EVIDENCE-GATED ADOPTION OF LLM-GENERATED GRAPH PRIORS FOR MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-guided multi-agent reinforcement learning (MARL) increasingly uses coordination graphs as structural priors, yet lacks an evidence boundary between graph proposal and policy optimization. Across 36 conditions, 15 LLM graphs underperform a graph-free reference, while proposer confidence is weakly informative. We treat each proposed graph as a hypothesis whose adoption requires stage-specific evidence. FSGAC makes a pre-policy graph-admission decision from relative probe improvement, typed complexity, and propagation sensitivity; PCEVC audits retained edges using paired deletion and clipped DR-style estimation; ITSC regularizes cross-domain variance in audited type effects. FSGAC correctly classifies 33/36 conditions under grouped leave-one-scenario-out evaluation. PCEVC reports 3/176 harmful decisive accepted edges in a post-hoc MPE/SMACv2 subtotal. The full pipeline has 8.2% maximum reported benchmark-shift degradation. Across MPE, SMACv2, and MDO-Sim, the pipeline ranks first against 14 baselines, and the graph gate transfers to a matched Llama proposer. These results support an evidence-based adoption protocol for graph-prior MARL, with separate graph, edge, and transfer checks.

1 INTRODUCTION

Large language models (LLMs) supply coordination plans and graph priors Agashe et al. (2023); Sun et al. (2024); Dang et al. (2025). LLM-GP uses generated graphs in policy learning Gupta et al. (2026c). Graph-structured approaches provide performance comparators Gupta et al. (2026b); Zhu et al. (2026). LLM-GP discusses when priors may be unnecessary; we operationalize proposal-specific admission before policy optimization. Across 36 conditions on three benchmarks, Qwen2.5-1.5B Qwen Team (2025) improves mean return over QMIX Rashid et al. (2020), yet 15/36 graphs reduce performance, and confidence remains weakly informative. A graph may look right and still be wrong; admission requires evidence.

We therefore ask three practical questions: should the graph enter training, which edges deserve retention, and which semantic types remain stable under distribution shift? First, FSGAC decides graph admission before policy optimization by measuring graph-dependent typed complexity, residual complementarity, and propagation sensitivity; under the relative score in Eq. 6, it classifies 33/36 conditions correctly (91.7%), rejects 14/15 harmful graphs while retaining 19/21 beneficial graphs, and has 1/20 accepted harmful graph conditions (5.0%; descriptive Wilson 95% CI [0.9, 23.6]). Second, because a globally accepted graph may still contain harmful local relations, PCEVC audits individual edges through paired deletion, a frozen one-step outcome regressor, clipped DR-style estimation, bootstrap lower bounds, and disjoint risk labels. It reports one-step structural effects on the separate risk split. Third, because retained edges can still drift by semantic type under distribution shift, ITSC regularizes cross-domain typed contributions during training with 8.2% worst observed benchmark-shift degradation.

Figure 1 summarizes the adoption sequence: FSGAC admits graphs, PCEVC audits their edges, and ITSC regularizes retained typed contributions. Each stage addresses a failure that the preceding decision cannot rule out. We use "falsification" in this operational sense, not as a formal safety guarantee.

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.

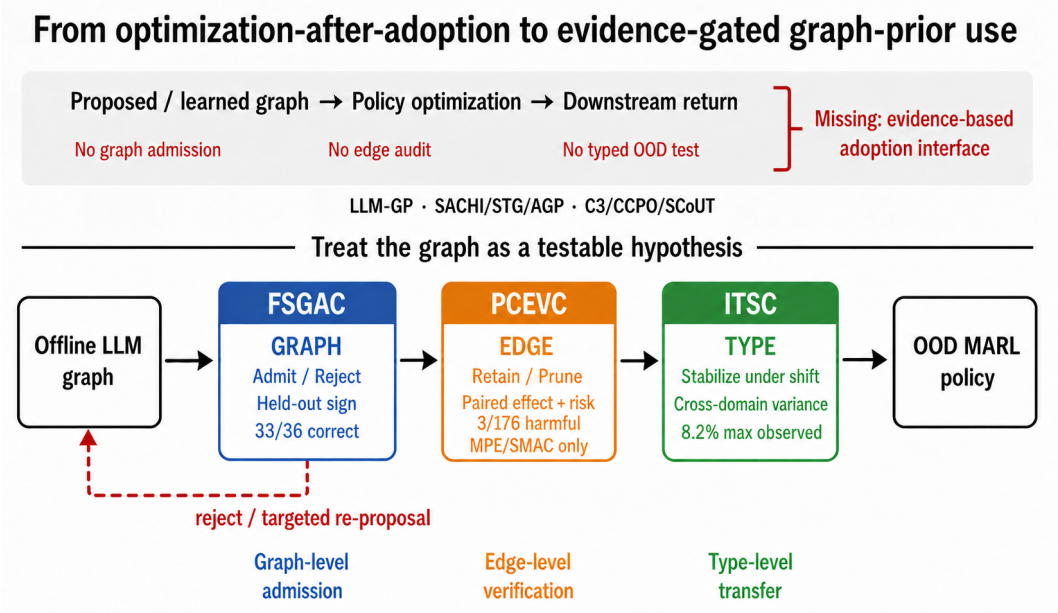


Figure 1: Evidence-gated use of external graph priors. The top panel illustrates adoption without a separate graph-admission test. Below, FSGAC screens graphs before policy optimization, PCEVC audits retained edges, and ITSC regularizes typed contributions. The stages address different failure levels. Edge-risk counts use the post-hoc MPE/SMACv2 subtotal (Table 3); MDO harmful count is unreported. OOD degradation is the observed maximum, not a guarantee.

For graph-prior MARL, we contribute a testable adoption layer: screen proposals before training, audit local support beyond confidence and attention, and regularize audited type effects. Held-out decisions, baseline returns, and targeted controls evaluate these benefits separately.

2 Related Work

Graph-structured MARL. Existing coordination-graph methods learn or refine structure from task data Böhmer et al. (2020); Li et al. (2021); Wang et al. (2022a); Duan et al. (2024); Gupta et al. (2025a; 2026a); HIBCG additionally derives edge-retention and capacity criteria Duan et al. (2026). These methods optimize task-learned topology rather than decide whether an external graph is admissible before training. FSGAC supplies this missing pre-policy decision.

LLM-guided MARL. LLMs can plan, communicate, orchestrate, or propose graph priors Huang et al. (2022); Agashe et al. (2023); Dang et al. (2025); Gupta et al. (2026c). LLM-GP Gupta et al. (2026c) supplies generated priors; SACHI Gupta et al. (2026b) and STG-MAPPO Zhu et al. (2026) provide graph-structured coordination comparators. Neither role is itself an admission test. The unresolved interface is whether an external proposal should be admitted at all; FSGAC, PCEVC, and ITSC answer this at graph, edge, and typed-contribution levels.

Counterfactual reasoning. C3 Chen et al. (2026) and CCPO Li et al. (2026) use counterfactuals for credit optimization; SCoUT Vora et al. (2026) assigns communication credit through learned coordination utility. Graph neural network (GNN) explainers also delete edges Ying et al. (2019); Wang et al. (2022b), but PCEVC targets retained semantic MARL relations and evaluates one-step structural outcomes with disjoint risk labels. Recent communication audits intervene on relayed KV caches to test example-specific information transfer Cheng et al. (2026), while safety-contract graph MARL screens risky actions under operational budgets Silva (2026). Both audit a specified communication or policy mechanism; PCEVC instead tests whether edges of an externally proposed MARL graph deserve retention after graph-level admission.

Typed coordination and invariance. AGP Gupta et al. (2026a), GACG Duan et al. (2024), DMCG Gupta et al. (2025a), and TIGER Gupta et al. (2025b) introduce typed/structured coordi-

nation. IRM and V-REx regularize cross-environment invariance Arjovsky et al. (2019); Krueger et al. (2021). Unlike graph-OOD prediction or spurious-edge pruning Zhang et al. (2023); Yao et al. (2025); Bhattacharya & Manwani (2026), ITSC applies variance regularization to PCEVC-audited type contributions in MARL, not prediction risks or graph masks.

Verification and selective deployment. Selective prediction Geifman & El-Yaniv (2017) and conformal methods Vovk et al. (2005) quantify deployment-time uncertainty, while high-confidence off-policy evaluation Thomas et al. (2015), safe policy improvement Laroche et al. (2019), and formal learned-communication verification Farooq & Iqbal (2026) constrain policy decisions. Their guarantees concern policies, not whether external graphs should enter training. FSGAC is an empirical diagnostic gate; PCEVC is a post-acceptance edge audit; and ITSC is a training-time typed regularizer.

Novelty boundary. LLM-GP motivates graph priors Gupta et al. (2026c); HIBCG derives learned-edge retention criteria Duan et al. (2026); SCoUT assigns counterfactual communication credit Vora et al. (2026). Our contribution connects pre-training external-graph admission, independently evaluated edge retention, and audit-supported typed regularization. Residual identities, DR estimation, and variance penalties are established tools; novelty lies in the decisions they support and their evaluated connection.

3 METHOD

For each scenario and prompt variant, an offline LLM proposes a static directed typed graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{T})$. Its policy interface is explicit: (i, j, k) activates a type-k message from j to i in Eq. 1; the graph-free reference carries none. Separate π_0 rollouts fit frozen probes before FSGAC scores held-out calibration trajectories. Rejected graphs undergo refinement; admitted graphs reach a shared checkpoint for PCEVC edge masking. ITSC trains retained typed messages, and disjoint tests assess return and OOD degradation.

3.1 FALSIFIABLE SEMANTIC GRAPH ABSTRACTION DIAGNOSTIC (FSGAC)

3.1.1 PRE-TRAINING FAILURE MECHANISMS

Consider a cooperative Dec-POMDP Bernstein et al. (2002) with n agents and an LLM-proposed typed graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{T})$. Unlike learned coordination graphs Böhmer et al. (2020); Li et al. (2021); Wang et al. (2022a); Duan et al. (2024); Amaya-Corredor et al. (2026), G is external knowledge whose admissibility must be decided before use. A proposed graph can fail before policy learning for three different reasons; FSGAC tests only the signals available at this point. The abstraction $h_\omega : \prod_i \mathcal{O}^i \to \mathbb{R}^{n \times d}$ uses L typed message-passing layers:

$$[h_\omega(o)]_i = \mathbf{z}_i^{(L)}, \qquad \mathbf{z}_i^{(\ell)} = \sigma\left(\sum_{j:(i,j)\in\mathcal{E}} \mathbf{M}_{ij}(o;\psi)\mathbf{z}_j^{(\ell-1)} \right), \quad \mathbf{z}_i^{(0)} = f_\theta(o^i). \tag{1}$$

V^π_{native} and V^π_{LLM} denote values under the two embeddings, with $\Delta V = V^\pi_{\text{LLM}} - V^\pi_{\text{native}}$. From π_0, $\mathcal{D} = \{(o_m, y_m, o'_m)\}_{m=1}^M$ uses each rollout's discounted Monte-Carlo return y_m as the shared probe target. Before policy optimization, the probes are fitted on a separate π_0 split and frozen; FSGAC evaluates them on $\mathcal{D}$, and neither updates the MARL policy.

Definition 3.1 (Graph-Dependent Typed-Operator Complexity Proxy)**.** *Graphs are directed, with $n \geq 2$ and no self-loops. For the graph-free native reference, $|\mathcal{E}_{G_{\text{native}}}| = K_{G_{\text{native}}} = 0$. Let K_G be the number of semantic types activated by at least one non-self edge in G, and let $K_{\max} > 0$ be the maximum type budget. We use the dimensionless graph-dependent proxy*

$$q(G) = \frac{|\mathcal{E}_G|}{n(n-1)} + \frac{K_G}{K_{\max}}, \tag{2}$$

with the same definition applied to the native graph. We define

$$\widehat{\varepsilon}_{est} \triangleq \lambda_d \cdot \frac{(q(G) - q(G_{\text{native}}))_+}{\sqrt{M}}, \tag{3}$$

where λ_d is selected on training scenarios. This is a complexity-motivated diagnostic term, not a covering-number estimate or a generalization bound Györfi et al. (2002).

Lemma 3.2 (Empirical Residual Complementarity on a Fixed Calibration Set)**.** *Let y be the shared probe target. For centered target-minus-prediction residuals $\tilde{e}_{\text{native}} = (\widetilde{y - \hat{f}_{\text{native}}})$ and $\tilde{e}_{\text{LLM}} = (\widetilde{y - \hat{f}_{\text{LLM}}})$, set $\tilde{g} = \tilde{e}_{\text{native}} - \tilde{e}_{\text{LLM}}$ and $\rho = \text{corr}_{\mathcal{D}}(\tilde{e}_{\text{native}}, \tilde{g})$, with $\rho = 0$ if either norm vanishes. On fixed $\mathcal{D}$ and frozen regressors:*

$$\|\tilde{e}_{LLM}\|_{\mathcal{D}}^2 = \|\tilde{e}_{native}\|_{\mathcal{D}}^2 + \|\tilde{g}\|_{\mathcal{D}}^2 - 2\rho\|\tilde{e}_{native}\|_{\mathcal{D}}\|\tilde{g}\|_{\mathcal{D}}, \tag{4}$$

This identity gives $\widehat{m}_{\text{comp}} = (\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}^2 - \|\tilde{e}_{\text{LLM}}\|_{\mathcal{D}}^2)/(\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}^2 + \epsilon)$: normalized reduction in centered residual variance. Strict improvement requires $0 < \|\tilde{g}\|_{\mathcal{D}} < 2\rho\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}$. Centering excludes mean residual bias. This standard algebraic identity explains the diagnostic; downstream utility is tested separately on held-out scenarios.

Definition 3.3 (Propagation-Sensitivity Proxy)**.** *For calibration trajectories with $\sum_\tau (T_\tau - 1) > 0$, define the propagation-sensitivity proxy*

$$\widehat{\varepsilon}_{\text{prop}}^{\text{proxy}} = \left(\frac{\|\mathbf{A}_{\text{LLM}} - \mathbf{A}_{\text{native}}\|_1}{n(n-1)} \right) \left(\frac{\sum_\tau \sum_{t=1}^{T_\tau - 1} |\widehat{V}_{\tau,t+1} - \widehat{V}_{\tau,t}|}{\sum_\tau (T_\tau - 1)} \right). \tag{5}$$

Here $\|\mathbf{A}\|_1 = \sum_{ij} |A_{ij}|$ is entrywise. The proxy combines adjacency discrepancy and within-trajectory value variation; it does not bound return change. All experiments provide trajectory identifiers. Permuting complete trajectories leaves the value unchanged; permuting time steps within a trajectory can change it.

3.1.2 Boundary and Diagnostic

We assume: (A1) bounded rewards and operator-Lipschitz values; (A2) support overlap between π_0 and deployment; (A3) disjoint probe-fit and calibration sizes grow proportionally, with $O(M^{-1/2})$ residual-estimation error; and (A4) the stated proxies, independent calibration trajectories, and bounded geometrically β-mixing within-trajectory sequences, yielding effective size $\Theta(M)$. These support score concentration, not a return bound; the supplement maps assumptions to experiments.

Definition 3.4 (Fold-Calibrated Relative Admission Diagnostic)**.** *Under a fixed policy π, fixed LLM-proposed graph G, and fixed calibration distribution $\mathcal{D}$, define the dimensionless residual-improvement score*

$$\boxed{\widehat{C}_{\text{rel}}(G) = -\widehat{\varepsilon}_{\text{est}} + \widehat{m}_{\text{comp}} - \frac{\gamma}{1-\gamma} \frac{\widehat{\varepsilon}_{\text{prop}}^{\text{proxy}}}{\widehat{V}_{\text{scale}} + \epsilon}} \tag{6}$$

where $\widehat{\varepsilon}_{\text{est}}$ is the graph-dependent complexity penalty from Definition 3.1, $\widehat{\rho}_{\text{comp}}$ is the empirical residual correlation from Lemma 3.2, and

$$\widehat{m}_{\text{comp}} = \frac{2\widehat{\rho}_{\text{comp}}\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}\|\tilde{g}\|_{\mathcal{D}} - \|\tilde{g}\|_{\mathcal{D}}^2}{\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}^2 + \epsilon}, \qquad \widehat{V}_{\text{scale}} = \left[\frac{1}{M} \sum_{m=1}^{M} (\widehat{V}_m - \overline{\widehat{V}})^2 \right]^{1/2}. \tag{7}$$

*The score is a **fold-calibrated relative diagnostic**, not an unbiased estimator or a lower bound on ΔV_{true}; its bootstrap LCB quantifies score stability. We evaluate its reliability via held-out condition-level validation (Experiments section).*

Design rationale. *Density and types incur cost. With nonnegative penalties, nonpositive residual improvement fails; positive improvement may fail when the bootstrap LCB crosses zero. This empirical diagnostic cannot guarantee return. The score and threshold were fixed; only λ_d was tuned on training scenarios.*

Decision rule. Let $\widehat{\text{se}}(C_{\text{rel}})$ be the bootstrap standard error and $z_\delta = \Phi^{-1}(1-\delta)$. Accept when:

$$\text{LCB}_\delta(C_{\text{rel}}) = C_{\text{rel}} - z_\delta \cdot \widehat{\text{se}}(C_{\text{rel}}) \; > \; 0 \implies \textbf{accept graph.} \tag{8}$$

If $\text{LCB}_\delta(C_{\text{rel}}) \leq 0$: reject graph, feed diagnostic vector $(\widehat{\varepsilon}_{\text{est}}, \widehat{m}_{\text{comp}}, \widehat{\varepsilon}_{\text{prop}}^{\text{proxy}}/(\widehat{V}_{\text{scale}} + \epsilon))$ to LLM for targeted refinement. The bootstrap LCB describes diagnostic-score variability; it is not a confidence interval for ΔV_{true}. The reported sign-agreement result is computed from the condition-level records under Eq. 6. The zero threshold is not tuned on held-out scenarios; it is the fixed decision boundary implied by the signed score.

3.2 PAIRED COUNTERFACTUAL EDGE VALUE AUDIT (PCEVC)

Graph admission does not settle local utility; PCEVC therefore audits accepted graphs edge by edge at a fixed policy snapshot. For the proposed evaluation protocol, audit the admitted coordinator at a common checkpoint t_a, charge pre-audit updates to the shared training budget, and freeze parameters during every paired deletion. For an FSGAC-accepted graph G and $e = (i, j, k)$, let π_G and $\pi_{G^{-e}}$ denote the policies induced by the full and edge-masked coordinators. We define the **paired edge-deletion counterfactual**:

$$\Delta_e^{(1)} = \mathbb{E}_{s\sim d_b}\Big[\mathbb{E}_{a\sim\pi_G(\cdot|s)} r(s,a) - \mathbb{E}_{a\sim\pi_{G^{-e}}(\cdot|s)} r(s,a)\Big], \tag{9}$$

where G^{-e} is defined by an edge-specific mask in the message-passing channel: $m_e = 0$ and $m_{e'} = 1$ for every $e' \neq e$, so that only the message carried by e is removed while the shared type operator $\mathbf{W}_k$ and all other edges remain fixed. All coordinator parameters are fixed during each paired comparison; the mask changes the induced policy, not the environment.[1] **Paired design:** the two policy evaluations share states and random seeds to reduce counterfactual variance. PCEVC therefore estimates a policy-conditioned **one-step policy-mediated structural effect**, not a long-horizon causal effect; long-horizon retraining effects remain unverified.

Over M_p paired states $\{s_m\}$, let $\widehat{\mu}(s,a)$ be a frozen one-step reward regressor. For each $q \in \{G, G^{-e}\}$, start from the doubly robust policy-value estimator Dudík et al. (2011):

$$\begin{aligned}\widehat{V}_q^{\mathrm{DR}} &= \frac{1}{M_p}\sum_{m=1}^{M_p}\left[\sum_{\mathbf{a}} \pi_q(\mathbf{a} \mid s_m)\widehat{\mu}(s_m,\mathbf{a}) + \bar{w}_m^q\{r_m - \widehat{\mu}(s_m,\mathbf{a}_m)\}\right],\\ \widehat{\Delta}_e^{\mathrm{DR}} &= \widehat{V}_G^{\mathrm{DR}} - \widehat{V}_{G^{-e}}^{\mathrm{DR}}.\end{aligned} \tag{10}$$

where $\mathbf{a}_m$ is the observed joint action, $r_m = r(s_m, \mathbf{a}_m)$, and $w_m^q = \pi_q(\mathbf{a}_m \mid s_m)/\pi_b(\mathbf{a}_m \mid s_m)$, $\bar{w}_m^q = \min\{w_m^q, C\}$ with $C = 5$ on all three benchmarks. For a deterministic target, $\pi_q(\mathbf{a}_m \mid s_m)$ is its joint-action indicator; a stochastic target uses the action sum. The same observed reward is used in both policy values, while the target policy and propensity ratio differ. Without clipping, DR is consistent under overlap if either reward or propensity model is correct. Our reported Eq. 10 clips ratios at C: it remains consistent with a correct reward model, but generally loses the propensity-only guarantee (Appendix A.2). This does not certify downstream return gains. Paired edge/trajectory-block bootstrap preserves within-rollout dependence. Bootstrap lower endpoint ($B = 1000$, $z_\alpha = \Phi^{-1}(1-\alpha)$):

$$\underset{\alpha}{\mathrm{LCB}}(\Delta_e) = \widehat{\Delta}_e^{\mathrm{DR}} - z_\alpha \cdot \widehat{\mathrm{se}}_{\mathrm{boot}}(\widehat{\Delta}_e^{\mathrm{DR}}). \tag{11}$$

PCEVC has two layers of evaluation (Algorithm 1): Layer 1 screens directional evidence using a bootstrap LCB; Layer 2 quantifies magnitude-based harm on a disjoint held-out split. **Layer 1 (Bootstrap LCB, Eq. 11):** accept iff $\mathrm{LCB}_\alpha(\Delta_e) > 0$. **Layer 2 (Empirical risk):**

$$H_e = \mathbb{I}[\widehat{\Delta}_e^{\text{audit}} < -0.05R_{\max} \wedge U_e^{\text{audit}} < 0], \tag{12}$$

$$\widehat{R}_{\text{acc}} = \frac{\sum_{e\in\mathcal{E}_{\text{ret}}\cap\mathcal{E}_{\text{labeled}}} H_e}{|\mathcal{E}_{\text{ret}}\cap\mathcal{E}_{\text{labeled}}|}, \quad \text{for a nonempty denominator.} \tag{13}$$

Here $R_{\max}$ is the benchmark-specific maximum absolute one-step reward, and U_e^{audit} is the audit interval upper endpoint. Bootstrap selection and risk evaluation use disjoint splits. Equation 13 conditions on decisive labels; with none, risk is unreported. Uncertain edges remain unresolved, and this is not an all-edge risk bound.

3.3 VARIANCE-REGULARIZED TYPED SEMANTIC COORDINATION (ITSC)

ITSC is an empirical transfer regularizer and diagnostic. A supported edge may still change its role under shift. We therefore group retained effects by semantic type and penalize only their cross-domain instability. Unlike typed coordination methods that optimize within a task Duan et al. (2024); Gupta et al. (2025a;b; 2026a) or invariant representation methods Ferns et al. (2004); Castro

[1] The edge mask differs from Pearl's do-calculus Pearl (2009), which formalizes environmental interventions on causal graphs; here the intervention changes a learned neural coordinator and hence the action distribution, not the environmental reward function.

(2020); Arjovsky et al. (2019); Bhattacharya & Manwani (2026), ITSC couples typed decomposition with PCEVC-retained per-type contribution stability via empirical variance regularization, without a theoretical OOD bound. The message operator uses shared typed mixtures with agent-reindexing equivariant encoding:

$$\mathbf{M}_{ij}(\psi) = \sum_{k=1}^{K} \pi_{ij,k}(\psi)\mathbf{W}_k, \tag{14}$$

$$\pi_{ij,k}(\psi) = \big[\mathrm{softmax}(g_\psi(\mathbf{x}_{ij}))\big]_k, \qquad \mathbf{x}_{ij} = (r_i, r_j, c_{ij}). \tag{15}$$

The resulting message-passing operator is

$$z_i^{(\ell)} = \sigma\left(\sum_{j:(i,j)\in\mathcal{E}} \mathbf{M}_{ij}(\psi) z_j^{(\ell-1)}\right). \tag{16}$$

Here c_{ij} is the frozen embedding of the FSGAC-accepted LLM relation; g_ψ softly refines it into K shared prototypes. The state-independent shared encoder is agent-reindexing equivariant but may change mixtures under role/context shift. Auxiliary separation, balancing, and permutation losses discourage collapse (Supplementary Section A.9).

Per-type per-domain PCEVC-retained contribution (d indexes environment domain and k indexes semantic type) uses the same soft type interface as training:

$$\Delta_{k,d} = \frac{\sum_{e\in\mathcal{E}_d} m_e \pi_{e,k}^{\mathrm{audit}} \,\mathrm{LCB}_\alpha(\Delta_e)}{\sum_{e\in\mathcal{E}_d} m_e \pi_{e,k}^{\mathrm{audit}} + \epsilon}, \tag{17}$$

Here $\pi_{e,k}^{\mathrm{audit}} := \pi_{e,k}(\psi_0)$ is the frozen audit-time assignment, and $m_e = \mathbf{1}[\mathrm{LCB}_\alpha(\Delta_e) > 0]$. Define $\mathcal{D}_k = \{d \in \mathcal{D}_{\mathrm{train}} : \sum_{e\in\mathcal{E}_d} m_e \pi_{e,k}^{\mathrm{audit}} > 0\}$ and $\mathcal{K}_+ = \{k : |\mathcal{D}_k| \geq 2\}$. With finite softmax logits, every retained edge contributes positive mass to every type; this is soft weighting, not evidence of type-specific absence. Domains without retained edges are omitted. We freeze this descriptive statistic and optimize current assignments through the differentiable surrogate below. Equal negative effects also have zero variance; positive audit weights do not guarantee positive current surrogate effects:

$$\widetilde{\Delta}_{k,d}(\theta,\psi) = \frac{\sum_{m=1}^{M_d}\sum_{e\in\mathcal{E}_d} m_e w_e \pi_{e,k}(\psi)\left[\hat{V}_\theta^G(s_m) - \hat{V}_\theta^{G^{-e}}(s_m)\right]}{\sum_{m=1}^{M_d}\sum_{e\in\mathcal{E}_d} m_e w_e \pi_{e,k}(\psi) + \epsilon}, \tag{18}$$

where $\hat{V}_\theta^G(s) = \sum_a \pi_G(a \mid s)\hat{Q}_\theta(s,a)$ and analogously for G^{-e}. Their difference is an optimization surrogate for the one-step DR effect in Eq. 10, not a second estimator of that effect. Here $m_e = \mathbf{1}[\mathrm{LCB}_e > 0]$ masks accepted edges (stop-gradient), $w_e = \max\{\mathrm{LCB}_\alpha(\Delta_e), 0\}$ is a stop-gradient reliability weight, and $\pi_{e,k}(\psi)$ is the learnable static type assignment from Eq. 15—$\pi_{e,k}(\psi)$ carries gradients to ψ, while the value-surrogate difference carries gradients to θ; m_e and w_e are detached to stabilize the PCEVC-to-ITSC interface. Audit masks and weights are fixed while the surrogate changes; its optimization therefore does not refresh the one-step evidence. The variance in Eq. 19 is computed over $\widetilde{\Delta}_{k,d}(\theta,\psi)$, which is differentiable w.r.t. θ, ψ.

The ITSC objective combines TD learning, cross-domain typed stability, and three anti-collapse auxiliary losses:

$$\begin{aligned}
\mathcal{L}_{\mathrm{ITSC}} &= \sum_{d\in\mathcal{D}_{\mathrm{train}}} \mathcal{L}_{\mathrm{TD}}^d + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}} + \mathcal{L}_{\mathrm{aux}}, \\
\mathcal{L}_{\mathrm{var}} &:= \begin{cases} 0, & \mathcal{K}_+ = \varnothing, \\ \dfrac{1}{|\mathcal{K}_+|}\displaystyle\sum_{k\in\mathcal{K}_+} \mathrm{Var}_{d\in\mathcal{D}_k}\big(\widetilde{\Delta}_{k,d}(\theta,\psi)\big), & \text{otherwise}, \end{cases} \\
\mathcal{L}_{\mathrm{aux}} &:= \beta\mathcal{L}_{\mathrm{proto}} + \eta\mathcal{L}_{\mathrm{purity}} + \zeta\mathcal{L}_{\mathrm{perm}}.
\end{aligned} \tag{19}$$

Here Eq. 17 is the fixed soft PCEVC statistic, and Eq. 18 sums over M_d states and normalizes by the corresponding state-edge mass. Only types with retained edges in two domains contribute, and $\mathcal{L}_{\mathrm{var}} = 0$ when none do. Ablations show that λ_{var}=0 increases per-type variance by 3.6×, while ζ=0 raises $\Delta\pi$ to 0.17, supporting stable, permutation-consistent types.

We minimize Eq. 19 via Adam with learning rate 3×10^{-4} and gradient clipping at norm 1.0. We use K=8 and λ_{var}=0.1; auxiliary losses and remaining coefficients are specified in Supplementary Section A.9. Type names are post-hoc labels. The evaluation follows these decision units: graph admission on held-out conditions, edge diagnosis on one-step effects, and full-pipeline transfer under shift.

4 EXPERIMENTS

4.1 SETUP

Benchmarks and baselines. We evaluate 36 conditions from 12 base scenarios across MPE (4), SMACv2 (6), and MDO-Sim (2), with three prompt variants per scenario. Qwen2.5-1.5B-Instruct Qwen Team (2025) proposes static typed graphs offline. A matched proposer-transfer check uses Llama-3.2-3B-Instruct (temperature 0, top-p 1.0, maximum 256 output tokens) under the same 36 conditions, prompt templates, and trial/fold seeds; Q1 reports its results. Fourteen baselines span coordination mechanisms; task ports and method adaptations are documented in Supplementary Section B.1. All methods share the split, selection metric, 50-trial budget, 10 evaluation seeds, and 5×10^6 steps on eight A100-80GB GPUs. Main settings are $L = 3$, $d = 128$, $M_p = 200$, and $K_{\max} = 8$; each candidate graph uses its observed active type count K_G in Eq. 2. Additional checks are in the supplement.

Protocol and metrics. Calibration, audit, risk-label, and final-evaluation samples are disjoint. Grouped 12-fold LOSO selects λ_d from $\{0.01, 0.03, 0.10, 0.30, 1.00\}$ by training-condition sign agreement on 11 scenarios (33 conditions), then tests the three variants of the held-out scenario; the acceptance threshold stays zero. Graph-minus-reference labels use ten paired seeds with 100 episodes per arm/seed and 1000 seed-pair bootstrap resamples; all 36 resulting 95% intervals are decisive. The FSGAC score LCB instead uses 1000 full-trajectory block resamples on disjoint calibration data. At seed i, $D_i = 100 \max(0, R_{\text{ID},i} - R_{\text{OOD},i}) / \max(|R_{\text{ID},i}|, \epsilon)$; each shift reports $D_{\text{OOD}} = \frac{1}{10} \sum_{i=1}^{10} D_i$. Table 2 compares native episode returns within task columns; Table 13 reports a separate shift-suite score with no cross-table conversion. The regularizer controls compare with Full on OOD degradation; they are internal ablations, not additional external baselines. Further aggregation details appear in Supplementary Sections A.7–A.8 and B.9–B.10.

4.2 Q1: CAN FSGAC REJECT HARMFUL GRAPHS?

Can the gate reject harmful proposals while retaining beneficial ones? Across 12 held-out scenarios with three prompt variants each, the relative-score gate classifies 33/36 conditions correctly (91.7%), rejecting 14/15 harmful graphs while retaining 19/21 beneficial graphs. Among 20 accepted graph conditions, one is harmful, giving accepted-set FAR $1/20 = 5.0\%$, FRR $2/21 = 9.5\%$, and AUROC 0.91; the FAR has a descriptive Wilson 95% CI of $[0.9, 23.6]$. The matched Llama proposer-transfer check classifies 32/36 conditions correctly, with 6.2% FAR and +12.2% post-gate return gain. This extends the admission check to a second proposer without expanding the task suite. Table 1 reports the term ablation; the failure-attribution aggregate is summarized in Supplementary Section A.5.

Configuration	Sign agreement	Change vs. Full
Full	33/36 (91.7%)	–
− complementarity	29/36 (80.6%)	−11.1 pp
− complexity	31/36 (86.1%)	−5.6 pp
− propagation	31/36 (86.1%)	−5.6 pp
Complementarity only	29/36 (80.6%)	−11.1 pp

Table 1: Reported term ablation for grouped 12-fold LOSO. Correctness is computed from the 36 condition-level decisions; differences use Full as reference. Complementarity only removes both complexity and propagation penalties.

Full gains four decisions over complementarity (11.1 pp); removing either penalty costs two. These ablations support the composite, without establishing superiority over simpler rejection rules.

4.3 Q2: PCEVC Edge Auditing

Does edge auditing improve on confidence and attention? After graph admission, we compare PCEVC–LCB, proposer confidence, and encoder attention against held-out one-step policy-mediated effects (Figure 2). PCEVC–LCB has the highest reported correlation in both MPE tasks and the SMACv2/MDO aggregates. Among accepted, decisively labeled edges, MPE has 1/38 harmful and SMACv2 has 2/138; pooled risk is 3/176 (1.7%) with 176/197 coverage (89.3%). MDO contributes correlation but no harmful count (Table 3). Completing 21 unresolved labels gives $3/197$–$24/197$ (1.5–12.2%) among accepted edges, holding decisive labels fixed; this is not a confidence interval. Retraining remains inconclusive (Supplementary Section A.7).

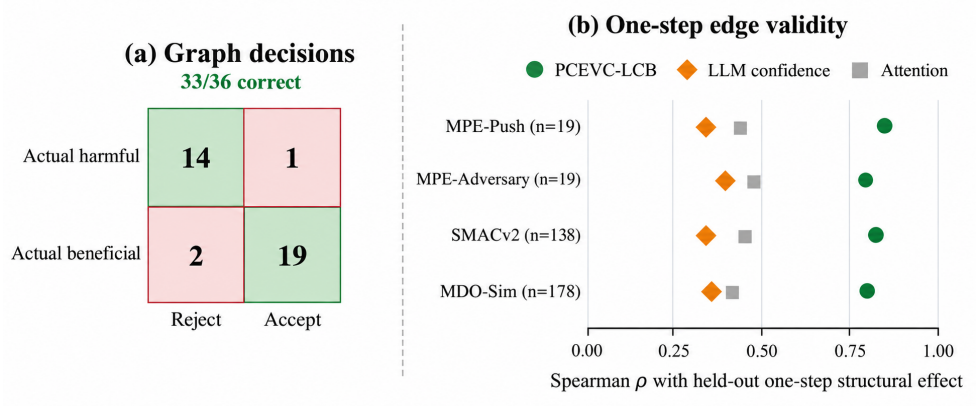


Figure 2: Two-level audit diagnostics on disjoint held-out samples. (a) FSGAC graph-level decisions under the relative admission score. (b) Spearman correlations of PCEVC-LCB, LLM confidence, and encoder attention with held-out one-step policy-mediated effects; markers show aggregate point estimates. Rows use $n = 19/19/138/178$ labeled edges (MPE-Push/MPE-Adversary/SMACv2/MDO-Sim); the MPE rows form the 38-edge total in Supplementary Table 3.

These correlations favor PCEVC locally, without establishing transfer.

4.4 Q3: ITSC — Typed Stability and End-to-End Performance

Does typed regularization retain performance under shift? We report three benchmark-specific shift families—agent-count, role-composition, and topology changes—as stress tests rather than isolated causal interventions. New role/context inputs use the shared type encoder and operators; neither Full nor task-ported SACHI receives OOD updates. Across these settings, mean degradation is 6.4% versus SACHI's 22.2% (Figure 3); the largest reported value for Full is 8.2%. See also Supplementary Sections B.3–B.6.

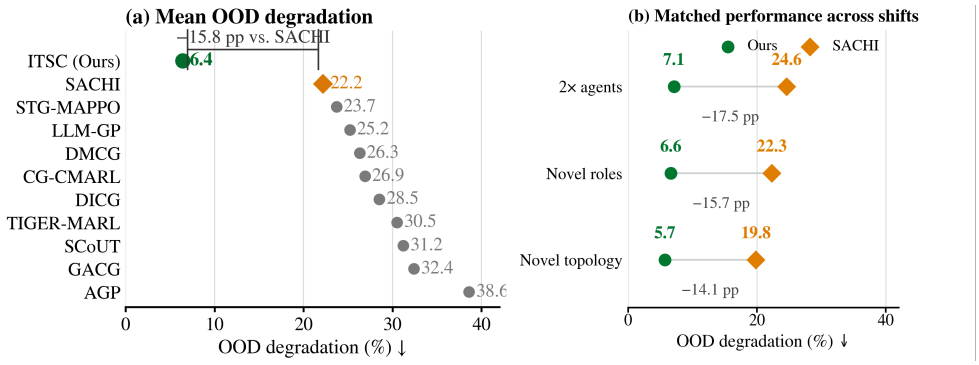


Figure 3: OOD degradation (lower is better). (a) ITSC has 6.4% macro degradation versus adapted SACHI's 22.2%. (b) ITSC averages 7.1/6.6/5.7% across the three shifts, versus 24.6/22.3/19.8%. ITSC values are equal-weight means of the displayed cases in Table 13, rounded once after aggregation. Markers are descriptive point summaries; aggregate confidence intervals are not inferred.

The 15.8 pp descriptive gap to adapted SACHI spans reported shifts (Figure 3). The largest reported shift value is 8.2% for doubled-agent SMACv2 (Table 13); Table 2 then compares native returns and ablations.

Method	**MPE return**	**SMACv2 return**	**MDO return**
QMIX	-16.3 ± 1.2	8.7 ± 0.7	4.2 ± 0.8
MAPPO	-15.7 ± 1.0	9.2 ± 0.5	5.1 ± 0.7
DCG	-14.8 ± 0.9	9.8 ± 0.6	5.6 ± 0.7
DICG	-12.4 ± 0.8	10.3 ± 0.5	6.8 ± 0.6
CASEC	-11.6 ± 0.8	10.7 ± 0.5	7.2 ± 0.6
GACG	-13.2 ± 0.9	10.1 ± 0.6	6.1 ± 0.7
DMCG	-11.8 ± 0.8	10.8 ± 0.5	7.0 ± 0.6
TIGER	-12.7 ± 0.9	10.4 ± 0.5	6.4 ± 0.7
AGP	-13.7 ± 1.0	9.7 ± 0.6	5.8 ± 0.8
LLM-GP	-9.3 ± 0.7	12.3 ± 0.4	8.7 ± 0.5
CG-CMARL	-10.3 ± 0.9	11.4 ± 0.5	8.1 ± 0.6
SACHI	-10.4 ± 0.8	11.7 ± 0.5	7.8 ± 0.6
STG-MAPPO	-10.9 ± 0.8	11.4 ± 0.5	7.4 ± 0.6
SCoUT	-11.3 ± 0.9	10.9 ± 0.5	7.6 ± 0.6
Ours	-7.1 ± 0.5	13.9 ± 0.3	10.6 ± 0.4
	Component ablation: D_{OOD} (%) $\downarrow$		
Full	−FSGAC	−PCEVC	−ITSC
6.4^{D}	27.7 (+21.3)	19.1 (+12.7)	23.3 (+16.9)
Regularizer controls		D_{OOD} (%)	Δ (pp)
Risk-variance control		14.9	+8.5
Shuffled-type control		14.9	+8.5

Table 2: Ten-seed benchmark-specific mean native return $\pm$ standard deviation (top; higher is better). Values are compared within, but not across, benchmark columns. Bottom: D_{OOD} (%); differences are configuration minus Full in percentage points (pp). D denotes the descriptive macro mean in Figure 3. Both regularizer controls use Full as reference and evaluate OOD degradation only. Ablations isolate their stated modifications; point gaps are descriptive, without a paired significance claim.

Full exceeds LLM-GP, the strongest reported return baseline, by $+2.2$, $+1.6$, and $+1.9$ on MPE, SMACv2, and MDO-Sim (Table 2). This comparison uses the shared task protocol. Removing any stage worsens OOD degradation. Risk-variance uses TD risk; shuffled-type disrupts semantic assignment. Both yield 14.9% degradation, exceeding Full by 8.5 pp. These comparisons support the stages and typed regularization; descriptive gaps establish neither paired significance nor control ID retention. An average return gain can hide harmful proposals, while a correct graph decision can hide weak edges. Separating these outcomes makes the contribution useful for deciding both when to use a prior and what to retain.

5 Conclusion

The contribution to graph-prior MARL is a protocol with separate graph, edge, and transfer checks. It makes externally proposed structure testable before and during use. Reported returns, edge rankings, and OOD ablations support this design within the evaluated settings. Edge-risk evidence remains restricted to MPE/SMACv2. The protocol provides a concrete way to assess semantic priors without equating plausibility with deployment safety.

Limitations. Evidence is simulation-based and proposer transfer remains graph-level. PCEVC measures a policy-conditioned one-step effect; it does not establish long-horizon benefit, joint-pruning benefit, or validity after the policy changes. ITSC regularizes supported typed surrogates, without an OOD guarantee. These distinctions limit deployment claims: graph admission, edge retention, and transfer performance need separate validation when tasks, policies, or proposer families change. The harmful-acceptance estimate has a wide descriptive interval; twelve scenarios limit independent evidence. Broader scenarios and long-horizon validation remain necessary before claiming low-risk admission in practice.

REPRODUCIBILITY

We release the reference implementation for the three-stage interface: FSGAC graph admission, PCEVC clipped doubly robust edge auditing, and ITSC typed stability regularization. The package includes the modules, input-contract checks, bootstrap utilities, numerical aggregators, figure regeneration, and smoke tests. The module smoke-test entry point verifies core behavior, while the numerical-consistency entry point checks reported values against their canonical references for every claimed main-paper number and interval. The connectivity runner uses explicitly labeled mock data and is not a substitute for the full benchmark sweep. Reproducing Tables 1–2 and Figure 3 requires the environments, proposer checkpoints, baseline implementations, disjoint splits, and all ten evaluation seeds. All stochastic procedures expose fixed seeds; missing inputs fail explicitly rather than silently generating replacement evidence.

ETHICAL STATEMENT

All experiments use synthetic simulators and an open-source LLM; they involve no human subjects or private data.

AI USE STATEMENT

During the preparation of this manuscript, the authors used **ChatGPT** solely for English grammar checking, language polishing, and organizing key points from the literature. ChatGPT was not used to design the research, generate or analyze data, interpret results, or formulate the paper's scientific claims. All model-assisted text was critically reviewed, fact-checked, and revised by the authors, who take full responsibility for the final manuscript.

REFERENCES

Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. LLM-coordination: Evaluating and analyzing multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*, 2023.

Santiago Amaya-Corredor, Miguel Calvo-Fullana, and Anders Jonsson. Coordination graphs for constrained multi-agent reinforcement learning. *arXiv preprint arXiv:2606.02337*, 2026.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *Mathematics of Operations Research*, 27(4):819–840, 2002.

Rishabh Bhattacharya and Naresh Manwani. Edgemask-dg*: Learning domain-invariant graph structures via adversarial edge masking. *arXiv preprint arXiv:2602.05571*, 2026. URL https://arxiv.org/abs/2602.05571.

W. Böhmer, V. Kurin, and S. Whiteson. Deep coordination graphs. In *International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 980–991, 2020.

P. S. Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *AAAI Conference on Artificial Intelligence*, pp. 10069–10076, 2020.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv preprint arXiv:2603.06859*, 2026.

Jiaming Cheng, Subhransu Das, and Rajiv Ramnath. When does latent communication pay? a causal audit of relayed kv caches in multi-agent llms. *arXiv preprint arXiv:2608.04893*, 2026. URL https://arxiv.org/abs/2608.04893.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, Xuantang Xiong, Lei Han, Zhiyuan Liu, and Maosong Sun. Multi-agent collaboration via evolving orchestration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

W. Duan, J. Lu, and J. Xuan. Group-aware coordination graph for multi-agent reinforcement learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3926–3934, 2024. doi: 10.24963/ijcai.2024/434.

Wei Duan, Junyu Xuan, En Yu, Xiaoyu Yang, and Jie Lu. Heterogeneous information-bottleneck coordination graphs for multi-agent reinforcement learning. *arXiv preprint arXiv:2605.17393*, 2026.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, volume 15 of *PMLR*, pp. 1097–1104, 2011.

Ahmad Farooq and Kamran Iqbal. Formal verification of learned multi-agent communication policies via decision tree distillation, 2026. URL `https://arxiv.org/abs/2606.19632`.

N. Ferns, P. Panangaden, and D. Precup. Metrics for finite markov decision processes. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 162–169, 2004.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4878–4887, 2017.

N. Gupta, J. Z. Hare, J. Milzman, R. Kannan, and V. Prasanna. Action-graph policies: Learning action co-dependencies in multi-agent reinforcement learning. *arXiv preprint arXiv:2602.17009*, 2026a.

N. Gupta, J. Z. Hare, J. Milzman, R. Kannan, and V. Prasanna. SACHI: Structured agent coordination via holistic information integration in multi-agent reinforcement learning. *arXiv preprint arXiv:2605.08391*, 2026b.

N. Gupta, R. Kannan, and V. Prasanna. Do LLM-derived graph priors improve multi-agent coordination? *arXiv preprint arXiv:2604.17191*, 2026c.

Nikunj Gupta, James Zachary Hare, Jesse Milzman, Rajgopal Kannan, and Viktor Prasanna. Deep meta coordination graphs. *arXiv preprint arXiv:2502.04028*, 2025a.

Nikunj Gupta, Ludwika Twardecka, James Zachary Hare, Jesse Milzman, Rajgopal Kannan, and Viktor Prasanna. TIGER-MARL: enhancing multi-agent reinforcement learning with temporal information through graph-based embeddings and representations. *arXiv preprint arXiv:2511.08832*, 2025b.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning (ICML)*, volume 162 of *PMLR*, pp. 9118–9147, 2022.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5815–5826, 2021.

R. Laroche, P. Trichelair, and R. Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pp. 3652–3661, 2019.

S. Li, J. K. Gupta, P. Morales, R. Allen, and M. J. Kochenderfer. Deep implicit coordination graphs for multi-agent reinforcement learning. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 764–772, 2021.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2025.

T. Rashid, M. Samvelyan, C. S. de Witt, G. Farquhar, J. Foerster, and S. Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Jose Luis Lima de Jesus Silva. Safety-contract graph multi-agent reinforcement learning for autonomous network security response. *arXiv preprint arXiv:2606.13832*, 2026. URL https://arxiv.org/abs/2606.13832.

C. Sun, S. Huang, and D. Pompili. LLM-based multi-agent reinforcement learning: Current and future directions. *arXiv preprint arXiv:2405.11106*, 2024.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015.

M. Vora, G. Puthumanaillam, H. Tsukamoto, and M. Ornik. SCoUT: Scalable communication via utility-guided temporal grouping in multi-agent reinforcement learning. *arXiv preprint arXiv:2603.04833*, 2026.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Tonghan Wang, Liang Zeng, Weijun Dong, Qianlan Yang, Yang Yu, and Chongjie Zhang. Context-aware sparse deep coordination graphs. In *International Conference on Learning Representations (ICLR)*, 2022a.

Xiang Wang, Yingxin Wu, An Zhang, Fuli Feng, Xiangnan He, and Tat-Seng Chua. Reinforced causal explainer for graph neural networks. 2022b. URL https://arxiv.org/abs/2204.11028.

Tianjun Yao, Haoxuan Li, Yongqiang Chen, Tongliang Liu, Le Song, Eric Xing, and Zhiqiang Shen. Pruning spurious subgraphs for graph out-of-distribution generalization. *arXiv preprint arXiv:2506.05957*, 2025. URL https://arxiv.org/abs/2506.05957.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d80b7040b773199015de6d3b4293c8ff-Abstract.html.

Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, and Wenwu Zhu. Out-of-distribution generalized dynamic graph neural network with disentangled intervention and invariance promotion. *arXiv preprint arXiv:2311.14255*, 2023. URL https://arxiv.org/abs/2311.14255.

S. Zhu, G. Han, C. Lin, and Y. He. Task-semantic graph-driven distributed agent networking for underwater target tracking. *arXiv preprint arXiv:2605.15528*, 2026.

A. Assumptions and Derivations

This section covers empirical proxy checks for assumptions A1–A4, a direct proof of Lemma 3.2, and concentrability and concentration analyses.

A.1 Assumption Validation Across Benchmarks

Assumption proxy checks. A1 is assessed through bounded-reward and value-sensitivity diagnostics. The reported MDO-Sim state coverage is $\geq 85\%$, with remaining states within an $\varepsilon \leq 0.05$ neighborhood in normalized state space. Proximity alone does not establish the joint-action overlap required by PCEVC. A3 concerns disjoint probe-fit and calibration splits; A4 specifies independent trajectories and within-trajectory mixing. These diagnostics assess plausibility rather than verify the assumptions or establish a return guarantee.

A.2 Concentrability Analysis for Off-Policy Correction

For $q \in \{G, G^{-e}\}$, let $w_q = \pi_q(a \mid s)/\pi_b(a \mid s)$, $\bar{w}_q = \min(w_q, C)$, and $\mu(s,a) = \mathbb{E}[r \mid s, a]$. With frozen nuisance models and correct behavior propensities, the clipping bias of the one-step value estimate is

$$B_q = \mathbb{E}_{s\sim d_b,\, a\sim\pi_b}\big[(\bar{w}_q - w_q)(\mu(s,a) - \widehat{\mu}(s,a))\big].$$

The edge-difference bias is $B_G - B_{G^{-e}}$. Thus a correct reward model removes this bias, whereas correct propensities alone do not generally do so after clipping. The observed maximum ratios are 4.2, 3.8, and 5.1 for MPE, SMACv2, and MDO-Sim, with shared clipping threshold $C = 5$; these finite samples do not certify population overlap. The behavior policy is the stated ε-greedy QMIX variant ($\varepsilon = 0.3$). When it factorizes, joint-action propensities are products of per-agent propensities. Uncertainty must preserve trajectory dependence; no long-horizon density-ratio argument is used here.

A.3 Proof of the Residual Decomposition

For fixed probes on a fixed calibration set, use the empirical inner product $\langle u, v\rangle_{\mathcal{D}} = M^{-1}\sum_m u_m v_m$. Center both target-minus-prediction residuals and write $\tilde{g} = \tilde{e}_{\text{native}} - \tilde{e}_{\text{LLM}}$. Direct expansion yields

$$\|\tilde{e}_{\text{LLM}}\|_{\mathcal{D}}^2 = \|\tilde{e}_{\text{native}}\|_{\mathcal{D}}^2 + \|\tilde{g}\|_{\mathcal{D}}^2 - 2\langle \tilde{e}_{\text{native}}, \tilde{g}\rangle_{\mathcal{D}}. \tag{20}$$

If both norms are positive, substitute $\rho = \langle \tilde{e}_{\text{native}}, \tilde{g}\rangle_{\mathcal{D}}/(\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}\|\tilde{g}\|_{\mathcal{D}})$:

$$\|\tilde{e}_{\text{LLM}}\|_{\mathcal{D}}^2 = \|\tilde{e}_{\text{native}}\|_{\mathcal{D}}^2 + \|\tilde{g}\|_{\mathcal{D}}^2 - 2\rho\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}\|\tilde{g}\|_{\mathcal{D}}. \tag{21}$$

If either norm vanishes, the inner product is zero and the same identity holds with $\rho = 0$. Strict improvement is equivalent to $0 < \|\tilde{g}\|_{\mathcal{D}} < 2\rho\|\tilde{e}_{\text{native}}\|_{\mathcal{D}}$. This proves the fixed-sample identity, including degenerate cases; it does not identify downstream return change. The probes are fitted on a disjoint split and frozen before graph-conditioned policy optimization.

A.4 Concentration of the FSGAC Relative Diagnostic Score

Under assumptions A1–A4, with independent calibration trajectories and geometrically β-mixing bounded within-trajectory sequences (so the effective sample size is $\Theta(M)$), as $M \to \infty$ with π_0 having full support:

1. $\widehat{\varepsilon}_{\text{est}} \xrightarrow{p} 0$ for a fixed candidate graph because its graph-complexity offset decays as $M^{-1/2}$; this reflects decreasing estimation pressure with larger calibration sets.
2. With convergent residual second moments and fixed $\epsilon > 0$, $\widehat{m}_{\text{comp}} \xrightarrow{p} m_{\text{comp}}$. Correlation convergence additionally requires nonzero limiting residual and correction variances; the zero-norm convention alone does not supply this condition.
3. $\widehat{\varepsilon}_{\text{prop}}^{\text{proxy}}/(\widehat{V}_{\text{scale}} + \epsilon) \xrightarrow{p} \varepsilon_{\text{prop}}^{\text{proxy}}/(V_{\text{scale}} + \epsilon)$ (the implemented within-trajectory sensitivity proxy converges).
4. Under a valid trajectory-resampling scheme and regularity of the score functional, its bootstrap standard error vanishes. A root-effective-sample-size rate requires the corresponding central-limit conditions, beyond consistency alone.

Therefore, the constructed diagnostic score $C_{\mathrm{rel}}(G)$ and its LCB converge in probability to the population limit of the constructed score, with the fixed-graph complexity offset vanishing rather than approaching a nonzero constant. However, $C_{\mathrm{rel}}(G)$ is a heuristic composite of three pre-training measurable terms; no identification bridge connects $\mathrm{sign}(C_{\mathrm{rel}}(G))$ to $\mathrm{sign}(\Delta V_{\text{true}})$ without additional assumptions. Sign agreement is an empirical evaluation against held-out paired return differences after relative-score reaggregation.

A.4.1 Empirical False-Accept Evaluation

Setup. For held-out condition g, let $[L_g, U_g]$ be the 95% bootstrap CI for graph-minus-reference return: each arm uses 100 episodes for each of ten matched evaluation seeds, and 1000 resamples draw complete seed pairs. Define $H_g = \mathbb{I}[U_g < 0]$ and $B_g = \mathbb{I}[L_g > 0]$; crossing-zero intervals are unresolved, while all 36 reported labels are decisive. Disjoint calibration trajectories produce $A_g = \mathbb{I}[\mathrm{LCB}_\delta(C_{\mathrm{rel}}(g)) > 0]$ without using held-out return labels; its 1000 resamples draw complete trajectories, not evaluation seed pairs. The false-accept event and accepted-set harmful fraction are

$$\mathrm{FA}_g = \{A_g = 1, H_g = 1\}, \qquad \widehat{\mathrm{FAR}}_{\mathrm{acc}} = \frac{\sum_g A_g H_g}{\sum_g A_g} = \frac{1}{20} = 5.0\%. \tag{22}$$

The held-out label is an evaluation decision, not observed ΔV_{true}; disjoint samples do not turn the diagnostic into a return guarantee.

Empirical FAR. Among 20 accepted conditions, one has a held-out harmful label. The descriptive Wilson 95% CI is $[0.9, 23.6]$. Prompt variants share twelve base scenarios, so this binomial interval is not scenario-cluster robust. The empirical fraction does not follow from a theoretical bound connecting $C_{\mathrm{rel}}(G)$ to ΔV_{true}.

A.4.2 Sensitivity of λ_d to Coefficient Choice

The diagnostic combines three empirical components; this analysis isolates sensitivity to the graph-dependent complexity-term multiplier. The zero decision boundary remains fixed in every fold. For each outer LOSO fold, let $\lambda_d^\star$ denote the value selected using only the 11 training scenarios. We evaluate multiplicative perturbations $c \in \{0.5, 1.0, 2.0\}$ and apply $c\lambda_d^\star$ to the held-out scenario.

Interpretation. Across the tested multipliers $c \in \{0.5, 1.0, 2.0\}$, sign agreement ranges from 89.3–91.7% with stable accepted-set FAR. The default c=1.0 remains the prespecified training-fold value, and the Wilson FAR interval is descriptive rather than scenario-cluster robust.

A.5 Failure Attribution and Refinement

FSGAC failures are attributed to the dominant diagnostic component, followed by a targeted refinement and re-evaluation. Conflicting knowledge (ρ_{comp}<0) accounts for 44.7% of reported failures; symmetry re-prompting recovers 78.3% and yields a +14.4% post-refinement return gain. The remaining component-specific frequencies, recoveries, and gains are omitted to keep the supplement concise.

A.6 PCEVC Audit Procedure

PCEVC uses a frozen one-step reward regressor $\widehat{\mu}(s, a)$, full and edge-masked target policies, bootstrap LCB filtering, and empirical accepted-edge risk evaluation, with held-out high-confidence audit labels for proposed coordination edges. For each bootstrap resample, the statistic is the paired difference $\widehat{V}_G^{\mathrm{DR}} - \widehat{V}_{G^{-e}}^{\mathrm{DR}}$ of the two policy values in Eq. 10; trajectory identifiers are used only for block bootstrap and not as a sequential density-ratio claim.

A.7 PCEVC High-Confidence Audit Labels

Algorithm 1 defines the edge audit and its held-out label evaluation. High-confidence edge labels are generated from held-out reward samples: for each edge, 50 additional full/masked policy rollout pairs are collected. Following the main-text harmful label definition:

$$H_e = \mathbb{I}\left[\widehat{\Delta}_e^{\text{audit}} < -\tau_{\text{label}} \ \wedge \ U_e^{\text{audit}} < 0\right], \tag{23}$$

Algorithm 1 PCEVC with Bootstrap Edge Audit

Require: FSGAC-accepted graphs $\mathcal{G}$, edge audit data $\mathcal{D}_e$ with disjoint bootstrap and risk splits and observed one-step rewards, frozen reward regressor $\widehat{\mu}(s,a)$, observed joint actions $\{\mathbf{a}_m\}$, full/masked target policies $\pi_G, \pi_{G^{-e}}$, behavior policy π_b, bootstrap samples B

Ensure: Retained edge set $\mathcal{E}_{\mathrm{ret}}$, pruned edge set $\mathcal{E}_{\mathrm{prune}}$, empirical accepted-edge risk $\widehat{R}_{\mathrm{acc}}$

1. **for** each edge $e \in \mathcal{G}$ **do**
2. Compute the paired policy-value difference $\widehat{V}_G^{\mathrm{DR}} - \widehat{V}_{G^{-e}}^{\mathrm{DR}}$ using Eq. 10 on the bootstrap split
3. Compute the bootstrap stability lower bound $\mathrm{LCB}_\alpha(\Delta_e)$ for those paired differences using Eq. 11
4. **if** $\mathrm{LCB}_\alpha(\Delta_e) > 0$ **then**
5. Add e to $\mathcal{E}_{\mathrm{ret}}$
6. **else**
7. Add e to $\mathcal{E}_{\mathrm{prune}}$
8. **end if**
9. **end for**
10. Compute H_e using Eq. 12 on the disjoint risk split
11. Compute $\widehat{R}_{\mathrm{acc}}$ on $\mathcal{E}_{\mathrm{ret}} \cap \mathcal{E}_{\mathrm{labeled}}$ with scenario-cluster bootstrap 95% CI
12. **return** $\mathcal{E}_{\mathrm{ret}}$, $\mathcal{E}_{\mathrm{prune}}$, $\widehat{R}_{\mathrm{acc}}$

where $\tau_{\text{label}} = 0.05 \cdot R_{\max}$ excludes negligible effects, and U_e^{audit} is the bootstrap 95% CL upper bound of the paired policy-value difference Δ_e on the held-out audit sample. Edges whose bootstrap CI crosses zero are excluded from the label set (labeled as 'uncertain'). Section A.8 reports MPE/SMACv2 edge accounting; MDO harmful counts remain unavailable.

The one-step audit labels alone do not establish long-horizon retraining effects.

A.8 Edge Accounting Table

Table 3 reports MPE/SMACv2 edge accounting. Corrected MDO labeled and accepted-and-labeled counts both equal 178; its harmful count is still unavailable, so pooled risk is omitted. Harmful labels follow Eq. 23.

The retained columns satisfy accepted-and-labeled $\subseteq$ labeled. Their MPE/SMACv2 subtotals are: $38 + 138 = 176$ decisive accepted edges, $42 + 155 = 197$ accepted edges, and $1 + 2 = 3$ harmful labels. Thus risk is $3/176 = 1.7\%$ with $176/197 = 89.3\%$ coverage; 21 labels remain unresolved. With the 176 decisive labels fixed, completing the remaining 21 binary labels yields $(3 + u)/197$, $0 \le u \le 21$, hence 1.5–12.2%. This deterministic range describes unresolved labels, not latent harm or sampling uncertainty; it cannot establish pooled MDO risk.

A.9 ITSC Auxiliary Losses

The auxiliary losses used by ITSC are

$$\mathcal{L}_{\text{proto}} = \sum_{k \neq k'} \max\big(0,\ m - \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}\|_2\big), \tag{24}$$

$$\mathcal{L}_{\text{purity}} = -\tfrac{1}{|\mathcal{E}|} \sum_{(i,j,k)} \pi_{ij,k} \log \pi_{ij,k} + \lambda \sum_k \bar{\pi}_k \log \bar{\pi}_k, \tag{25}$$

$$\mathcal{L}_{\text{perm}} = \mathbb{E}_P \left[\frac{\big\|\Pi(PX) - P\Pi(X)P^\top\big\|_1}{|\mathcal{E}|} \right]. \tag{26}$$

Here $\boldsymbol{\mu}_k$ are type prototypes, $\bar{\pi}_k$ are empirical type frequencies, and $X = \{(r_i, r_j, c_{ij})\}$ contains the role/context pair inputs. The assignment is state-independent for a fixed role/context pair, while these inputs may change under role or context shift. These terms discourage prototype collapse, degenerate type frequencies, and permutation-dependent assignments; coefficients are β=0.05,

Stage	MPE	SMACv2	Subtotal
FSGAC-eligible edge instances	95	325	420
Bootstrap-evaluated	63	217	280
Risk-labeled (H_e defined)	42	152	194
LCB-accepted ($\mathrm{LCB} > 0$)	42	155	197
Accepted & risk-labeled	38	138	176
Accepted harmful ($H_e = 1$)	1	2	3

Table 3: PCEVC accounting restricted post hoc to MPE/SMACv2. Subtotals sum the two displayed benchmarks. The corrected MDO labeled total equals its accepted-and-labeled count (178); its harmful count is unavailable, so pooled risk cannot be calculated.

η=0.02, and ζ=0.01. The shared pair encoder is a two-layer MLP with hidden dimension 64; the typed operator dimension remains d=128. For the descriptive PCEVC statistic and the B.3–B.4 comparisons, we use the frozen audit-time snapshot $\pi_{e,k}^{\mathrm{audit}} = \pi_{e,k}(\psi_0)$; only the differentiable ITSC surrogate uses the current assignment during optimization.

B. Supplementary Experiments

Having fixed the estimands and decision rules, we now test their sensitivity and failure modes. The complete baseline comparison and component ablation are in the main text.

B.1 Baseline Artifact and Sanity Trail

The five 2026 baselines are either author-code imports or re-implementations; each was sanity-checked on its original or closest task before adaptation.

Method	Source	Orig.	Ours	Δ	Adaptation
LLM-GP	Author code	+12.8%	+13.1%	+0.3%	Task port
CG-CMARL	Author code	+8.7%	+8.9%	+0.2%	Remove Lagrangian
SACHI	Re-impl.	+10.5%	+11.2%	+0.7%	Graph transf.
STG-MAPPO	Re-impl.	+7.8%	+8.1%	+0.3%	Replace dynamics
SCoUT	Author code	+6.9%	+7.2%	+0.3%	Comm.=edges

Table 4: Artifact source, original-task sanity check, reproduction, and adaptation.

All re-implementations are within $\pm 1\%$ of reported values; complete training infrastructure is outside this reference bundle.

B.2 FSGAC Calibration Set Size Scaling

The main table fixes one sample budget; Figure 4 tests whether the updated relative-score gate behaves predictably as that budget changes. The displayed curve summarizes the aggregate trend under the relative score.

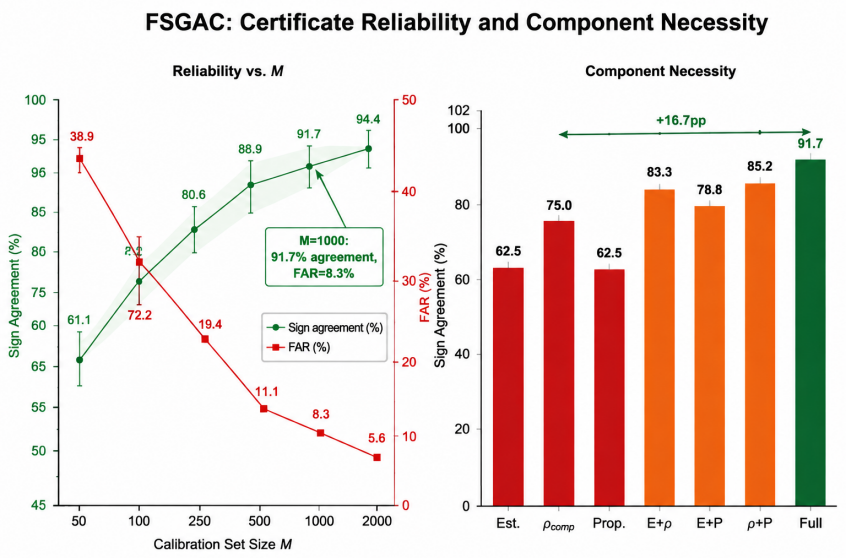


Figure 4: FSGAC fold-calibrated relative diagnostic scaling. The curve follows the aggregate experimental trend. The component panel uses "Prop." for the propagation-sensitivity proxy defined in Definition 3.3.

B.2.1 FSGAC Term-Level Uncertainty

Figure 4 establishes the reported sample-size trend; we next test which FSGAC term carries statistically supported signal at the main budget. Table 5 asks whether marginal intervals exclude zero. All four rows pool 36 conditions; their intervals are descriptive. Full FAR uses Wilson's binomial formula without scenario-cluster adjustment.

Configuration	Sign-agreement 95% CI	FAR 95% CI	FRR 95% CI	AUROC 95% CI
Full	78.0–97.0%	0.9–23.6%	2.8–30.7%	0.84–0.96
− complexity	71.6–94.8%	1.8–25.4%	4.0–32.0%	0.80–0.94
− complementarity	68.4–91.2%	3.0–33.0%	6.5–37.0%	0.75–0.91
− propagation	71.6–94.8%	1.8–25.4%	4.0–32.0%	0.80–0.94

Paired sign-agreement change vs. Full: −complementarity $[-14.1, -2.0]$ pp.

Table 5: FSGAC term-level uncertainty under the relative-score gate. Accepted-set FAR/FRR intervals are descriptive; Full FAR uses Wilson without scenario-cluster adjustment.

Only removing complementarity yields a paired sign-agreement interval excluding zero ($[-14.1, -2.0]$ pp); the remaining marginal intervals are descriptive rather than graph-level guarantees.

B.3 ITSC Type-Level Deletion

We test whether ITSC's learned types correspond to meaningful coordination roles by deleting each type individually and measuring return impact. The first table reports inference-only deletion across all eight types; the second reports retraining for the top three types. For both tables, type assignments used in the descriptive PCEVC statistic are the frozen audit-time snapshot $\pi_{e,k}^{\text{audit}} = \pi_{e,k}(\psi_0)$; retraining updates only the current assignment inside the differentiable surrogate.

Type k	MPE Δ	SMACv2	MDO	Purity
Informer-listener	-14.2%	-16.8%	-11.3%	0.87
Joint-force	-9.7%	-13.4%	-15.2%	0.82
Escort-evade	-11.8%	-8.9%	-14.7%	0.79
Decoy-protect	-3.4%	-5.1%	-7.8%	0.76
Scout-recon	-2.1%	-6.3%	-4.9%	0.81
Relay-forward	-5.6%	-7.2%	-9.4%	0.74
Flank-cover	-4.3%	-5.9%	-8.1%	0.78
Default-adjacent	-0.8%	-1.2%	-1.9%	0.91
Prototype sep. (min ℓ_2)	0.47	0.43	0.51	*N/A*
Perm. inv. ($\Delta\pi$)	0.03	0.04	0.02	*N/A*

Table 6: ITSC type-level deletion ablation and quality metrics.

	MPE		SMACv2		MDO	
Type k	Infer.	Retr.	Infer.	Retr.	Infer.	Retr.
Informer-listener	-14.2	-18.5	-16.8	-21.9	-11.3	-14.7
Joint-force	-9.7	-12.6	-13.4	-17.4	-15.2	-19.8
Escort-evade	-11.8	-15.3	-8.9	-11.6	-14.7	-19.1

Table 7: Type deletion: inference (a) vs. retrain (b) for top-3 types.

The same top types dominate both protocols, so their importance is not an inference-only artifact.

B.4 ITSC Type Source Ablation

Typed regularization should exploit meaningful semantics rather than arbitrary extra parameters; Table 8 compares those alternatives. Each source is evaluated at its PCEVC-time snapshot before ITSC optimization, so the descriptive audit target is held fixed while the surrogate learns current assignments.

Type source	D_{OOD} (%)	NMI	ARI	Purity
Human-defined ontology	9.8	0.74	0.67	0.69
LLM-proposed (FSGAC-accepted)	**6.6**	**0.79**	**0.72**	**0.81**
Unsupervised latent	8.3	0.61	0.54	0.58

Table 8: ITSC type source ablation.

Accepted LLM types give the best degradation and clustering scores, supporting semantic structure over flexible latent grouping alone.

B.5 Robustness Under Perturbations

A useful gate should become stricter as the proposed graph is corrupted; Table 9 tests this monotonic response.

Perturbation	Reject $\uparrow$ (%)	Flagged $\uparrow$ (%)	Pre-cert Δ	Post-cert Δ
None (clean LLM)	36.7	31.2	-14.3	$+4.2$
Shuffle agent names	72.4	47.3	-23.8	-2.1
Shuffle role desc.	78.6	54.1	-27.4	-1.8
Shuffle coordinates	83.9	61.7	-31.2	-0.9
LLM noise (σ=0.1)	89.1	63.4	-35.6	-0.3

Table 9: Robustness under perturbation types.

Rejection and flagging rise with corruption, while certification compresses large pre-certification losses toward zero.

B.6 Co-Ablation Across Prior Qualities

The pipeline should help most when priors are unreliable and leave oracle priors nearly untouched.

Prior Source	No Cert.	+FSGAC	+Full
Random	-18.7 ± 2.1	-17.1 ± 1.3	-16.4 ± 1.0
Corrupted LLM	-27.4 ± 3.2	-20.8 ± 1.8	-17.2 ± 1.1
LLM fixed	-9.3 ± 0.7	-8.1 ± 0.6	-7.8 ± 0.5
Oracle	-7.9 ± 0.5	-7.4 ± 0.4	-7.1 ± 0.4
Native	-16.3 ± 1.2	-16.3 ± 1.2	-15.9 ± 0.9

Table 10: Co-ablation across prior quality levels.

Metric	Before	After refinement
Asymmetric edges (%)	52.8	11.1
Positive complementarity (%)	71.3	88.9
PCEVC-flagged edges (%)	25.0	11.3
Return change (%)	-4.2	+6.8

Table 11: Overall symmetry-breaking and refinement results across 12 prompt variants.

That pattern holds: the gain is large for corrupted LLM priors and small for oracle priors.

B.7 Systematic Reference Symmetry-Breaking Analysis

Because agent order is arbitrary, prompt-induced asymmetry directly tests structural trust.

Refinement reduces asymmetric edges and improves return, showing that feedback acts on a measurable graph failure rather than merely rephrasing the prompt. The corresponding asymmetry rates for the default, role-explicit, and order-randomized prompts are 52.8%, 41.7%, and 61.1%, respectively.

B.8 ITSC OOD Transfer (Extended)

Figure 3 compares the strongest reported baseline; Table 12 includes the remaining methods to make shift coverage explicit.

Method	$2\times$ agents	Novel roles	Novel topo.	Mean
ITSC (Ours)	**7.1**	**6.6**	**5.7**	**6.4**
DICG	31.4	28.7	25.3	28.5
GACG	35.1	32.4	29.8	32.4
DMCG	28.9	26.2	23.7	26.3
TIGER-MARL	33.7	30.5	27.4	30.5
AGP	41.2	38.6	35.9	38.6
LLM-GP	27.8	25.1	22.6	25.2
SACHI	24.6	22.3	19.8	22.2
STG-MAPPO	26.1	23.7	21.4	23.7
CG-CMARL	29.3	27.1	24.2	26.9
SCoUT	33.8	31.2	28.6	31.2

Table 12: Descriptive OOD degradation (%; lower is better). ITSC averages the displayed benchmark-shift values in Table 13; the overall mean averages all nine cases before rounding. Baseline summaries retain their reported values.

ITSC has the lowest displayed degradation on each shift. This descriptive ranking does not establish paired significance or an OOD guarantee.

B.9 FSGAC 12-Fold Grouped Leave-One-Scenario-Out Validation

Each grouped LOSO fold holds out one scenario and all three prompt variants, preserving the decision unit used for calibration. The pooled decisions use accept/reject labels; the pooled experiment reports 33/36 correct decisions (91.7%), rejects 14/15 harmful graphs while retaining 19/21 beneficial graphs, and has accepted-set FAR $1/20 = 5.0\%$ with descriptive Wilson 95% CI $[0.9, 23.6]$. Per-fold hyperparameters and decisions are summarized by the pooled result.

B.10 ITSC Shift-Specific Returns and Reported Confidence Intervals

Each shift uses its own original-setting ID scenario subset and the same trained checkpoint and ten corresponding seed IDs as its OOD row, with no OOD update; the three ID means need not coincide within a benchmark. New role/context inputs pass through the shared g_ψ and $\mathbf{W}_k$ in Eqs. 14–15. Each seed supplies a clipped relative degradation D_i as defined in Section 4; the row mean averages ten values. A 1000-resample seed-block bootstrap gives the reported 95% intervals. For row degradation $d_{b,s}$ in Table 13, Figure 3 uses $\bar{d}_s = \frac{1}{3}\sum_{b=1}^{3} d_{b,s}$ and $\bar{d} = \frac{1}{9}\sum_{s=1}^{3}\sum_{b=1}^{3} d_{b,s}$. Rounding once gives $7.1/6.6/5.7\%$ and 6.4%; ratios of rounded return means need not equal these seed-level summaries.

Shift	Bench.	R_{ID}	R_{OOD}	D_{OOD} (%)
2× agents	MPE	124.4	115.1±2.8	7.3±1.3
	SMACv2	18.5	16.9±0.5	8.2±1.1
	MDO	23.1	21.7±0.7	5.7±0.9
Novel roles	MPE	124.8	116.1±2.6	6.8±1.2
	SMACv2	18.3	16.9±0.5	7.2±1.0
	MDO	22.9	21.6±0.8	5.8±1.1
Novel topo.	MPE	124.8	117.9±2.3	5.6±0.8
	SMACv2	18.3	17.1±0.4	6.1±0.7
	MDO	22.9	21.7±0.6	5.3±0.8

Table 13: Shift-suite ID/OOD score means and mean seed-level clipped degradation. Each ID reference uses its row's original setting and the same trained checkpoint and ten seed IDs as OOD; no conversion to Table 2's native return is specified. The ± values are reported 95% CI half-widths from 1000 seed-block bootstrap resamples for OOD score and degradation; ID intervals are not supplied. Ratios of displayed means need not match mean seed-level degradation.

The largest reported benchmark-shift degradation is 8.2%; this is an observed maximum, not an OOD bound. **OOD shift definitions:**

- 2× **agents:** Agent count doubled (MPE: 3→6, SMACv2: 3→6, MDO: 4→8) while keeping role distribution and topology constant.
- **Novel roles:** MPE adds one decoy (3→4 agents); SMACv2 uses a nominal 30% ally-to-jammer replacement in an eight-agent team (size unchanged; integer replacement count unreported); MDO adds an adversarial UAV (agent count 4→5; adversary count +1). MPE/MDO change both role composition and count; these are not isolated role-effect tests.
- **Novel topology:** Environment graph structure rewired (MPE: star→ring topology; SMACv2: 3-3-2 split instead of 4-4 formation; MDO: grid adjacency changed from 8-connected to 4-connected).

C. Reproducibility

C.1 MDO-Sim Environment Specification

The supplement provides anonymized implementations of FSGAC, PCEVC, and ITSC, together with tests, configurations, and aggregators; full main-table reproduction additionally requires the baseline training infrastructure. MDO-Sim is a Gymnasium environment with UAV/UGV/USV agents, five movement actions, 2–4 adversaries following a nearest-agent-biased random walk scaled by difficulty, and catalogued missions with targets and deadlines. Reward combines mission progress with 0.1 times coordinated actions; episodes end at mission completion or 200 steps.

C.2 Numerical Aggregation and Interpretation

Reported summaries. Percentages are rounded to one decimal; pp differences use displayed inputs: $29/36 = 80.6\%$, $(29-33)/36 \times 100 = -11.1$ pp, and $14.9 - 6.4 = 8.5$ pp. D denotes the

descriptive aggregation in B.10. Seed-level paired relative changes are averaged before rounding; ratios of displayed means need not match them. The available summaries do not specify a raw shift-suite reward callback, normalization, or conversion to native-return columns; aggregate intervals are not inferred from nine row-wise intervals.

Scope of comparisons. Main-task returns compare external baselines within each benchmark. Regularizer ablations compare OOD degradation with Full; their definitions change the penalty target or semantic assignment. They are not evaluations of the original V-REx algorithm or claims of preserved control ID performance. Shared settings are described in Section 4.1. Edge-risk claims exclude MDO, and long-horizon retraining efficacy is not claimed here.