δ is a safety tolerance threshold.

III. Proposed Approach

We propose FMEA-Net, a safety-aware learning architecture that combines (i) an explicit and interpretable cause $\to$ failure mode $\to$ effect risk chain as an inductive bias and (ii) online calibration of this structured risk model under deployment mismatch. Structured expert priors improve interpretability and sample efficiency, but can be brittle under distribution shift; FMEA-Net preserves these benefits while reducing brittleness through experience-driven semantic adaptation.

At each time step, agent i receives a local observation and outputs both a control action and a compact safety summary from FMEA-Net. This safety context captures likely hazard causes, possible failure modes, and the severity of downstream effects. The policy is conditioned on this context during training and execution, enabling coverage-oriented behavior with reduced severity-weighted risk. During training, the scalar risk score from FMEA-Net serves as an auxiliary learning signal for safer behavior. At deployment, we freeze the control policy and adapt only the risk-semantics module, thereby separating stable control from recalibrated risk perception under changing terrain, sensing quality, or interaction conditions.

A. FMEA-Net: Learnable Structured Risk Semantics

FMEA-Net is a differentiable structured module with three stages: (1) a cause recognizer that maps observations to a belief over hazard causes, (2) a structured propagation stage that maps causes to failure modes and then to effects using learnable conditional tables initialized from an expert prior, and (3) a risk aggregator that outputs both a severity-weighted scalar risk and a compact safety context used for conditioning.

From observations to a cause belief The first stage converts raw local observations into a categorical belief over a discrete set of human-interpretable hazard causes (e.g., slippery terrain, loose gravel, obstacle-dense corridor, low visibility/noise, communication-interference regions). We implement this with a neural encoder that outputs a normalized distribution: $\mathbf{p}_C^i(t) = \text{softmax}\big(f_\phi(o_t^i)\big)$. This design is sensor-agnostic: it only requires that available sensing can be mapped to a cause-belief distribution.

Structured propagation through causes, modes, and effects Rather than collapsing observations into an opaque scalar

penalty, FMEA-Net explicitly models why risk arises (causes), how it manifests (failure modes), and what consequences it can produce (effects). We represent this mechanism using two learnable conditional tables initialized from an expert-inspired prior and refined by calibration: one mapping causes to modes and one mapping modes to effects. Given the cause belief, we propagate beliefs through the chain:

$$\mathbf{p}_M^i(t) = \mathbf{p}_C^i(t) W_{CM}, \qquad \mathbf{p}_E^i(t) = \mathbf{p}_M^i(t) W_{ME}. \quad (3)$$

This explicit factorization preserves interpretability and enables auditing at each stage (cause, mode, effect), instead of exposing only a single scalar penalty.

Severity-weighted risk score and safety context To obtain an actionable scalar signal, we assign each effect a nonnegative severity weight (derived from the prior) and compute a severity-weighted expectation under the effect belief:

$$\rho_t^i = \langle \mathbf{p}_E^i(t), \mathbf{s} \rangle. \quad (4)$$

FMEA-Net also produces a compact safety context vector used to condition the actor-critic; in practice, this context is built from interpretable quantities such as mode/effect beliefs, the scalar risk score, and optional confidence features from the cause recognizer. During training, we incorporate the risk score via severity-aware reward shaping to improve credit assignment when unsafe events are rare:$\tilde{r}_t^i = r_t^i - \beta\, \rho_t^i$, where β controls the task-safety trade-off.

Online Calibration and Deployment-Time Adaptation Expert priors may become inaccurate in new environments, for example due to terrain changes, sensor degradation, or shifts in interaction density, causing mismatch between the prior tables and the true hazard-to-failure relations. To address this, we adapt only the risk-semantics parameters using weak supervision from onboard diagnostics, such as IMU signatures for slip, impact patterns for collision, localization residual spikes for drift, and communication logs for dropouts. Updates are triggered only when the evidence is sufficiently confident: $\mathbb{I}_{\text{update}} = \mathbb{1}\left[q_t^i > \tau_{\text{conf}}\right]$. When triggered, we apply bounded gradient updates to reduce prediction–reality mismatch while anchoring the learned semantics to the prior. The calibration objective combines a data-fit term with a prior regularizer:

$$\mathcal{L}_{\text{cal}} = \mathcal{L}_{\text{CE}}\big(\mathbf{p}_Y^i(t), \hat{\mathbf{y}}_t^i\big) + \lambda_{\text{reg}}\, D_{\text{KL}}(P_\Theta \,\|\, P_{\text{prior}}). \quad (5)$$

For stability, we maintain a calibration buffer with both normal and failure-associated events, clip updates to avoid abrupt semantic changes, and separate timescales so that semantic updates remain small and frequent while policy learning follows standard MAPPO.

Finally, we combine FMEA-Net with MAPPO by conditioning the actor-critic on safety context and training with a shaped reward. At deployment, we freeze the control policy and calibrate only the risk semantics, enabling stable control with adaptive risk perception under prior mismatch. We study multi-agent safe coverage under uncertain and misspecified hazard semantics, with emphasis on the decision layer: learning and calibrating structured risk semantics for safe exploration and shielding. Experiments use discretized grid maps with occupancy, traversability, and severity-coded hazards, focusing on semantics-aware risk modeling and control rather than end-to-end perception or continuous-time dynamics.

IV. Experimental Results

We evaluate the proposed neuro-symbolic semantics calibration framework (FMEA-Net) against strong PPO-family MARL baselines and address three research questions:

RQ1 (SafetyEfficiency Trade-off): Can safety-conditioned policies achieve high safe coverage while reducing violations and severity-weighted risk exposure?

RQ2 (Robustness to Mismatch): Does online semantic calibration mitigate performance degradation when the symbolic safety prior is corrupted or hazard statistics shift?

RQ3 (Interpretability and Alignment): Do the learned semantics correct the specific mis-specified entries in the prior while remaining stable (i.e., avoiding uncontrolled drift)?

A. Evaluation Protocol, Suites, and Metrics

Our evaluation isolates the effect of structured risk semantics and online semantic calibration. Unless stated otherwise, all numbers are reported as mean $\pm$ standard deviation over multiple training seeds, with each seed evaluated on a fixed set of test maps per suite.

TABLE I
Evaluation suites used in Section IV.

Suite	Description
ID (in-distribution)	Train/test share hazard statistics; the prior P_0 matches the environment.
UM (unseen maps)	Test on held-out maps/seeds not seen during training (same hazard classes).
PM (prior mismatch)	Corrupt the symbolic prior P_0 (permutations, wrong links, wrong severity scaling).
SN (sensor noise)	Increase observation/actuation noise at test time while freezing the policy.

Task metric (safe coverage). Safe coverage $\mathcal{G}_{\text{safe}} \in [0, 100]$ measures the fraction of traversable area covered.

Safety metrics (violations and risk exposure). We report: (i) violations V (count of unsafe events during execution), and (ii) severity-weighted cumulative risk $\sum_t \rho_t$ computed from the per-step risk score produced by the semantics module (higher-severity effects contribute more). We emphasize $\sum_t \rho_t$ because it captures both frequency and consequence of unsafe behavior.

Semantic calibration indicators. To verify that semantic learning is doing meaningful work (and not merely acting as another penalty), we monitor (i) whether specific corrupted links in $P_0(M \mid C)$ and $P_0(E \mid M)$ move toward their corrected values (Tables VII,VIII), and (ii) whether updates remain stable via prior-anchoring (qualitatively reflected by smooth convergence in Fig. 2). The diagnostic signal is not used as an action input to the policy and is not an oracle supervisor. It is a noisy, event-driven self-supervision cue available only after an interaction (e.g., slip signature, collision impulse, wheel-IMU discrepancy), used solely to calibrate semantic parameters. The control policy executes under the

same partial observations as baselines; calibration updates are gated by confidence to prevent drift under noisy diagnostics.

B. Experimental Setup

We evaluate across map sizes, team sizes (N agents), and hazard prevalence to stress coordination under long horizons and heterogeneous risk. Unless otherwise stated, baselines share the same MAPPO training budget and backbone capacity; methods differ only in how risk is represented and updated (unstructured scalar penalty vs. structured semantics; fixed prior vs. calibrated semantics). Values differ across tables because each table uses a different held-out test map set (and seeds), so compare methods within the same table.

C. Compared Methods

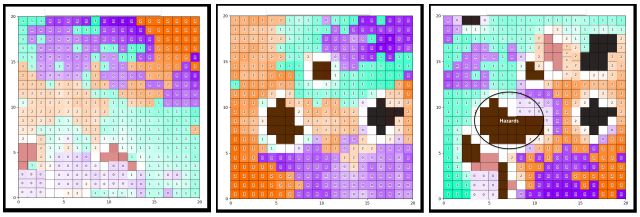
MAPPO (task-only). Standard MAPPO trained only for the task objective; safety is not explicitly modeled. **MAPPO + scalar risk penalty.** MAPPO with an additional hand-designed scalar risk penalty, without structured causemodeeffect semantics. **MACPO.** A constrained/safe MARL baseline representative of CMDP-style optimization using a constraint/cost signal. **Static-FMEA.** MAPPO conditioned on a structured prior $P_0(M \mid C), P_0(E \mid M)$, but the semantics are fixed (no test-time correction). **FMEA-Net (ours).** MAPPO conditioned on a differentiable semantics module initialized from P_0, with online calibration that corrects mis-specified links while anchoring to the prior to prevent drift.

TABLE II
COMPARISON MAP SIZE= 20×20, N=5, MEDIUM DENSITY).

Method	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
MAPPO [13]	**93.56 ± 1.69**	1.78 ± 0.14	0.96 ± 0.29
MAPPO + scalar risk penalty	87.10 ± 1.37	1.02 ± 0.11	0.70 ± 0.27
MACPO [14]	86.90 ± 1.64	0.57 ± 0.23	0.53 ± 0.05
Static-FMEA	84.98 ± 1.47	0.71 ± 0.28	0.83 ± 0.27
Ours	91.50 ± 1.71	**0.52 ± 0.17**	**0.51 ± 0.16**

The unstructured scalar penalty reduces violations but can induce conservative behavior (coverage drop), while fixed structured semantics can reduce risk but may misguide decisions if the prior is imperfect. Our calibrated semantics retain high coverage while minimizing both violations and severity-weighted exposure, consistent with the goal of semantic safety conditioning rather than pure penalty tuning.

Hyperparameters. Unless stated otherwise, we use the following values, horizon $T_{\max} = 2500$, $\Delta t = 0.5$, and maximum speed 2.0, with a fixed map per episode, 20 training maps (seed base 123) sampled cyclically, and coverage threshold 0.90. We enable FMEA-Net when `risk_mode` starts with `fmea_net` and set $\alpha_{\text{risk}} = 0.2$, $\lambda_{\text{KL}} = 0.05$, batch size 256, and one calibration update per iteration, optimized with Adam (10^{-3}) and an event replay buffer of size 50000. MAPPO uses $\gamma = 0.99$, 5 epochs per update, batch size 1024, actor/critic learning rates $3{\times}10^{-4}$ and 10^{-3}, and two-layer MLPs (width 256, ReLU) for actor and centralized critic; episodes terminate at $T_{\max}$ (optionally early at the safe-coverage threshold).



(a) (b) (c)

Fig. 1. **Hazard-density sweep.** Representative 20×20 test maps with 3 agents under (a) low, (b) medium, and (c) high hazard density. Hazard severity is color-coded as pink (low risk), brown (medium risk), and black (high risk). The mean $(C_{\text{safe}}, \sum_t \rho_t)$ values are: low $(91.20 \pm 1.10,\ 0.70 \pm 0.09)$, medium $(89.60 \pm 1.15,\ 1.12 \pm 0.14)$, and high $(85.10 \pm 1.50,\ 1.85 \pm 0.23)$.

Network architecture and module sizes. Each agent encodes its local observation with a shared CNN: $\text{Conv2D}(32, 3 \times 3, \text{ReLU})$, $\text{MaxPool}(2 \times 2)$, $\text{Conv2D}(64, 3 \times 3, \text{ReLU})$, followed by flattening and $\text{FC}(128, \text{ReLU})$ to obtain an embedding z_t^i. A cause-belief head maps z_t^i through $\text{FC}(64, \text{ReLU})$ and $\text{FC}(|\mathcal{C}|, \text{softmax})$ to produce $p(C \mid o_t^i)$. FMEA-Net performs differentiable cause-mode-effect propagation using two linear (no-bias) tables W_{CM} and W_{ME}, and aggregates effect beliefs into a scalar risk via a fixed severity vector; the actor conditions on a safety context g_t^i formed by concatenating the relevant beliefs and risk. The decentralized actor is an MLP with three $\text{FC}(128, \text{ReLU})$ layers and a softmax output (optionally a GRU with hidden size 128 for recurrent policies). Under CTDE, the centralized critic pools per-agent features/context (e.g., mean pooling) and applies three $\text{FC}(128, \text{ReLU})$ layers followed by a linear value head.

D. Main Results Across Evaluation Suites

Table III summarizes performance under the four suites (Table I). Three trends are consistent across suites:

(1) Structured semantics improve safety signals beyond scalar penalties. Relative to MAPPO (task-only), adding any explicit safety mechanism reduces V and $\sum_t \rho_t$, but the form of safety matters. Unstructured scalar penalties reduce violations but often trade away coverage, indicating that the agent learns "avoid everything" rather than why hazards matter.

(2) Fixed priors help only when the prior is correct. Static-FMEA is strong in ID/UM because the prior is aligned. However, in PM (prior corruption), fixed semantics become actively harmful: the policy is conditioned on incorrect causal links and cannot repair them, leading to elevated violations and risk exposure.

(3) Online semantic calibration is most valuable under mismatch and noise. Under PM, our method improves $\mathcal{G}_{\text{safe}}$ by $+5.91$ points and reduces cumulative risk by $\approx 57\%$ relative to Static-FMEA (from 2.31 to 0.99), while also lowering violations by $\approx 58\%$ (from 2.79 to 1.18). Under SN, calibration remains beneficial even with a frozen policy: compared to Static-FMEA, we gain $+5.80$ points in $\mathcal{G}_{\text{safe}}$ and reduce both V and $\sum_t \rho_t$ by $\approx 41\%$ (from 3.08 to 1.82 and from 2.41 to 1.43, respectively). In ID, we reduce cumulative risk by $\approx 68\%$ (from 1.54 to 0.50) relative to MAPPO (task-only) with only a 1.36 point reduction in $\mathcal{G}_{\text{safe}}$, reflecting a favorable safety–efficiency shift.

TABLE III

MAIN RESULTS ACROSS EVALUATION SUITES. METRICS: SAFE COVERAGE $\mathcal{G}_{\text{SAFE}} \uparrow$, VIOLATIONS $V \downarrow$, AND SEVERITY-WEIGHTED CUMULATIVE RISK $\sum_t \rho_t \downarrow$. VALUES ARE MEAN $\pm$ STD OVER SEEDS. BEST IS **BOLD**; SECOND BEST IS UNDERLINED.

Method	ID			UM			PM			SN		
	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
MAPPO (task-only)	**93.20 ± 1.08**	1.78 ± 0.14	1.54 ± 0.53	86.53 ± 1.41	2.64 ± 0.21	2.06 ± 0.64	84.01 ± 1.72	3.19 ± 0.27	2.58 ± 0.78	80.18 ± 2.06	4.04 ± 0.33	3.14 ± 0.96
MAPPO + scalar risk penalty	90.12 ± 1.15	1.02 ± 0.11	0.77 ± 0.37	86.97 ± 1.36	1.68 ± 0.17	1.25 ± 0.49	85.09 ± 1.64	2.14 ± 0.22	1.66 ± 0.63	81.31 ± 1.98	2.83 ± 0.29	2.13 ± 0.81
MACPO [14]	88.03 ± 1.29	0.57 ± 0.23	0.53 ± 0.24	84.21 ± 1.52	1.09 ± 0.12	0.91 ± 0.38	80.47 ± 1.88	1.63 ± 0.16	1.29 ± 0.52	76.79 ± 2.21	2.04 ± 0.19	1.67 ± 0.66
Static-FMEA	92.01 ± 1.02	0.71 ± 0.28	0.57 ± 0.26	88.77 ± 1.28	1.26 ± 0.13	1.04 ± 0.41	82.52 ± 1.95	2.79 ± 0.25	2.31 ± 0.72	79.12 ± 2.34	3.08 ± 0.28	2.41 ± 0.83
Ours	91.84 ± 0.97	**0.52 ± 0.17**	**0.50 ± 0.23**	**89.58 ± 1.16**	**1.01 ± 0.11**	**0.83 ± 0.34**	**88.43 ± 1.37**	**1.18 ± 0.13**	**0.99 ± 0.41**	**84.92 ± 1.73**	**1.82 ± 0.18**	**1.43 ± 0.58**

TABLE IV

SCALABILITY OF OUR METHOD (MEAN $\pm$ STD).

Map	N	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
10×10	3	95.1±0.7	0.42±0.05	0.36±0.10
	5	95.8±0.8	0.55±0.06	0.34±0.11
	8	95.2±0.9	0.76±0.08	0.39±0.13
15×15	3	92.7±1.0	0.88±0.09	0.62±0.20
	5	93.4±1.1	1.05±0.10	0.57±0.18
	8	92.9±1.2	1.32±0.12	0.66±0.22
20×20	3	90.4±1.3	1.38±0.13	0.88±0.30
	5	91.6±1.4	1.62±0.15	0.81±0.28
	8	91.0±1.6	1.95±0.18	0.85±0.31
	12	90.2±1.8	2.35±0.22	0.96±0.35
30×30	3	86.2±1.9	2.10±0.21	1.90±0.70
	5	88.7±2.0	2.55±0.25	1.75±0.68
	8	90.0±2.1	3.05±0.30	1.85±0.75
	12	89.5±2.2	3.60±0.35	2.02±0.82

E. Scalability with Map Size and Number of Agents

Table IV evaluates the scalability of the proposed method. As map size increases, coverage becomes more challenging due to longer horizons and partial observability, while risk per step also rises because agents must traverse longer hazard-exposed paths. Increasing team size generally improves parallelism, leading to higher $\mathcal{G}_{\text{safe}}$ on larger maps, but also increases interaction complexity, which is reflected in gradually higher V and risk per step. Importantly, performance degrades smoothly rather than collapsing, indicating that semantic conditioning remains effective at larger scales.

TABLE V

HAZARD DENSITY SWEEP. PERFORMANCE UNDER INCREASING HAZARD PREVALENCE (FIXED MAP SIZE/TEAM SIZE).

Density	Method	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
Low	MAPPO + scalar risk penalty	**92.80 ± 0.95**	1.05 ± 0.10	1.03 ± 0.38
	Static-FMEA	89.10 ± 1.00	0.82 ± 0.08	0.85 ± 0.32
	Ours	91.20 ± 1.10	**0.68 ± 0.07**	**0.70 ± 0.28**
Medium	MAPPO + scalar risk penalty	84.90 ± 1.45	2.05 ± 0.18	1.78 ± 0.60
	Static-FMEA	86.70 ± 1.35	1.72 ± 0.15	1.58 ± 0.54
	Ours	**89.60 ± 1.15**	**1.18 ± 0.11**	**1.12 ± 0.42**
High	MAPPO + scalar risk penalty	77.40 ± 1.95	3.55 ± 0.30	2.70 ± 0.90
	Static-FMEA	79.20 ± 1.85	3.18 ± 0.27	2.45 ± 0.84
	Ours	**85.10 ± 1.50**	**2.05 ± 0.18**	**1.85 ± 0.68**

F. Robustness to Hazard Density

We vary hazard density from low to high while keeping map size and team size fixed. Fig. 1 shows three representative maps. As density increases, all methods exhibit lower $\mathcal{G}_{\text{safe}}$ and higher V and $\sum_t \rho_t$. However, the performance gap widens in dense regimes: unstructured penalties either under-penalize rare severe events, increasing $\sum_t \rho_t$, or over-penalize broadly, reducing $\mathcal{G}_{\text{safe}}$. Static-FMEA performs well only when its fixed semantics remain aligned with the environment. By contrast, FMEA-Net degrades more gracefully, maintaining lower violations and cumulative exposure while preserving safe coverage.

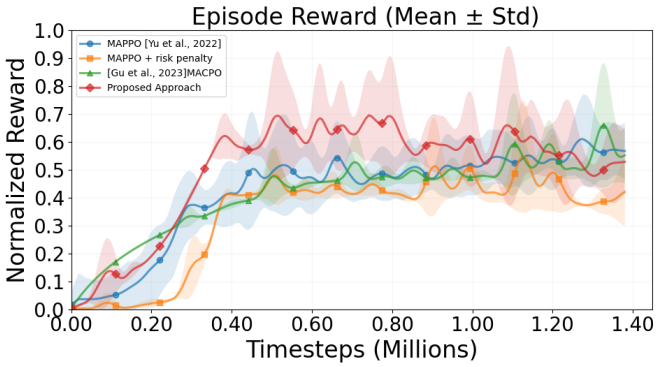


Fig. 2. Training curves (mean $\pm$ std over seeds) showing normalized episode reward versus environment timesteps.

G. Robustness to Symbolic-Prior Corruption (Risk Mismatch)

We next isolate semantic mismatch by introducing structured corruption into the symbolic prior. Table VI follows an increasing difficulty order: sparse link flips are local and partly recoverable, wrong severity scaling distorts the consequence model, row or column permutations disrupt global semantic alignment, and joint corruption combines all errors and is therefore the most severe. The resulting monotonic degradation shows that the benchmark genuinely stresses semantic consistency rather than merely adding generic noise.

TABLE VI

SENSITIVITY TO SYMBOLIC-PRIOR CORRUPTION FOR OUR METHOD.

Mismatch type	$\mathcal{G}_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
Random link corruption (sparse flips)	87.62 ± 1.58	2.41 ± 0.21	0.91 ± 0.34
Wrong severity scaling $s(e)$	85.14 ± 1.73	2.89 ± 0.25	1.06 ± 0.39
Row permutation in $P_0(M \mid C)$	82.08 ± 1.96	3.64 ± 0.30	1.31 ± 0.46
Column permutation in $P_0(E \mid M)$	80.41 ± 2.05	4.06 ± 0.34	1.47 ± 0.51
Joint corruption (worst-case)	77.93 ± 2.27	4.78 ± 0.40	1.76 ± 0.62

H. Semantic Calibration and Training Stability

A key objective is to ensure that semantic learning repairs only the mis-specified entries without destabilizing policy optimization. Tables VII and VIII show this behavior: calibration corrects the targeted mismatches, while the remaining rows remain unchanged. This is important for interpretability, since the safety gains can be linked to specific causal corrections rather than drift toward an arbitrary risk model.

TABLE VIII

RISK-MODEL MISMATCH AND ONLINE CORRECTION IN MODE-TO-EFFECT SEMANTICS. ONLY THE MIS-SPECIFIED ENTRIES (BOLD) DIFFER BETWEEN THE PRIOR AND CALIBRATED TABLES.

(c) Prior $P_0(E \mid M)$

Mode M	Delay	Cov. loss	Damage	Abort
Slip	**0.25**	**0.60**	0.10	0.05
Collision	**0.45**	0.10	**0.25**	0.20
Stall	0.15	0.65	0.10	0.10
Loc. drift	0.10	0.75	0.10	0.05
Comm. drop	0.15	**0.60**	0.05	**0.20**

(d) Calibrated $P_\phi(E \mid M)$

Mode M	Delay	Cov. loss	Damage	Abort
Slip	**0.65**	**0.20**	0.10	0.05
Collision	**0.15**	0.10	**0.55**	0.20
Stall	0.15	0.65	0.10	0.10
Loc. drift	0.10	0.75	0.10	0.05
Comm. drop	0.15	**0.25**	0.05	**0.55**

TABLE IX

ABLATION STUDY (HIGH HAZARD DENSITY MAP).

Components				Metrics		
MAPPO	FMEA tables	Sev. weights	Online calib.	$C_{\text{safe}} \uparrow$	$V \downarrow$	$\sum_t \rho_t \downarrow$
✓	×	×	×	83.20 ± 1.70	3.10 ± 0.28	1.18 ± 0.42
✓	✓	×	×	85.10 ± 1.62	2.55 ± 0.23	1.02 ± 0.36
✓	✓	✓	×	86.80 ± 1.55	2.10 ± 0.20	0.87 ± 0.31
✓	✓	✓	✓	88.00 ± 1.48	1.75 ± 0.18	0.75 ± 0.21

TABLE VII

RISK-MODEL MISMATCH AND ONLINE CORRECTION IN CAUSE-TO-MODE SEMANTICS. BOLD ENTRIES INDICATE CORRECTED MISMATCHES; ALL OTHERS REMAIN UNCHANGED.

(a) Prior $P_0(M \mid C)$ **(mis-specified)**

Cause C	Slip	Collision	Stall	Loc. drift	Comm. drop
Slippery terrain	**0.20**	**0.65**	0.05	0.07	0.03
Loose gravel/uneven	0.20	0.10	0.55	0.10	0.05
Obstacle-dense corridor	**0.50**	**0.35**	0.05	0.08	0.02
Low visibility/noise	0.05	0.10	0.05	0.70	0.10
Comm interference zone	0.05	0.05	0.05	**0.55**	**0.30**
Low battery/high load	0.05	0.05	0.75	0.10	0.05

(b) Calibrated $P_\theta(M \mid C)$

Cause C	Slip	Collision	Stall	Loc. drift	Comm. drop
Slippery terrain	**0.70**	**0.15**	0.05	0.07	0.03
Loose gravel/uneven	0.20	0.10	0.55	0.10	0.05
Obstacle-dense corridor	**0.15**	**0.70**	0.05	0.08	0.02
Low visibility/noise	0.05	0.10	0.05	0.70	0.10
Comm interference zone	0.05	0.05	0.05	**0.15**	**0.70**
Low battery/high load	0.05	0.05	0.75	0.10	0.05

I. Ablations and Sensitivity

We report component ablations (Table IX). The ablations isolate the contribution of online calibration, prior anchoring, and safety-context conditioning.

Overall, the results support the three research questions: (i) semantic safety conditioning improves the safety–efficiency trade-off by reducing violations and severity-weighted exposure without sacrificing coverage; (ii) online calibration is most beneficial under prior mismatch and sensor noise, where fixed semantics become brittle; and (iii) the learned semantics correct targeted mis-specified links while preserving interpretability through constrained updates.

V. CONCLUSION

In this work, we addressed the brittleness of Safe Multi-Agent Reinforcement Learning (MARL) under epistemic uncertainty, where discrepancies between training priors and deployment reality lead to catastrophic failure. We proposed FMEA-Net, an architecture that decouples behavioral control from risk perception. By explicitly modeling safety as a differentiable causal chain (Cause $\rightarrow$ Mode $\rightarrow$ Effect), the FMEA-Net provides an interpretable interface for agents to reason about environmental hazards. Our results confirm that this approach achieves the resilience of adaptive systems while retaining the parameter-space stability of robust control, significantly outperforming static baselines in environments with shifting hazard dynamics. Future work will focus on extending the causal graph to handle continuous, multi-modal sensor fusion and investigating distributed calibration protocols where agents share semantic updates to accelerate team-wide adaptation in large-scale deployments.

REFERENCES

[1] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theories and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] H. Wang, J. Qin, and Z. Kan, "Shielded planning guided data-efficient and safe reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3808–3819, 2024.

[3] S. Gu, J. Kuba, G. Chen, Y. Du, Y. Yang, and A. Knoll, "Safe multi-agent reinforcement learning for multi-robot control," *Artificial Intelligence*, vol. 324, p. 104014, 2023.

[4] L. Zhang, L. Li, W. Wei, H. Song, Y. Yang, and J. Liang, "Scalable constrained policy optimization for safe multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[5] L. De Raedt, S. Dumančić, R. Manhaeve, and G. Marra, "From statistical relational to neuro-symbolic artificial intelligence," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

[6] B. C. Colelough and W. Regli, "Neuro-symbolic ai in 2024: A systematic review," *arXiv preprint arXiv:2501.05435*, 2025.

[7] L. Keller, D. Tanneberg, and J. Peters, "Neuro-symbolic imitation learning: Discovering symbolic abstractions for skill learning," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 6519–6526.

[8] N. Jansen, B. Könighofer, S. Junges, A. Serban, and R. Bloem, "Safe reinforcement learning using probabilistic shields," in *31st International Conference on Concurrency Theory (CONCUR 2020)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020, pp. 3–1.

[9] W.-C. Yang, G. Marra, G. Rens, and L. De Raedt, "Safe reinforcement learning via probabilistic logic shields," *arXiv preprint arXiv:2303.03226*, 2023.

[10] Y. Emam, A. Notomista, M. Glotfelter, A. L. Thomaz, and M. Egerstedt, "Safe reinforcement learning using robust control barrier functions," *IEEE Robotics and Automation Letters*, 2022.

[11] I. D. Wijegunawardana, S. B. P. Samarakoon, M. V. J. Muthugala, and M. R. Elara, "Fmea-based coverage-path-planning strategy for floor-cleaning robots," *Advanced Intelligent Systems*, vol. 5, no. 11, p. 2300260, 2023.

[12] A. Martinez-Seras, A. Andres, and J. Del Ser, "Advancing towards safe reinforcement learning over sparse environments with out-of-distribution observations: Detection and adaptation strategies," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–10.

[13] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in neural information processing systems*, vol. 35, pp. 24 611–24 624, 2022.

[14] S. Gu, J. Grudzien Kuba, Y. Chen, Y. Du, L. Yang, A. Knoll, and Y. Yang, "Safe multi-agent reinforcement learning for multi-robot control," *Artificial Intelligence*, vol. 319, p. 103905, 2023.