BEYOND GLOBAL SYMMETRY: STATE-CONDITIONED GRAPH-STRUCTURED PARTIAL EQUIVARIANCE FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symmetry provides an effective inductive bias for reinforcement learning by enabling generalization across equivalent states, actions, and entities. However, in structured environments such as multi-agent systems and articulated robots, a fixed global symmetry can become unreliable, because dynamics variations, actuation constraints, and interaction changes may break symmetry selectively across entities and states. To address this, we propose Graph-Structured Partial Equivariance (GSPE), a framework that represents candidate symmetry relations using a state-conditioned symmetry graph. Each edge estimates the reliability of applying an equivariant inductive bias between two entities under the current state. Based on this framework, we introduce Graph-Conditioned Policy Decomposition (GCPD), which decomposes policy computation into graph-conditioned equivariant and residual components, preserving reliable symmetries while enabling adaptive symmetry breaking when relations become unreliable. Experiments on MuJoCo, MPE, and three-robot real-world deployment show that GSPE consistently improves robustness and generalization under structured symmetry violations compared with baselines.

1 INTRODUCTION

Reinforcement learning (RL) enables continuous and multi-agent control (Schulman et al., 2017; Haarnoja et al., 2018; Lowe et al., 2017), but remains vulnerable to distribution shifts (Cobbe et al., 2019). Symmetry supports generalization by sharing across equivalent states, actions, and entities (Bronstein et al., 2021). Yet hard equivariance can be restrictive when that symmetry is locally incorrect (Wang et al., 2022c; 2023).

A weakened rear-right leg can disrupt some relations while leaving others useful (Figure 1). A global gate cannot express this split. Even one score per leg ties together relations that can differ. Moreover, a fixed failure does not imply fixed reliability: contact, load, and compensatory motion change with the state. A static edge mask cannot follow these changes. This motivates two complementary choices: *edge-wise* scores distinguish relations within a state, and *state-wise* scores track each relation as the system evolves. Existing adaptive approaches relax representations, route whole models, score transformed transitions, or gate receiving nodes (Park et al., 2025; Chang et al., 2026; Yu et al., 2024; McClellan et al., 2025). Graph-Structured Partial Equivariance (GSPE) instead estimates each relation's current reliability. Graph-Conditioned Policy Decomposition (GCPD) combines shared equivariant computation with relation-specific residuals gated by these scores.

Our contributions are threefold. First, GSPE formalizes state-dependent relation reliability, and GCPD turns it into selective symmetry breaking with a conditional policy-equivariance bound. Second, MPE navigation (Lowe et al., 2017) and MuJoCo Walker2d/Ant (Todorov et al., 2012; Towers et al., 2025) test control under structured perturbations and frozen transfer across failure strengths and identities; independent diagnostics examine physical compatibility, within-condition variation, and the graph's functional role. Third, a frozen CoopNavi-n=3 policy succeeds on three Unitree Go2 robots under seen-identity, delayed-onset, and unseen-identity mobility attenuation.

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.

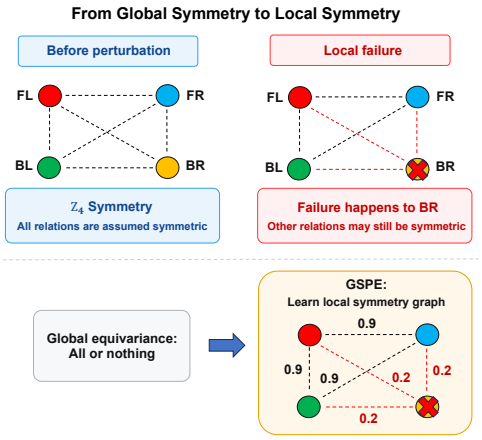


Figure 1: Relation-level motivation. A quadruped can admit candidate Z_4 limb correspondences before perturbation. After a local failure, relations incident to the affected limb are directly disrupted, while other relations may remain reliable or change through coupled dynamics. Global equivariance makes one system-wide decision, whereas our GSPE predicts state-conditioned reliability for each candidate relation.

2 Related Work

Approximate equivariance. Relaxations use residuals (Finzi et al., 2021), relaxed sharing (Wang et al., 2022c), transformation subsets (Romero & Lohit, 2022), or tuned regularizers (Kim et al., 2023). Theory distinguishes correct, incorrect, and extrinsic equivariance (Wang et al., 2023); detection estimates input-wise symmetry (Urbano & Romero, 2024). Control uses augmentation and mirror consistency (Mittal et al., 2024); ApproxEqui-SAC relaxes convolutions with static parameters (Park et al., 2025). These approaches lack state-wise relation scores.

Partial symmetry in MARL. Exact equivariance is encoded in MDP homomorphisms, actor–critics, and cooperative graph policies (van der Pol et al., 2020; Wang et al., 2022a;b; Mondal et al., 2022; van der Pol et al., 2022; McClellan et al., 2024; Chen & Zhang, 2024). PSE bounds symmetry-augmented Q discrepancy by $(\epsilon + \gamma\delta)/(1-\gamma)$ under bounded reward/transition mismatch (Yu et al., 2024). PEnGUiN blends equivariant and free updates with one receiving-node score $\alpha_i(s)$ (McClellan et al., 2025); PE-SAC routes whole actor/critic models with state- or state–action gates (Chang et al., 2026). These resolutions leave heterogeneous pair-wise reliability and a corresponding policy-equivariance guarantee unresolved.

Graph-based RL. Graphs represent pair-wise payoffs (Boehmer et al., 2020), dynamic/contextual coordination (Li et al., 2021; Wang et al., 2022d), agent interactions (Jiang et al., 2020), morphology (Wang et al., 2018), and 3D geometric priors (Chen et al., 2023). Their edges encode interaction or structure, rather than group-compatible estimates of current symmetry-relation reliability.

3 Problem Setup and Partial Equivariance

3.1 Entities, Actions, and Candidate Relations

Consider a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with entity set $V = \{1, \ldots, n\}$, state $s = (c, x_1, \ldots, x_n)$, shared context c, and factored action space $\mathcal{A} = \prod_{i \in V} \mathcal{A}_i$. Define $V_A = \{i \in V : \mathcal{A}_i \text{ is not a singleton}\}$: these are the entities with actual control degrees of freedom. An entity without an actuator has a dummy singleton action space and is excluded from policy-equivariance sums; for example, Ant's torso provides context, while its four legs are action-bearing entities. For a fixed task, an environment condition δ indexes perturbation type, strength, and affected identity, determining P_δ, R_δ. We suppress this index when discussing one MDP. It is an

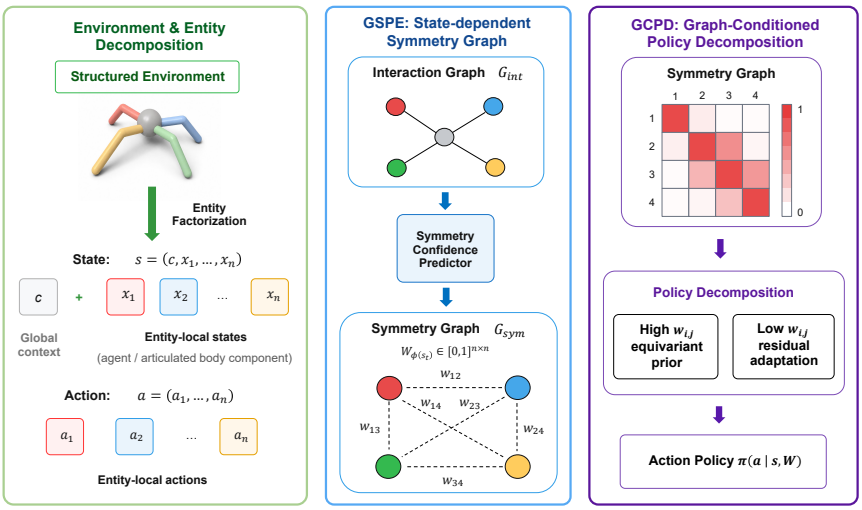


Figure 2: Overview of Graph-Structured Partial Equivariance (GSPE) and Graph-Conditioned Policy Decomposition (GCPD). Given an entity-factored environment, GSPE learns a state-dependent symmetry graph $W_\phi(s_t)$ from transition and reward consistency during training. The learned graph provides the operational reliability of candidate relations to GCPD, which combines shared equivariant computation with symmetry-conditioned residual adaptation. The same symmetry graph is provided as detached relational context to the centralized critic during SAC optimization.

analysis index, not a privileged policy input. Let a finite group G act on states, factored actions, and entity indices through

$$T_g^S : \mathcal{S} \to \mathcal{S}, \quad T_g^A : \mathcal{A} \to \mathcal{A}, \quad P_g \in \{0,1\}^{n \times n}, \tag{1}$$

where $P_g e_i = e_{gi}$, $T_{g,i}^S : \mathcal{X}_i \to \mathcal{X}_{gi}$, $T_{g,i}^A : \mathcal{A}_i \to \mathcal{A}_{gi}$ are the induced local state and action maps. A transformation $g \in G$ is an exact MDP symmetry if

$$R(T_g^S s, T_g^A a) = R(s, a), \tag{2}$$

$$P(T_g^S B \mid T_g^S s, T_g^A a) = P(B \mid s, a) \tag{3}$$

for every measurable $B \subseteq \mathcal{S}$. Under these conditions, an optimal policy can be chosen equivariant (van der Pol et al., 2020; Wang et al., 2022a). Let E_{sym} be a nonempty undirected candidate-relation set closed under the node action, and define

$$E_{\text{sym}}^{\rightarrow} = \{(i,j), (j,i) : \{i,j\} \in E_{\text{sym}}\}.$$

Each undirected relation has one shared confidence $w_{ij} = w_{ji}$. The directed expansion is used only for message passing: throughout, m_{ij} denotes a message sent from j to receiving node i, and $\mathcal{N}_{\text{sym}}(i) = \{j : \{i,j\} \in E_{\text{sym}}\}$. Assume that G preserves V_A, with $i \neq gi \Rightarrow \{i, gi\} \in E_{\text{sym}}$. Write a generic symmetry graph as $W(s) = W(s)^\top = [w_{ij}(s)]$, whose off-diagonal support lies in E_{sym}, with $w_{ii}(s) = 1$ for $i \in V_A$.

3.2 Graph-Structured Partial Equivariance

Definition 1 (Graph-structured partial equivariance)**.** *A symmetry graph is G-admissible if*

$$W(T_g^S s) = P_g W(s) P_g^\top \qquad \forall\, s \in \mathcal{S},\ g \in G. \tag{4}$$

For policy marginal π_i and discrepancy D*, define*

$$d_{i,g}^\pi(s) = \mathsf{D}\big(\pi_{gi}(\cdot \mid T_g^S s), (T_{g,i}^A)_\# \pi_i(\cdot \mid s)\big)\,. \tag{5}$$

Here $\#$ denotes the pushforward of a probability distribution. D *may be KL divergence or, for Gaussian policies, a weighted mean–variance (or mean–log-standard-deviation) distance. For state*

and nonidentity-group sampling distributions d_S and q_G, fix $\epsilon_0 > 0$. Let

$$\mathcal{E}_{\mathrm{GSPE}}(\pi, W) = \mathbb{E}_{\substack{s\sim d_S \\ g\sim q_G}} \left[\frac{\sum_{i\in V_A} w_{i,gi}(s) d^{\pi}_{i,g}(s)}{\sum_{i\in V_A} w_{i,gi}(s) + \epsilon_0} \right]. \tag{6}$$

We call π an ε-GSPE policy relative to W if $\mathcal{E}_{\mathrm{GSPE}}(\pi, W) \le \varepsilon$.

High-confidence relations receive greater weight, whereas low-confidence relations may deviate. This weighted criterion alone admits a trivial zero off-diagonal graph. We therefore train the graph with a separate, detached auxiliary target rather than minimizing this criterion over its weights; the loss and its nonzero-target condition are given below.

4 METHOD

GSPE distinguishes the state-derived physical interaction graph $\mathcal{G}_{\mathrm{int}}(s)$ from the symmetry graph $\mathcal{G}_{\mathrm{sym}}(s) = (V, E_{\mathrm{sym}}, W_\phi(s))$, whose confidences lie in $[0, 1]$. Fixed maps C^x_i, C^a_i express local states and actions in compatible canonical coordinates. GCPD uses relation reliability to gate non-equivariant residuals (Figure 2). Future transitions and rewards supervise the graph only during training; deployment uses the current state.

4.1 LEARNING THE STATE-DEPENDENT SYMMETRY GRAPH

An exact MDP symmetry preserves both dynamics and rewards. We use local transition and reward consistency as an operational proxy for each candidate correspondence. For a replay transition (s_t, a_t, r_t, s_{t+1}), define canonical local variables

$$\xi_{i,t} = C^x_i(x_{i,t}), \qquad \widetilde{a}_{i,t} = C^a_i(a_{i,t}). \tag{7}$$

A shared predictor estimates the canonical local change. Its normalized residual and the resulting transition defect are

$$\begin{aligned} \widehat{\Delta\xi}_{i,t} &= f_\omega(\xi_{i,t}, \widetilde{a}_{i,t}), & \rho_{i,t} &= \left(\Delta\xi_{i,t} - \widehat{\Delta\xi}_{i,t}\right) \oslash b^P, \\ \Delta\xi_{i,t} &= \xi_{i,t+1} - \xi_{i,t}, & \epsilon^P_{ij,t} &= d_P(\rho_{i,t}, \rho_{j,t}), \end{aligned} \tag{8}$$

where $\oslash$ denotes coordinate-wise division, $b^P \in \mathbb{R}^{d_\xi}_{>0}$ is a running coordinate scale, and d_P is a normalized robust distance.

To test reward symmetry without handcrafted local rewards, a scalar predictor f_ν is trained on a canonical transition descriptor $z_t = z(s_t, a_t, s_{t+1})$. Let $\Pi_{ij} z_t$ consistently exchange the current-state, action, and next-state slots of entities i and j. The reward-function defect is

$$\epsilon^R_{ij,t} = \frac{|f_\nu(z_t) - f_\nu(\Pi_{ij} z_t)|}{b^R}, \tag{9}$$

where b^R is a running reward scale. We combine the two defects into the soft relation target

$$\bar{w}_{ij,t} = \exp\left(-\tau_P \epsilon^P_{ij,t} - \beta_{\mathrm{rew}} \tau_R \epsilon^R_{ij,t}\right). \tag{10}$$

The dynamics and reward predictors are trained by one-step regression.

Before sampling a_t, the actor-independent scorer encodes the current-state node features $X_t = \chi(s_t)$ and the state-derived interaction adjacency $M^{\mathrm{int}}_t = \mathrm{Adj}(\mathcal{G}_{\mathrm{int}}(s_t))$:

$$H^W_t, h^W_{\mathcal{G},t} = \mathrm{Enc}^W_\phi\left(X_t, M^{\mathrm{int}}_t\right), \qquad H^W_t = (h^W_{1,t}, \ldots, h^W_{n,t}). \tag{11}$$

For candidate edge $\{i, j\}$, define the symmetric feature

$$p_{ij,t} = \left[h^W_{i,t} + h^W_{j,t}, |h^W_{i,t} - h^W_{j,t}|, |\xi_{i,t} - \xi_{j,t}|, h^W_{\mathcal{G},t}, e_{ij}\right], \tag{12}$$

where $e_{ij} = e_{ji}$ describes the candidate relation. The scorer and its loss are

$$w_{ij,t} = \sigma\left(g_\phi(p_{ij,t})\right), \tag{13}$$

$$\mathcal{L}_{\mathrm{graph}} = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|E_{\mathrm{sym}}|} \sum_{\{i,j\}\in E_{\mathrm{sym}}} \left(w_{ij,t} - \mathrm{sg}[\bar{w}_{ij,t}]\right)^2\right]. \tag{14}$$

Here sg is stop-gradient. Only the graph encoder and scorer parameters ϕ receive gradients from $\mathcal{L}_{\text{graph}}$; the target is detached from both auxiliary predictors.

For fixed targets, let $\mu = \mathbb{E}_{(s,t),e}[\bar{w}_{e,t}]$, with e uniform over nonempty E_{sym}. A constant off-diagonal prediction $c \in (0,1)$ is feasible for the shared sigmoid scorer, and $\mathcal{L}_{\text{graph}}(c) - \mathcal{L}_{\text{graph}}(0) = c^2 - 2c\mu < 0$ whenever $0 < c < \min\{1, 2\mu\}$. Thus, if $\mu > 0$, the all-zero graph is not a loss optimum. This fixed-target comparison does not guarantee convergence or exclude near-zero predictions; detached RL gradients prevent the actor from reducing its regularizer by directly lowering confidences.

For graph equivariance, we require the encoder and interaction mask to be node-equivariant, node features to transform by node permutation in canonical channel coordinates, pooled context to be invariant, and relation descriptors to be tied on group orbits. Consequently $p_{gi,gj}(T_g^S s) = p_{ij}(s)$: the shared scorer satisfies Equation 4. Pair-feature symmetry separately gives $w_{ij} = w_{ji}$; it alone does not establish group equivariance. The diagonal on V_A is fixed to one.

For all RL updates, we use

$$W^{\text{det}}(s) = \text{sg}[W_\phi(s)], \quad w_{ij}^{\text{det}}(s) = \left[W^{\text{det}}(s)\right]_{ij}. \tag{15}$$

Importantly, the learned confidence should not be interpreted as an exact measurement of MDP symmetry. Instead, it is an operational criterion for deciding whether equivariant coupling remains a useful inductive bias for policy computation.

4.2 Graph-Conditioned Policy Decomposition

The actor computes $H, h_{\mathcal{G}} = \text{Enc}_\theta(X, A)$, where A is the interaction adjacency. For the receiver-first channel $(i,j) \in E_{\text{sym}}^{\rightarrow}$, GCPD adds a symmetry-breaking correction to a shared equivariant message:

$$m_{ij} = m_\theta^{\text{eq}}(h_i, h_j, e_{ij}) + \left(1 - w_{ij}^{\text{det}}\right) m_\theta^{\text{res}}(h_i, h_j, r_{ij}). \tag{16}$$

The shared branch uses canonical features and group-orbit-tied e_{ij}, satisfying $m_{gi,gj}^{\text{eq}}(T_g^S s) = T_g^h m_{ij}^{\text{eq}}(s)$, where T_g^h is the induced message-space action. The residual receives ordered features $r_{ij} = [h_i, h_j, h_i - h_j, b_i^{\text{recv}}, b_j^{\text{send}}, b_{ij}^{\text{edge}}]$; the learned role embeddings are not tied across group orbits. "Unrestricted" refers to the absence of an equivariance constraint, not to unbounded outputs. The factor $1 - w_{ij}^{\text{det}}$ gates this correction, while the equivariant branch remains active. Summing incoming messages gives node representations and Gaussian parameters

$$h_i' = U_\theta\left(h_i, \sum_{j \in \mathcal{N}_{\text{sym}}(i)} m_{ij}\right),$$
$$(\widetilde{\mu}_i, \widetilde{\ell}_i) = \text{Dec}_\theta(h_i', h_{\mathcal{G}}), \qquad \widetilde{\ell}_i = \log \widetilde{\sigma}_i. \tag{17}$$

Sampled canonical actions are mapped to environment coordinates with $(C_i^a)^{-1}$. When $w_{ij} = 1$, the corresponding residual is inactive; as w_{ij} decreases, only that relation receives additional free capacity. Let $q_W(s) = ((\widetilde{\mu}_i(s), \widetilde{\ell}_i(s)))_{i=1}^n$ be the Gaussian-parameter map and let T_g^Π be its induced group action (the node permutation in canonical coordinates).

We define the graph-equivariance defect as

$$\varepsilon_W(s,g) := \left\| W(T_g^{\mathcal{S}} s) - P_g W(s) P_g^\top \right\|_F. \tag{18}$$

Proposition 1 (GCPD policy-equivariance bound)**.** *Fix norms on messages and Gaussian parameters (denoted $\|\cdot\|_\Pi$). Assume the candidate set is group-closed; the encoder, shared message branch, aggregation, update, pooling, and decoder are equivariant; $\|m_{ij}^{\text{res}}\| \le B_{\text{res}}$ uniformly on the considered states and transforms; and the downstream map is L_0-Lipschitz with respect to summed message perturbations. Define $\Delta_{ij,g}^{\text{res}}(s) = \|m_{gi,gj}^{\text{res}}(T_g^S s) - T_g^h m_{ij}^{\text{res}}(s)\|$. With $L = L_0$ and $L_W = L_0 B_{\text{res}} \sqrt{|E_{\text{sym}}^{\rightarrow}|}$,*

$$\left\| q_W(T_g^{\mathcal{S}} s) - T_g^\Pi q_W(s) \right\|_\Pi \le L \sum_{(i,j) \in E_{\text{sym}}^{\rightarrow}} (1 - w_{ij}) \Delta_{ij,g}^{\text{res}} + L_W \varepsilon_W(s,g). \tag{19}$$

Consequently, an equivariant scorer satisfies $\varepsilon_W(s, g) = 0$, while high relation confidence suppresses non-equivariant residual computation.

For fixed condition δ, the group-augmented training law defines the Bayes auxiliary-target graph

$$w^\star_{e,\delta}(s) = \mathbb{E}[\bar{w}_{e,t} \mid s_t = s, \delta], \qquad W^\star_\delta(s) = [w^\star_{ij,\delta}(s)]. \tag{20}$$

with the same diagonal convention. Its state marginal is d_δ; this training distribution need not equal a deployed policy's occupancy. The graph predicts auxiliary targets, not return-optimal couplings. Appendix A.2 quantifies the approximation gaps of global gates and static graphs.

Proposition 2 (Frozen-controller return bound)**.** *Fix controller parameters, an environment condition, and the common initial-state distribution. Let $\mathcal{W}_1$ be the 1-Wasserstein distance under the action metric. Assume finite returns and first action moments, $\mathcal{W}_1(\pi_{W_\phi}(\cdot \mid s), \pi_{W^\star_\delta}(\cdot \mid s)) \le L_F \|W_\phi(s) - W^\star_\delta(s)\|_F$, and $Q^{\pi_{W^\star_\delta}}(s, \cdot)$ is L_Q-Lipschitz under the same action metric. For normalized discounted occupancy $d_{\pi_{W_\phi}}$,*

$$|J(\pi_{W_\phi}) - J(\pi_{W^\star_\delta})| \le \frac{L_Q L_F}{1-\gamma} \sqrt{\mathbb{E}_{s\sim d_{\pi_{W_\phi}}} \|W_\phi(s) - W^\star_\delta(s)\|^2_F}. \tag{21}$$

Both bounds are conditional on the stated assumptions; they do not assert that an unconstrained neural network satisfies them globally. Proofs and distribution qualifications are in Appendix A.

4.3 Centralized Graph-Conditioned Critic

The detached symmetry graph conditions two independently parameterized centralized critics. Each encodes canonical node state–action features, weights relational messages by w^{det}_{ij}, and applies mean–maximum pooling. The pooled representation is concatenated with the original state and action vectors in the environment's fixed entity order before an unconstrained MLP readout. This ordered pathway enables value asymmetries beyond invariant pooling: graph features do not impose global invariance on the complete Q-function. Inputs contain no privileged failure information, and RL gradients do not update the graph scorer.

4.4 Learning Objectives

We use standard twin-critic Soft Actor-Critic (Haarnoja et al., 2018), conditioning the actor and critics on detached current-state graphs. The only nonstandard actor term is

$$\mathcal{L}_\pi := \mathcal{L}^{\mathrm{SAC}}_\pi + \lambda_{\mathrm{sym}} \mathcal{L}^{\mathrm{new}}_{\mathrm{sym}}. \tag{22}$$

Let Γ be a generator set and $Z_g(s) = |V_A|$. This fixed-count normalizer preserves the absolute strength of each confidence weight. We use the transformed-state equivariance loss

$$\begin{aligned} \ell^{\mathrm{eq}}_{i,g}(s) = {} & \|\widetilde{\mu}_{gi}(T^S_g s) - \widetilde{\mu}_i(s)\|^2_2 \\ & + \lambda_\sigma \|\widetilde{\ell}_{gi}(T^S_g s) - \widetilde{\ell}_i(s)\|^2_2 \end{aligned} \tag{23}$$

$$\mathcal{L}^{\mathrm{new}}_{\mathrm{sym}} = \mathbb{E}_{\substack{s\sim\mathcal{D} \\ g\sim\mathrm{Unif}(\Gamma)}} \left[\frac{\sum_{i\in V_A} \mathrm{sg}(w_{i,gi}(s)) \ell^{\mathrm{eq}}_{i,g}(s)}{Z_g(s)} \right]. \tag{24}$$

Unlike same-state action agreement, this objective compares a state with its group transform and is therefore an equivariance regularizer. Each policy pass uses its own detached graph. We sample only generators: a Z_4 quarter turn for Ant, a Z_2 bilateral exchange for Walker2d, and a random agent transposition for MPE. Every indexed state, landmark, contact, and action block is transformed consistently.

The SAC target uses only the environment's global reward. Auxiliary modules minimize $\mathcal{L}_{\mathrm{aux}} = \lambda_{\mathrm{graph}}\mathcal{L}_{\mathrm{graph}} + \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dyn}} + \lambda_{\mathrm{rew}}\mathcal{L}_{\mathrm{rew}}$. Stop-gradient targets and W^{det} isolate the scorer, predictors, actor, and critics.

Algorithm 1 gives the complete update sequence; d_t marks termination. The dynamics predictor fits canonical one-step changes with domain-specific robust regression; the reward predictor fits observed global rewards by MSE on original transitions. Permuted descriptors are used only to construct reward-consistency targets. During scorer warm-up, its loss coefficient is zero and confidences remain 0.8; domain-specific ramps and predictor normalization are in Appendix D.4. Appendix C specifies every auxiliary/SAC loss and its gradient ownership.

Algorithm 1 Training GSPE with GCPD

Initialize actor θ, scorer ϕ, predictors (ω, ν), twin critics (ψ_1, ψ_2), target critics, entropy temperature, and replay $\mathcal{D}$; initialize off-diagonal confidences to 0.8
for each environment step **do**
  Predict $W_\phi(s_t)$ from the current state; sample $a_t \sim \pi_\theta(\cdot \mid s_t, \mathrm{sg}[W_\phi(s_t)])$
  Execute a_t; store $(s_t, a_t, r_t, s_{t+1}, d_t)$ in $\mathcal{D}$
  for each scheduled gradient update after replay learning starts **do**
    Sample minibatch $\mathcal{B}$; update ω with $\mathcal{L}_{\mathrm{dyn}}$ and ν with $\mathcal{L}_{\mathrm{rew}}$
    Form transition/reward defects and detached targets $\bar{W}$ (Equations 8–10)
    Update only ϕ with $\mathcal{L}_{\mathrm{graph}}$ using the warm-up/ramp schedule (Appendix D.4)
    Recompute detached graphs for current and next minibatch states
    Update ψ_1, ψ_2 with the soft Bellman loss
    Sample $g \sim \mathrm{Unif}(\Gamma)$; transform all indexed state blocks and compute detached $W_\phi(T_g^S s)$
    Update only θ with $\mathcal{L}_\pi$; update entropy temperature
    Soft-update target critics
  end for
end for

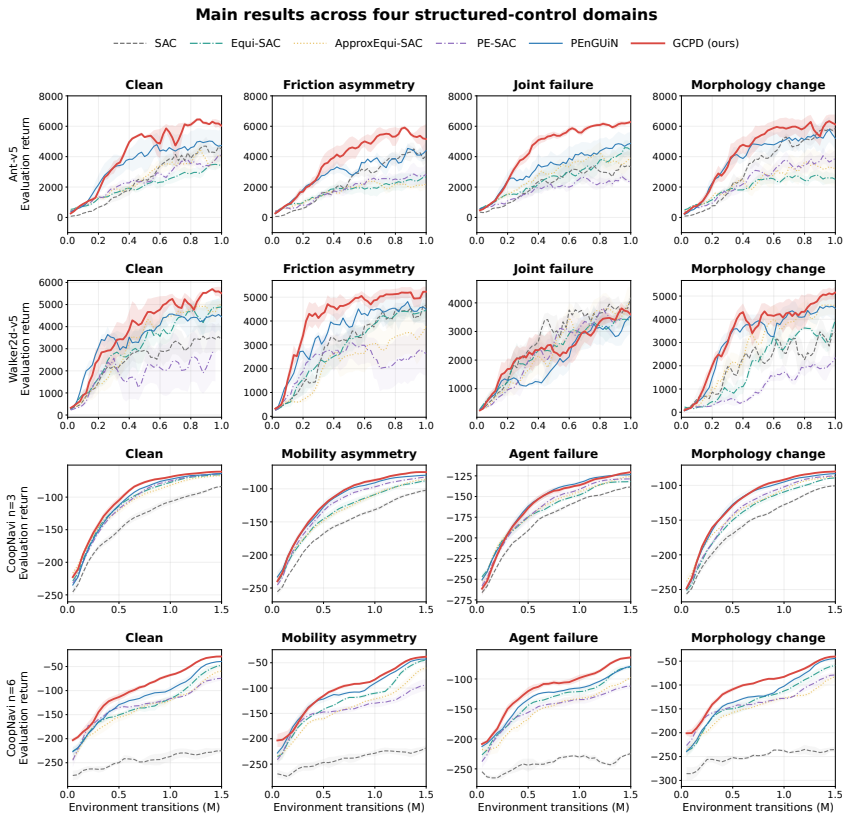


Figure 3: Final return under clean and structured symmetry-breaking conditions. Each method is trained from scratch with the same interaction budget. Full numerical values are reported in the Supplementary Material.

Deployment. At each control step, recompute the graph from the current observation and evaluate the frozen GCPD actor. Future transitions, rewards, predictors, target construction, and critics are not used; only graph values change online, not network parameters.

5 EXPERIMENTS

We evaluate GCPD through four questions: (1) Does it improve control under localized symmetry breaking? (2) Does the symmetry graph vary within a fixed perturbation, and does graph–state matching matter? (3) What do state dependence and relation-wise scoring add over global or static scores? (4) Is GSPE robust to imperfect priors and unseen failure identities?

5.1 EXPERIMENTAL SETUP

Benchmarks and perturbations. We evaluate MuJoCo Ant-v5 and Walker2d-v5 and MPE cooperative navigation with $n \in \{3, 6\}$ agents, covering limb relations, bilateral morphology, and agent permutations at two graph sizes. Each domain includes a nominal (*clean*) condition and three localized perturbations affecting friction or mobility, actuation, and morphology. Here, *clean* denotes the absence of externally injected perturbations rather than exact symmetry; in MuJoCo, contacts, joint limits, and phase-dependent dynamics can still induce local symmetry violations. Within each standard perturbed condition, the affected entity is fixed, but its identity is not provided to the policy as privileged input.

Baselines and evaluation. We compare GCPD with unconstrained SAC (Haarnoja et al., 2018), globally equivariant Equi-SAC, the static relaxation ApproxEqui-SAC (Park et al., 2025), the state-dependent whole-model router PE-SAC (Chang et al., 2026), and PEnGUiN-style Node-Gated SAC. All methods use matched interaction budgets and SAC optimization. Main-comparison return is averaged over the last five evaluations and four seeds (error bars: standard error). Frozen graph diagnostics use gate-independent state regimes, an independent cloned-state response oracle, and condition-mean, shuffled, or delayed graphs. Complete protocols are in the Supplementary Material.

Real-robot setup. We deploy a frozen CoopNavi-n=3 GCPD policy on three Unitree Go2 quadrupeds. Each robot is localized online using onboard range sensing (Wang et al., 2026), and a central host receives their planar positions and velocities, constructs the policy observation, and dispatches motion commands to all three robots. To match the simulated mobility-asymmetry condition, we multiply the attainable planar motion of one robot by $\alpha = 0.55$. The perturbed identity and perturbation onset are not provided to the policy. Hardware, communication, state–action mapping, and safety details are provided in the Supplementary Material.

5.2 RESULTS: ANSWERS TO THE FOUR QUESTIONS

Q1: Does GCPD improve control? Figure 3 shows that GCPD achieves the highest mean final return in 15 of 16 domain–condition combinations. It performs best in all four nominal environments, suggesting that relation-wise adaptation is useful even without injected perturbations. Walker2d joint failure is the only exception, where SAC has a slightly higher but statistically overlapping mean.

Q2: Does the graph track the current state? Figure 4 tests three complementary roles. Graph confidence aligns with a cloned-state control-response oracle independent of the auxiliary predictors (AUROC 0.906; AUPRC 0.275, or $5.0\times$ higher than random prevalence). Under a fixed perturbation, Ant exhibits broad demand-dependent modulation, whereas CoopNavi-n=6 selectively reduces confidence on perturbation-incident relations when the affected agent must move quickly. Across eight frozen Ant checkpoints, condition-mean, state-shuffled, and delayed graphs reduce mean return by about 539, 1586, and 1032–2066, respectively.

Q3: What is gained by state- and relation-wise modeling? Matched variants isolate state dependence and relation granularity: Static-Global and Dynamic-Global use one system-level score, Static-Edge fixes each relation score, and Dynamic-Node shares a state-dependent score across incoming relations. GT-SAC replaces symmetry-conditioned coupling with unconstrained multi-head graph attention and receives no symmetry-reliability supervision. Edge-GNN retains dynamic edge scoring but replaces symmetry-conditioned coupling with generic weighted message passing. A privileged perturbation-location graph marks known-fault incident edges but cannot track state-dependent effects elsewhere. Table 1 shows that the full model reaches the best return. Both state

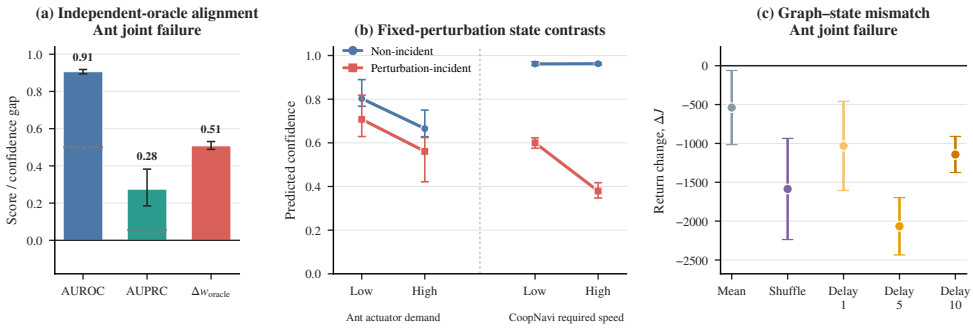


Figure 4: Mechanistic validation of symmetry graph. (a) Agreement with an independent cloned-state response oracle; dashed segments denote random-ranking baselines. (b) Confidence changes across states under a fixed perturbation. (c) Return changes across eight frozen checkpoints under condition-mean, state-shuffled, or delayed graphs. Error bars denote 95% confidence intervals.

Table 1: Ant joint-failure ablation. Values are final return, mean ± standard error over eight seeds.

Variant	Ant joint failure
Static-Global	4684±679
Dynamic-Global	5140±624
Static-Edge	5206±561
Dynamic-Node	4694±445
GT-SAC	4521±470
Edge-GNN	5535±489
Perturbation graph	5316±236
GCPD (ours)	**6165±184**

Table 2: Robustness to false candidate relations on Ant joint failure: final return, mean ± standard error over eight seeds.

Nominal false ratio	Equi-SAC	PEnGUiN	GCPD (ours)
0%	4049±680	4949±880	**6165±184**
12.5%	3537±642	5050±458	**5994±124**
31.25%	2974±308	4640±516	**5969±178**
50%	2161±551	3600±386	**5749±233**

dependence and relation granularity contribute within this controlled family; fault location alone misses contact- and compensation-dependent reliability.

Q4: Is GSPE robust and transferable? We first inject false relations into the candidate symmetry graph while retaining all valid relations. Table 2 shows that GCPD remains robust across all corruption levels. Graph diagnostics show that GCPD assigns lower confidence to false relations, whereas PEnGUiN does not reliably distinguish them from valid ones. We next test zero-shot transfer without fine-tuning. On Ant, GCPD trained only at $\alpha = 0.05$ achieves the highest mean return when evaluated at $\alpha \in \{0.4, 0.7, 1.0\}$. On CoopNavi-$n$=6, GCPD trained only with failed pair $\{A_0, A_1\}$ achieves the highest mean across all identity-overlap groups, although Equi-SAC and PEnGUiN also transfer stably.

Finally, we deploy a frozen CoopNavi-n=3 policy on three Unitree Go2 robots under known-identity, delayed-onset, and unseen-identity mobility perturbations. GCPD succeeds in all 15 trials, while its graph adapts to both perturbation onset and identity (complete protocols in the Supplementary Material).

6 Conclusion

GSPE represents the state-dependent reliability of candidate equivariant relations, and GCPD uses it to combine shared equivariant computation with relation-specific residual capacity. Predefined discrete relations and canonicalization maps preclude directly discovering missing correspondences or continuous symmetries such as $\mathrm{SE}(2)$ or $\mathrm{SE}(3)$. Its short-horizon transition–reward signal is a local proxy for long-horizon MDP symmetry. Future work includes continuous and learned symmetries, multi-step probes, and sparse distributed inference.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related literature, and assist in proving mathematical claims. OpenAI Codex was used to adapt the supplied LaTeX manuscript to the ICLR 2027 template, assist with compilation and layout checks, and condense the introduction and discussion, and restore the local-failure motivation. It also assisted in clarifying definitions, assumptions and the training pipeline, explaining the fixed-target non-collapse argument, and reorganizing the supplementary material. The authors are responsible for the final text, references, mathematical arguments, and other content of this work.

REPRODUCIBILITY STATEMENT

Appendix A gives the theoretical assumptions and proofs. The appendices also provide auxiliary targets and optimization details, architecture and domain instantiations, experimental protocols, hyperparameters, and additional results, including the real-robot deployment.

REFERENCES

Wendelin Boehmer, Vitaly Kurin, and Shimon Whiteson. Deep Coordination Graphs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 980–991. PMLR, 2020. URL https://proceedings.mlr.press/v119/boehmer20a.html.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges, 2021. URL https://arxiv.org/abs/2104.13478.

Junwoo Chang, Minwoo Park, Joohwan Seo, Roberto Horowitz, Jongmin Lee, and Jongeun Choi. Partially Equivariant Reinforcement Learning in Symmetry-Breaking Environments. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dRDcVyobhH.

Dingyang Chen and Qi Zhang. $\mathrm{E}(3)$-Equivariant Actor-Critic Methods for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7640–7666. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24az.html.

Runfa Chen, Jiaqi Han, Fuchun Sun, and Wenbing Huang. Subequivariant Graph Reinforcement Learning in 3D Environments. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 4545–4565. PMLR, 2023. URL https://proceedings.mlr.press/v202/chen23i.html.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying Generalization in Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 2019.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual Pathway Priors for Soft Equivariance Constraints. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30037–30049. Curran Associates, Inc., 2021.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018.

Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkxdQkSYDB.

Hyunsu Kim, Hyungi Lee, Hongseok Yang, and Juho Lee. Regularizing Towards Soft Equivariance Under Mixed Symmetries. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16712–16727. PMLR, 2023. URL https://proceedings.mlr.press/v202/kim23p.html.

Sheng Li, Jayesh K. Gupta, Peter Morales, Ross E. Allen, and Mykel J. Kochenderfer. Deep Implicit Coordination Graphs for Multi-Agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 764–772. International Foundation for Autonomous Agents and Multiagent Systems, 2021. URL https://www.ifaamas.org/Proceedings/aamas2021/pdfs/p764.pdf.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6379–6390. Curran Associates, Inc., 2017.

Joshua McClellan, Naveed Haghani, John Winder, Furong Huang, and Pratap Tokekar. Boosting Sample Efficiency and Generalization in Multi-Agent Reinforcement Learning via Equivariance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 41132–41156, 2024. doi: 10.52202/079017-1301. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4830a9b95a2f63fc4b3fe09abc18f045-Abstract-Conference.html.

Joshua McClellan, Greyson Brothers, Furong Huang, and Pratap Tokekar. PEnGUiN: Partially Equivariant Graph NeUral Networks for Sample Efficient MARL. *Reinforcement Learning Journal*, 6:2637–2651, 2025. URL https://rlj.cs.umass.edu/2025/papers/Paper310.html.

Mayank Mittal, Nikita Rudin, Victor Klemm, Arthur Allshire, and Marco Hutter. Symmetry Considerations for Learning Task Symmetric Robot Policies. In *2024 IEEE International Conference on Robotics and Automation*, pp. 7433–7439. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611493.

Arnab Kumar Mondal, Vineet Jain, Kaleem Siddiqi, and Siamak Ravanbakhsh. EqR: Equivariant Representations for Data-Efficient Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15908–15926. PMLR, 2022.

Jung Yeon Park, Sujay Bhatt, Sihan Zeng, Lawson L. S. Wong, Alec Koppel, Sumitra Ganesh, and Robin Walters. Approximate Equivariance in Reinforcement Learning. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4177–4185. PMLR, 2025.

David W. Romero and Suhas Lohit. Learning Partial Equivariances from Data. In *Advances in Neural Information Processing Systems*, volume 35, pp. 36466–36478. Curran Associates, Inc., 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A Physics Engine for Model-Based Control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A Standardized Interface for Reinforcement Learning Environments. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Alonso Urbano and David W. Romero. Self-Supervised Detection of Perfect and Partial Input-Dependent Symmetries. In *Proceedings of the Geometry-Grounded Representation Learning and Generative Modeling Workshop*, volume 251 of *Proceedings of Machine Learning Research*, pp. 113–131. PMLR, 2024. URL https://proceedings.mlr.press/v251/urbano24a.html.

Elise van der Pol, Daniel E. Worrall, Herke van Hoof, Frans A. Oliehoek, and Max Welling. MDP Homomorphic Networks: Group Symmetries in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4199–4210. Curran Associates, Inc., 2020.

Elise van der Pol, Herke van Hoof, Frans A. Oliehoek, and Max Welling. Multi-Agent MDP Homomorphic Networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=H7HDG--DJF0.

Dian Wang, Robin Walters, and Robert Platt. $\mathrm{SO}(2)$-Equivariant Reinforcement Learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=7F9cOhdvfk_.

Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant Q Learning in Spatial Action Spaces. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1713–1723. PMLR, 2022b.

Dian Wang, Xupeng Zhu, Jung Yeon Park, Mingxi Jia, Guanang Su, Robert Platt, and Robin Walters. A General Theory of Correct, Incorrect, and Extrinsic Equivariance. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40006–40029. Curran Associates, Inc., 2023.

Liansheng Wang, Xinke Zhang, Chenhui Li, Dongjiao He, Yihan Pan, and Jianjun Yi. Super-lio: A robust and efficient lidar-inertial odometry system with a compact mapping strategy. *IEEE Robotics and Automation Letters*, 11(3):2666–2673, 2026.

Rui Wang, Robin Walters, and Rose Yu. Approximately Equivariant Networks for Imperfectly Symmetric Dynamics. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23078–23091. PMLR, 2022c. URL https://proceedings.mlr.press/v162/wang22aa.html.

Tingwu Wang, Renjie Liao, Jimmy Ba, and Sanja Fidler. NerveNet: Learning Structured Policy with Graph Neural Networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1sqHMZCb.

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*, 2022d. URL https://openreview.net/forum?id=wQfgfb8VKTn.

Xin Yu, Rongye Shi, Pu Feng, Yongkai Tian, Simin Li, Shuhao Liao, and Wenjun Wu. Leveraging Partial Symmetry for Multi-Agent Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17583–17590, 2024. doi: 10.1609/aaai.v38i16.29709. URL https://ojs.aaai.org/index.php/AAAI/article/view/29709.

Supplementary material. Appendix A gives assumptions and proofs; Appendix B gives architectures and canonicalization; Appendix C details training and gradient isolation; Appendix D gives experimental protocols; Appendix E reports additional results; Appendix F contains robot deployment and safety details.

A Assumptions, Approximation Results, and Proofs

A.1 Relation-Level Motivation

Figure 1 illustrates the relation-level motivation; the following lemmas analyze the limitations of global and static scores.

A.2 Graph-Estimation Approximation Lemmas

The graph $W^\star_\delta$ uses the same fixed diagonal convention. Lemmas S.1–S.2 show that a global gate incurs $\mathrm{Var}_e[w^\star_{e,\delta}(s)]$, while a condition-specific static graph incurs

$$\frac{1}{|E_{\mathrm{sym}}|}\sum_e \mathrm{Var}_{s\sim d_\delta}\left[w^\star_{e,\delta}(s)\right].$$

Fix an environment condition δ, let $m = |E_{\mathrm{sym}}|$, let V_A denote the action-bearing entities defined in Section 3.1, and define

$$\begin{aligned}\mu_e(s) &= \mathbb{E}[\bar{w}_{e,t} \mid s_t = s, \delta],\\ v_e(s) &= \mathrm{Var}(\bar{w}_{e,t} \mid s_t = s, \delta).\end{aligned} \tag{A.1}$$

At a fixed state, write $\mathbb{E}_s[\cdot] = \mathbb{E}[\cdot \mid s_t = s, \delta]$. The edge-wise and global-gate risks are

$$\begin{aligned}\mathcal{R}_{\mathrm{edge}}(s;\{w_e\}) &= \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}} \mathbb{E}_s[(w_e(s) - \bar{w}_{e,t})^2],\\ \mathcal{R}_{\mathrm{global}}(s;\lambda) &= \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}} \mathbb{E}_s[(\lambda(s) - \bar{w}_{e,t})^2].\end{aligned} \tag{A.2}$$

Lemma S.1 (Global gate's relation-estimation approximation gap)**.** *The best edge predictor is $w^\star_e(s) = \mu_e(s)$, the best global gate is $\lambda^\star(s) = m^{-1}\sum_e w^\star_e(s)$, and*

$$\mathcal{R}^\star_{\mathrm{global}}(s) - \mathcal{R}^\star_{\mathrm{edge}}(s) = \mathrm{Var}_{e\sim\mathrm{Unif}(E_{\mathrm{sym}})}[w^\star_e(s)]. \tag{A.3}$$

Proof. The conditional bias–variance decomposition gives

$$\mathcal{R}_{\mathrm{edge}}(s;\{w_e\}) = \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}}\left[v_e(s) + (w_e(s) - \mu_e(s))^2\right]. \tag{A.4}$$

Thus each edge is minimized by $w^\star_e = \mu_e$. For a global gate, the same decomposition replaces w_e by λ; differentiating the resulting quadratic gives $\lambda^\star = \bar{\mu} = m^{-1}\sum_e \mu_e$. The conditional-variance terms cancel when subtracting the two optimal risks, leaving $m^{-1}\sum_e(\mu_e - \bar{\mu})^2 = \mathrm{Var}_e[w^\star_e]$. □

For a fixed condition, define the state-dependent and static risks

$$\begin{aligned}\mathcal{R}_{\mathrm{state}} &= \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}} \mathbb{E}[(w_e(s_t) - \bar{w}_{e,t})^2 \mid \delta],\\ \mathcal{R}_{\mathrm{static}} &= \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}} \mathbb{E}[(c_e - \bar{w}_{e,t})^2 \mid \delta].\end{aligned} \tag{A.5}$$

Lemma S.2 (Static graph's state-variation approximation gap)**.** *The best state-dependent predictor is $w^\star_e(s) = \mu_e(s)$, the best static edge is $c^\star_e = \mathbb{E}_{s\sim d_\delta}[w^\star_e(s)]$, and*

$$\mathcal{R}^\star_{\mathrm{static}} - \mathcal{R}^\star_{\mathrm{state}} = \frac{1}{m}\sum_{e\in E_{\mathrm{sym}}} \mathrm{Var}_{s\sim d_\delta}[w^\star_{e,\delta}(s)]. \tag{A.6}$$

Proof. Conditioning first on s gives, for each edge,

$$\mathbb{E}[(w_e(s_t) - \bar{w}_{e,t})^2 \mid \delta] = \mathbb{E}_{s\sim d_\delta}[v_e(s) + (w_e(s) - \mu_e(s))^2]. \tag{A.7}$$

The pointwise optimum is $w_e^\star = \mu_e$. For a static c_e, write $\bar{\mu}_e = \mathbb{E}_{s\sim d_\delta}\mu_e(s)$. The law of total variance yields

$$\begin{aligned}\mathbb{E}[(c_e - \bar{w}_{e,t})^2 \mid \delta] = \mathbb{E}_{s\sim d_\delta}[v_e(s)] \\ + \mathrm{Var}_{s\sim d_\delta}[\mu_e(s)] + (c_e - \bar{\mu}_e)^2.\end{aligned} \tag{A.8}$$

Hence $c_e^\star = \bar{\mu}_e$. Subtracting the pointwise optimum and averaging over edges proves Equation A.6. □

A.3 Proof of Proposition 1: GCPD Policy Equivariance

We make explicit the assumptions used in Proposition 1. The candidate edge set is closed under the group action. The state encoder, equivariant message branch, aggregation, node update, pooling, and decoder are G-equivariant. The residual messages are uniformly bounded by a constant $B_{\mathrm{res}} < \infty$, and the composition of aggregation, update, pooling, and decoding is L_0-Lipschitz with respect to the sum of its message perturbations.

For a directed candidate edge $(i, j) \in E^{\rightarrow}_{\mathrm{sym}}$, write

$$w_{ij} = [W(s)]_{ij}, \qquad w'_{ij} = [W(T_g^{\mathcal{S}} s)]_{gi,gj}. \tag{A.9}$$

We also define

$$\begin{aligned}W_g(s) &:= P_g W(s) P_g^\top, \\ \varepsilon_W(s, g) &:= \left\| W(T_g^{\mathcal{S}} s) - W_g(s) \right\|_F.\end{aligned} \tag{A.10}$$

Because $[W_g(s)]_{gi,gj} = w_{ij}$, $\varepsilon_W(s, g)$ measures the failure of the predicted graph to transform equivariantly.

Recall that the GCPD message has the form

$$m_{ij}(s) = m_{ij}^{\mathrm{eq}}(s) + (1 - w_{ij}) m_{ij}^{\mathrm{res}}(s). \tag{A.11}$$

Equivariance of the shared branch gives

$$m_{gi,gj}^{\mathrm{eq}}(T_g^{\mathcal{S}} s) = T_g^h m_{ij}^{\mathrm{eq}}(s). \tag{A.12}$$

Consequently,

$$\begin{aligned}&m_{gi,gj}(T_g^{\mathcal{S}} s) - T_g^h m_{ij}(s) \\ &= (1 - w_{ij}) \Big[m_{gi,gj}^{\mathrm{res}}(T_g^{\mathcal{S}} s) - T_g^h m_{ij}^{\mathrm{res}}(s) \Big] \\ &\quad + (w_{ij} - w'_{ij}) m_{gi,gj}^{\mathrm{res}}(T_g^{\mathcal{S}} s).\end{aligned} \tag{A.13}$$

Define the residual equivariance defect

$$\Delta_{ij,g}^{\mathrm{res}}(s) = \left\| m_{gi,gj}^{\mathrm{res}}(T_g^{\mathcal{S}} s) - T_g^h m_{ij}^{\mathrm{res}}(s) \right\|. \tag{A.14}$$

Using $w_{ij} \in [0, 1]$ and the residual-message bound gives

$$\begin{aligned}&\left\| m_{gi,gj}(T_g^{\mathcal{S}} s) - T_g^h m_{ij}(s) \right\| \\ &\quad \le (1 - w_{ij}) \Delta_{ij,g}^{\mathrm{res}}(s) + B_{\mathrm{res}} |w_{ij} - w'_{ij}|.\end{aligned} \tag{A.15}$$

Because the group action permutes the directed candidate edges, Cauchy–Schwarz yields

$$\begin{aligned}&\sum_{(i,j)\in E^{\rightarrow}_{\mathrm{sym}}} |w_{ij} - w'_{ij}| \\ &\le \sqrt{|E^{\rightarrow}_{\mathrm{sym}}|} \left(\sum_{(i,j)\in E^{\rightarrow}_{\mathrm{sym}}} |w_{ij} - w'_{ij}|^2 \right)^{1/2} \\ &\le \sqrt{|E^{\rightarrow}_{\mathrm{sym}}|}\, \varepsilon_W(s, g).\end{aligned} \tag{A.16}$$

Summing Equation A.15 over directed candidate edges and applying the downstream Lipschitz bound therefore gives

$$\begin{aligned}&\left\|q_W(T_g^{\mathcal{S}}s) - T_g^{\Pi}q_W(s)\right\|_{\Pi}\\ &\quad\le L_0 \sum_{(i,j)\in E_{\mathrm{sym}}^{\rightarrow}} (1-w_{ij})\Delta_{ij,g}^{\mathrm{res}}(s)\\ &\qquad + L_0 B_{\mathrm{res}}\sqrt{|E_{\mathrm{sym}}^{\rightarrow}|}\,\varepsilon_W(s,g).\end{aligned} \tag{A.17}$$

Setting

$$L = L_0, \qquad L_W = L_0 B_{\mathrm{res}}\sqrt{|E_{\mathrm{sym}}^{\rightarrow}|} \tag{A.18}$$

proves Proposition 1. In particular, a graph-equivariant scorer satisfies $W(T_g^{\mathcal{S}}s) = P_g W(s) P_g^{\top}$ and hence $\varepsilon_W(s,g) = 0$. If all relevant relation confidences equal one, the residual term also vanishes and the policy recovers exact equivariance.

A.4 Proof of Proposition 2: Frozen-Controller Return

Fix the controller parameters θ and write

$$\widehat{\pi} = \pi_{\widehat{W}}, \qquad \pi^{\star} = \pi_{W^{\star}}. \tag{A.19}$$

Both policies are evaluated in the same fixed environment condition, have finite returns and first action moments under the chosen action metric, use the same initial-state distribution, and differ only in the graph supplied to the same graph-conditioned controller. Let $d_{\widehat{\pi}}$ denote the normalized discounted state occupancy of $\widehat{\pi}$, and abbreviate

$$\widehat{\pi}_s = \widehat{\pi}(\cdot \mid s), \qquad \pi_s^{\star} = \pi^{\star}(\cdot \mid s). \tag{A.20}$$

The performance-difference identity gives

$$\begin{aligned}&J(\widehat{\pi}) - J(\pi^{\star})\\ &= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\widehat{\pi}}}\left[\mathbb{E}_{a\sim\widehat{\pi}_s}Q^{\pi^{\star}}(s,a) - \mathbb{E}_{a\sim\pi_s^{\star}}Q^{\pi^{\star}}(s,a)\right].\end{aligned} \tag{A.21}$$

By assumption, $Q^{\pi^{\star}}(s,\cdot)$ is L_Q-Lipschitz under the action metric associated with $\mathcal{W}_1$. Kantorovich–Rubinstein duality therefore implies

$$\begin{aligned}&\left|\mathbb{E}_{a\sim\widehat{\pi}_s}Q^{\pi^{\star}}(s,a) - \mathbb{E}_{a\sim\pi_s^{\star}}Q^{\pi^{\star}}(s,a)\right|\\ &\qquad\le L_Q\mathcal{W}_1\left(\widehat{\pi}_s,\pi_s^{\star}\right).\end{aligned} \tag{A.22}$$

The assumed graph sensitivity of the frozen controller gives

$$\mathcal{W}_1\left(\widehat{\pi}_s,\pi_s^{\star}\right) \le L_F\left\|\widehat{W}(s) - W^{\star}(s)\right\|_F. \tag{A.23}$$

Combining Equations A.21–A.23 yields

$$\begin{aligned}&|J(\widehat{\pi}) - J(\pi^{\star})|\\ &\quad\le \frac{L_Q L_F}{1-\gamma}\mathbb{E}_{s\sim d_{\widehat{\pi}}}\left[\left\|\widehat{W}(s) - W^{\star}(s)\right\|_F\right].\end{aligned} \tag{A.24}$$

Finally, Cauchy–Schwarz gives

$$\begin{aligned}&\mathbb{E}_{s\sim d_{\widehat{\pi}}}\left[\left\|\widehat{W}(s) - W^{\star}(s)\right\|_F\right]\\ &\quad\le\sqrt{\mathbb{E}_{s\sim d_{\widehat{\pi}}}\left[\left\|\widehat{W}(s) - W^{\star}(s)\right\|_F^2\right]} = \sqrt{\mathcal{R}_W}.\end{aligned} \tag{A.25}$$

Therefore,

$$\left|J(\pi_{\widehat{W}}) - J(\pi_{W^{\star}})\right| \le \frac{L_Q L_F}{1-\gamma}\sqrt{\mathcal{R}_W}, \tag{A.26}$$

which proves Proposition 2.

This result compares two graphs under the same frozen controller parameters. It is an upper bound on return sensitivity, not a lower bound on regret and not a guarantee that the learned GCPD policy achieves higher return. Moreover, $W^{\star}$ denotes the Bayes-optimal auxiliary-target graph rather than a graph chosen to maximize long-horizon return.

A.5 Implications and Limitations of the Bounds

The two lemmas concern graph prediction, not control. To connect their quantities to Proposition 2, suppose the symmetric adjacency stores each undirected edge twice and let $\bar{w}^\star(s) = m^{-1}\sum_e w_e^\star(s)$. Let $W_\lambda(s)$ place $\lambda(s)$ on every candidate edge and let W_c place c_e on edge e; both matrices are symmetric and zero outside the candidate-edge mask, with the same unit diagonal on V_A. For any evaluation distribution d,

$$\mathbb{E}_d\|W_\lambda - W^\star\|_F^2 = 2m\,\mathbb{E}_d[\mathrm{Var}_e(w_e^\star(s)) + (\lambda(s) - \bar{w}^\star(s))^2], \tag{A.27}$$

$$\mathbb{E}_d\|W_c - W^\star\|_F^2 = 2\sum_{e\in E_{\mathrm{sym}}} [\mathrm{Var}_d(w_e^\star(s)) + (c_e - \mathbb{E}_d w_e^\star(s))^2]. \tag{A.28}$$

Thus Lemmas S.1 and S.2 identify irreducible terms in the graph error supplied to Proposition 2, provided d is the same discounted occupancy as in that proposition and $d_\delta = d$ (or both distribution shifts are controlled). They make the best graph-error certificate available to the restricted class looser. Because Proposition 2 is only an upper bound for fixed controller parameters, this does not lower-bound regret or prove that GCPD's realized return is strictly higher. The graph–state mismatch intervention provides the direct functional evidence that matching W to the current state matters for control. Moreover, the two approximation lemmas characterize the restricted global and static symmetry-graph classes used in our ablations; they are not a taxonomy of prior partial-symmetry methods such as PSE or PEnGUiN.

B Architecture, Canonicalization, and Domain Graphs

B.1 Observation Factorization and Canonicalization

Ant. The official 105-D state is retained without a fault identifier or simulator parameter. Root pose and velocity are broadcast as shared context. Joint positions, joint velocities, and contact-force statistics are gathered into the four leg-local blocks described in Section B.4. The action is split into two torques per leg. The quarter-turn transform rotates planar root quantities and consistently cycles joint, contact/body-indexed, and action blocks. Fixed index permutations and sign maps then place all four leg blocks in the same canonical convention.

Walker2d. The official 17-D state is split into root context and six joint-local position–velocity pairs. Three fixed bilateral maps align right/left hip, knee, and foot semantics. The Z_2 transform applies the same exchange to every joint-, contact-, and action-indexed slot. The policy receives no explicit left-foot failure flag.

MPE. Each environment observation is represented as an ordered stack of agent-local observations. Canonicalization consistently permutes all agent axes, including other-agent slots and the agent axis of agent–landmark tensors; landmark slots follow the stated landmark action, if any. The same permutation is applied to action/history blocks and to current-state, action, and next-state portions of the reward descriptor, preventing an inconsistent partial permutation from creating an artificial reward defect.

B.2 MPE Graph Construction

For cooperative navigation with n agents, each agent is represented as one node,

$$V_{\mathrm{MPE}} = \{v_1, \ldots, v_n\}. \tag{B.1}$$

A node contains the agent-local observation together with any shared or broadcast team context used by the policy. Local actions are represented in a common agent-centric coordinate system. The interaction graph follows the environment's observation or communication neighborhood; if all agents observe one another, this graph is complete.

All exchangeable agent pairs are candidate symmetry relations:

$$E_{\text{sym}}^{\text{MPE}} = \{\{i, j\} : 1 \le i < j \le n\} = E(K_n). \tag{B.2}$$

Thus the three-agent and six-agent tasks contain three and fifteen candidate symmetry confidences, respectively. Canonicalization is the identity after consistently reordering agent-indexed observation and action blocks. The transition predictor is shared across agents, and the reward permutation target exchanges a candidate pair in the full team transition descriptor while leaving global context unchanged. Because K_n is invariant to every agent relabeling, this candidate set is closed under S_n.

The exact local observation partition, interaction-edge weighting, and communication radius used in the final experiments are specified in Section D.1.

B.3 WALKER2D GRAPH CONSTRUCTION

Walker2d is represented by one root node and six articulated-joint nodes:

$$V_{\text{Walker}} = \{v_{\text{root}}, v_{\text{Rhip}}, v_{\text{Rknee}}, v_{\text{Rfoot}}, v_{\text{Lhip}}, v_{\text{Lknee}}, v_{\text{Lfoot}}\}. \tag{B.3}$$

Each articulated node receives the root height, root angle, and root velocity context together with its local joint position and velocity, its normalized kinematic depth, and an endpoint indicator. Every articulated node outputs one torque; the root node is not actuated.

The interaction graph contains self-loops and the fixed kinematic chains

$$\text{root}-\text{Rhip}-\text{Rknee}-\text{Rfoot}, \qquad \text{root}-\text{Lhip}-\text{Lknee}-\text{Lfoot}. \tag{B.4}$$

In the non-static encoder, a weak state-dependent similarity edge is also added when local joint features are strongly aligned. This interaction adjacency is distinct from the symmetry graph.

The candidate symmetry graph contains three bilateral relations,

$$E_{\text{sym}}^{\text{Walker}} = \{\{v_{\text{Rhip}}, v_{\text{Lhip}}\}, \{v_{\text{Rknee}}, v_{\text{Lknee}}\}, \{v_{\text{Rfoot}}, v_{\text{Lfoot}}\}\}. \tag{B.5}$$

The state and action canonicalization maps reorder branch coordinates using fixed index maps (and any convention-specific sign maps) so that corresponding left and right actuators share a common semantic orientation. The three-edge set maps to itself under the Z_2 exchange. Consequently, Walker2d can assign different confidence to the hip, knee, and foot correspondences rather than collapsing the entire leg pair into one gate.

B.4 ANT GRAPH CONSTRUCTION

The Ant implementation contains exactly five articulated-body nodes,

$$V_{\text{Ant}} = \{v_{\text{torso}}, v_{\text{RB}}, v_{\text{FL}}, v_{\text{FR}}, v_{\text{LB}}\}. \tag{B.6}$$

Each leg node represents a complete hip–ankle kinematic branch rather than an individual joint or limb segment. It contains the branch-local hip and ankle positions and velocities, contact information aggregated from the associated body segments, and the broadcast torso/root pose and velocity. Each leg node outputs two actions: hip and ankle torques. The torso node has no local actuator action.

In the current implementation, each node first forms a 24-dimensional semantic feature consisting of root pose, root velocity, two local joint positions, two local joint velocities, mean contact wrench, and three contact-force statistics, followed by five shared type and kinematic metadata features expressed in the node's canonical frame; no absolute leg-identity feature is used. For the torso, unavailable local joint and action entries are filled with zeros.

The interaction graph begins with self-loops and a torso–leg star:

$$A_{ii}^{\text{base}} = 1, \qquad A_{\text{torso},\ell}^{\text{base}} = A_{\ell,\text{torso}}^{\text{base}} = 0.35. \tag{B.7}$$

Here i ranges over all five nodes and ℓ over all four legs; all other entries of A^{base} are zero. Let ξ_ℓ denote the seven-dimensional actuator-aligned vector containing two joint positions, two joint

velocities, and three contact statistics for leg ℓ. We apply the same weak similarity rule to all six distinct, unordered leg pairs. With

$$\kappa_{ij} = \left| \left\langle \frac{\xi_i}{\|\xi_i\|_2 + \varepsilon}, \frac{\xi_j}{\|\xi_j\|_2 + \varepsilon} \right\rangle \right|, \tag{B.8}$$

the additional interaction weight is

$$A^{\mathrm{sim}}_{ij}(s) = 0.10 \ \mathrm{clip}\left(\frac{\kappa_{ij} - 0.85}{0.15}, 0, 1 \right). \tag{B.9}$$

Both message directions receive the same similarity weight. The similarity adjacency is zero on self-loops and on edges incident to the torso. The encoder adjacency is the element-wise maximum

$$A(s) = \max\left\{ A^{\mathrm{base}}, A^{\mathrm{sim}}(s) \right\}. \tag{B.10}$$

Thus, every leg pair uses the same state-dependent construction, without an additional pair-specific weight or fixed leg–leg offset.

The Ant symmetry graph is distinct from this interaction graph. Its vertices are the four articulated-leg nodes, and its candidate edge set is the complete graph

$$\begin{aligned} V_{\mathrm{leg}} &= \{\mathrm{RB}, \mathrm{FL}, \mathrm{FR}, \mathrm{LB}\}, \\ E^{\mathrm{Ant}}_{\mathrm{sym}} &= \binom{V_{\mathrm{leg}}}{2} = E(K_4). \end{aligned} \tag{B.11}$$

It therefore contains six state-dependent symmetry confidences. The torso participates in the actor interaction graph and centralized critic but has no candidate symmetry edge. For transition supervision, ξ_ℓ is the actuator-aligned local vector above. For reward supervision, the descriptor concatenates current and next torso context with all four legs' current canonical features, actions, and next canonical features; each target is obtained by consistently exchanging one candidate leg pair. The complete four-leg graph is closed under the Z_4 rotation.

B.5 Network Architectures

The implementation contains five parameter groups with deliberately separated optimization pathways.

Actor interaction encoder. A shared node encoder maps entity features to hidden states. A message-passing stack operates on $\mathcal{G}_{\mathrm{int}}(s)$, and invariant pooling produces the global feature $h_{\mathcal{G}}$. Canonical node features and a group-closed interaction rule ensure $A(T^S_g s) = P_g A(s) P_g^\top$; shared message maps then give $h_{gi}(T^S_g s) = h_i(s)$ in the canonical channel convention, while $h_{\mathcal{G}}$ is invariant.

Graph scorer. An independent group-compatible encoder Enc^W_ϕ produces h^W_i and invariant $h^W_{\mathcal{G}}$. For every undirected candidate edge, the MLP g_ϕ consumes

$$\left[h^W_i + h^W_j, |h^W_i - h^W_j|, |\xi_i - \xi_j|, h^W_{\mathcal{G}}, e_{ij} \right] \tag{B.12}$$

and outputs one logit followed by a sigmoid. The same scorer is shared across all candidate edges, and e_{ij} is tied on group orbits. Thus $p_{gi,gj}(T^S_g s) = p_{ij}(s)$ and the scorer is graph-equivariant, not merely symmetric in i, j.

Auxiliary predictors. The dynamics predictor f_ω is shared across candidate entities and predicts canonical local increments. The reward predictor f_ν is a scalar MLP operating on a canonicalized global transition descriptor. Neither predictor is needed at deployment.

GCPD actor. Each directed candidate relation is processed by a shared equivariant message branch and an ordered residual branch. The equivariant branch uses group-orbit-tied relation descriptors, whereas the residual branch receives a residual-only ordered role descriptor r_{ij} that is not tied across group orbits. Concretely, $r_{ij} = [h_i, h_j, h_i - h_j, b^{\mathrm{recv}}_i, b^{\mathrm{send}}_j, b^{\mathrm{edge}}_{ij}]$, where the three

b terms are learned ordered role embeddings used only by the residual. With receiver-first indexing, m_{ij} travels from j to i; these embeddings distinguish receiver, sender, and ordered edge. The combined messages are summed over j at receiving node i, after which a shared decoder outputs canonical Gaussian means $\widetilde{\mu}_i$ and log standard deviations $\widetilde{\ell}_i$. Inverse canonicalization maps the sampled local actions back to environment coordinates.

Twin critics. The two critics use independent parameters; within each critic, the node encoder and message maps are shared across entities and relations. For $k \in \{1, 2\}$, let c be invariant global context and let e_{ij} denote the candidate-relation descriptor. Critic k computes

$$\begin{aligned}
u_i^{(k)} &= \phi_{\psi_k}\left([C_i^x(x_i), C_i^a(a_i), c]\right), \\
r_{ij}^{(k)} &= \left[u_i^{(k)}, u_j^{(k)}, |u_i^{(k)} - u_j^{(k)}|, e_{ij}\right], \\
m_{ij}^{Q,(k)} &= w_{ij}^{\mathrm{det}}\, \chi_{\psi_k}(r_{ij}^{(k)}), \\
\widetilde{u}_i^{(k)} &= U_{\psi_k}\left(u_i^{(k)}, \sum_{j \in \mathcal{N}_{\mathrm{sym}}(i)} m_{ij}^{Q,(k)}\right), \qquad \text{(B.13)} \\
z_{\mathrm{pool}}^{(k)}(s, a; W^{\mathrm{det}}) &= \mathrm{pool}_{i \in V}\, \widetilde{u}_i^{(k)}, \\
Q_{\psi_k}(s, a; W^{\mathrm{det}}) &= \rho_{\psi_k}\left(\left[z_{\mathrm{pool}}^{(k)}(s, a; W^{\mathrm{det}}), \mathrm{vec}_{\mathrm{ord}}(s), \mathrm{vec}_{\mathrm{ord}}(a)\right]\right), \qquad \text{(B.14)}
\end{aligned}$$

Here pool concatenates mean and maximum pooling, and ρ_{ψ_k} is an unconstrained MLP readout. $\mathrm{vec}_{\mathrm{ord}}$ flattens the original state or action in the environment's fixed input order, preserving which blocks belong to each entity without sorting or invariant pooling. It uses only the observations and actions already available to the model; no failure identity, failure strength, or other privileged information is added.

The ordered state–action pathway allows the readout to represent value asymmetries that invariant pooling would remove. Thus the graph provides relational features without imposing global invariance on the complete Q-function. Lower edge confidence attenuates relational messages; it does not itself relax the invariance of the pooled graph branch. The graph is detached, so RL gradients do not update the graph scorer. No actor, graph-scorer, or auxiliary-predictor parameters are shared with either critic.

The actor and auxiliary graph encoders use two message-passing layers. Their input projections and message layers are followed by $\tanh$ nonlinearities and residual LayerNorm updates. The GCPD actor uses hidden width 256 on Ant and Walker2d and 128 on both MPE tasks. The centralized graph-conditioned critic uses hidden width 256 on Ant, 128 on Walker2d, and 64 for the residual critic component on MPE. All scorer and auxiliary-predictor MLPs use two hidden layers unless noted otherwise. ReLU is used in the actor heads, scorer, predictors, and critic. Table S2 records the final capacity audit for the Ant comparison, and Section D.4 gives all optimization settings.

B.6 Causality, Replay, and Reset Handling

The target $\bar{w}_{ij,t}$ is constructed only after transition (s_t, a_t, r_t, s_{t+1}) has entered replay. It supervises $W_\phi(s_t)$ during training, but s_{t+1} and r_t are never inputs to the deployed scorer. At test time the actor computes $W_\phi(s_t)$ from the current observation and fixed parameters.

For vectorized MuJoCo training, an environment can reset one component immediately after termination. We store the environment-provided final observation as s_{t+1} and never pair a terminal state with the first observation of the reset episode. True terminal transitions set the Bellman mask to zero. Time-limit truncations retain the final observation and are handled with the same time-limit convention for all methods. Auxiliary transition and reward targets exclude any terminal-to-reset pseudo-transition. The first transition after reset uses a fresh local history; no state from the preceding episode is carried into graph supervision.

B.7 Computational Complexity

Let $n = |V|$, $m_{\text{int}} = |E_{\text{int}}|$, $m_{\text{sym}} = |E_{\text{sym}}|$, and h be hidden width. Each of the two deployed two-layer encoders (the actor interaction encoder and the independent auxiliary graph encoder) costs

$$O\big(nh^2 + m_{\text{int}}h\big) \tag{B.15}$$

with sparse aggregation, while pair scoring and GCPD relation messages cost

$$O\big(m_{\text{sym}}h^2\big)\,. \tag{B.16}$$

The centralized critic has the same edge-linear scaling. In the evaluated domains, $m_{\text{sym}} \in \{3, 6, 15\}$, so the relation module is small compared with the dense actor and critic MLPs. Auxiliary predictors increase training cost but are discarded at deployment. Deployed inference contains the actor encoder, auxiliary graph encoder and scorer, and GCPD actor; the critics, target networks, reward predictor, transition predictor, and target construction are absent.

C Auxiliary Target and Optimization Details

C.1 Transition Defect

Let V_{sym} denote the entities incident to candidate symmetry edges. For each entity, canonical local features and actions are

$$\xi_{i,t} = C_i^x(x_{i,t}), \qquad \widetilde{a}_{i,t} = C_i^a(a_{i,t}). \tag{C.1}$$

A shared one-step predictor estimates the canonical local increment,

$$\widehat{\Delta\xi}_{i,t} = f_\omega(\xi_{i,t}, \widetilde{a}_{i,t}), \qquad \Delta\xi_{i,t} = \xi_{i,t+1} - \xi_{i,t}, \tag{C.2}$$

and is trained with

$$\mathcal{L}_{\text{dyn}} = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|V_{\text{sym}}|}\sum_{i\in V_{\text{sym}}}\left\|\widehat{\Delta\xi}_{i,t} - \Delta\xi_{i,t}\right\|_2^2\right]. \tag{C.3}$$

To account for heterogeneous coordinate magnitudes, for each coordinate q we compute a minibatch statistic

$$\widehat{b}^P_{q,\mathcal{B}} = \max\left(b_{\min}, \frac{1}{|\mathcal{B}||V_{\text{sym}}|}\sum_{(t,i)\in\mathcal{B}\times V_{\text{sym}}}|\Delta\xi_{i,t,q}|\right), \tag{C.4}$$

and optionally maintain an exponential moving average

$$b^P_q \leftarrow \beta_P b^P_q + (1-\beta_P)\widehat{b}^P_{q,\mathcal{B}}. \tag{C.5}$$

The normalized prediction residual is

$$\rho_{i,t} = \left(\Delta\xi_{i,t} - \widehat{\Delta\xi}_{i,t}\right) \oslash b^P. \tag{C.6}$$

We use the coordinate-wise SmoothL1 function

$$\ell_\delta(u) = \begin{cases} \dfrac{u^2}{2\delta}, & |u| \le \delta, \\ |u| - \dfrac{\delta}{2}, & |u| > \delta, \end{cases} \tag{C.7}$$

and define the transition defect

$$\epsilon^P_{ij,t} = \frac{1}{d_\xi}\sum_{q=1}^{d_\xi}\ell_{\delta_P}\left(\rho_{i,t,q} - \rho_{j,t,q}\right). \tag{C.8}$$

The corresponding transition-consistency target is

$$\bar{w}^P_{ij,t} = \exp\left(-\tau_P\epsilon^P_{ij,t}\right). \tag{C.9}$$

Comparing normalized residuals rather than raw state changes removes motion already explained by the shared model and makes the target more sensitive to relation-specific dynamics deviations.

C.2 Reward Permutation Defect

We use a canonicalized transition descriptor of the form

$$z_t = [c_t, c_{t+1}, \xi_{1:n_s,t}, \widetilde{a}_{1:n_s,t}, \xi_{1:n_s,t+1}], \tag{C.10}$$

where $n_s = |V_{\mathrm{sym}}|$. A shared scalar predictor estimates the environment reward,

$$\widehat{r}_t = f_\nu(z_t), \tag{C.11}$$

and is trained by

$$\mathcal{L}_{\mathrm{rew}} = \mathbb{E}_{\mathcal{D}}\left[(f_\nu(z_t) - r_t)^2\right]. \tag{C.12}$$

For candidate relation $\{i, j\}$, the operator Π_{ij} consistently exchanges the current-state, action, and next-state blocks of entities i and j while leaving the shared context unchanged. We compute the reward scale

$$\widehat{b}^R_{\mathcal{B}} = \max\left(b^R_{\min}, \frac{1}{|\mathcal{B}|}\sum_{t\in\mathcal{B}}|r_t|\right), \tag{C.13}$$

with an optional exponential moving average

$$b^R \leftarrow \beta^{\mathrm{ema}}_R b^R + (1 - \beta^{\mathrm{ema}}_R)\,\widehat{b}^R_{\mathcal{B}}. \tag{C.14}$$

The reward permutation defect is

$$\epsilon^R_{ij,t} = \frac{|f_\nu(z_t) - f_\nu(\Pi_{ij} z_t)|}{b^R}, \tag{C.15}$$

and the reward-consistency target is

$$\bar{w}^R_{ij,t} = \exp\left(-\tau_R \epsilon^R_{ij,t}\right). \tag{C.16}$$

The final target combines both signals:

$$\bar{w}_{ij,t} = \bar{w}^P_{ij,t}\left(\bar{w}^R_{ij,t}\right)^{\beta_{\mathrm{rew}}} = \exp\left(-\tau_P \epsilon^P_{ij,t} - \beta_{\mathrm{rew}}\tau_R \epsilon^R_{ij,t}\right). \tag{C.17}$$

Setting $\beta_{\mathrm{rew}} = 0$ recovers transition-only supervision. The target evaluates invariance of the learned global reward function; it does not require entity-local reward decomposition.

C.3 Generator-Based Policy Equivariance Loss

Let Γ generate the candidate group. We set $w_{ii}(s) = 1$ for $i \in V_A$ and require every moved correspondence $\{i, gi\}$ to be a candidate edge. Let $T^S_{g,i} : \mathcal{X}_i \to \mathcal{X}_{gi}$ and $T^A_{g,i} : \mathcal{A}_i \to \mathcal{A}_{gi}$ denote the induced local maps. The canonical maps satisfy

$$C^x_{gi} \circ T^S_{g,i} = C^x_i, \qquad C^a_{gi} \circ T^A_{g,i} = C^a_i, \tag{C.18}$$

so local features are comparable and policy equivariance is direct equality of canonical Gaussian parameters. The induced local action maps are signed permutations and commute with the actor's elementwise $\tanh$ squashing. Writing $\widetilde{\ell}_i = \log\widetilde{\sigma}_i$, we optimize

$$\begin{aligned}\ell^{\mathrm{eq}}_{i,g}(s) = {} & \|\widetilde{\mu}_{gi}(T^S_g s) - \widetilde{\mu}_i(s)\|_2^2 \\ & + \lambda_\sigma \|\widetilde{\ell}_{gi}(T^S_g s) - \widetilde{\ell}_i(s)\|_2^2,\end{aligned} \tag{C.19}$$

$$\mathcal{L}^{\mathrm{new}}_{\mathrm{sym}} = \mathbb{E}_{\substack{s\sim\mathcal{D} \\ g\sim\mathrm{Unif}(\Gamma)}}\left[\frac{\sum_{i\in V_A}\mathrm{sg}(w_{i,gi}(s))\,\ell^{\mathrm{eq}}_{i,g}(s)}{Z_g(s)}\right], \tag{C.20}$$

$$Z_g(s) = |V_A|. \tag{C.21}$$

The two policy passes evaluate s and $T^S_g s$ with their respective detached graphs. The fixed-count normalizer prevents a single transposition's confidence from canceling between numerator and denominator, so low confidence genuinely weakens the regularizer. This fixed diagonal also penalizes changes in the canonical policy parameters of MPE agents fixed by a sampled transposition. The loss therefore updates the actor but not the graph scorer. Zero loss enforces generator equivariance on fixed factors and positively weighted moved correspondences. It implies full-group equivariance only when every policy component is covered and the generator equalities compose; for nonzero errors, composition can accumulate error with group word length.

Domain transformations. For Ant, $\Gamma = \{r\}$ with quarter-turn cycle $r = (\mathrm{RB\ FL\ FR\ LB})$. In the MuJoCo body-frame xy convention, the hip anchors are $\mathrm{RB} = (+,-)$, $\mathrm{FL} = (+,+)$, $\mathrm{FR} = (-,+)$, and $\mathrm{LB} = (-,-)$, which fixes this counterclockwise cycle. The transform rotates any observed planar root position, composes the root orientation with the corresponding yaw quaternion, rotates vector-valued root linear and angular velocities, and cycles all leg, body-contact, and two-action blocks; rotates each contact wrench's force and torque components; and applies the fixed joint-coordinate sign maps before canonicalization. For Walker2d, $\Gamma = \{q\}$, where q exchanges right/left hip, knee, and foot state and action blocks. For MPE, each minibatch samples a random agent transposition; these transpositions generate S_n. The joint-state transformer applies the induced map to every occurrence of an index, not only the focal node: agent–agent and agent–landmark tensors, landmark-associated slots, contact features, action histories, and action-indexed blocks are transformed consistently. Landmark identities remain fixed unless the specified group action also permutes them. In Ant, the single generator directly weights the four cycle relations; the two opposite-leg edges in $E(K_4)$ still gate GCPD residuals but do not appear directly in Equation C.20.

C.4 Gradient Isolation

The graph scorer uses an auxiliary encoder whose parameters are independent of the actor. Let

$$H_t^W, h_{\mathcal{G},t}^W = \mathrm{Enc}_\phi^W(X_t, \mathbf{A}_t) \tag{C.22}$$

and define the symmetric pair feature

$$p_{ij,t} = \left[h_{i,t}^W + h_{j,t}^W, |h_{i,t}^W - h_{j,t}^W|, |\xi_{i,t} - \xi_{j,t}|, h_{\mathcal{G},t}^W, e_{ij}\right]. \tag{C.23}$$

The scorer receives

$$w_{ij,t} = \sigma\left(g_\phi(p_{ij,t})\right). \tag{C.24}$$

The unit diagonal on V_A is inserted deterministically and is not a scorer output. Thus graph supervision updates only the auxiliary graph encoder and scorer parameters ϕ. The graph loss is

$$\mathcal{L}_{\mathrm{graph}} = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|E_{\mathrm{sym}}|}\sum_{\{i,j\}\in E_{\mathrm{sym}}}\left(w_{ij,t} - \mathrm{sg}[\bar{w}_{ij,t}]\right)^2\right], \tag{C.25}$$

so gradients from $\mathcal{L}_{\mathrm{graph}}$ do not update the dynamics or reward predictors through the target. When the encoder is node-equivariant, pooled context is invariant, and $e_{gi,gj} = e_{ij}$ on group orbits, the shared scorer obeys $W_\phi(T_g^S s) = P_g W_\phi(s) P_g^\top$. These conditions are stronger than the symmetry $w_{ij} = w_{ji}$ and are enforced by the group-compatible encoder in Appendix B.5. RL gradients cannot update the scorer through the detached graph. For a fixed nonnegative target distribution with positive mean μ, the feasible constant predictor $c \in (0, \min\{1, 2\mu\})$ has loss smaller than the zero predictor by $2c\mu - c^2$. This comparison holds with auxiliary targets fixed and does not guarantee convergence or rule out near-zero confidence during optimization. If all targets vanish, the nonzero-target argument does not apply.

For reinforcement-learning updates we define

$$W^{\mathrm{det}}(s) = \mathrm{sg}[W_\phi(s)]. \tag{C.26}$$

All actor, critic, Bellman-target, and transformed-state regularizer evaluations use W^{det}; the transformed actor pass uses $W^{\mathrm{det}}(T_g^S s)$. Forward values are unchanged, but RL gradients cannot update ϕ through either pass.

The resulting parameter ownership is summarized below.

Objective	Updated parameters
$\mathcal{L}_{\mathrm{dyn}}$	dynamics predictor ω
$\mathcal{L}_{\mathrm{rew}}$	reward predictor ν
$\mathcal{L}_{\mathrm{graph}}$	graph encoder and scorer ϕ
$\mathcal{L}_\pi$ (including $\mathcal{L}_{\mathrm{sym}}^{\mathrm{new}}$)	actor and shared encoder θ
$\mathcal{L}_Q$	twin critics ψ_1, ψ_2

The target critics and soft Bellman target are evaluated without gradient, as in standard SAC.

C.5 Soft Actor–Critic Objectives

For compactness, write $W_s = W^{\det}(s)$ and $W_{s'} = W^{\det}(s')$. Given a replay transition (s, a, r, s', d) and $a' \sim \pi_\theta(\cdot \mid s', W_{s'})$, the soft Bellman target is

$$y = r + \gamma(1-d)\Big[\min_{k\in\{1,2\}} Q_{\bar{\psi}_k}(s', a'; W_{s'}) - \alpha \log \pi_\theta(a' \mid s', W_{s'})\Big], \tag{C.27}$$

where d is the terminal indicator and $\bar{\psi}_k$ denotes target-critic parameters. The twin critic loss is

$$\mathcal{L}_Q = \sum_{k=1}^{2} \mathbb{E}_{(s,a,r,s',d)\sim\mathcal{D}} \left[\left(Q_{\psi_k}(s, a; W_s) - \mathrm{sg}[y] \right)^2 \right]. \tag{C.28}$$

For $s \sim \mathcal{D}$ and $a \sim \pi_\theta(\cdot \mid s, W_s)$, the standard actor objective is

$$\mathcal{L}_\pi^{\mathrm{SAC}} = \mathbb{E}\left[\alpha \log \pi_\theta(a \mid s, W_s) - \min_{k\in\{1,2\}} Q_{\psi_k}(s, a; W_s) \right]. \tag{C.29}$$

GCPD optimizes

$$\mathcal{L}_\pi = \mathcal{L}_\pi^{\mathrm{SAC}} + \lambda_{\mathrm{sym}} \mathcal{L}_{\mathrm{sym}}^{\mathrm{new}}, \tag{C.30}$$

where the transformed-state term is defined in Equation C.20. The graph supplied to every actor and critic evaluation is detached as specified in Section C.4.

C.6 Training Procedure

Algorithm 1 in the main text gives the training pipeline. Predictor fitting, detached target construction, scorer updates, critic updates, actor/temperature updates, and target-network updates follow that order. The domain-specific update schedules and scorer warm-up are given in Appendix D.4.

At deployment, the predictor and target-construction modules are not required. At each control step, the auxiliary graph encoder and scorer compute $W_\phi(s_t)$ from the current observation, after which the GCPD actor samples the action. Network parameters remain fixed; only the graph values are recomputed online.

D Experimental Details

D.1 Environments and Perturbations

Table S1 specifies the four benchmark domains. A perturbation identity is fixed for an entire training and evaluation condition. Its identity and coefficient are not appended to the policy observation. Therefore a method may infer its effect only from ordinary state variables and their consequences.

Ant-v5. The four leg labels are front-left (FL), front-right (FR), right-back (RB), and left-back (LB). Friction asymmetry creates two low-friction and two high-friction legs. Joint failure attenuates both commands of the fixed RB leg. Morphology change scales the front-left branch while leaving the other three branches unchanged. Episodes use the standard 1000-step Ant-v5 horizon.

Walker2d-v5. The left-foot perturbation is fixed across seeds. The recovery protocol uses a nonzero action multiplier of 0.30 rather than deleting the actuator, which preserves a meaningful but asymmetric control problem. The morphology condition scales the left-leg and left-foot body and geometry offsets and lengths. Episodes use the standard 1000-step Walker2d-v5 horizon.

Cooperative navigation. We use PIC-compatible continuous-action simple-spread environments. The world radius is 1.5, action sensitivity is 5, and the horizon is 50 steps. There are as many landmarks as agents. The centralized reward is the sum of assignment-distance and collision terms. The $n = 3$ condition perturbs one fixed agent; the $n = 6$ condition perturbs two fixed agents so the larger graph contains both incident and non-incident relations. Initial positions and landmarks are

Table S1: Domain and perturbation protocol. Multipliers are relative to the unmodified simulator. "Action" denotes the command applied after the policy output and before simulator integration.

Domain	Observation / action	Friction or mobility	Failure	Morphology
Ant-v5	official 105-D state / 8 torques	FL, LB friction $\times 0.55$; FR, RB $\times 1.15$	both RB torques $\times 0.05$	FL limb planar offsets and length $\times 1.25$
Walker2d-v5	official 17-D state / 6 torques	left-foot friction $\times 0.55$	left-foot actuator $\times 0.30$	left leg and foot geometry $\times 1.25$
CoopNavi-n=3	agent observations / 2-D action per agent	fixed agent 0 action $\times 0.55$	fixed agent 0 action $\times 0.15$	agent 0: size $\times 1.7$, mass $\times 2.0$
CoopNavi-n=6	agent observations / 2-D action per agent	fixed agents $0, 1$ action $\times 0.65$	fixed agents $0, 1$ action $\times 0.15$	agents 0, 1: size $\times 1.7$, mass $\times 2.0$

randomized by the environment seed. Perturbation identity remains fixed during training and the main comparison, and is varied only in the frozen identity-transfer sweep below.

The real-robot study reuses a frozen CoopNavi-n=3 checkpoint and is therefore a deployment case study rather than a fifth training domain. No hardware data are used for policy training or fine-tuning. Section F.1 gives the state–action interface and safety protocol, and Section F.2 reports the deployment results.

D.2 Baselines

SAC. The unconstrained baseline uses a two-hidden-layer Gaussian MLP actor and two centralized MLP critics.

Equi-SAC. Every actor and critic linear map is projected onto the candidate group representation: the four-leg cyclic representation for Ant, bilateral exchange for Walker2d, and agent permutation for MPE. It therefore enforces the supplied symmetry globally.

ApproxEqui-SAC. This baseline uses the same group representations but augments the projected layers with learnable relaxation channels. The relaxation is architectural and static after training; it does not predict a relation-wise graph.

PE-RL-SAC (PE-SAC in figures). This implementation of Partially Equivariant RL contains equivariant and unconstrained branches and predicts one state-level routing scalar. Its gate is trained from disagreement between one-step predictors. All candidate relations receive the same routing decision for a given state.

PEnGUiN-style Node-Gated SAC. This baseline replaces GCPD's relation-wise confidence with one state-dependent coefficient $\alpha_i(s)$ per receiving node. The same coefficient controls every relation entering node i, so the model can adapt locally across entities but cannot assign different validity to relations (i, j) and (i, k) in the same state. It uses the same entity factorization, candidate graph, SAC protocol, and matched interaction budget as GCPD.

Edge-GNN. The controlled dynamic-edge baseline uses the same entity factorization, independent two-layer auxiliary graph encoder, scorer width, auxiliary target construction and losses, and capacity-matching protocol as GCPD. It retains canonicalization for constructing matched auxiliary targets but removes actor-side canonical equivariant coupling, the transformed-state equivariance regularizer, and the equivariant/residual decomposition, using $m_{ij} = w_{ij}^{\mathrm{det}} m_\theta(h_i, h_j)$ instead. It retains the same graph-gradient isolation and tests whether generic dynamic edge weighting is sufficient.

Symmetry-graph ablations. Static-Global learns one scalar $\sigma(b)$; Dynamic-Global predicts one state-dependent scalar $\lambda(s)$; Static-Edge learns one scalar $\sigma(b_{ij})$ for each candidate edge; Dynamic-Node predicts one state-dependent score for each receiving node; and GT-SAC replaces the GCPD branches with unconstrained multi-head graph attention. Edge-GNN retains the graph encoder and

Table S2: Final Ant-v5 capacity audit against GCPD (ours). Counts include the online actor, twin critics, and trainable auxiliary modules.

Method	Main hidden width	Gate / predictor width	Actor+Q_1+Q_2	Auxiliary	Matched total
SAC	782	–	2,112,200	0	2,112,200
Equi-SAC	672	–	2,116,818	0	2,116,818
ApproxEqui-SAC	668	–	2,098,892	0	2,098,892
PE-SAC	448	128 / 256	1,863,655	243,922	2,107,577
PEnGUiN	256	node scorer included	2,110,068	included	2,110,068
GCPD (ours)	256	scorer / predictors included	2,109,808	included	2,109,808

Table S3: Task-specific main-network hidden widths outside the exact Ant audit. Widths are selected separately because symmetry-induced parameter sharing makes equal width different from equal capacity.

Domain	SAC	Equi	Approx.	PE	PEnGUiN	GCPD actor	GCPD critic
Walker2d-v5	512	448	448	320	256	256	128
CoopNavi-$n = 3$	128	128	128	128	128	128	64 residual
CoopNavi-$n = 6$	128	128	128	128	128	128	64 residual

scorer but replaces partial equivariance with unrestricted weighted messages. The perturbation-location graph assigns low confidence to relations incident to the known perturbed entity and high confidence elsewhere. It uses privileged identity but is not the Bayes-optimal, state-dependent validity $w^{\star}_{ij}(s)$.

D.3 Capacity Matching

We match trainable capacity rather than forcing a common hidden width across architectures with different parameter sharing. The target is the clean Ant GCPD (ours) online model, including the trainable scorer and auxiliary modules contained in the actor object and both critics. The target critics are excluded for every method because they are frozen copies of the online critics; the scalar entropy temperature is also excluded. For PE-SAC, its trainable dynamics predictor is included in the count.

The relative deviations from GCPD are $+0.11\%$, $+0.33\%$, -0.52%, -0.11%, and $+0.012\%$ for SAC, Equi-SAC, ApproxEqui-SAC, PE-SAC, and PEnGUiN, respectively. Walker2d and MPE use the same nearest-count rule within each domain, while preserving the domain-specific graph and action heads. Capacity is never matched by changing the interaction budget, replay size, number of gradient updates, or evaluation schedule.

D.4 Hyperparameters

Final GCPD auxiliary configuration. Table S5 reports the implementation used for the experiments, including domain-specific target normalization and optimization schedules.

The auxiliary optimizer is updated once per gradient update after replay learning begins. Consequently, Walker2d performs four auxiliary updates per vector step, corresponding to one update per real transition. During the gate-supervision warm-up, the predictor is trained but the graph-loss coefficient is zero. Because the scorer output layer is zero-initialized with bias $\mathrm{logit}(0.8)$, the actor therefore uses $w_{ij} = 0.8$ on candidate off-diagonal relations during this period. Ant and Walker2d then linearly increase the graph-loss coefficient to 0.05 by 300k and 500k transitions, respectively; MPE switches to 0.05 after 50k transitions.

Ant normalizes each canonical transition coordinate by its current minibatch mean absolute change, lower-bounded by 0.05, and applies SmoothL1 with $\delta_P = 1$. Walker2d instead compares the pre-

Table S4: Common SAC and evaluation hyperparameters. Walker2d counts steps and update schedules in real environment transitions despite using four vector environments.

Hyperparameter	Ant-v5	Walker2d-v5	CoopNavi-$n = 3$	CoopNavi-$n = 6$
Environment transitions	1.0M	1.0M	1.5M	1.5M
Vector environments	1	4	1	1
Gradient updates / vector step	1	4	1	1
Replay capacity	300,000	1,000,000	1,000,000	1,000,000
Learning starts	10,000	10,000	2,000	2,000
Batch size	256	256	256	256
Discount γ	0.99	0.99	0.95	0.95
Target update τ	0.005	0.005	0.01	0.01
Actor learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Critic learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
Actor update period	2	2	2	2
Initial entropy coefficient	0.2	0.2	0.2	0.2
Entropy autotuning	yes	yes	yes	yes
Evaluation interval	20,000	20,000	50,000	50,000
Episodes / evaluation	5	5	50	5

Table S5: Final auxiliary scorer configuration used by GCPD. All step counts denote real environment transitions. "No augmentation" means that transformed transitions are evaluated by the reward predictor only to construct a detached consistency target; they are not added to its supervised training set.

Hyperparameter	Ant-v5	Walker2d-v5	CoopNavi-$n = 3$	CoopNavi-$n = 6$
Initial off-diagonal confidence	0.8	0.8	0.8	0.8
Graph-loss coefficient	0.05	0.05	0.05	0.05
Predictor-loss coefficient	0.05	0.05	0.05	0.05
Policy-symmetry coefficient	0.1	0.1	0.1	0.1
Transition temperature τ_P	2.0	2.0	2.0	2.0
Reward temperature τ_R	1.0	1.0	1.0	1.0
Reward exponent β_{rew}	0.25	0.20	0.25	0.25
Scorer hidden layers	1×256	1×256	2×128	2×128
Transition/reward predictor layers	2×256	2×256	2×128	2×128
Auxiliary learning rate	3×10^{-4}	5×10^{-5}	3×10^{-4}	3×10^{-4}
Gate-supervision warm-up	100k	100k	50k	50k
Full graph-loss coefficient reached at	300k	500k	50k	50k
SmoothL1 threshold δ_P	1.0	not used	1.0	1.0
Transition-scale floor	0.05	0.10	0.02	0.02
Transition-scale EMA	none	none	0.99	0.99
Transition-scale initialization	minibatch statistic	minibatch statistic	0.1	0.1
Reward-scale floor	1.0	1.0	1.0	1.0
Reward-scale EMA	none	none	none	none
Expected predictor outliers	0	0	1	2
Reward-permuted augmentation	no	no	no	no

dicted responses of the original and bilaterally exchanged transitions using a normalized mean absolute discrepancy with scale floor 0.10; no SmoothL1 or running transition scale is used. MPE initializes $b^P = 0.1$ and updates it as

$$b^P \leftarrow 0.99 b^P + 0.01 \max(0.02,\, \mathrm{mean}_{t,i}\, |\Delta\xi_{i,t}|)\,. \tag{D.1}$$

For MPE, the expected-outlier count affects only robust fitting of the shared transition predictor. With per-agent prediction error $e^{\mathrm{pred}}_{i,t}$, let $\mathcal{H}^{(k)}_t$ contain the $n-k$ agents with the smallest errors. The

implemented predictor loss is

$$\mathcal{L}_{\mathrm{dyn}} = \frac{1}{|\mathcal{B}|(n-k)} \sum_{t \in \mathcal{B}} \sum_{i \in \mathcal{H}_t^{(k)}} e_{i,t}^{\mathrm{pred}}, \qquad k = \begin{cases} 1, & n = 3, \\ 2, & n = 6. \end{cases} \tag{D.2}$$

This trimming does not prescribe which relations receive low confidence; all agents are retained when constructing the relation target. Section E.6 reports a controlled sensitivity analysis of this count.

The reward predictors are trained by MSE on the original replay transitions. Pair-permuted descriptors are used only in additional forward passes to construct detached reward-consistency targets. All RL gradients to the scorer and auxiliary predictors are disabled, as described in Section C.4.

D.5 Evaluation Protocol

Every method is trained independently in every domain–condition cell. Main comparisons use four independent seeds and deterministic mean-action evaluation. Within each seed, final return is the average of the final five scheduled evaluations. Tables report the mean and standard error across seeds. Learning curves first aggregate evaluation episodes within a seed and then aggregate seeds at matched environment-transition counts. No training episode is reused as an evaluation episode.

The Ant relation-ablation and imperfect-prior tables instead report the scheduled evaluation at the 1M-transition endpoint, consistently for all methods within each table.

The same interaction budget, update-to-data ratio, replay protocol, termination handling, and evaluation resets are used across methods in a domain. For Walker2d, four vector environments perform 250K vector steps, and each vector step is followed by four gradient updates, so the total interaction budget remains 1M and UTD is approximately one.

Graph diagnostics use frozen final checkpoints and disjoint evaluation transitions. Panel (a) uses 50,000 percentile-bootstrap resamples over four checkpoint-level estimates. Panel (b) uses four independently trained frozen checkpoints in both domains. Regime thresholds are computed separately within each checkpoint; state–edge samples are first averaged within checkpoint, and Table S16 reports mean $\pm$ standard error across checkpoint seeds. Figure 4(b) uses checkpoint-level percentile bootstrap confidence intervals. Panel (c) uses eight independently trained frozen checkpoints with ten reset-matched episodes each. Means and standard errors aggregate checkpoint averages; paired 95% return-change intervals use Student-t with seven degrees of freedom.

D.6 Reproducibility Notes

This subsection summarizes experiment-specific implementation and evaluation details.

- Perturbation identity is fixed within each training condition and omitted from the policy observation; only the frozen identity sweep varies it after training.
- All methods use equal real-transition budgets and matched SAC update schedules within a domain.
- Ant baseline widths are verified by the explicit nearest-count audit in Table S2.
- Vector-environment terminal observations are stored correctly; terminal-to-reset pseudo-transitions are excluded.
- Final return averages the final five evaluations within each seed before computing mean and standard error across seeds.
- Mechanistic analyses use frozen checkpoints and disjoint evaluation transitions; no graph diagnostic updates policy parameters.
- The real-robot policy is frozen and receives no hardware fine-tuning. Task completion and graph diagnostics are reported as separate endpoints, with each hardware episode treated as one unit.
- The independent oracle is calibrated on clean states and does not use either learned auxiliary predictor.

Table S6: Final evaluation return, mean ± standard error over four seeds.

Domain	Condition	SAC	Equi-SAC	ApproxEqui-SAC
Ant-v5	Clean	4379.1±310.0	3449.2±238.2	3789.8±492.5
	Friction asymmetry	4115.6±273.0	2499.0±415.4	2156.3±262.8
	Joint failure	3424.1±918.2	4301.6±456.7	3646.8±609.5
	Morphology change	5764.5±72.3	2580.6±366.6	3213.5±1090.9
Walker2d-v5	Clean	3470.8±612.4	4869.4±387.4	4805.7±345.5
	Friction asymmetry	4374.4±68.6	4365.6±335.1	3457.5±945.8
	Joint failure	**3848.9±207.9**	3408.4±242.5	3716.4±471.6
	Morphology change	3275.1±373.2	3381.6±208.0	4605.7±425.0
CoopNavi-$n = 3$	Clean	−87.8±1.9	−66.5±0.8	−68.2±2.0
	Mobility asymmetry	−105.6±2.4	−90.5±2.7	−88.1±1.5
	Agent failure	−140.7±1.4	−132.2±1.5	−128.1±2.7
	Morphology change	−102.9±0.7	−91.5±1.3	−88.9±1.8
CoopNavi-$n = 6$	Clean	−227.4±7.7	−54.8±6.2	−68.9±8.7
	Mobility asymmetry	−224.5±6.8	−51.4±3.0	−72.5±6.5
	Agent failure	−233.4±2.6	−84.6±4.9	−105.4±3.1
	Morphology change	−239.1±6.1	−66.5±6.4	−94.2±9.2

Domain	Condition	PE-SAC	PEnGUiN	GCPD
Ant-v5	Clean	3815.2±281.6	4897.9±906.3	**6196.6±178.6**
	Friction asymmetry	2780.8±622.1	4220.5±239.1	**5235.6±648.7**
	Joint failure	2555.0±130.0	4837.3±730.8	**6229.6±123.1**
	Morphology change	3865.9±454.6	5424.0±843.0	**6185.7±543.7**
Walker2d-v5	Clean	2434.7±1108.9	4493.8±713.9	**5625.8±137.3**
	Friction asymmetry	2665.7±939.5	4445.7±578.2	**5060.6±301.4**
	Joint failure	3459.2±429.4	3235.6±538.3	3721.7±188.9
	Morphology change	2074.8±367.0	4529.7±461.9	**5087.9±243.4**
CoopNavi-$n = 3$	Clean	−65.1±1.1	−65.0±1.1	**−61.6±0.5**
	Mobility asymmetry	−84.1±2.2	−80.4±1.6	**−75.2±1.2**
	Agent failure	−129.0±2.0	−124.0±1.2	**−122.7±1.0**
	Morphology change	−87.3±1.2	−84.2±1.8	**−80.8±0.5**
CoopNavi-$n = 6$	Clean	−77.9±6.3	−43.1±0.9	**−31.0±1.0**
	Mobility asymmetry	−100.0±11.0	−45.5±1.6	**−40.9±0.8**
	Agent failure	−115.0±4.5	−85.1±1.8	**−67.6±2.0**
	Morphology change	−84.9±6.8	−48.9±2.9	**−44.7±1.6**

- The identity sweep uses six methods, four seeds, all 15 double-failure pairs, 30 episodes per cell, and no fine-tuning (10,800 episodes total).

E Additional Experiments and Ablations

E.1 Full Final-Return Results

Table S6 gives the numerical values plotted in the compact main-paper figure. GCPD ranks first in 15 of the 16 domain–condition cells; Walker2d joint failure is the exception.

E.2 Zero-Shot Failure-Strength Transfer

Policies are trained only with RB action scale $\alpha_{\text{train}} = 0.05$ and evaluated without updates at four scales. Here $\alpha = 1$ restores the nominal actuator and lower α denotes more severe attenuation. Table S7 shows that GCPD gives the strongest mean return at every unseen scale and remains competitive at the training scale.

Table S7: Zero-shot Ant transfer across RB actuator-failure strengths. All policies are trained only at $\alpha = 0.05$. Values are mean ± standard error across four frozen checkpoints; each checkpoint is evaluated for five deterministic episodes at each test strength. The checkpoint seed is the statistical unit.

Method	$\alpha = 1.0$	$\alpha = 0.7$	$\alpha = 0.4$	$\alpha = 0.05$
SAC	476.9±196.2	1593.9±341.2	2570.5±665.5	3150.2±1030.4
Equi-SAC	1448.6±232.3	2233.5±268.9	3293.2±384.2	3953.4±429.9
ApproxEqui-SAC	1101.4±765.1	1952.7±1210.0	1903.1±934.2	4050.3±580.1
PE-SAC	−73.1±42.0	96.5±81.4	735.4±119.8	2700.2±181.3
PEnGUiN	3401.8±1074.5	2867.7±884.8	3472.4±931.9	4491.0±770.2
GCPD	**3555.1±1298.8**	**4708.7±861.3**	**5086.8±281.4**	**6442.8±205.3**

Table S8: Frozen zero-shot transfer across CoopNavi-n=6 double-failure identities. Training exposes only $\{A_0, A_1\}$ at $\alpha = 0.15$. Values are mean ± standard error across four frozen checkpoints. Within each checkpoint, returns are first averaged over pairs in the identity group and over 30 episodes per pair. The checkpoint seed is the statistical unit.

Method	Seen $\{0, 1\}$	One unseen	Two unseen
SAC	−223.21±1.48	−243.25±2.80	−252.47±3.41
Equi-SAC	−78.30±4.50	−78.02±3.54	−76.80±3.50
ApproxEqui-SAC	−92.54±2.48	−94.55±1.70	−97.29±2.50
PE-SAC	−106.83±5.79	−107.64±5.15	−107.83±4.61
PEnGUiN	−75.51±2.16	−74.76±2.74	−73.86±3.22
GCPD (ours)	**−70.12±1.01**	**−69.25±1.10**	**−68.17±1.65**

E.3 Frozen Failure-Identity Transfer

For CoopNavi-n=6, all six methods use four seeds and are trained only with failed pair $\mathcal{F}_{\text{train}} = \{A_0, A_1\}$ at action multiplier $\alpha_{\text{train}} = 0.15$. We freeze every checkpoint and exhaustively evaluate all $\binom{6}{2} = 15$ double-failure pairs at $\alpha_{\text{test}} = 0.15$, using 30 episodes per seed–pair cell and no fine-tuning. "Seen" is exactly $\mathcal{F}_{\text{train}}$; "one unseen" contains one training and one new identity (eight pairs); and "two unseen" contains neither training identity (six pairs). In total, the sweep contains $6{\times}4{\times}15{\times}30 = 10{,}800$ frozen-evaluation episodes.

GCPD has the highest mean return in all three identity-overlap groups. Relative to the seen pair, SAC drops by 20.04 and 29.26 points for one- and two-unseen groups, respectively, which is consistent with fixed-location memorization. Equi-SAC and PEnGUiN remain stable across identities, while GCPD obtains the highest mean return in every identity-overlap group. We therefore conclude that GCPD combines strong absolute performance with identity transfer, rather than claiming that the compared symmetry-aware baselines fail to generalize.

E.4 State Dependence and Relation Granularity

Static-Global, Dynamic-Global, Static-Edge, and GCPD form the controlled 2×2 comparison of state dependence and relation granularity. Dynamic-Node predicts one state-dependent score per receiving node and shares it across all incoming relations. GT-SAC replaces symmetry-conditioned coupling with unconstrained multi-head graph attention and receives no symmetry-validity supervision. Edge-GNN keeps the learned relation scorer but uses generic weighted message passing instead of the equivariant/residual decomposition. The perturbation-location graph uses privileged fault identity but remains a static binary graph. All variants use the same interaction budget and closely matched capacity.

GCPD has the highest mean return, exceeding global, static, node-wise, generic-attention, and privileged-location alternatives. This supports both state dependence and edge-level granularity, while showing that fault location alone does not capture state-dependent relation validity.

Table S9: Ant joint-failure ablation, final return (mean ± standard error over eight seeds).

Variant	Final return
Static-Global	4684±679
Dynamic-Global	5140±624
Static-Edge	5206±561
Dynamic-Node	4694±445
GT-SAC	4521±470
Edge-GNN	5535±489
Perturbation-location graph	5316±236
GCPD	**6165±184**

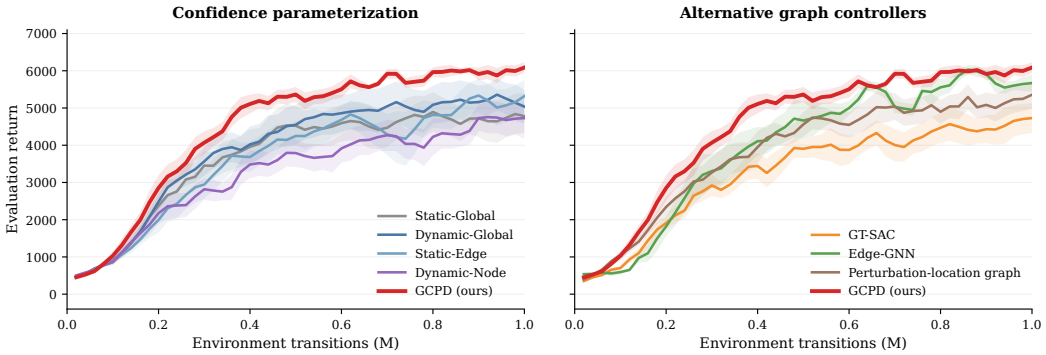


Figure S1: Complete Ant joint-failure learning curves for the ablations reported in Table S9. The left panel varies state dependence and relation granularity; the right panel compares generic graph controllers and the privileged static perturbation-location graph. Lines use a centered three-evaluation moving average, and shading denotes standard error across the checkpoint set used by each comparison.

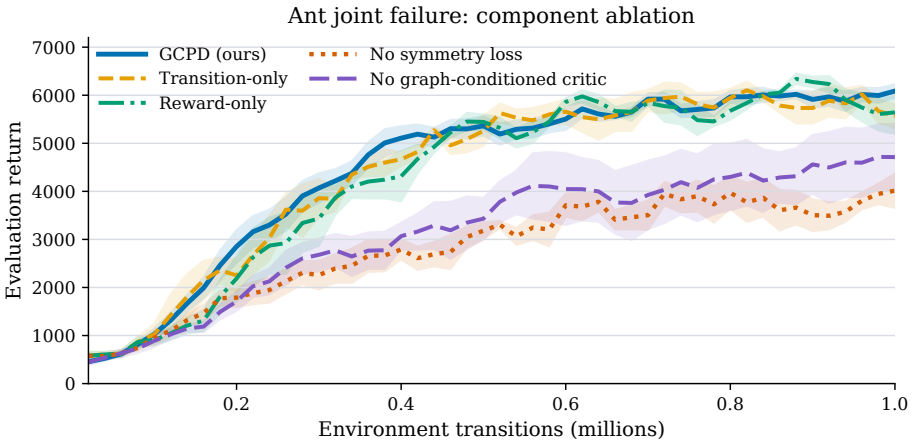


Figure S2: Complete Ant joint-failure learning curves for the component ablation in Table S10. Lines use a centered three-evaluation moving average, and shading denotes standard error across the corresponding checkpoint set.

E.5 Component Ablation

We separately remove the transition or reward term from the graph target, the policy-symmetry loss, and graph context from the centralized critic on Ant joint failure. The remaining actor, critic, scorer, training budget, and SAC optimization protocol are unchanged.

Table S10: Ant joint-failure component ablation over four seeds. Values are final return (mean ± standard error).

Variant	Final return
GCPD (ours)	**6165±184**
Transition-only	5724±282
Reward-only	5799±246
No symmetry loss	3751±205
No graph-conditioned critic	4646±709

Table S11: Sensitivity to predictor trimming. Final return averages the last five evaluations within each seed. Values are mean ± standard error over four seeds; bold rows use the default count for that domain.

Domain	k	Final return	Confidence gap
CoopNavi-n=3	0	−126.63±1.52	0.026±0.003
CoopNavi-n=3	**1**	**−124.62±1.61**	**0.668±0.002**
CoopNavi-n=3	2	−125.68±2.14	−0.165±0.002
CoopNavi-n=6	0	−75.05±2.83	0.034±0.002
CoopNavi-n=6	1	−77.98±3.31	0.034±0.004
CoopNavi-n=6	**2**	**−72.24±1.57**	**0.552±0.010**

The complete model has the highest mean return. Either transition or reward consistency alone retains substantial control performance, indicating partially redundant auxiliary signals rather than making either one individually indispensable. Removing the policy-symmetry loss causes the largest degradation, while removing graph context from the critic also reduces mean return and increases variability.

E.6 Predictor-Trimming Sensitivity in CoopNavi

The expected-outlier count k removes the k largest per-agent errors only when fitting the shared transition predictor; it neither identifies the affected agents nor removes any agent from the relation target. We vary $k \in \{0, 1, 2\}$ under agent failure while keeping the architecture, 1.5M-transition budget, and all other hyperparameters fixed. The confidence gap is the mean confidence on healthy–healthy relations minus that on perturbation-incident relations, so a positive value indicates the desired localization.

Return changes modestly across counts, but relation localization is sensitive to predictor contamination: $k = 1$ for $n = 3$ and $k = 2$ for $n = 6$ produce large positive gaps, whereas misspecified counts do not. Thus k is a robust-fitting inductive bias rather than a supplied fault identity, but its value should be reported and tuned when the expected number of local outliers changes.

E.7 Walker2d Generic-Residual Control

The final Walker2d implementation uses the original/exchanged response discrepancy described in Section D.4. To test whether the results depend on that domain-specific construction, we replace only the transition target with the generic pairwise residual used by the common formulation. A full next-observation predictor gives $\widehat{\Delta s}_t = \hat{s}_{t+1} - s_t$. For bilateral actuator i, define the normalized position–velocity residual

$$u_{i,t} = \left[\frac{\Delta q_{i,t} - \widehat{\Delta q}_{i,t}}{b_q}, \frac{\Delta \dot{q}_{i,t} - \widehat{\Delta \dot{q}}_{i,t}}{b_{\dot{q}}} \right], \tag{E.1}$$

where each scale uses an EMA coefficient 0.99 and floor 0.02. For bilateral pair $\{i, j\}$, the replacement target is

$$\epsilon^{P,\mathrm{gen}}_{ij,t} = \frac{1}{2} \sum_{c=1}^{2} \ell_1(u_{i,t,c} - u_{j,t,c}), \qquad \bar{w}^{P,\mathrm{gen}}_{ij,t} = \exp(-\tau_P \epsilon^{P,\mathrm{gen}}_{ij,t}). \tag{E.2}$$

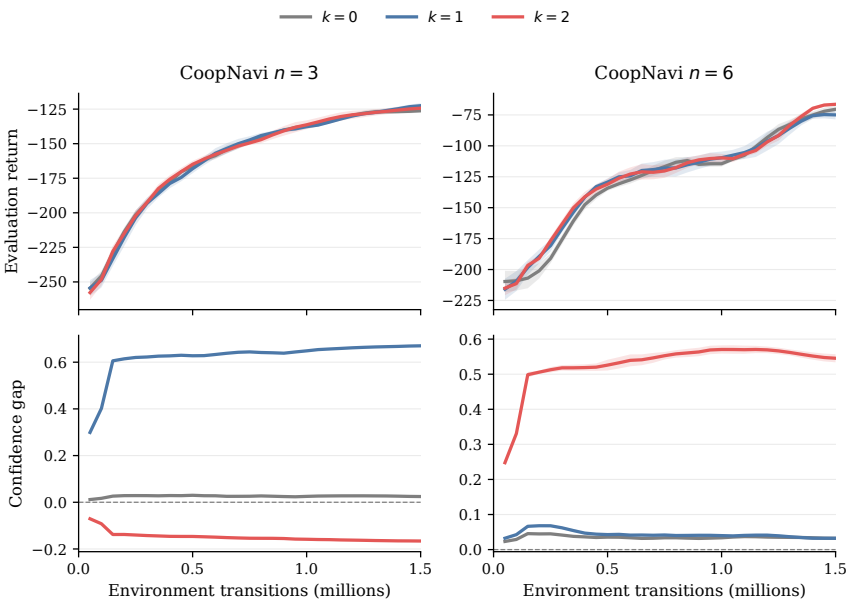


Figure S3: Complete predictor-trimming sensitivity curves under CoopNavi agent failure. The top row reports evaluation return and the bottom row reports the healthy-minus-incident confidence gap. Lines use a centered three-evaluation moving average, and shading denotes standard error over four seeds.

Table S12: Walker2d transition-target control. Values are final return, mean ± standard error over four matched seeds; within each seed we average the last five evaluations. Both rows in Table S12 are obtained from a separate paired rerun and do not reuse the checkpoints reported in Table S6.

Transition target	Clean	Friction	Joint failure	Morphology
Walker response	**5412.9±336.3**	3860.7±166.1	2412.8±843.3	**4448.4±324.0**
Generic residual + SmoothL1	4926.8±482.5	**5122.6±106.1**	**3564.3±763.7**	3083.1±827.1

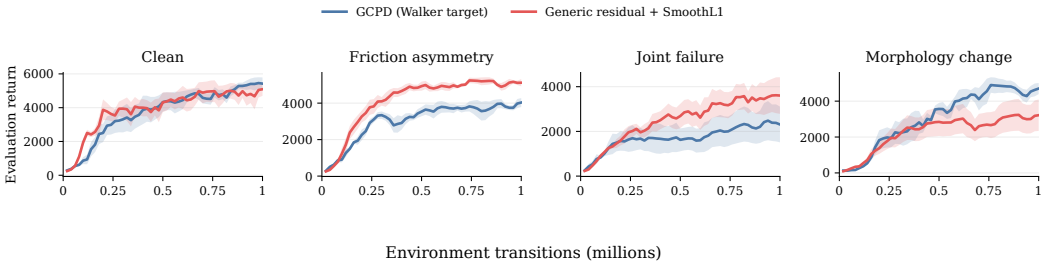


Figure S4: Complete Walker2d learning curves for the transition-target control. The two variants differ only in transition-target construction. Lines use a centered three-evaluation moving average, and shading denotes standard error over four matched seeds.

The scalar reward-permutation target, actor and critic, graph scorer, loss weights, and one-million-transition protocol remain unchanged.

The generic target improves friction asymmetry and joint failure, while the Walker-specific response target is stronger in clean and morphology conditions. Hence the generic GSPE target remains competitive, but the choice of observable physical-consistency probe affects which violation types are localized most effectively.

Table S13: Robustness to imperfect candidate relations on Ant joint failure over eight seeds. The first column reports the mean fraction of the fixed four-relation false torso–leg pool activated across seeds, rather than the fraction of false relations among all active candidates.

Mean activated false-pool fraction	Equi-SAC	PEnGUiN	GCPD
0%	4049±680	4949±880	**6165±184**
12.5%	3537±642	5050±458	**5994±124**
31.25%	2974±308	4640±516	**5969±178**
50%	2161±551	3600±386	**5749±233**

Table S14: Held-out prediction of the learned relation target. Entries are mean ± standard error across four independently trained frozen checkpoints; the checkpoint seed is the statistical unit. N_{tr} is the total number of held-out transitions. ECE uses ten equal-width confidence bins.

Domain	Condition	N_{tr}	MSE	MAE
Ant-v5	Clean	19,900	0.021±0.001	0.065±0.002
Ant-v5	Joint failure	13,725	0.030±0.005	0.101±0.014
CoopNavi-n=6	Clean	3,760	0.00024±0.00008	0.0055±0.0001
CoopNavi-n=6	Agent failure	3,760	0.022±0.002	0.093±0.004

Domain	Condition	R^2	Spearman	ECE
Ant-v5	Clean	0.755±0.009	0.675±0.011	0.010±0.003
Ant-v5	Joint failure	0.708±0.043	0.770±0.027	0.019±0.005
CoopNavi-n=6	Clean	0.635±0.068	0.403±0.025	0.0007±0.0002
CoopNavi-n=6	Agent failure	0.818±0.015	0.793±0.009	0.048±0.008

E.8 Robustness to Imperfect Candidate Relations

Terminology clarification In main-paper Table 2, "Nominal false ratio" is shorthand for the *mean activated false-pool fraction* defined below; it does not denote the fraction of false relations among all active candidate relations.

Detailed Setup Let $\mathcal{F}$ denote the fixed pool of four invalid torso–leg relations. For seed q, let $\mathcal{A}_q(c) \subseteq \mathcal{F}$ be the subset activated at corruption setting c. We report

$$\bar{r}(c) = \frac{1}{S}\sum_{q=1}^{S} \frac{|\mathcal{A}_q(c)|}{|\mathcal{F}|}. \tag{E.3}$$

Each relation is counted once as an undirected relation, although it is instantiated in both message directions. For seed slots $q \bmod 4$, the activated counts are $(0,0,0,0)$, $(1,0,1,0)$, $(2,1,1,1)$, and $(2,2,2,2)$ at the four tested settings. These correspond to mean activated false-pool fractions of 0%, 12.5%, 31.25%, and 50%, respectively. All six valid leg–leg candidate relations remain present.

GCPD has the highest mean return at every tested ratio. Its edge scorer also separates injected relations from valid candidates: at 31.25% and 50% corruption, the mean valid-minus-false confidence gaps are 0.70 and 0.65, compared with only 0.13 and 0.09 for PEnGUiN. The corresponding PEnGUiN false-relation confidence remains near 0.60, whereas GCPD lowers it to 0.12 and 0.18. Thus GCPD learns to suppress false relations, while a node-wise gate cannot reliably isolate them from valid relations incident to the same node.

E.9 Held-Out Confidence Prediction

We first test whether the scorer fits its own auxiliary target on independent transitions. This is a necessary optimization check but is not independent evidence of physical validity.

Table S15: Independent one-step physical alignment underlying main-paper Figure 4(a). Entries are mean ± standard error across four frozen checkpoints. AP lift is computed checkpoint-wise as AUPRC divided by oracle-invalid prevalence.

Domain	Condition	Prev.	AUROC	AUPRC	AP lift	Δw
Ant	Clean	0.051±0.003	0.922±0.012	0.334±0.031	6.70±0.94	0.534±0.049
Ant	Joint failure	0.055±0.011	0.906±0.007	0.275±0.061	4.99±0.53	0.508±0.013
Coop-6	Clean	0.082±0.030	0.608±0.076	0.179±0.050	2.94±1.19	0.008±0.005
Coop-6	Agent failure	0.576±0.003	0.854±0.004	0.822±0.004	1.43±0.01	0.344±0.008

Calibration. For each bin B_k, we plot mean predicted confidence against mean target confidence and compute

$$\mathrm{ECE} = \sum_k \frac{|B_k|}{N} \left| \frac{1}{|B_k|} \sum_{u \in B_k} w_u - \frac{1}{|B_k|} \sum_{u \in B_k} \bar{w}_u \right|. \tag{E.4}$$

The low ECE and positive rank correlations show that the scorer generalizes its supervision beyond replay samples used for fitting.

E.10 Independent Counterfactual Oracle

The independent oracle clones a simulator state, applies small positive and negative control perturbations to one canonical entity at a time, advances one simulator step, and records the resulting canonical control-response signature. Pair defect is the robust normalized distance between two entity signatures. Feature scales and the invalidity threshold are calibrated on clean states only; an edge is oracle-invalid when its defect exceeds the clean 95th percentile. The oracle never calls f_ω, f_ν, or the learned target.

For the main Ant joint-failure case, AUPRC is 5.00 times its prevalence baseline. The confidence separation is

$$\Delta w = \mathbb{E}[w_{ij} \mid \text{oracle-compatible}] - \mathbb{E}[w_{ij} \mid \text{oracle-invalid}]. \tag{E.5}$$

The oracle evaluates the same baseline action from a cloned simulator state. For every local action coordinate k of entity i, we apply symmetric perturbations $a_k^{\pm} = \mathrm{clip}(a_k \pm \epsilon)$ and advance each clone by one simulator step. We use $\epsilon = 0.15$ in Ant and $\epsilon = 0.20$ in CoopNavi. The resulting response signature is

$$z_i(s,a) = \mathrm{concat}_{k \in \mathcal{A}_i} \frac{\psi_i(s_{i,k}^+) - \psi_i(s_{i,k}^-)}{a_{i,k}^+ - a_{i,k}^-}. \tag{E.6}$$

For Ant, $\psi_i = [C_i^x x_i', v_{\mathrm{root}}', r']$; for CoopNavi, $\psi_i = [p_i', v_i', \bar{p}', \bar{v}', r_i', r']$. For response coordinate ℓ, clean calibration data define

$$b_\ell = \max\{Q_{0.75}(|z_{i,\ell} - z_{j,\ell}|)\,, b_{\min}\}\,, \tag{E.7}$$

where $b_{\min}$ is 0.05 times the median positive coordinate scale, lower bounded by 10^{-4}. Pair defect is the quantile-normalized RMS distance

$$D_{ij}(s) = \left[\frac{1}{K} \sum_{\ell=1}^{K} \left(\frac{z_{i,\ell} - z_{j,\ell}}{b_\ell} \right)^2 \right]^{1/2}. \tag{E.8}$$

An edge is oracle-invalid when D_{ij} exceeds the clean calibration 95th percentile. Calibration uses 400 cloned clean states per domain, corresponding to 2,400 Ant and 6,000 CoopNavi state–edge samples. Calibration states are excluded from evaluation. This oracle is used in main-paper Figure 4(a) and never calls either learned auxiliary predictor.

E.11 Fixed-Perturbation State Regimes

All regime labels are computed after rollout and are not gate inputs. Ant states are divided at within-checkpoint medians of fault-leg contact load, joint speed, actuator demand, body angular speed, and a turn/recovery score. Stance versus swing is defined from foot contact. For MPE, Hungarian

Table S16: Fixed-condition regime contrasts underlying main-paper Figure 4(b). Entries are mean ± standard error across four frozen checkpoints; each checkpoint is split at its own regime median before aggregation. The analysis contains 4,158 Ant and 12,000 CoopNavi state–edge samples. The checkpoint seed is the statistical unit.

Domain	Regime	Relation type	Low regime	High regime
Ant joint failure	Actuator demand	Non-incident	0.803 ± 0.029	0.665 ± 0.032
		Perturbation-incident	0.708 ± 0.048	0.561 ± 0.048
CoopNavi-n=6 agent failure	Required speed	Non-incident	0.961 ± 0.005	0.962 ± 0.003
		Perturbation-incident	0.600 ± 0.015	0.379 ± 0.022

assignment defines the restricted agent's target; regimes use required speed, target distance, critical-target ownership, crowding, and action demand. Perturbation-incident edges contain a perturbed entity; non-incident edges connect only unperturbed entities.

Ant exhibits broad state modulation: high demand reduces non-incident and incident confidence by 0.138 and 0.147, respectively. MPE exhibits selective modulation: high required speed reduces incident-edge confidence by 0.221, while non-incident confidence is essentially unchanged.

E.12 Graph–State Mismatch Intervention

We evaluate eight independently trained, frozen Ant joint-failure checkpoints under six inference modes without retraining (Figure 4(c)). For each checkpoint, the condition-mean graph averages $W_\phi(s)$ over five disjoint calibration episodes. Shuffling independently draws a graph from another calibration state in the same condition, preserving its marginal distribution while breaking graph–state alignment. Delay k uses the current graph for the first k steps and then uses $W_\phi(s_{t-k})$.

Returns are measured over ten reset-matched episodes per checkpoint; current, condition-mean, shuffled, and delayed modes use matched resets within each checkpoint. To isolate the graph's immediate effect on the controller from subsequent trajectory divergence, we additionally evaluate every intervention on states from the current-graph trajectory. Only the current-graph action is executed in the environment. For mode m, the same-state action discrepancy is

$$\Delta_a^{(m)} = \mathbb{E}_{\tau \sim d_{\text{current}}} \left[\frac{1}{T_\tau} \sum_{t=0}^{T_\tau - 1} \left\| \mu(s_t; \widetilde{W}_t^{(m)}) - \mu(s_t; W_\phi(s_t)) \right\|_2 \right]. \tag{E.9}$$

For delayed modes, $\widetilde{W}_t^{(m)} = W_\phi(s_{t-k})$ along the same current-graph trajectory. Return and action shift are first averaged within each checkpoint, then summarized as mean ± standard error across the eight checkpoints, with $\text{SEM} = \text{SD}_{\text{checkpoints}}/\sqrt{8}$. For return changes, we first compute each checkpoint's paired difference $\Delta J_k = J_k^{\text{intervention}} - J_k^{\text{current}}$. The two-sided 95% Student-t interval is $\overline{\Delta J} \pm t_{0.975,7}\, \text{SD}(\Delta J_k)/\sqrt{8}$, where $t_{0.975,7} \approx 2.365$; the checkpoint, rather than an individual episode, is the statistical unit.

Table S17: Ant joint-failure graph interventions over eight frozen checkpoints, corresponding to Figure 4(c). Return and same-state action shift are reported as mean ± standard error across eight independently trained checkpoints. Each checkpoint is evaluated on ten reset-matched episodes. For ΔJ, brackets denote a paired 95% Student-t confidence interval across checkpoint-level return differences. Same-state action shift compares counterfactual actions at states from the current-graph trajectory.

Graph	Return	ΔJ vs. current [95% CI]	Same-state action ℓ_2 shift
Current state	6249.0 ± 92.1	$0.0\ [0.0, 0.0]$	0.000 ± 0.000
Condition mean	5710.3 ± 207.5	$-538.7\ [-1016.0, -61.3]$	0.362 ± 0.017
State shuffled	4662.9 ± 250.4	$-1586.1\ [-2236.7, -935.4]$	0.404 ± 0.019
Delay 1	5217.2 ± 192.0	$-1031.8\ [-1605.4, -458.2]$	0.309 ± 0.015
Delay 5	4183.5 ± 162.0	$-2065.5\ [-2434.2, -1696.8]$	0.463 ± 0.022
Delay 10	5106.8 ± 123.9	$-1142.2\ [-1374.4, -910.1]$	0.331 ± 0.026

All five substitutions reduce mean return relative to the current-state graph, and all five 95% confidence intervals for ΔJ lie below zero. Delay 5 produces the largest mean decrease (2065.5), followed by state shuffling (1586.1). Each substitution also changes the action at the same physical state, with mean ℓ_2 shifts of 0.309–0.463. These interventions support a functional role for graph–state alignment across the eight frozen checkpoints.

F Real-Robot Deployment and Case Study

F.1 Real-Robot Deployment Protocol

Platform and common state interface. We deploy the frozen CoopNavi-n=3 GCPD checkpoint on three Unitree Go2 quadrupeds, denoted A_0, A_1, A_2. Each robot runs LiDAR–inertial localization against the same prebuilt map and transmits timestamped planar position, yaw, and planar velocity to a central host. Onboard odometry is produced at 20 Hz; the host consumes the latest valid state and executes policy inference and command transmission at 10 Hz. The host rejects stale or mutually inconsistent states before enabling motion.

We keep a single physical scale across arena sizes: 1 simulation unit equals 3.33 m. Thus a centered $20\,\mathrm{m} \times 20\,\mathrm{m}$ arena maps to $[-3, 3]^2$, whereas the available $7\,\mathrm{m} \times 7\,\mathrm{m}$ hardware arena occupies the compact subregion $[-1.05, 1.05]^2$. Positions, velocities, and the three surveyed goal locations are transformed into this common simulation frame before constructing the exact 3×14 CoopNavi observation. Goals remain an unordered set, as in the simulator; the controller is not given a fixed robot–goal assignment.

Figure S5 summarizes the physical deployment.

Frozen checkpoint and action conversion. The deployed checkpoint is the seed-3 CoopNavi-n=3 model trained under mobility asymmetry on agent A_0 with multiplier $\alpha = 0.55$. We freeze all network parameters, including the state-conditioned graph scorer, and provide neither perturbation identity nor onset as an input.

The simulator action is a normalized two-dimensional world-frame control, not a calibrated Go2 velocity. We therefore use the explicit deployment map

$$\mathbf{v}_{i,\mathrm{des}}^{\mathrm{world}} = 0.45\,\mathbf{a}_i\ \mathrm{m\,s^{-1}}, \tag{F.1}$$

followed by a planar speed limit of $0.60\,\mathrm{m\,s^{-1}}$, an acceleration limit of $0.25\,\mathrm{m\,s^{-2}}$, a jerk limit of $0.75\,\mathrm{m\,s^{-3}}$, and a first-order low-pass time constant of 0.35 s. A separate bounded yaw controller turns each robot toward its commanded travel direction, with a maximum yaw rate of $1.0\,\mathrm{rad\,s^{-1}}$. Finally, the world-frame planar command is rotated into the robot body frame before transmission.

Mobility-attenuation conditions. We evaluate three conditions: (i) trained-identity immediate, which attenuates A_0 from the first enabled control frame; (ii) trained-identity delayed, which attenuates A_0 after 3.0 s of cumulative hardware-active control; and (iii) alternate-identity immediate,

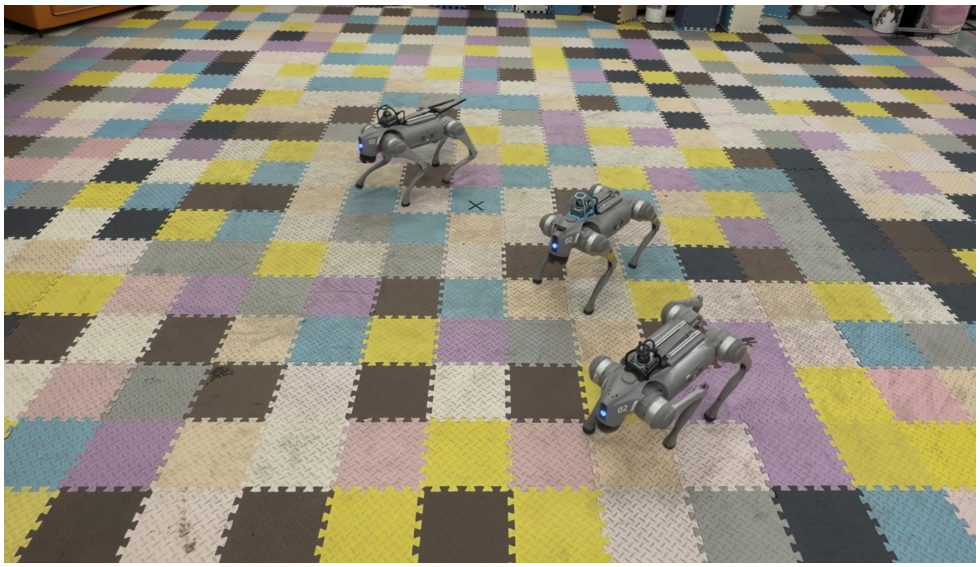


Figure S5: Real-robot deployment setup for the frozen CoopNavi-n=3 controller. The three Unitree Go2 platforms instantiate A_0, A_1, and A_2 in the common mapped arena; the marked goals, if shown, form an unordered three-target set. The mobility intervention is applied in the command sender, while the policy and graph scorer remain frozen.

which instead attenuates A_2. The third condition changes the perturbed identity from the one used during checkpoint training.

The selected robot's nominal planar command is first multiplied by 0.55 at the final sender stage. Real Go2 locomotion also has a low-speed execution region that is absent from MPE. For every robot, commands with planar norm at most $0.03\,\mathrm{m\,s^{-1}}$ are set to zero, while commands with norm between 0.03 and $0.15\,\mathrm{m\,s^{-1}}$ are direction-preservingly raised to $0.15\,\mathrm{m\,s^{-1}}$. For the attenuated robot, the 0.55-scaled command is sent continuously whenever its norm is at least $0.15\,\mathrm{m\,s^{-1}}$. If its scaled norm lies between 0.03 and $0.15\,\mathrm{m\,s^{-1}}$, the sender transmits $0.15\,\mathrm{m\,s^{-1}}$ on 15 uniformly distributed ticks and an explicit zero on the remaining 5 ticks of each 20-tick period. Packets are still sent on every tick, so this commanded duty cycle does not create a communication watchdog timeout. Yaw is not attenuated, and a reached-target latch overrides all shaping with a three-axis zero command. This hybrid rule approximates the intended command-level mobility ratio without claiming that measured physical velocity is exactly 0.55 of the nominal velocity.

Safety, completion, and logging. Before each trial, every robot must pass a stationary localization check and the three zero-command paths must acknowledge successfully. During a trial, the host enforces a $0.30\,\mathrm{m}$ minimum pair distance, a map geofence, a live $0.5\,\mathrm{s}$ dead-man signal, and $1.0\,\mathrm{s}$ limits on state age and three-robot receive-time skew. Robot-local relays independently stop on command loss, and an operator retains a physical emergency stop. A robot is latched after it enters a $0.30\,\mathrm{m}$ target radius and is released only if it leaves a $0.40\,\mathrm{m}$ radius. An episode succeeds when all three unordered targets are covered continuously for $0.5\,\mathrm{s}$ within the $120\,\mathrm{s}$ horizon.

For every formal episode we record the state, raw action, nominal and transmitted command, perturbation phase, target latches, and the full 3×3 dynamic symmetry graph $W_\phi(s_t)$. Graph summaries exclude startup, safety-inhibited, post-sender, and target-latched samples. Because adjacent $10\,\mathrm{Hz}$ samples are strongly correlated, the episode, rather than the control tick, is the reporting unit.

F.2 Real-Robot Deployment Case Study

Task completion. Table S18 summarizes the fixed reporting set. The frozen checkpoint completes all 15 real-robot episodes without hardware fine-tuning: $5/5$ under the trained perturbation identity, $5/5$ when the same perturbation begins only after motion is underway, and $5/5$ when the perturbation

is moved to the previously unseen identity A_2. The latter two blocks are particularly informative because neither onset nor identity is provided to the policy. The result therefore tests the complete localization–communication– inference–control loop, rather than replaying simulator states through the actor.

Table S18: Three-Go2 frozen-policy deployment. "Immediate" begins at the first enabled control frame; "delayed" begins after 3.0 s of cumulative hardware-active control.

Condition	Agent	Onset	Success
Trained identity, immediate	A_0	immediate	$5/5$
Trained identity, delayed	A_0	3.0 s	$5/5$
Alternate identity, immediate	A_2	immediate	$5/5$
Total	–	–	$\mathbf{15/15}$

A substantial simulation-to-real scale gap. The hardware result is not obtained by reproducing MPE dynamics. In simulation, the actor controls a normalized force-like input under the update $v_{t+1} = 0.75v_t + 0.5a_t$. On hardware, the same actor output is interpreted through a velocity heuristic, actuator dead zones, low-pass and acceleration constraints, body-frame conversion, legged locomotion dynamics, localization noise, and wireless latency. Moreover, the 7 m arena covers only $[-1.05, 1.05]^2$ at the fixed physical scale, rather than expanding to fill the simulator extent. The $0.45\,\mathrm{m\,s^{-1}}$ policy scale and the low-speed duty rule are consequently interface calibrations, not dynamics matching or policy fine-tuning. Viewed against this distribution shift, the $15/15$ completion result indicates that the frozen policy retains useful robustness well beyond the simulator's exact state–action response.

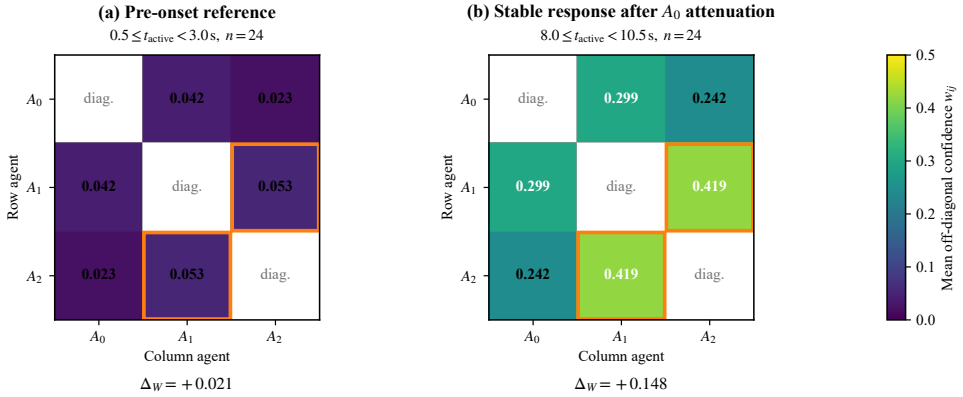


Figure S6: Before/after symmetry graphs for an illustrative completed A_0 delayed-onset deployment episode. Each panel is the mean $W_\phi(s_t)$ over an equal 2.5 s retained window: (a) the pre-onset reference, $0.5 \le t_{\text{active}} < 3.0$ s, and (b) a stable post-onset response window, $8.0 \le t_{\text{active}} < 10.5$ s ($n = 24$ ticks per panel). The intervention begins at $t_{\text{active}} = 3.0$ s. Orange outlines mark the only relation not incident to perturbed A_0, $A_1 \leftrightarrow A_2$. Diagonal identity entries are masked and both panels use the same off-diagonal color scale. The local-relation gap increases from $+0.021$ to $+0.148$; this single-run visualization is a mechanism example, not a claim of perfect trial-wise detection.

Retention of local symmetry breaking. We use $W_\phi(s_t)$ as a mechanism probe, not as a binary fault classifier or a task-success criterion. For a trial perturbing A_k, let $w_{\text{healthy},t}$ be the only agent-pair edge not incident to A_k, and let $w_{\text{inc},1,t}$ and $w_{\text{inc},2,t}$ be the two perturbation-incident edges. We define the local-relation gap over a retained tick set $\mathcal{T}$ as

$$\Delta_W(\mathcal{T}) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left[w_{\text{healthy},t} - \frac{w_{\text{inc},1,t} + w_{\text{inc},2,t}}{2} \right], \tag{F.2}$$

where a positive value is the expected local symmetry-breaking ordering: the relation that does not touch the attenuated robot remains more reliable on average. For the aggregate episode statistics below, $\mathcal{T}$ is the complete retained active-control portion of an episode; Figure S6 instead uses the two explicitly stated $2.5\,\mathrm{s}$ subwindows.

Figure S6 visualizes one completed trained-identity delayed-onset episode using equal-duration windows. Before onset, the three edge means lie in a narrow 0.023–0.053 band. In the stable post-onset window, both A_0-incident edges (0.299 and 0.242) are below the healthy A_1–A_2 edge (0.419), and the window-level gap $\Delta_W(\mathcal{T})$ increases from $+0.021$ to $+0.148$. The relevant evidence is this relative local separation, not absolute suppression of every A_0-incident edge.

Using the complete retained active-control portion of each of the same 15 successful episodes, $9/15$ show a positive Δ_W, with episode-mean $\Delta_W = +0.0166$. The healthy edge is the highest-confidence edge on an episode-mean 39.1% of retained ticks, compared with the 33.3% random-ranking reference for three edges. The positive ordering appears in $3/5$ trained-identity immediate trials and $4/5$ delayed-onset trials. It is also present in $2/5$ alternate-identity trials, showing that the graph response is not hard-coded to the training-time identity. As a complete-episode existence-of-response check, the largest gap in the three respective conditions is $+0.214$, $+0.144$, and $+0.150$, with healthy-edge Top-1 rates of 78.1%, 52.1%, and 61.8%.

The hardware separation is, as expected, weaker and noisier than the matched simulator reference (episode gap $+0.187$ and healthy-edge Top-1 rate 68.5%). This attenuation is consistent with the much larger state–action and physical-scale mismatch described above: command shaping changes the realized motion, and the other robots' closed-loop compensation also changes incident-edge evidence. Nevertheless, the positive aggregate gap, the majority of direction-correct episodes, and clear responses under all three conditions show that the learned graph retains the ability to express *local*, relation-specific symmetry breaking after transfer to hardware. We therefore interpret the real-robot graph as evidence of partially preserved local symmetry-breaking capability, rather than as a perfectly calibrated online detector. Importantly, task completion remains $15/15$ even where the diagnostic margin is small, consistent with GCPD using graded graph-conditioned residual capacity rather than requiring a discrete failure label.