EPIPLAN: TEST-TIME POLICY-SPACE ROLLOUT FOR EPIGRAPH-CONDITIONED SAFE MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Epigraph-conditioned safe reinforcement learning trains a controller across task-budget levels, but deployment typically collapses this structure to the policy at a learned critic root. In distributed epigraph control, local critic roots can be aggregated into a public team boundary, but executing only that boundary member does not compare the joint feedback trajectories induced by nearby boundary-relative choices. We introduce *Prospective Epigraph-Space Planning* (EPIPLAN), which treats the frozen controller as a one-dimensional family of complete team policies and selects a member at deployment without retraining or optimizing joint-action sequences. From the same augmented state, EPIPLAN rolls out one-sided displacements, recomputes the public root along each branch, uses frozen task and safety critics as terminal estimates, and executes the lowest-cost predicted-feasible member while retaining its displacement. We formalize this actual-receding family and its deployment epigraph; under stated approximation and margin conditions, our analysis bounds finite-realization error, selected-policy feasibility residual, and task loss in terms of continuation, calibration, and coordinate resolution. Across 18 native-eligible frozen checkpoints in six settings drawn from five cooperative tasks, EPIPLAN lowers episode unsafety relative to paired public-root execution, while every environment-mean reward loss remains within a prespecified 15% tolerance. An exact-marginal permutation control that preserves each checkpoint's offset histogram but breaks state–member pairing raises expected episode unsafety by 10.38 percentage points ($p = .0001$), showing that the gain depends on contextual assignment rather than offset frequencies alone.

1 INTRODUCTION

Safe multi-agent deployment must preserve team utility while preventing trajectory-level violations. This tension appears in multi-robot navigation, aerial coordination, and warehouse automation Sartoretti et al. (2019); Park et al. (2024); Li et al. (2021a): a locally acceptable action may still induce a poor joint continuation. Penalty and expected-cost methods shape training Achiam et al. (2017); Zhao et al. (2023), but an expectation can conceal unsafe episodes. Runtime shields and barrier controllers instead constrain individual actions Alshiekh et al. (2018); Zhang et al. (2019; 2023). We ask a complementary question: can a frozen controller expose alternative coordinated behaviors that deployment can compare?

Epigraph reinforcement learning provides this structure by conditioning a policy on a task-budget coordinate So & Fan (2023). DEF-MARL extends the construction to teams: each agent inverts a local safety critic, and the maximum of the local roots defines a public coordinate Zhang et al. (2025a). The public-root realization studied here immediately executes the policy at this learned boundary. The root locates the current boundary, but it does not compare the joint feedback trajectories generated by nearby members. Epigraph training thus produces a behavior family while public-root deployment uses only its boundary member.

To use this family at deployment, we introduce *Prospective Epigraph-Space Planning* (EPIPLAN). EPIPLAN interprets the coordinate as an index over closed-loop team policies rather than a scalar discarded after root finding. It anchors one-sided displacements at the public boundary; each actual-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

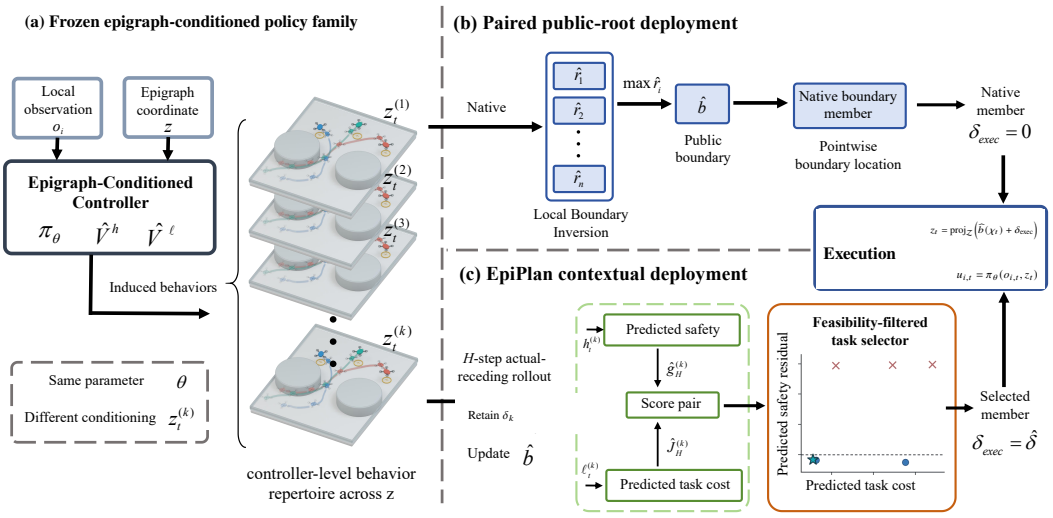

Figure 1: **EpiPlan deployment architecture.** (a) A frozen epigraph-conditioned actor and its task and safety critics represent a family of coordinated behaviors across the task-budget coordinate z. (b) The paired public-root realization aggregates local boundary estimates and executes the boundary member $\delta_{\text{exec}} = 0$; it is distinct from the published local-root/no-communication evaluation. (c) EPIPLAN rolls boundary-relative members from the same augmented state, retaining each displacement while updating the public boundary and recurrent states. It then executes the lowest-cost predicted-feasible member through the same controller. Predicted feasibility is a planning score, not observed trajectory safety.

receding member retains its displacement while recomputing the root along its own continuation. Candidate branches copy the same physical and recurrent states, so they represent complete feedback policies rather than open-loop action sequences. EPIPLAN executes the lowest-cost predicted-feasible member while keeping the actor and critics frozen (Figure 1).

A finite prospective evaluation turns this family into a decision rule. For each candidate, EPIPLAN simulates an H-step joint continuation and uses the frozen task and safety critics as terminal estimates. A design-calibrated safety score filters the family, and the task score ranks admitted members. The search remains one-dimensional in the displacement while each candidate preserves team feedback and recurrent state. Short rollouts and terminal values are familiar from model-predictive planning Chua et al. (2018); Janner et al. (2019); Sikchi et al. (2022); here they compare complete policies already represented by the controller.

We formalize the actual-receding family through a deployment epigraph and analyze its finite realization. The resulting bounds separate continuation, calibration, and coordinate errors in the selected member's feasibility residual and task loss. A complementary result transports local root error to the public boundary under explicit approximation, crossing-margin, and inactive-projection conditions.

The evaluation tests whether family selection improves frozen deployment and why. Across native-eligible DEF-MARL checkpoints spanning the evaluated settings, EPIPLAN consistently reduces episode unsafety relative to paired public-root execution; hierarchical confidence intervals support these improvements, while mean reward losses remain within the prespecified utility budget. A matched fixed-displacement control explains only part of the advantage. An exact-marginal permutation preserves each checkpoint's offset distribution but breaks the state–member correspondence; safety then degrades reliably under reassignment. These comparisons show that the additional gain arises from context-dependent member selection rather than from offset frequency alone.

Our contributions are:

- We formulate post-training deployment as a decision over the state-indexed policy family induced by an epigraph-conditioned controller, and define its boundary-relative actual-receding members.

- We develop a finite selector that compares joint continuations with frozen terminal critics and design-only calibration, requiring neither retraining nor optimization over joint-action sequences.
- We characterize the induced deployment epigraph and derive conditional bounds that separate continuation, calibration, and coordinate errors in the selected member's feasibility and task performance, together with a public-boundary root-transport result.
- We evaluate 18 native-eligible frozen checkpoints across six settings and show that contextual family selection lowers episode unsafety; fixed and exact-marginal controls attribute the additional gain to matching policy members with deployment contexts.

2 Related Work

Safe learning and multi-agent control. Safe reinforcement learning has been studied through constrained optimization, runtime intervention, and certificate-based control Brunke et al. (2022); Garg et al. (2024). CPO controls expected cumulative cost with a trust-region update Achiam et al. (2017), while Lagrangian methods enforce an expected reward or cost budget Tessler et al. (2019). MAFoCoPS extends constrained policy-space optimization to multi-agent settings Zhao et al. (2023). MAPPO and HAPPO provide strong cooperative policy-optimization foundations Yu et al. (2022); Kuba et al. (2022), and broader studies have adapted constrained learning specifically to multi-robot control Gu et al. (2023). Expectation-level constraints nevertheless differ from the maximum trajectory violation considered here. Shielding rejects unsafe actions Alshiekh et al. (2018), MAMPS invokes a model-predictive backup Zhang et al. (2019), and graph control barrier functions learn distributed certificates and controllers Zhang et al. (2023; 2025b). These approaches optimize or modify actions and policies. EpiPlan instead works within the structured policy space already represented by an epigraph-conditioned controller.

Epigraph safe control. Reach-avoid reinforcement learning encodes safety and liveness within a dynamic programming objective Hsu et al. (2021). EFPPO instead reformulates stabilize–avoid control through a task-budget epigraph, avoiding a Lagrangian saddle point So & Fan (2023). DefMarl derives a multi-agent decomposition in which local safety boundaries combine through a public maximum and condition a shared actor Zhang et al. (2025a). We inherit that inner actor, boundary decomposition, and safety-side ordering in the exact setting. EpiPlan retains the public coordinate after training and uses its induced closed-loop policies as the domain of a horizon-calibrated epigraph.

Conditioned policy spaces. Universal value function approximators represent outcomes as functions of both state and goal Schaul et al. (2015), while goal-conditioned policies expose reusable behaviors that can be composed by a planner Nasiriany et al. (2019). These works show how conditioning variables can organize related behaviors. More directly, model-predictive controllers plan over learned latent actions Li et al. (2021b), compact trajectory codes Jiang et al. (2023), and instance-specific latent policy coordinates Chalumeau et al. (2023). Constraint-conditioned safe policies support zero-shot deployment at different budgets Yao et al. (2023); Guo et al. (2025). These precedents make clear that planning over a conditioned-policy coordinate is not new by itself. Our narrower distinction is that the coordinate and its one-sided order are inherited from an epigraph safe controller: retaining a displacement while recomputing local roots and their public maximum induces a coupled actual-receding team-policy family.

Finite model prediction. Probabilistic dynamics models support online trajectory optimization under model uncertainty Chua et al. (2018), while model-based policy optimization uses short simulated rollouts to limit compounding model error Janner et al. (2019). LOOP combines such rollouts with a learned continuation value Sikchi et al. (2022), and distributed online rollout coordinates local routing decisions through communication Weber et al. (2024). Their candidate domains are typically actions, action sequences, or local decisions. EpiPlan evaluates complete team policies indexed by the public epigraph coordinate. Its search dimension therefore does not grow with the joint action space, while finite continuations and terminal values retain the coordination represented by the conditioned actor.

3 PROBLEM AND EPIGRAPH PRELIMINARIES

Consider n agents with joint state x, local observations $o_i = O_i(x)$, joint action $u = (u_1, \dots, u_n)$, and deterministic dynamics $x^+ = f(x, u)$. Let $\ell(x, u) = -r(x, u)$ be team task cost and let $\bar{h}_{ic}(x) \le 0$ denote the continuous physical safety constraint c for agent i. Following DEF-MARL, the safety critic is trained with the separated signal

$$h^{\nu}_{ic}(x) = \bar{h}_{ic}(x) + \nu\sigma(\bar{h}_{ic}(x)), \qquad \sigma(a) = \begin{cases} 1, & a > 0, \\ -1, & a \le 0, \end{cases} \tag{1}$$

where the convention at zero matches the released implementation. We include time and all recurrent actor and critic states in an augmented Markov state χ. For a finite episode of length T, define

$$\begin{aligned} J_{\gamma,\Pi}(\chi) &= \sum_{t=0}^{T-1} \gamma^t \ell(x_t, u_t), \\ G_{\mathrm{phys},\Pi}(\chi) &= \max_{0\le t\le T} \max_{i,c} \bar{h}_{ic}(x_t), \\ G_{\nu,\Pi}(\chi) &= \max_{0\le t\le T} \max_{i,c} h^{\nu}_{ic}(x_t). \end{aligned} \tag{2}$$

The analysis is pointwise in χ. Experimental rates average the event $\mathbf{1}\{G_{\mathrm{phys},\Pi} > 0\}$ over held-out initial states. This event is unchanged when computed from $G_{\nu,\Pi}$, whereas the physical analysis uses the continuous functional $G_{\mathrm{phys},\Pi}$ because the binary event is not generally Lipschitz.

For $0 \le \xi \le \nu$, the exact DEF-MARL epigraph reference is

$$\begin{aligned} \Phi^{\star}_{\nu,\xi}(\chi, z) &= \inf_{\Pi} \max\{G_{\nu,\Pi}(\chi) + \xi, J_{\gamma,\Pi}(\chi) - z\}, \\ b_{\mathrm{EF},\nu,\xi}(\chi) &= \inf\{z : \Phi^{\star}_{\nu,\xi}(\chi, z) \le 0\}. \end{aligned} \tag{3}$$

When the robust problem is feasible and attained, $b_{\mathrm{EF},\nu,\xi} = \min_{\Pi: G_{\nu,\Pi} \le -\xi} J_{\gamma,\Pi}$. The margin ξ therefore acts on the separated critic signal rather than as a physical clearance threshold. With the safe-side convention in Eq. (1), $0 \le \xi \le \nu$ leaves the exact physical feasible set unchanged. Under the decomposition conditions of DEF-MARL, the team boundary is the maximum of local boundaries.

The learned controller contains an actor $\pi_\theta(o_i, z)$, a centralized task critic $\widehat{V}^{\ell}$, and local safety critics $\widehat{V}^{h}_{ic}$ trained for h^{ν}_{ic}. Define

$$\widehat{q}_i(\chi, z) = \max_{c} \widehat{V}^{h}_{ic}(\chi, z) + \xi. \tag{4}$$

Let $\widehat{r}_i$ denote a local boundary estimate satisfying $\widehat{q}_i(\chi, \widehat{r}_i) \simeq 0$. The corresponding public coordinate is

$$\widehat{b}(\chi) = \max_{i} \widehat{r}_i(\chi), \qquad u_i = \pi_\theta(o_i, \widehat{b}(\chi)). \tag{5}$$

The distributed epigraph decomposition prescribes this maximum as the team-level boundary.

4 EPIGRAPH-SPACE CONSTRAINED REFINEMENT

4.1 ACTUAL-RECEDING POLICY FAMILY

Let $\mathcal{Z} = [z_{\min}, z_{\max}]$ be the admissible coordinate interval, with Euclidean projection $\mathrm{proj}_{\mathcal{Z}}(a) := \arg\min_{z\in\mathcal{Z}}(z - a)^2$. Let $\Delta = [0, \delta_{\max}]$. For every displacement $\delta \in \Delta$, define the executed coordinate along its own continuation as

$$z^{\delta}_t = \mathrm{proj}_{\mathcal{Z}}\Big(\widehat{b}(\chi^{\delta}_t) + \delta\Big). \tag{6}$$

Let $\Pi^{\mathrm{RC},\chi}_{\theta,\delta}$ denote the complete continuation that applies $\pi_\theta(O(x^{\delta}_t), z^{\delta}_t)$ and propagates all recurrent states. The state-indexed family is

$$\mathfrak{P}^{\mathrm{RC}}_{\theta}(\chi) = \{\Pi^{\mathrm{RC},\chi}_{\theta,\delta} : \delta \in \Delta\}. \tag{7}$$

Each member therefore represents a complete continuation rather than one action or a fixed-z rollout. The public boundary is recomputed at every step, while the displacement is retained.

For each member, let $J_{\mathrm{RC}}(\chi, \delta)$ denote its discounted task functional $J_{\gamma, \Pi^{\mathrm{RC},\chi}_{\theta,\delta}}$ and let $G_{\mathrm{RC}}(\chi, \delta)$ denote its continuous physical residual $G_{\mathrm{phys}, \Pi^{\mathrm{RC},\chi}_{\theta,\delta}}$.

The ideal constrained refinement and its operator are

$$\delta^\star_\tau(\chi) \in \arg\min_{\delta \in \Delta} J_{\mathrm{RC}}(\chi, \delta) \quad \text{s.t.} \quad G_{\mathrm{RC}}(\chi, \delta) \le \tau,$$
$$\mathcal{R}_\tau[\mathfrak{P}^{\mathrm{RC}}_\theta](\chi) = \Pi^{\mathrm{RC},\chi}_{\theta,\delta^\star_\tau}. \tag{8}$$

If the native member $\delta = 0$ is feasible, this ideal operator cannot have higher task cost than native within the induced family.

4.2 Prospective Refinement Epigraph

Introduce an outer task level y, distinct from the actor input z, and let $g_\tau(\chi, \delta) = G_{\mathrm{RC}}(\chi, \delta) - \tau$. Define

$$\Psi_\theta(\chi, y) = \min_{\delta \in \Delta} \max\{g_\tau(\chi, \delta), J_{\mathrm{RC}}(\chi, \delta) - y\},$$
$$y^\star_\tau(\chi) = \min_y y \quad \text{s.t.} \quad \Psi_\theta(\chi, y) \le 0. \tag{9}$$

Theorem 1 (Refinement–epigraph equivalence)**.** *If Δ is compact, the functionals are continuous, and the feasible set in Eq.* (8) *is nonempty, then $y^\star_\tau = J_{\mathrm{RC}}(\chi, \delta^\star_\tau)$. Every inner witness at $y^\star_\tau$ solves Eq.* (8)*.*

The statement follows because $\Psi_\theta(\chi, y) \le 0$ exactly when some family member satisfies both $G_{\mathrm{RC}} \le \tau$ and $J_{\mathrm{RC}} \le y$. This outer problem is defined over the closed-loop policy family induced by the original epigraph coordinate.

4.3 Horizon-Calibrated Epigraph Realization

Let $\Delta_L \subset \Delta$ be a finite coordinate family. A prospective H-step continuation defines

$$\widehat{J}_H(\chi, \delta) = \sum_{t=0}^{H-1} \gamma^t \ell(\widehat{x}^\delta_t, \widehat{u}^\delta_t) + \gamma^H \widehat{V}^\ell(\widehat{\chi}^\delta_H, z^\delta_H). \tag{10}$$

The safety estimate $\widehat{S}^\nu_H$ combines the prospective h^ν signals and terminal $\widehat{V}^h$ through DEF-MARL's discounted maximum recursion; below we write it as $\widehat{S}_H$ for brevity. A horizon-indexed calibration map $\mathcal{C}_H$, shared in form and design split across horizons, maps design-only trajectory scores to the fixed threshold

$$c_H = \mathcal{C}_H(\mathcal{D}_{\mathrm{cal}}; \widehat{S}_H, \widehat{J}_H). \tag{11}$$

The design-only map searches fixed score quantiles and selects c_H lexicographically by safe task success, unsafety, task success, reward, and displacement; the complete specification is given in the supplement. The pair (H, c_H) forms a single estimator. The horizon determines the trajectory information available to the score, while $\mathcal{C}_H$ maps that information to the common physical feasibility convention. The calibration centers the learned feasibility residual

$$\widehat{g}_H(\chi, \delta) = \widehat{S}_H(\chi, \delta) - c_H. \tag{12}$$

The calibrated level c_H remains distinct from the physical threshold τ and ξ remains the inherited critic-space margin. The gap between $\widehat{g}_H$ and g_τ reflects critic and calibration error together with the mismatch between discounted and hard trajectory safety. Thus the horizon-indexed object is the calibrated score pair $\widehat{\mathcal{E}}^{\mathcal{C}}_H = (\widehat{g}_H, \widehat{J}_H)$, rather than the uncalibrated safety score $\widehat{S}_H$ alone.

The resulting *horizon-calibrated prospective epigraph* is

$$\widehat{\Psi}^{\mathcal{C}}_{H,L}(\chi, y) = \min_{\delta \in \Delta_L} \max\{\widehat{g}_H(\chi, \delta), \widehat{J}_H(\chi, \delta) - y\}. \tag{13}$$

When a predicted-feasible member exists, the outer minimizer is equivalently the member with smallest $\widehat{J}_H$ among $\widehat{g}_H \le 0$. Otherwise, EPIPLAN selects the member with minimum estimated feasibility residual. This alternative lies outside the feasible epigraph and is not covered by the safety conclusion below.

Operationally, calibration is frozen using design trajectories. At test time, EPIPLAN copies the physical and recurrent state, rolls every $\delta \in \Delta_L$ for H steps while recomputing the public boundary, selects the lowest predicted-cost member among $\widehat{g}_H \le 0$ (or the minimum-residual member if none is predicted feasible), and executes the selected displacement with continued root recomputation. The full operational specification, state propagation, and tie rules are in the supplement.

5 APPROXIMATION AND SAFETY ANALYSIS

Scope of the analysis. The results below separate three roles. The deployment-epigraph identity and exact-limit proposition define the reference object; the finite-realization theorem bounds the selector actually implemented; and boundary transport gives an epigraph-specific structural explanation under stronger assumptions. The max nonexpansiveness and nested-grid rate are standard supporting facts. None of the unknown uniform error envelopes is estimated by our experiments, and the empirical selector uses the untightened branch. Thus the main theoretical claim is approximation-aware selection, not certified collision-free neural control.

The supplement proves that exact root recomputation agrees in value with the robust Bellman solution when reachable subproblems are attained and projection is inactive (and in trajectory under uniqueness). This exact-limit result is a reference point, not an assumption asserted for the learned controller.

Horizon–coordinate approximation. Assume that on Δ the true task and feasibility residuals are L_J- and L_g-Lipschitz. Let Δ_L have covering radius ρ_L and suppose

$$\sup_{\delta\in\Delta_L} |\widehat{J}_H - J_{\mathrm{RC}}| \le \varepsilon_J(H), \quad \sup_{\delta\in\Delta_L} |\widehat{g}_H - g_\tau| \le \varepsilon_g(H). \tag{14}$$

The envelopes admit the structural decomposition

$$\begin{aligned} \varepsilon_J(H) &\le \varepsilon^J_{\mathrm{model}}(H) + \varepsilon^J_{\mathrm{proto}}(H) + \gamma^H \varepsilon^{\mathrm{tail}}_{V_\ell}(H), \\ \varepsilon_g(H) &\le \varepsilon^g_{\mathrm{model}}(H) + \varepsilon^g_{\mathrm{proto}}(H) + \gamma^H \varepsilon^{\mathrm{tail}}_{V_h}(H) \\ &\quad + \varepsilon_{\mathrm{sem}}(H) + \varepsilon_{\mathcal{C}}(H). \end{aligned} \tag{15}$$

Exact prospective dynamics make the prefix model error vanish. The term $\varepsilon_{\mathrm{proto}}$ measures the mismatch between training and actual-receding continuations. The term $\varepsilon_{\mathrm{sem}}$ separates the discounted h^ν recursion from the continuous physical trajectory residual, while $\varepsilon_{\mathcal{C}}$ measures horizon-indexed calibration error. These terms need not decrease monotonically with H.

Theorem 2 (Uniform finite-realization approximation)**.** *Under Eq.* (14)*, for every task level y,*

$$\begin{aligned} &|\widehat{\Psi}^{\mathcal{C}}_{H,L}(\chi, y) - \Psi_\theta(\chi, y)| \\ &\qquad \le \max\{\varepsilon_J(H), \varepsilon_g(H)\} + \max\{L_J, L_g\}\rho_L. \end{aligned} \tag{16}$$

The maximum is nonexpansive in the vector infinity norm. Minimization over the same finite family preserves the error, while the nearest member contributes the covering term. The bound places temporal continuation, calibrated feasibility, and coordinate resolution within one prospective epigraph.

Theorem 3 (Finite-realization selection guarantee)**.** *Let $\delta^\circ \in \Delta$ satisfy $g_\tau(\chi, \delta^\circ) \le -\varepsilon_g(H) - L_g\rho_L$. If* EPIPLAN *uses its feasible branch, then*

$$\begin{aligned} g_\tau(\chi, \widehat{\delta}) &\le \varepsilon_g(H), \\ J_{\mathrm{RC}}(\chi, \widehat{\delta}) &\le J_{\mathrm{RC}}(\chi, \delta^\circ) + 2\varepsilon_J(H) + L_J\rho_L. \end{aligned} \tag{17}$$

If the selection rule instead requires $\widehat{g}_H \le -\varepsilon_g(H)$ and the comparator has one additional $\varepsilon_g(H)$ of margin, the first conclusion strengthens to $g_\tau \le 0$. If $\delta = 0$ is predicted feasible, the native comparison has no coordinate-covering term and satisfies $J_{\mathrm{RC}}(\widehat{\delta}) \le J_{\mathrm{RC}}(0) + 2\varepsilon_J(H)$.

Table 1: Held-out performance of EPIPLAN. Each row averages three frozen base-policy checkpoints retained by a native-only eligibility rule, with 128 paired test episodes per checkpoint. Corridor uses seeds 0/2/3; the supplement reports every cohort and exclusion. Fixed is the design-selected constant displacement on the same episodes. Agent unsafe follows the DEF-MARL convention and uses the paired zero-displacement public-boundary rollout as native. Signed changes are EPIPLAN minus native, so positive ΔReward and ΔJoint are better.

	Episode unsafe			Agent unsafe			
Environment	Native	Fixed	EPIPLAN	Native	EPIPLAN	ΔReward	ΔJoint
Corridor	.2005	.0625	.0208	.1059	.0130	$-.0028$	$+.0964$
Target	.2943	.2083	.1536	.1363	.0642	$-.0087$	$+.1224$
Target ($N = 8$)	.4401	.3828	.2370	.1022	.0456	$-.0350$	$+.0156$
Spread	.1849	.1536	.0755	.0608	.0269	$+.0047$	$+.0286$
ConnectSpread	.1589	.0781	.0521	.1033	.0330	$+.0088$	$+.0547$
Line (5 obs.)	.2396	.2344	.1302	.0868	.0521	$-.0087$	$+.0156$

Theorem 3 gives a conditional guarantee for the discounted functional J_{RC} under uniform neural error bounds. If $|\ell_t| \le L_\ell$, its difference from the undiscounted task cost $C_{\mathrm{ret}} = -R_{\mathrm{undisc}} = \sum_t \ell_t$ is bounded by $L_\ell \sum_{t=0}^{T-1}(1-\gamma^t)$. This conversion can be loose, so the experiments evaluate undiscounted reward preservation directly and use the untightened empirical branch without estimating the uniform constants. Consequently, Eq. (17) does not certify collision-free execution for the reported neural policies.

Boundary uncertainty and coordinate resolution. An epigraph-specific transport result in the supplement assumes a two-sided linear crossing margin m, uniform score error ε_h, numerical root residual ε_r, and inactive projection. It gives both local and public-root error at most $\eta = (\varepsilon_h + \varepsilon_r)/m$; a selected displacement at least η is then physically safe. These strong, unmeasured conditions explain why a one-sided family is natural but do not certify the deployed selector. For adjacent-midpoint grids with largest initial gap g_0, $|\Delta_L| = 1 + (K_0 - 1)2^L$ and $\rho_L \le g_0 2^{-(L+1)}$, so only the coordinate-covering term decreases geometrically.

Depth H and refinement level L govern distinct approximation axes. Larger H replaces more of the terminal proxy with member-specific dynamics, but $\varepsilon_g(H)$ can retain a discounted-to-hard semantic floor and need not decrease monotonically. If both error envelopes vanish, $\rho_L \to 0$, and the ideal optimizer is unique and stably approachable by strictly feasible members, the selected coordinate and outer task level converge to the ideal operator. Full proofs and the required feasible-set stability condition are in the supplement.

6 EXPERIMENTS

6.1 SETUP

Environments. We evaluate six settings drawn from five MPE-style tasks: CORRIDOR, TARGET, LINE, SPREAD, and CONNECTSPREAD Lowe et al. (2017); Zhang et al. (2025a). They cover navigation, cooperative goal reaching, formation, coverage, and connectivity-preserving coverage. Each setting uses three native-eligible frozen checkpoints, not an unfiltered sample of training seeds. The historical design-only gate required normalized progress $\ge .8$, joint success $\ge .1$, utility-failure rate $\le .9$, and native episode unsafety $\ge .01$. It was applied before inspecting EPIPLAN outcomes and deliberately conditions the study on task-capable policies with safety headroom. We therefore evaluate post-training repairability, not training success or robustness over all attempted seeds.

Protocol. The primary realization uses $H = 32$ within episodes of length $T = 128$ and the adjacent-midpoint family Δ_L described above. Each checkpoint uses 64 design episodes for c_H and 128 disjoint paired test episodes; paired runs share initial states and rollout keys. We use $\nu = .5$, physical threshold $\tau = 0$, and the released boundary $\xi = .21$, with a matched $\xi = .4$ sensitivity test. The coordinate is selected initially and retained under Eq. (6). A 15% native-relative reward-loss budget is used only as a held-out empirical criterion, not as a theoretical or online constraint.

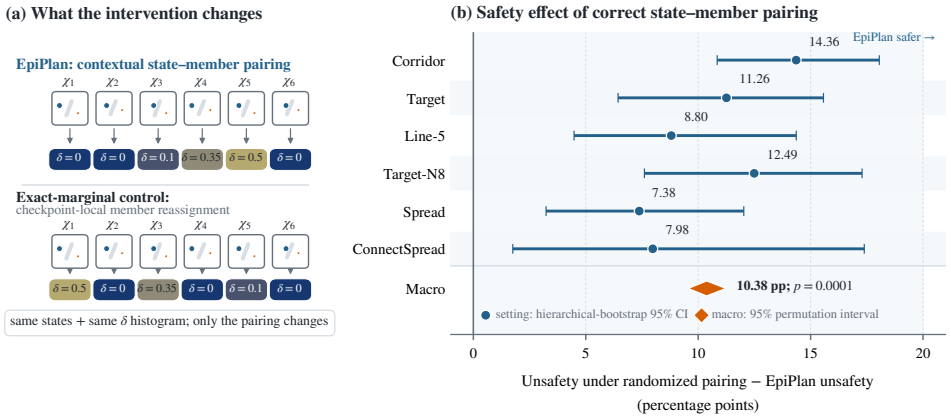

Figure 2: **Exact-marginal reassignment isolates contextual member selection.** (a) The retrospective control preserves each checkpoint's held-out states, candidate outcomes, and exact displacement histogram while reassigning state–member pairs. (b) Randomized-minus-EPIPLAN episode unsafety; positive values favor EPIPLAN. Circles show hierarchical-bootstrap 95% intervals (3×128 test episodes per setting), and the diamond shows the macro 95% permutation interval. The 10,000-permutation macro gap is 10.38 percentage points ($p = .0001$).

The paired native comparator recomputes and communicates the public maximum of local roots and applies zero displacement. This isolates the effect of EPIPLAN, but it is not the local-root/no-communication execution used in the published DEF-MARL evaluation; we therefore call it the public-root DEF-MARL realization. The primary outcome gathers the selected member from the paired vectorized candidate frontier rather than claiming bitwise equivalence to a scalar execute-only replay.

Metrics. **Episode unsafe** records whether any agent violates a constraint; **agent unsafe** is its agent-level counterpart and complements DEF-MARL's safe rate. **Reward** is undiscounted cumulative environment reward. **Task success** requires completion of at least $2/3$ and normalized goal progress of at least .8; **joint success** additionally requires trajectory safety. We report proportions and hierarchical-bootstrap intervals over policies and episodes.

6.2 SAFETY–UTILITY IMPROVEMENT

Primary results. Table 1 shows lower episode unsafety for every retained checkpoint, with all hierarchical 95% intervals for absolute reduction above zero. Agent unsafety also falls in all six settings. Relative to the design-selected constant, EPIPLAN is safer at every setting-level aggregate and has higher aggregate joint success in five settings; individual checkpoint comparisons need not improve. All reward changes satisfy the empirical budget. The results include eight-agent Target, two additional cooperative tasks, and a Line clutter shift from three training obstacles to five test obstacles.

Distribution-matched offset control. To distinguish contextual selection from merely using a favorable mixture of offsets, we permute the held-out EPIPLAN offset labels within each checkpoint. This preserves the exact marginal offset histogram and every archived candidate outcome while destroying only state–offset pairing. The safety advantage is positive in all six settings (absolute reduction .0738–.1436) and averages .1038; 10,000 within-checkpoint permutations give a 95% null interval $[.0955, .1124]$ for this reduction and one-sided $p = .0001$. Figure 2 visualizes the intervention and its setting-level effects. Relative reward is no worse in any setting and macro joint success increases by .0753. This exact-marginal test is a retrospective mechanism diagnostic, not a deployable baseline; a design-histogram randomizer gives the same qualitative conclusion without using held-out labels.

Public-boundary anchoring. An independent $H = 4$ audit separates the inherited public boundary from coordinate selection. Public aggregation lowers unsafety from .2839 to .1823 on Corridor and .3099 to .2839 on Target; refinement further reaches .1042 and .2005. Table 1 uses this stronger public comparator.

6.3 Operating Points and Components

Global-margin control. A separate $H = 4$ cohort tunes DEF-MARL's global margin ξ on design data. Its unsafe/task-cost pairs are .0573/.2747 on Corridor and .1719/.3034 on Target, versus .1107/.2688 and .2305/.2752 for EPIPLAN. This descriptive control is independent; Table 1 provides the same-cohort fixed comparison.

Conservative-anchor robustness. The $\xi = .4$ baseline already operates conservatively and leaves less safety headroom than the released $\xi = .21$ setting used in the primary study. Under this stronger anchor, EPIPLAN still reduces episode unsafety from .0417 to .0156 on Corridor and from .1927 to .1068 on Target. All six checkpoints improve, with 95% intervals [.0052, .0573] and [.0443, .1302]. Reward changes ($-.0021$, $-.0121$) remain within budget, while joint success changes by $-.0182$ and $+.0651$.

Secondary robustness and operational audits. Under a five-checkpoint hidden-gain audit, safety-benefit retention is $1.00/.67$ on Corridor and $.58/.74$ on Target at execution gains $.9/1.1$, with reward loss below 13%; this supports only a narrow actuation-mismatch claim. The prespecified $H = 32$ gives the best Target safety–utility point, although intermediate horizons are nonmonotone. Across the six primary settings, predicted-feasible coverage is 83.9–99.7%, all positive coordinates are exercised, and connected consensus reproduces centralized selection on every audited trajectory. Runtime is 1.29–1.39$\times$ native with compute $O(K_L H C_{\text{step}})$ and no trainable parameters. Complete denominators, sensitivity tables, numerical discrepancies, and the equivalence proof are in the supplement; these diagnostics are not distribution-free guarantees.

7 Limitations and Conclusion

EPIPLAN opens a broader design space for deployment-time reasoning over structured policy families. Adaptive proposal mechanisms could replace the fixed candidate grid and progressively refine promising boundary-relative members according to predictive uncertainty and task value. Learned or ensemble dynamics could extend the rollout layer beyond exact simulators, while uncertainty-calibration modules could automatically tighten feasibility thresholds under model shift. The same policy-family perspective could support multiple coupled safety budgets, hierarchical conditioning coordinates, and asynchronous decentralized selection through compressed consensus messages. More broadly, it may apply whenever training exposes an ordered, low-dimensional family of feedback policies whose members can be evaluated contextually at deployment.

Epigraph training exposes such a family of cooperative policies through a public task-budget coordinate. EPIPLAN turns this latent structure into a deployment-time decision space, evaluates complete feedback-policy members through finite joint continuations, and selects a predicted-feasible member with favorable task value. Across three frozen checkpoints in each of six multi-agent settings, it improves the realized safety–utility operating point without retraining or high-dimensional joint-action optimization. The distribution-matched control further shows that the gain comes from matching policy members to their initial contexts, rather than simply favoring more conservative coordinates. Together, these results establish contextual policy-family selection as a practical new use of epigraph-conditioned controllers at deployment.

Reproducibility statement

The supplement gives complete proofs, cohort construction, hyperparameters, per-checkpoint results, operational diagnostics, and protocol identities. The anonymous artifact contains frozen configurations, checkpoint and source-file hashes, raw candidate arrays, aggregation code, and tests needed to regenerate the reported tables and audits.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017. URL http://proceedings.mlr.press/v70/achiam17a.html.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2669–2678. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11797. URL https://doi.org/10.1609/aaai.v32i1.11797.

Lukas Brunke, Melissa Greeff, Adam W. Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P. Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annu. Rev. Control. Robotics Auton. Syst.*, 5:411–444, 2022. doi: 10.1146/ANNUREV-CONTROL-042920-020211. URL https://doi.org/10.1146/annurev-control-042920-020211.

Felix Chalumeau, Shikha Surana, Clément Bonnet, Nathan Grinsztajn, Arnu Pretorius, Alexandre Laterre, and Thomas D. Barrett. Combinatorial optimization with policy adaptation using latent space search. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/18d3a2f3068d6c669dcae19ceca1bc24-Abstract-Conference.html.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 4759–4770, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Kunal Garg, Songyuan Zhang, Oswin So, Charles Dawson, and Chuchu Fan. Learning safe control for multi-robot systems: Methods, verification, and open challenges. *Annu. Rev. Control.*, 57: 100948, 2024. doi: 10.1016/J.ARCONTROL.2024.100948. URL https://doi.org/10.1016/j.arcontrol.2024.100948.

Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois C. Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artif. Intell.*, 319:103905, 2023. doi: 10.1016/J.ARTINT.2023.103905. URL https://doi.org/10.1016/j.artint.2023.103905.

Zijian Guo, Weichao Zhou, Shengao Wang, and Wenchao Li. Constraint-conditioned actor-critic for offline safe reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=nrRkAAAufl.

Kai-Chieh Hsu, Vicenc Rubies-Royo, Claire J. Tomlin, and Jaime F. Fisac. Safety and liveness guarantees through reach-avoid reinforcement learning. In Dylan A. Shell, Marc Toussaint, and M. Ani Hsieh (eds.), *Robotics: Science and Systems XVII, Virtual Event, July 12-16, 2021*, 2021. doi: 10.15607/RSS.2021.XVII.077. URL https://doi.org/10.15607/RSS.2021.XVII.077.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12498–12509, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html.

Zhengyao Jiang, Tianjun Zhang, Michael Janner, Yueying Li, Tim Rocktäschel, Edward Grefenstette, and Yuandong Tian. Efficient planning in a compact latent action space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cA77NrVEuqn.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

Jiaoyang Li, Andrew Tinka, Scott Kiesel, Joseph W. Durham, T. K. Satish Kumar, and Sven Koenig. Lifelong multi-agent path finding in large-scale warehouses. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 11272–11281. AAAI Press, 2021a. doi: 10.1609/AAAI.V35I13.17344. URL https://doi.org/10.1609/aaai.v35i13.17344.

Tianyu Li, Roberto Calandra, Deepak Pathak, Yuandong Tian, Franziska Meier, and Akshara Rai. Planning in learned latent action spaces for generalizable legged locomotion. *IEEE Robotics and Automation Letters*, 6(2):2682–2689, 2021b. doi: 10.1109/LRA.2021.3062342.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6379–6390, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html.

Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. Planning with goal-conditioned policies. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 14814–14825, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c8cc6e90ccbff44c9cee23611711cdc4-Abstract.html.

Soohyun Park, Chanyoung Park, and Joongheon Kim. Learning-based cooperative mobility control for autonomous drone-delivery. *IEEE Trans. Veh. Technol.*, 73(4):4870–4885, 2024. doi: 10.1109/TVT.2023.3330460. URL https://doi.org/10.1109/TVT.2023.3330460.

Guillaume Sartoretti, Justin Kerr, Yunfei Shi, Glenn Wagner, T. K. Satish Kumar, Sven Koenig, and Howie Choset. PRIMAL: pathfinding via reinforcement and imitation multi-agent learning. *IEEE Robotics Autom. Lett.*, 4(3):2378–2385, 2019. doi: 10.1109/LRA.2019.2903261. URL https://doi.org/10.1109/LRA.2019.2903261.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1312–1320. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/schaul15.html.

Harshit Sikchi, Wenxuan Zhou, and David Held. Learning off-policy with online planning. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1622–1633. PMLR, 2022. URL https://proceedings.mlr.press/v164/sikchi22a.html.

Oswin So and Chuchu Fan. Solving stabilize-avoid via epigraph form optimal control using deep reinforcement learning. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.085. URL https://doi.org/10.15607/RSS.2023.XIX.085.

Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkfrvsA9FX.

Jamison W. Weber, Dhanush R. Giriyan, Devendra R. Parkar, Dimitri P. Bertsekas, and Andréa W. Richa. Distributed online rollout for multivehicle routing in unmapped environments. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (eds.), *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pp. 1910–1918. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 2024. doi: 10.5555/3635637.3663054. URL https://dl.acm.org/doi/10.5555/3635637.3663054.

Yihang Yao, Zuxin Liu, Zhepeng Cen, Jiacheng Zhu, Wenhao Yu, Tingnan Zhang, and Ding Zhao. Constraint-conditioned policy optimization for versatile safe reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/29906cbd165b78991da2c4dbabc2a04b-Abstract-Conference.html.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre M. Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html.

Songyuan Zhang, Kunal Garg, and Chuchu Fan. Neural graph control barrier functions guided distributed collision-avoidance multi-agent control. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2373–2392. PMLR, 2023. URL https://proceedings.mlr.press/v229/zhang23h.html.

Songyuan Zhang, Oswin So, Mitchell Black, Zachary Serlin, and Chuchu Fan. Solving Multi-Agent Safe Optimal Control with Distributed Epigraph Form MARL. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.027.

Songyuan Zhang, Oswin So, Kunal Garg, and Chuchu Fan. GCBF+: A neural graph control barrier function framework for distributed safe multiagent control. *IEEE Trans. Robotics*, 41:1533–1552, 2025b. doi: 10.1109/TRO.2025.3530348. URL https://doi.org/10.1109/TRO.2025.3530348.

Wenbo Zhang, Osbert Bastani, and Vijay Kumar. MAMPS: Safe multi-agent reinforcement learning via model predictive shielding, 2019. URL https://arxiv.org/abs/1910.12639. arXiv:1910.12639.

Youpeng Zhao, Yaodong Yang, Zhenbo Lu, Wengang Zhou, and Houqiang Li. Multi-agent first order constrained optimization in policy space. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/7b64c47dcb067efd6be5eee854c14835-Abstract-Conference.html.