GROUP-CONTEXT SYMMETRY BREAKING: TURNING ENVIRONMENTAL ASYMMETRIES INTO A LEARNABLE GROUP CONTEXT FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Symmetry lets reinforcement learning agents share experience across related states and actions, but environmental asymmetries make that sharing biased. How can a broken symmetry remain a useful structural prior? We propose Group-Context Symmetry Breaking (GSB), which represents the asymmetry in a learnable context that transforms with the task's group. GSB enforces joint equivariance over state, action, and context, so an asymmetric environment becomes a fixed-context slice of an equivariant family. Real transitions teach the model which slice to use. We derive bounds linking context displacement under the group, model error, and symmetry deviation. Any model equivariant in state and action alone must pay at least half the largest deviation, so GSB provably approximates the environment better whenever its context family's approximation error falls below that floor. In soft actor–critic, the learned model generates experience in locomotion, and in manipulation it corrects real transitions carried along the group orbit while the policy reads the learned context. We compare GSB with strict, relaxed, and partially equivariant priors, including partially equivariant SAC built for symmetry-breaking environments, and with unstructured SAC. Under controlled symmetry breaking, these priors fall below even unstructured SAC on most tasks, whereas GSB attains the highest mean return of all methods at the end of training on the Hexapod, Ant, and Humanoid tasks and the most consistent return on Walker2d. On two manipulation tasks where agents train on one half of the goal domain under a constant lateral force, GSB attains the best held-out return against every baseline and ranks first in every seed. GSB's correction of the transported transitions at the learned context adds 4.35 and 1.76 in held-out return, each larger than GSB's lead over the strongest baseline, so the lead rests on this correction.

1 INTRODUCTION

Group symmetry is an effective structural prior for reinforcement learning (RL), because it lets an agent share a law of dynamics and reward across related states and actions (van der Pol et al., 2020; Mondal et al., 2020; Bronstein et al., 2021). The agent **trains less yet learns more**, since a collected transition also teaches the unseen transitions on its symmetry orbit. Equivariant RL uses this sharing to improve sample efficiency in control and manipulation (van der Pol et al., 2020; Wang et al., 2022a; Zhu et al., 2022; Tangri et al., 2024). For an exactly symmetric Markov decision process (MDP), an optimal policy can be chosen equivariant, so the structural restriction preserves optimality (Zinkevich & Balch, 2001; Ravindran & Barto, 2003). However, this guarantee rests on exact symmetry. Without it, related state–action pairs obey different laws, so symmetry reuse introduces systematic error. An uneven load on one limb, for example, changes the dynamics without changing the robot's nominal morphology (Figure 1a). Exact symmetry priors then spread the error, while unstructured learners give up the sharing. This paper asks *how a broken symmetry can remain a useful structural prior*.

Under symmetry breaking, the error of an exact symmetry prior reaches both value estimates and learned models, and more data cannot remove it. Chang et al. (2026) show that group-invariant Bellman backups propagate local violations into value estimates elsewhere, degrading learning in

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.

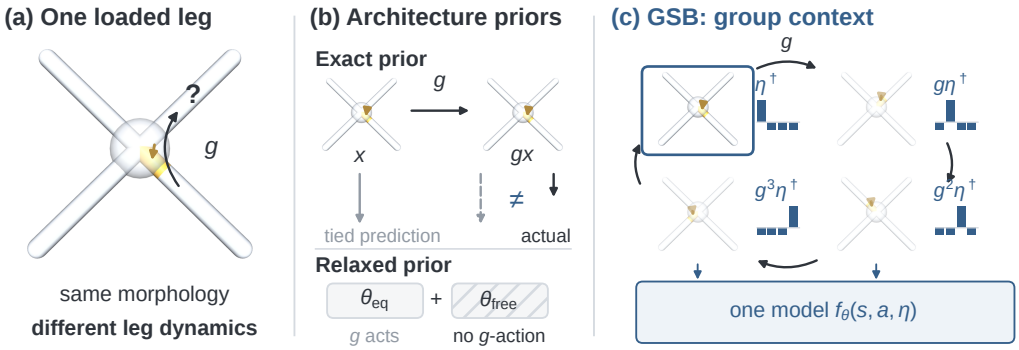

Figure 1: A group context lets one model share structure across asymmetric environments. (a) A constant load on one leg (ochre) breaks the Ant's rotational symmetry. The group moves the load to the neighbouring leg. (b) Exact priors tie predictions across the orbit despite the load, while relaxed priors hold the asymmetry in free weights on which the group does not act. (c) GSB feeds a context that rotates with the robot into one jointly equivariant model. The outlined member is the observed environment, and the four drawn members are the leg positions the load can occupy.

grid-world and manipulation tasks. Wang et al. (2023b) establish an approximation-error lower bound for models constrained to an equivariance that the target violates. Park et al. (2025) analyse the resulting cost of imposing exact symmetry in approximately symmetric tasks. A symmetry prior that survives the breaking must therefore represent the task's *asymmetry*, including its direction and magnitude, although the environment supplies neither explicitly.

Existing RL approaches lack an input that carries the asymmetry and transforms with the group. Architectural relaxations hold the asymmetry in parameters on which the group does not act (Finzi et al., 2021a; Romero & Lohit, 2022; van der Ouderaa et al., 2023; Park et al., 2025; Chang et al., 2026). Approximate-invariance analyses bound the asymmetry's cost but keep it out of the model (Park et al., 2025; Wang et al., 2023b). Context-conditioned RL encodes the task without prescribing its group action (Lee et al., 2020; Rakelly et al., 2019; Zintgraf et al., 2020). Each of them therefore gives up exact symmetry or leaves the asymmetry unmodelled.

We propose *Group-Context Symmetry Breaking* (GSB), which turns the environment's asymmetry into a learnable *group context*. With it, one exactly equivariant model shares structure across asymmetric environments. GSB treats the asymmetry as a context η, an input that the group moves and that conditions the transition and reward model. Our hypothesis is that exact equivariance holds on the extended space (s, a, η), with each environment a fixed-context *slice* $\eta = \eta^\dagger$ (Figure 1c). GSB learns this context from real transitions, without knowing the asymmetry's direction or magnitude.

Within one environment, GSB reuses a transition by transforming it with a group element while holding the context at its learned value, with a provable error bound (Figure 2). Joint transport of a state–action pair $x = (s, a)$ by a group element g would also move η into another context slice. GSB instead predicts at $g \cdot x$ in the original environment with η held fixed, and we bound its error by how far g moves the context and by the model error.

GSB brings its learned model and context into control in locomotion and manipulation through soft actor–critic (SAC; Haarnoja et al., 2018). In locomotion, the learned slice supplies model-generated experience (Sutton, 1991; Janner et al., 2019). In manipulation, the learned slice corrects real transitions carried along the group orbit, and the learned context enters a jointly equivariant policy (Figure 3).

Contributions. Our contributions show that a broken symmetry remains a useful structural prior once its asymmetry is learned as a group context.

We propose GSB to keep an exact group symmetry in asymmetric environments. GSB learns the asymmetry as a group context, so one transition and reward model stays exactly equivariant in state, action, and context, and each asymmetric environment becomes one fixed-context slice.

GSB enables symmetry reuse and provably avoids the error that exact priors propagate. We prove that every exact symmetry prior pays at least half the largest deviation, whereas GSB pays only its approximation error and bounds the error of every reused transition (Propositions 1 and 2).

GSB attains state-of-the-art returns under controlled symmetry breaking. On four locomotion tasks that we build to break symmetry with a known load, GSB has the highest mean return of all methods on our own Hexapod, the Ant, and the Humanoid, and the most consistent return on Walker2d.

GSB achieves transfer to mirrored conditions never seen in training. Trained under a constant lateral force on one half of the goal domain in FetchSlide and FetchPush, GSB attains the best return on the mirrored half, a lead that rests on correcting mirrored transitions at the learned context.

2 Related Work

We distinguish GSB from prior work by where each approach represents the environment's asymmetry. Exact-equivariant reinforcement learning builds the task's symmetry group into policies, value functions, or joint state–action representations (Cohen & Welling, 2016; van der Pol et al., 2020; Mondal et al., 2020; Wang et al., 2022a; 2023a). Exact equivariance also extends to offline RL (Tangri et al., 2024) and to equivariant diffusion policies (Wang et al., 2024a). These approaches impose exact equivariance on the learned function. GSB imposes joint equivariance on the model and its context inputs, allowing each fixed-context environment to break the nominal symmetry.

Architectural relaxations admit the asymmetry inside the network, in parameters or pathways on which the group does not act. Residual pathways, relaxed group and steerable convolutions in reinforcement learning agents, learned partial and layer-wise equivariance, and input-dependent routing restore expressiveness through departures from exact equivariance (Finzi et al., 2021a; Park et al., 2025; Romero & Lohit, 2022; van der Ouderaa et al., 2023; Chang et al., 2026). Approximately equivariant dynamics models relax weight sharing to fit imperfectly symmetric systems and reveal where the symmetry breaks (Wang et al., 2022b; 2024b). GSB keeps the model exactly equivariant on the extended space, isolates the asymmetry in an environment-level input, and constrains how that input transforms with the group.

Outside reinforcement learning, equivariant networks already take symmetry-breaking inputs (Xie & Smidt, 2024; Lawrence et al., 2025). Smidt et al. (2021) use Curie's principle to learn such inputs for Euclidean equivariant networks, and Kaba & Ravanbakhsh (2023) let an equivariant multilayer perceptron break the symmetry of its input. GSB fits its transforming input from reinforcement learning transitions as one context per environment, and carries that context into a prediction-error bound and into control.

Data augmentation encourages symmetry through transformed samples but presumes the invariance it is meant to instil. Augmentation methods train on transformed copies of collected experience (Laskin et al., 2020; Yarats et al., 2021; 2022), regularising the model toward the assumed invariance without constraining the network. GSB fits its model and group context on real transitions, generates experience at the learned context, and corrects transported copies there rather than trusting them.

Environment-level descriptions bound the asymmetry or infer it as a latent parameter, but prescribe no group action for it. Approximate-invariance analyses bound reward and transition discrepancies and use them to control the resulting deviation in value (Park et al., 2025; Wang et al., 2023b), so the asymmetry enters as a task discrepancy rather than as an input. Hidden-parameter MDPs (Doshi-Velez & Konidaris, 2016; Killian et al., 2017) and context-conditioned RL (Lee et al., 2020; Rakelly et al., 2019; Zintgraf et al., 2020) infer a latent task parameter from transitions and feed it to the model or policy. GSB gives that parameter a prescribed group action, so the family is exactly equivariant, each environment is one of its slices, and that action fixes the parameter's dimension (Proposition 10). Other work discovers symmetries, bounds the generalisation of equivariant networks, and models symmetry-breaking bifurcations with equivariant flows (Dehmamy et al., 2021; Behboodi et al., 2022; Hendriks et al., 2025). GSB makes the learned task input the object of the approximate-invariance bounds, since its displacement under the group enters the deviation and prediction-error bounds.

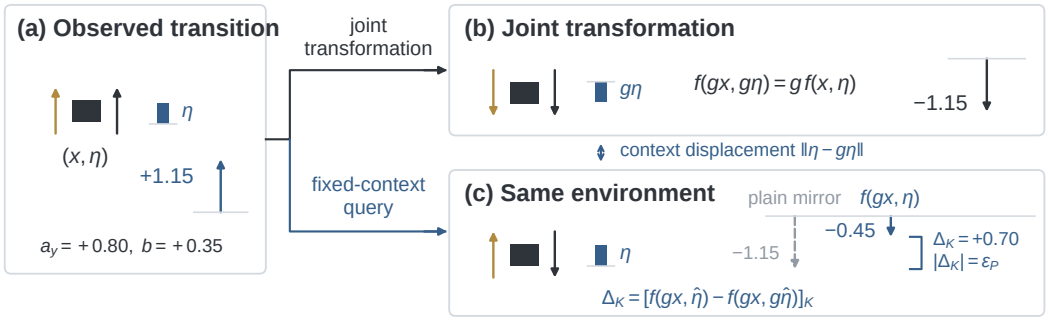

Figure 2: Joint transformation is exact, but reuse in the same environment holds the context fixed. In a one-step toy, action a_y (black) and force b (ochre) give $v' = a_y + b$ (blue arrow), with context $\eta = b$. Mirroring action and force gives -1.15 (b), whereas the original environment returns -0.45 at the mirrored input (c). The correction $\Delta_K = +0.70$, the model's shift on the corrected slots K (here v') from context $g\hat{\eta}$ to $\hat{\eta}$, moves the plain mirror to the true outcome and equals the deviation $\epsilon_P(x; g)$ of equation 2 under the absolute difference d_P. The context displacement $\|\eta - g\eta\|$ enters the bound of Section 3.3.

3 Method and Theory

3.1 Setup

We locate the effect of a broken symmetry within the transition model rather than the policy, and quantify the error it induces. Let $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a Markov decision process (MDP) with transition kernel $P(\cdot \mid s, a)$, reward $r(s, a)$, and discount $\gamma \in [0, 1)$. A finite group G acts on states and actions, written $g \cdot s$ and $g \cdot a$, and the pair $x = (s, a)$ has orbit $\{(g \cdot s,\, g \cdot a) : g \in G\}$. For a next-state distribution μ, $g_{\#}\mu$ denotes its push-forward, the distribution of $g \cdot s'$ when $s' \sim \mu$. Equivariant methods assume M to be *strictly G-invariant* (Appendix A),

$$P(g \cdot s' \mid g \cdot s,\, g \cdot a) = P(s' \mid s, a), \qquad r(g \cdot s,\, g \cdot a) = r(s, a) \tag{1}$$

for every $g \in G$ and every (s, a, s'). Their sample efficiency is the $|G|$-fold reuse this assumption licenses (van der Pol et al., 2020; Mondal et al., 2020; Wang et al., 2022a).

We measure how far the environment violates the symmetry on each orbit through the deviation fields

$$\epsilon_P(x; g) = d_P\big(P(\cdot \mid s, a),\, (g^{-1})_{\#}P(\cdot \mid g \cdot s,\, g \cdot a)\big)\,, \qquad \epsilon_r(x; g) = |r(s, a) - r(g \cdot s,\, g \cdot a)|\,. \tag{2}$$

The fields compare what the environment does at x with what it does at the transported pair, read back in the coordinates of x, for a discrepancy d_P specified in Appendix A. The field $\epsilon_P(x; g)$ is the error of treating the transported pair as evidence about x, and $\Delta = \sup_{x,g} \epsilon_P(x; g)$ is the largest such error a task admits. Δ is a property of the system rather than a statistic of the data, since a ground slope or an unequal actuator gain sets its value and further interaction leaves it unchanged.

We show that the deviation imposes a cost on every strictly equivariant model. A strictly equivariant model assigns orbit members the same law in common coordinates. When the environment's laws differ by $\epsilon_P(x; g)$, at least one model error therefore reaches $\epsilon_P(x; g)/2$, and the supremum gives a worst-case approximation floor of $\Delta/2$ (Proposition 1). The floor belongs to the model class, so more data or capacity within it cannot remove it. Orbit data augmentation meets the same conflict when it treats real and transformed transition laws as interchangeable (Lemma 5).

We escape the floor of strict equivariance by giving the model an extra input that carries the asymmetry, so that equivariance can relate environments with different asymmetries. What must this input carry, and can real transitions supply its value?

3.2 THE GROUP CONTEXT

We derive the added input from the group action itself, which admits a single vector that carries the breaking. A deployed system violates equation 1 through a cause that stays fixed while the agent acts, such as a load on one leg. A group element moves that cause to a corresponding place, such as the neighbouring leg. The cause's components along directions that every group element holds pointwise fixed are unchanged by this move, so they cannot break the symmetry (Lemma 8), and the breaking lives in the complement of these directions, the *breaking subspace* $\mathcal{V}$. One coordinate per dimension of $\mathcal{V}$ therefore suffices, and no lower-dimensional context can separate all the environments a family must cover (Proposition 10). We call this vector the *group context* η and build GSB upon it, since η transforms with the group and belongs to the environment, not to the model.

GSB replaces the single kernel P by a family $\mathcal{F} = \{(K_\theta, \bar{r}_\theta) : \theta \in \Theta\}$ over the extended space $\mathcal{S} \times \mathcal{A} \times \mathcal{V}$, pairing a Markov kernel over next states with a scalar reward model, where $\mathcal{V}$ carries the induced isometric G-action. Three conditions specify it. Each member is jointly equivariant in state, action, and context,

$$K_\theta(g \cdot s' \mid g \cdot s,\, g \cdot a,\, g \cdot \eta) = K_\theta(s' \mid s, a, \eta), \qquad \bar{r}_\theta(g \cdot s,\, g \cdot a,\, g \cdot \eta) = \bar{r}_\theta(s, a, \eta); \tag{3}$$

each member's prediction changes with its context no faster than fixed concave moduli allow, which later bounds the effect of moving the context (Section 3.3),

$$d_P(K_\theta(\cdot \mid x, \eta_1),\, K_\theta(\cdot \mid x, \eta_2)) \le \omega_P(\|\eta_1 - \eta_2\|), \qquad |\bar{r}_\theta(x, \eta_1) - \bar{r}_\theta(x, \eta_2)| \le \omega_r(\|\eta_1 - \eta_2\|); \tag{4}$$

and an environment lies in the GSB class when one member reproduces it at a bounded context,

$$\begin{aligned} \|\eta^\dagger\| &\le B_\eta, \\ \sup_{s,a} d_P\big(P(\cdot \mid s, a),\, K_{\theta^\dagger}(\cdot \mid s, a, \eta^\dagger)\big) &\le \delta_P, \\ \sup_{s,a} \big|r(s, a) - \bar{r}_{\theta^\dagger}(s, a, \eta^\dagger)\big| &\le \delta_r. \end{aligned} \tag{5}$$

The symmetry of equation 3 is exact on the extended space, and the observed environment is the *slice* of the family at $\eta = \eta^\dagger$. In Figure 2, the joint-transport edge connects (x, η) to $(g \cdot x, g \cdot \eta)$ and obeys equation 3, whereas the fixed-context edge ends at $(g \cdot x, \eta)$ within the original environment. The two endpoints differ only in the context, and equation 4 bounds the effect of that change on the model prediction. We learn the model and context from real transitions (Section 3.4), without being told the asymmetry's direction or magnitude. **Relaxed priors repair the model on the broken space. GSB repairs the space.**

3.3 QUANTIFYING THE GAIN

We prove that GSB approximates the environment better than every strictly equivariant model whenever its family's error falls below the floor that strict equivariance imposes. A group-fixed context defines a strictly equivariant subfamily within GSB (Proposition 4), and GSB extends it by letting the context vary. A1–A6 (Appendix A) fix the group action and d_P and restate equation 3–equation 5.

Proposition 1 (Cost of strict equivariance and the GSB margin)**.** *Under A1–A6, every strictly G-equivariant kernel Q satisfies $\sup_x d_P(P(\cdot \mid x), Q(\cdot \mid x)) \ge \Delta/2$, whereas the member $K_{\theta^\dagger}(\cdot \mid \cdot, \eta^\dagger)$ of equation 5 satisfies $\sup_x d_P(P(\cdot \mid x), K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger)) \le \delta_P$. Whenever $\delta_P < \Delta/2$, GSB approximates the environment better than every strictly equivariant kernel, by at least $\Delta/2 - \delta_P$.*

A load the group moves between limbs makes the environment a slice of a jointly equivariant family, so δ_P is only the model class's approximation error (Lemma 9). Appendix B proves Proposition 1 (Proposition 6, Corollary 7).

We bound the prediction error of reuse within the observed environment. Joint transport gives an exact equivariance identity. We carry it back to the observed slice by comparing contexts, at the cost of a context-modulus term that vanishes when the group fixes the context. The same argument bounds the environment's own breaking.

Proposition 2 (Deviation band and reuse error)**.** *Under A1–A6, for every x, every $g \in G$, and every learned slice $(\hat{\theta}, \hat{\eta})$ with $\hat{\theta} \in \Theta$ and $\|\hat{\eta}\| \le B_\eta$,*

$$\epsilon_P(x; g) \le 2\delta_P + \omega_P\big(\|\eta^\dagger - g^{-1} \cdot \eta^\dagger\|\big)\,, \tag{6}$$

$$d_P\big(P(\cdot \mid g \cdot x),\, K_{\hat{\theta}}(\cdot \mid g \cdot x, \hat{\eta})\big) \le \delta_P + e_{\hat{\theta},\hat{\eta}}(x) + \omega_P\big(\|\eta^\dagger - g^{-1} \cdot \eta^\dagger\|\big) + \omega_P\big(\|\hat{\eta} - g^{-1} \cdot \hat{\eta}\|\big)\,, \tag{7}$$

where $e_{\hat{\theta},\hat{\eta}}(x) = d_P(K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger), K_{\hat{\theta}}(\cdot \mid x, \hat{\eta}))$ is the error of the learned slice at the untransformed pair.

A single interaction thus supplies $|G|$ transitions whose error stays inside the band, because equation 7 presumes nothing about how $(\hat{\theta}, \hat{\eta})$ were obtained. This is the precise sense in which the agent trains on less real experience and learns from more. The bound consists of realizability, the learned slice's error, and two context-transport terms. Appendices B and D prove the two bounds (Propositions 3 and 11).

3.4 The GSB framework

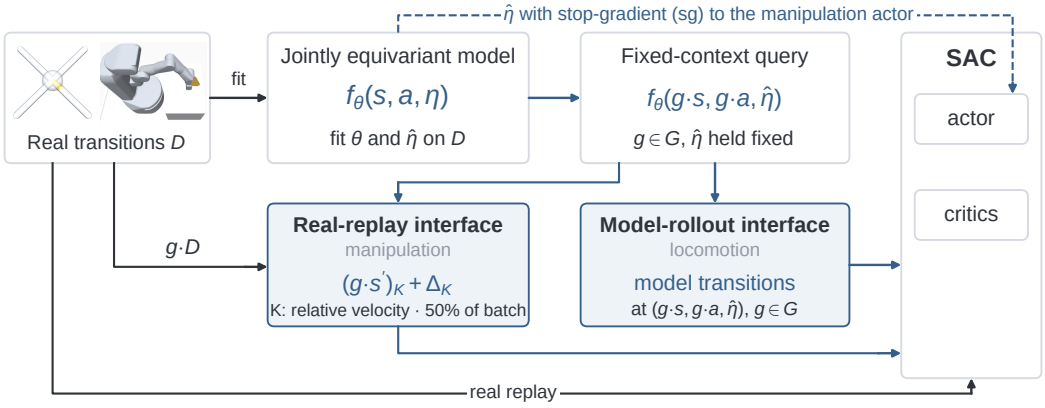

Figure 3: One learned model and context reach control through two interfaces. Real transitions fit the jointly equivariant model and a single context $\hat{\eta}$, and the model is queried at transformed inputs with $\hat{\eta}$ held fixed. The real-replay interface corrects the relative-velocity slots K of mirrored real transitions, and the model-rollout interface generates transitions for every group element. The manipulation actor also reads $\hat{\eta}$ through a stop-gradient, and SAC updates neither the model nor $\hat{\eta}$.

GSB uses one construction on both task families. Its prediction at the environment's own $\hat{\eta}$ for a transformed state–action pair handles the breaking, and the two interfaces of Figure 3 differ only in what they do with that prediction. In locomotion, the model and context share the objective

$$\min_{\theta,\ \eta:\ \|\eta\| \le B_\eta} \mathcal{L}_{\mathrm{dyn}}(\theta, \eta; \mathcal{D}_{\mathrm{real}}) + c_r\, \mathcal{L}_{\mathrm{rew}}(\theta, \eta; \mathcal{D}_{\mathrm{real}}), \tag{8}$$

where c_r is fixed before training and $\mathcal{L}_{\mathrm{dyn}}$ is the one-step prediction loss over the whitened state. A single context per environment is shared by the transition and reward heads and never made a function of the observation. The group action, not tuning, settles its dimension, which stays small. On the Ant, the context holds one number per leg, and dropping the group-fixed all-equal direction keeps three (Appendix H).

K_θ is a deterministic kernel (Appendix J) that predicts the state change in whitened coordinates, with a whitening that commutes with the group, and is made exactly equivariant by group averaging (Appendix E). A reward head shares its context. We use standard SAC twin critics without context inputs in both task families, and SAC updates neither the context nor the model.

We let the context reach control through the model's predictions in locomotion. A standard tanh-Gaussian actor reads the state alone and learns from real and model-generated experience. Standard rollouts cover the identity element. Symmetry expansion covers each remaining group element once.

These predictions are generated without gradients, so the learned slice shapes control through the training data. The synthetic-replay fraction follows a fixed cap on Hexapod and a cap that decays to zero on Ant, Humanoid, and Walker2d (Appendix G).

For manipulation, the context reaches control along two paths at once. The actor is jointly equivariant in state and context and receives a detached $\hat{\eta}$, so policy gradients cannot change the context. Instead of model rollouts, half of each SAC batch holds real transitions carried along the group orbit. To relabel them as experience of the training environment, the model corrects the relative-velocity coordinates of their measured next states by the difference between its predictions at $\hat{\eta}$ and at $g \cdot \hat{\eta}$. The correction acts only on the relative object–gripper velocity, and the other coordinates keep the mirrored measurement (Appendix H). The learned model and context therefore support control twice over, once through the corrected transitions and once through the context the actor reads.

4 Experiments

GSB attains the highest mean final-window return of all methods on Hexapod, Ant, and Humanoid, the most consistent return on Walker2d, and the best held-out return on both manipulation tasks.

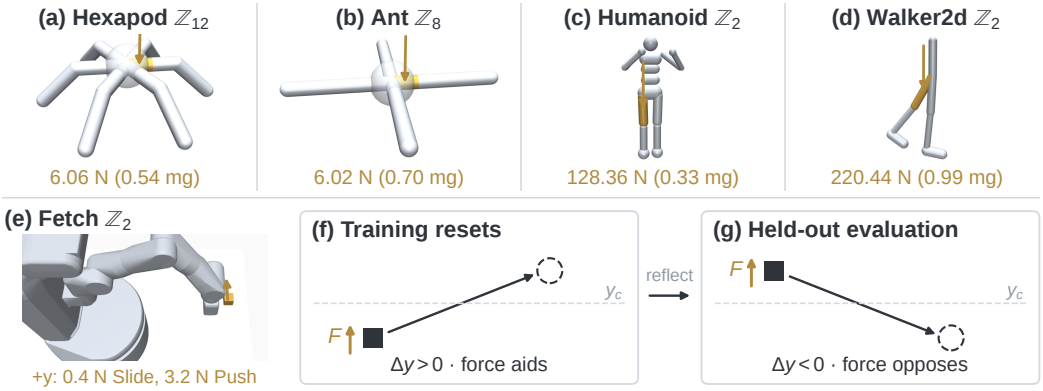

Figure 4: Controlled symmetry breaking in six tasks. A hidden constant force (ochre) acts at one body's centre of mass: a world-down load on one limb of each locomotion robot, given as F and as a multiple of the robot's weight mg (a–d), and a $+y$ force on the Fetch object (e). Manipulation trains on resets that require $+y$ object motion, which the force aids (f), and evaluates on their mirror images about y_c, which the force opposes (g). Hexapod and Ant use $\mathbb{Z}_{12}$ and $\mathbb{Z}_8$ because a full turn flips the sign of the torso quaternion (Appendix F).

4.1 Experimental setup

Tasks and comparisons. We introduce a constant world-down force at one body's centre of mass in our six-legged Hexapod and in the Ant, Humanoid, and Walker2d robots of OpenAI Gym (Brockman et al., 2016), all simulated in MuJoCo (Todorov et al., 2012). The force acts throughout training and evaluation, while its magnitude and location are hidden from the observation. The load on one body from an orbit of corresponding limbs breaks the declared symmetry. Each task keeps its original action space and reward function. All methods on a task share the same force. The locomotion budget is one million environment steps. Figure 4(a–d) shows the models, groups, and forces, and Appendix F gives the exact settings (Table 2). Both manipulation tasks come from the MuJoCo Fetch suite (Plappert et al., 2018) and apply a constant $+y$ force at the object's centre of mass, 0.4 N on FetchSlide and 3.2 N on FetchPush (Figure 4(e)). Training episodes place the goal on the $+y$ side of the object, so the force assists the required motion. The held-out half reflects object and goal under the task's $\mathbb{Z}_2$ symmetry while the force keeps its world direction, so there the same force opposes the required motion (Figure 4(f,g)). Every manipulation run lasts 500,000 environment steps. We compare GSB with a strictly equivariant multilayer perceptron (EMLP; Finzi et al., 2021b), residual pathway priors (RPP; Finzi et al., 2021a), and partially equivariant SAC (PE-SAC; Chang et al., 2026), which was built for symmetry-breaking environments, and with vanilla SAC (Haarnoja et al.,

Table 1: Final-window return after 1M locomotion or 500k manipulation steps, mean ± sample standard deviation over seeds 42–44. Higher is better. Manipulation returns sum negative goal distances. Best means are bold. †One Humanoid PE-SAC run has 9 of 11 final-window evaluations.

	Locomotion				Manipulation, held-out half	
Method	Hexapod	Ant	Humanoid	Walker2d	FetchSlide	FetchPush
Vanilla SAC	3101.1 ± 491.3	5066.7 ± 952.9	4539.2 ± 384.3	$\mathbf{3936.3} \pm 967.8$	-19.92 ± 1.30	-7.08 ± 1.57
EMLP	1294.3 ± 2301.5	3613.2 ± 339.1	2616.8 ± 260.0	2990.2 ± 550.2	-27.06 ± 1.89	-11.10 ± 2.55
RPP	3273.7 ± 561.3	4713.8 ± 1226.0	3268.0 ± 712.7	3562.2 ± 523.0	-14.30 ± 0.50	-7.32 ± 1.43
PE-SAC	3319.0 ± 348.1	4240.8 ± 777.7	4269.6 ± 547.2 †	3654.5 ± 591.8	-25.72 ± 1.10	-10.43 ± 2.13
GSB	$\mathbf{4415.8} \pm 732.5$	$\mathbf{5238.0} \pm 1549.1$	$\mathbf{4879.0} \pm 187.2$	3610.9 ± 146.7	$\mathbf{-12.52} \pm 0.14$	$\mathbf{-5.54} \pm 1.37$

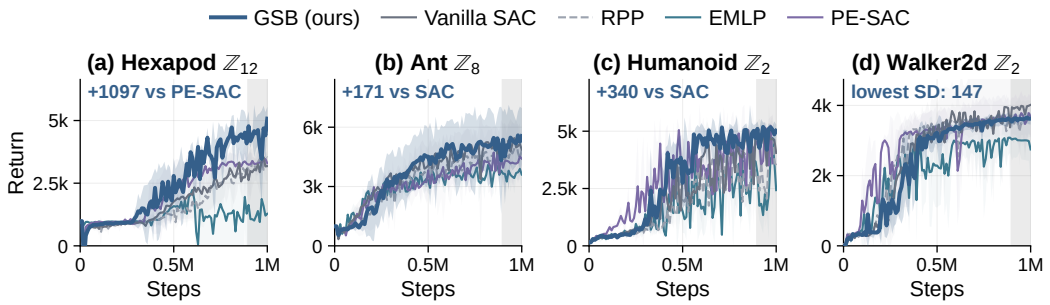

Figure 5: GSB leads on Hexapod, Ant, and Humanoid and is the most consistent on Walker2d. Lines are three-seed means with sample standard deviations shaded and the final window in gray. Labels give GSB's final-window lead over the strongest baseline and its Walker2d standard deviation, the lowest of all methods. The Humanoid PE-SAC curve ends at 980k steps.

2018) as the unstructured reference. RPP and PE-SAC permit departures from strict equivariance, so the comparison spans strict, relaxed, and unstructured alternatives.

Evaluation. The primary metric is the mean evaluation return over the final 100,000 environment steps of each run, averaged over the shared seeds 42–44, with the sample standard deviation across seeds as the spread. Manipulation evaluations score twenty held-out episodes (Appendices G, H).

4.2 Locomotion under controlled symmetry breaking

GSB attains the highest mean final-window return on Hexapod, Ant, and Humanoid (Table 1), spanning cyclic limb and bilateral symmetries. On Ant, Humanoid, and Walker2d, vanilla SAC outperforms all three symmetry priors. On Hexapod and Humanoid, GSB exceeds every baseline in every seed. GSB's widest margin is on Hexapod, 4415.8 against 3319.0 for PE-SAC. On Walker2d, GSB has the smallest across-seed standard deviation, 146.7 against 523.0–967.8. On Hexapod, GSB pulls ahead from about 0.3M steps and leads at every evaluation from 0.73M (Figure 5). GSB outperforms its strictly equivariant variant with a pinned context in every seed on Hexapod, with means of 4415.8 against 3693.9 (Appendix K).

4.3 Transfer across the broken symmetry

GSB transfers across the broken symmetry to the half the agent never practised, with the best held-out return on both tasks and the first rank in every seed (Table 1, Figure 6(c)). The held-out half is the unseen mirror image of the training half, so held-out control requires exactly the reuse that Proposition 2 bounds. On FetchSlide, every GSB run exceeds every baseline run. GSB leads every baseline, by at least 1.78 on FetchSlide and 1.54 on FetchPush. From 300k steps on FetchSlide and 360k on FetchPush, 60% and 72% of the budget, GSB's held-out mean already exceeds the strongest baseline's final-window return.

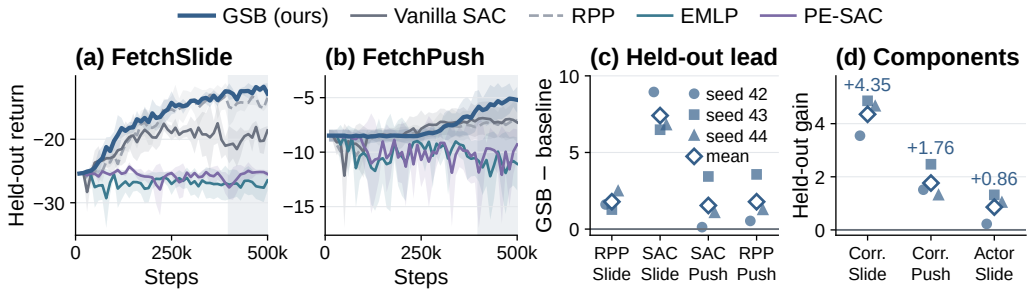

Figure 6: GSB transfers across the broken symmetry on both tasks. (a,b) Held-out return, three-seed mean $\pm$ sample standard deviation, final window shaded. (c) GSB's final-window held-out lead over the two strongest baselines, per seed. (d) Held-out gains of the correction over the plain mirror and of the jointly equivariant actor over a state-only actor. All 21 per-seed values are positive.

Where the gain comes from. Plain mirroring, which is orbit data augmentation, treats a transported transition as if the training environment had produced it. The deviation field of equation 2 measures the error of that premise, and averaging over the copies cannot remove it (Lemma 5). GSB corrects the mirrored next state in the slots K by $\Delta_K = [f(g{\cdot}s, g{\cdot}a, \hat{\eta}) - f(g{\cdot}s, g{\cdot}a, g{\cdot}\hat{\eta})]_K$, the learned model's estimate, at the learned context, of how far the true outcome lies from the plain mirror. The corrected slots carry only the learned slice's error (Appendix I). Because Δ_K vanishes at a group-fixed context (Proposition 4), the plain mirror is GSB's strictly equivariant special case. Against the plain mirror, GSB's correction adds 4.35 in held-out return on FetchSlide and 1.76 on FetchPush. On FetchSlide, GSB's jointly equivariant actor adds a further 0.86 over a state-only actor, and every gain is positive in every seed (Figure 6(d); Appendix I). Against training without transported transitions, the plain mirror loses 1.91 and 0.77 in every seed, while GSB's correction turns these losses into gains of 2.44 and 1.00 in every seed (Appendix I). The gains over the plain mirror exceed GSB's lead over the strongest baseline, so the transfer advantage rests on the correction.

Symmetry priors under breaking. Symmetry priors designed to improve on unstructured RL fall below vanilla SAC, which uses no symmetry at all, on most of the six tasks. EMLP, the only baseline enforcing exact symmetry, ranks last on every task, and vanilla SAC outranks both relaxed priors, including PE-SAC built for symmetry-breaking environments, on four of them. Imposed without a model of how the group moves the asymmetry, the symmetry thus becomes a handicap rather than an advantage. GSB turns it back into an advantage, with the best mean return of all methods on five tasks and the smallest across-seed standard deviation on four. A GSB run beats a run of each task's strongest baseline with probability 0.74 (95% bootstrap interval 0.57–0.89; Agarwal et al., 2021).

5 Conclusion

GSB turns environmental asymmetry into a learnable part of the symmetry prior. A group context fitted from real transitions makes an asymmetric environment one slice of a jointly equivariant family, and our analysis quantifies its gain over strict equivariance. Any strictly equivariant model pays at least half the largest deviation, while GSB's reuse at the learned context stays within a band set by realizability, model error, and context displacement (Propositions 1 and 2). Under controlled symmetry breaking, where the existing symmetry priors fall below even unstructured SAC on most tasks, GSB attains the highest mean final-window return of all methods on Hexapod, Ant, and Humanoid and the most consistent return on Walker2d. GSB also attains the best held-out return on both manipulation tasks in every seed, where its correction of the transported transitions at the learned context carries the gain. Across distinct morphologies and across the mirror to unseen conditions, a single method turns the breaking from a source of bias into structure the agent exploits. A broken symmetry thereby remains a useful structural prior.

Ethics Statement

This work is carried out entirely in physics simulation. It involves no human subjects, no personal or sensitive data, and no data collection from people. The controlled tasks add a constant force to standard simulated robots and to our own six-legged robot, and we foresee no direct harmful application of the method beyond the general considerations that apply to reinforcement learning for control.

Reproducibility Statement

Every method is trained on the same three seeds, 42, 43 and 44, and scored by the same rule: the mean evaluation return over the final window, which is $[900{,}000, 1{,}000{,}000]$ for locomotion and $[400{,}000, 500{,}000]$ for manipulation, inclusive of both ends. Appendix G gives the aggregation formula and the replay-fraction schedule used on each task. Appendix H lists the shared SAC hyperparameters, the architecture of every baseline, and the evaluation protocol, ten episodes every 10,000 steps for locomotion and twenty held-out episodes every 10,000 steps for manipulation. Appendix F specifies the loaded body, the group action and the exact force for each controlled locomotion task, and Appendices I and K list the components each matched manipulation and locomotion control retains. The assumptions behind the theoretical results are stated in Appendix A and every formal claim is proved in the appendix. The supplementary material contains a reference implementation of GSB, namely the group context, the jointly equivariant model, both control interfaces and the correction, together with checks of the structural statements it relies on, the Hexapod model and its symmetry group, and the settings of every experiment.

Use of AI

We used AI assistants to aid and polish the writing, to retrieve and discover related work, and to support research ideation and execution, including experiment code. The authors take full responsibility for the content.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, 2021.

Arash Behboodi, Gabriele Cesa, and Taco S. Cohen. A PAC-bayesian generalization bound for equivariant networks. In *Advances in Neural Information Processing Systems*, 2022.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv preprint arXiv:1606.01540*, 2016.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 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.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2990–2999, 2016.

Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic symmetry discovery with lie algebra convolutional network. In *Advances in Neural Information Processing Systems*, 2021.

Finale Doshi-Velez and George Konidaris. Hidden parameter Markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 1432–1440, 2016.

Marc Finzi, Gregory Benton, and Andrew Gordon Wilson. Residual pathway priors for soft equivariance constraints. In *Advances in Neural Information Processing Systems*, 2021a.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3318–3328, 2021b.

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*, 2018. URL https://arxiv.org/abs/1801.01290.

Fleur Hendriks, Ondřej Rokoš, Martin Doškář, Marc G. D. Geers, and Vlado Menkovski. Equivariant flow matching for symmetry-breaking bifurcation problems. *arXiv preprint arXiv:2509.03340*, 2025.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1906.08253.

Sékou-Oumar Kaba and Siamak Ravanbakhsh. Symmetry breaking and equivariant neural networks. *arXiv preprint arXiv:2312.09016*, 2023.

Taylor W. Killian, Samuel Daulton, George Konidaris, and Finale Doshi-Velez. Robust and efficient transfer learning with hidden parameter Markov decision processes. In *Advances in Neural Information Processing Systems*, 2017.

Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2004.14990.

Hannah Lawrence, Vasco Portilheiro, Yan Zhang, and Sékou-Oumar Kaba. Improving equivariant networks with probabilistic symmetry breaking. In *International Conference on Learning Representations*, 2025.

Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. Context-aware dynamics model for generalization in model-based reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*. PMLR, 2020.

Arnab Kumar Mondal, Pratheeksha Nair, and Kaleem Siddiqi. Group equivariant deep reinforcement learning. *arXiv preprint arXiv:2007.03437*, 2020.

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. URL https://proceedings.mlr.press/v258/park25d.html.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 2019.

Balaraman Ravindran and Andrew G. Barto. SMDP homomorphisms: An algebraic approach to abstraction in semi-Markov decision processes. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003. URL https://www.ijcai.org/Proceedings/03/Papers/145.pdf.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. In *Advances in Neural Information Processing Systems*, 2022.

Tess E. Smidt, Mario Geiger, and Benjamin Kurt Miller. Finding symmetry breaking order parameters with Euclidean neural networks. *Physical Review Research*, 3:L012002, 2021. doi: 10.1103/PhysRevResearch.3.L012002.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

Arsh Tangri, Ondrej Biza, Dian Wang, David Klee, Owen Howell, and Robert Platt. Equivariant offline reinforcement learning. *arXiv preprint arXiv:2406.13961*, 2024.

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, 2012.

Tycho F. A. van der Ouderaa, Alexander Immer, and Mark van der Wilk. Learning layer-wise equivariances automatically using gradients. In *Advances in Neural Information Processing Systems*, 2023.

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*, 2020.

Dian Wang, Robin Walters, and Robert Platt. $\mathrm{SO}(2)$-equivariant reinforcement learning. In *International Conference on Learning Representations*, 2022a.

Dian Wang, Jung Yeon Park, Neel Sortur, Lawson L. S. Wong, Robin Walters, and Robert Platt. The surprising effectiveness of equivariant models in domains with latent symmetry. In *International Conference on Learning Representations*, 2023a.

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*, 2023b.

Dian Wang, Stephen Hart, David Surovik, Tarik Kelestemur, Haojie Huang, Haibo Zhao, Mark Yeatman, Jiuguang Wang, Robin Walters, and Robert Platt. Equivariant diffusion policy. In *Conference on Robot Learning*, 2024a.

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, 2022b.

Rui Wang, Elyssa Hofgard, Han Gao, Robin Walters, and Tess Smidt. Discovering symmetry breaking in physical systems with relaxed group convolution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50599–50621, 2024b.

YuQing Xie and Tess Smidt. Equivariant symmetry breaking sets. *Transactions on Machine Learning Research*, 2024.

Denis Yarats, Ilya Kostrikov, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2004.13649.

Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2107.09645.

Xupeng Zhu, Dian Wang, Ondrej Biza, Guanang Su, Robin Walters, and Robert Platt. Sample efficient grasp learning using equivariant models. In *Robotics: Science and Systems*, 2022. doi: 10.15607/RSS.2022.XVIII.071.

Martin Zinkevich and Tucker Balch. Symmetry in Markov decision processes and its implications for single agent and multiagent learning. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 632–640, 2001. URL https://dl.acm.org/doi/10.5555/645530.655821.

Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep rl via meta-learning. In *International Conference on Learning Representations*, 2020.

A METRIC AND MEASURABILITY ASSUMPTIONS

The statements proved in Appendix B use the following assumptions. They concern the declared model class and the chosen discrepancy.

A1. Group action. G is finite. Each $g \in G$ acts on $\mathcal{S}$ and on $\mathcal{A}$ by a bimeasurable bijection, written $g \cdot s$ and $g \cdot a$, and we write $g \cdot x = (g \cdot s,\, g \cdot a)$ for $x = (s, a)$. On the context space $\mathcal{V}$, a finite-dimensional normed vector space, G acts linearly and isometrically, so $\|g \cdot \eta\| = \|\eta\|$ and $\|g \cdot \eta_1 - g \cdot \eta_2\| = \|\eta_1 - \eta_2\|$ for all g and all η_1, η_2.

A2. Push-forward. For a probability measure μ on $\mathcal{S}$ and a measurable $C \subseteq \mathcal{S}$, the push-forward is

$$(g_{\#}\mu)(C) = \mu(g^{-1} \cdot C). \tag{9}$$

Because each g acts bimeasurably and G is a group, $g_{\#}$ is a bijection on probability measures with

$$(g^{-1})_{\#} \circ g_{\#} = \mathrm{id}, \qquad g_{\#} \circ (g^{-1})_{\#} = \mathrm{id}. \tag{10}$$

The measure-level identity equation 9 is the definition used throughout. Displayed density identities elsewhere in the paper, such as Equation 1, are shorthand for it: if one additionally assumes a reference measure on $\mathcal{S}$ that every $g \in G$ preserves and with respect to which the kernels admit densities, then the measure-level identity is equivalent to the corresponding density identity holding for almost every next state. That reference measure is an extra assumption and not a consequence of A1. For kernels that admit no density, such as the near-deterministic transitions of the continuous-control tasks studied here, the density form has no literal reading and only equation 9 applies. All proofs below are carried out at the measure level.

A3. Discrepancy. d_P is a discrepancy on probability measures over $\mathcal{S}$ satisfying, for all μ, ν, λ and all $g \in G$,

$$d_P(\mu, \nu) = d_P(\nu, \mu), \qquad d_P(\mu, \nu) \le d_P(\mu, \lambda) + d_P(\lambda, \nu), \qquad d_P(g_{\#}\mu, g_{\#}\nu) = d_P(\mu, \nu). \tag{11}$$

All three are used below. The third is the only *metric-specific* ingredient: it is what has to be checked separately for each choice of d_P, and it holds for W_2 when every g acts as an isometry on $\mathcal{S}$ and for TV by invariance of total variation under measurable bijections.

A4. Joint equivariance. Every member of $\mathcal{F}$ satisfies Equation 3 at the measure level: for every $\theta \in \Theta$, every $g \in G$, every x, and every $\eta \in \mathcal{V}$ with $\|\eta\| \le B_\eta$,

$$K_\theta(\cdot \mid g \cdot x,\, g \cdot \eta) = g_{\#} K_\theta(\cdot \mid x, \eta), \qquad \bar{r}_\theta(g \cdot x,\, g \cdot \eta) = \bar{r}_\theta(x, \eta). \tag{12}$$

Because the action on $\mathcal{V}$ is isometric by A1, the admissible set $\{\eta : \|\eta\| \le B_\eta\}$ is preserved by the action, so the quantifier in equation 12 is closed under the group.

A5. Context moduli. $\omega_P, \omega_r : [0, \infty) \to [0, \infty)$ are nondecreasing and concave with $\omega_P(0) = \omega_r(0) = 0$, and Equation 4 holds for every $\theta \in \Theta$, every x, and every $\eta_1, \eta_2 \in \mathcal{V}$.

A6. Realizability. The observed MDP lies in the GSB class of Equation 5: there exist $\theta^\dagger \in \Theta$ and $\eta^\dagger \in \mathcal{V}$ with $\|\eta^\dagger\| \le B_\eta$ for which the two suprema in Equation 5 are at most δ_P and δ_r. The suprema range over all of $\mathcal{S} \times \mathcal{A}$, so the bounds hold at x and at $g \cdot x$ simultaneously for every $g \in G$.

A7. Second-order conditions for $d_P = W_2$. When d_P is the quadratic Wasserstein distance, $\mathcal{S}$ is a Polish space carrying its Borel σ-algebra, every $g \in G$ acts on $\mathcal{S}$ as an isometry, and every kernel considered takes values in $\mathcal{P}_2(\mathcal{S})$, the set of probability measures with finite second moment. Isometry of the action keeps push-forwards inside $\mathcal{P}_2(\mathcal{S})$ and gives the third identity of equation 11 for W_2; finiteness of second moments makes W_2 finite and the triangle inequality available.

B Proofs of the deviation band and the strict-equivariance floor

Throughout, $x = (s, a)$ ranges over $\mathcal{S} \times \mathcal{A}$ and g ranges over G. Recall the deviation fields of Equation 2, which are defined from the observed environment alone.

Proposition 3 (Context-section deviation band)**.** *Under A1–A6, for every x and every $g \in G$,*

$$\epsilon_P(x; g) \le 2\delta_P + \omega_P\big(\|\eta^\dagger - g^{-1}\cdot\eta^\dagger\|\big), \qquad \epsilon_r(x; g) \le 2\delta_r + \omega_r\big(\|\eta^\dagger - g^{-1}\cdot\eta^\dagger\|\big). \tag{13}$$

Proof. Fix x and g. The context $g^{-1}\cdot\eta^\dagger$ is admissible: the action on $\mathcal{V}$ is isometric by A1, so $\|g^{-1}\cdot\eta^\dagger\| = \|\eta^\dagger\| \le B_\eta$ by A6, which is what allows A4 to be invoked at that context. We first record the identity that converts a group action on the inputs into a movement of the context. Joint equivariance equation 12, applied at the context $g^{-1}\cdot\eta^\dagger$, gives

$$K_{\theta^\dagger}(\cdot \mid g\cdot x,\, \eta^\dagger) = K_{\theta^\dagger}\big(\cdot \mid g\cdot x,\, g\cdot(g^{-1}\cdot\eta^\dagger)\big) = g_{\#}K_{\theta^\dagger}(\cdot \mid x,\, g^{-1}\cdot\eta^\dagger). \tag{14}$$

Applying $(g^{-1})_{\#}$ to both sides and using equation 10,

$$(g^{-1})_{\#}K_{\theta^\dagger}(\cdot \mid g\cdot x,\, \eta^\dagger) = K_{\theta^\dagger}(\cdot \mid x,\, g^{-1}\cdot\eta^\dagger). \tag{15}$$

By the triangle inequality in equation 11,

$$\epsilon_P(x; g) \le \underbrace{d_P\big(P(\cdot \mid x),\, K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger)\big)}_{\text{(i)}} + \underbrace{d_P\big(K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger),\, (g^{-1})_{\#}P(\cdot \mid g\cdot x)\big)}_{\text{(ii)}}. \tag{16}$$

Term (i) is at most δ_P by A6. Splitting (ii) at $(g^{-1})_{\#}K_{\theta^\dagger}(\cdot \mid g\cdot x, \eta^\dagger)$ gives

$$\begin{aligned}\text{(ii)} \le{}& \underbrace{d_P\big(K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger),\, (g^{-1})_{\#}K_{\theta^\dagger}(\cdot \mid g\cdot x, \eta^\dagger)\big)}_{\text{(iii)}} \\ &+ \underbrace{d_P\big((g^{-1})_{\#}K_{\theta^\dagger}(\cdot \mid g\cdot x, \eta^\dagger),\, (g^{-1})_{\#}P(\cdot \mid g\cdot x)\big)}_{\text{(iv)}}.\end{aligned} \tag{17}$$

By equation 15, term (iii) equals $d_P(K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger), K_{\theta^\dagger}(\cdot \mid x, g^{-1}\cdot\eta^\dagger))$, which is at most $\omega_P(\|\eta^\dagger - g^{-1}\cdot\eta^\dagger\|)$ by A5. By invariance of d_P under push-forward and its symmetry, term (iv) equals $d_P(P(\cdot \mid g\cdot x), K_{\theta^\dagger}(\cdot \mid g\cdot x, \eta^\dagger))$, which is at most δ_P by A6 applied at the point $g\cdot x$. Adding the three bounds gives the first inequality of equation 13.

For the reward, joint invariance of $\bar r_\theta$ in Equation 3, applied at the context $g^{-1}\cdot\eta^\dagger$, gives $\bar r_{\theta^\dagger}(g\cdot x, \eta^\dagger) = \bar r_{\theta^\dagger}(x, g^{-1}\cdot\eta^\dagger)$. Hence

$$\begin{aligned}\epsilon_r(x; g) &\le \big|r(x) - \bar r_{\theta^\dagger}(x, \eta^\dagger)\big| + \big|\bar r_{\theta^\dagger}(x, \eta^\dagger) - \bar r_{\theta^\dagger}(g\cdot x, \eta^\dagger)\big| + \big|\bar r_{\theta^\dagger}(g\cdot x, \eta^\dagger) - r(g\cdot x)\big| \\ &\le \delta_r + \omega_r\big(\|\eta^\dagger - g^{-1}\cdot\eta^\dagger\|\big) + \delta_r,\end{aligned} \tag{18}$$

where the middle term used the displayed identity together with A5, and the two outer terms used A6 at x and at $g\cdot x$. This is the second inequality of equation 13. □

Two features of the proof are worth isolating. The two occurrences of δ_P come from realizability at the two distinct points x and $g\cdot x$, which is why the constant is $2\delta_P$ and not δ_P. The context enters only through (iii), so the right-hand side of equation 13 does not depend on x. When $g\cdot\eta^\dagger = \eta^\dagger$ the displacement vanishes and both bounds reduce to $2\delta_P$ and $2\delta_r$; if in addition $\delta_P = \delta_r = 0$, both deviations are zero for every x.

The next statement is the formal content of the strict-equivariance reference used in the main text.

Proposition 4 (Invariant context gives a strictly equivariant section)**.** *Assume A1, A2, and A4, and let $\eta_0 \in \mathcal{V}$ satisfy $\|\eta_0\| \le B_\eta$ and $g \cdot \eta_0 = \eta_0$ for every $g \in G$. Then $(s,a) \mapsto K_\theta(\cdot \mid s,a,\eta_0)$ is a strictly G-equivariant transition kernel and $(s,a) \mapsto \bar{r}_\theta(s,a,\eta_0)$ is a G-invariant reward function, for every $\theta \in \Theta$.*

Proof. Fix θ, g, and x. Since $g \cdot \eta_0 = \eta_0$, substituting η_0 into equation 12 gives

$$K_\theta(\cdot \mid g \cdot x, \eta_0) = K_\theta(\cdot \mid g \cdot x,\, g \cdot \eta_0) = g_{\#} K_\theta(\cdot \mid x, \eta_0), \tag{19}$$

which is strict G-equivariance of the section, and likewise $\bar{r}_\theta(g \cdot x, \eta_0) = \bar{r}_\theta(g \cdot x, g \cdot \eta_0) = \bar{r}_\theta(x, \eta_0)$. □

Lemma 5 (Averaging does not remove the transport bias)**.** *Assume A1–A3 and fix x. Write $T_g = (g^{-1})_{\#} P(\cdot \mid g \cdot x)$ for the law of the transported pair returned to the coordinates of x, so that $T_e = P(\cdot \mid x)$ and $\epsilon_P(x; g) = d_P(T_e, T_g)$. Then every law Q on next states satisfies*

$$\max_{g \in G} d_P(Q,\, T_g) \;\ge\; \tfrac{1}{2} \max_{g \in G} \epsilon_P(x; g). \tag{20}$$

Proof. Let $M = \max_{g \in G} d_P(Q, T_g)$, a maximum over a finite set that includes $g = e$. For every g the triangle inequality gives $\epsilon_P(x; g) = d_P(T_e, T_g) \le d_P(T_e, Q) + d_P(Q, T_g) \le 2M$. Taking the maximum over g gives equation 20. □

The transported copies of one pair therefore admit no common fit: whatever weight a learner places on them, some copy is left at least half the largest deviation at that pair away from the fitted law. The bias is a property of the environment at x rather than sampling noise, which is what Proposition 6 then globalises over the state–action space.

Proposition 6 (Strict-equivariance floor)**.** *Assume A1–A3. Let Q be any strictly G-equivariant transition kernel, that is, $Q(\cdot \mid g \cdot x) = g_{\#} Q(\cdot \mid x)$ for all x and g, and let q be any G-invariant reward function, that is, $q(g \cdot x) = q(x)$. Then*

$$\sup_x d_P(P(\cdot \mid x), Q(\cdot \mid x)) \ge \frac{1}{2} \sup_{x,g} \epsilon_P(x; g), \qquad \sup_x |r(x) - q(x)| \ge \frac{1}{2} \sup_{x,g} \epsilon_r(x; g). \tag{21}$$

Proof. Write $\beta = \sup_x d_P(P(\cdot \mid x), Q(\cdot \mid x))$, with values in $[0, \infty]$. Equivariance of Q gives $(g^{-1})_{\#} Q(\cdot \mid g \cdot x) = Q(\cdot \mid x)$ by equation 10. Fix x and g. By the triangle inequality,

$$\epsilon_P(x; g) \le d_P(P(\cdot \mid x), Q(\cdot \mid x)) + d_P\big((g^{-1})_{\#} Q(\cdot \mid g \cdot x),\, (g^{-1})_{\#} P(\cdot \mid g \cdot x)\big)\,. \tag{22}$$

The first term is at most β. By invariance of d_P under push-forward and its symmetry, the second equals $d_P(P(\cdot \mid g \cdot x), Q(\cdot \mid g \cdot x))$, also at most β. Hence $\epsilon_P(x; g) \le 2\beta$ for every x and g, and taking the supremum over x and g gives the first inequality of equation 21. The reward case is identical with $|r(x) - q(x)|$ in place of d_P and $q(g \cdot x) = q(x)$ in place of the equivariance of Q. □

Both inequalities are read in the extended non-negative reals. They are informative as finite lower bounds when β and the corresponding reward supremum are finite, which A7 provides for $d_P = W_2$. Proposition 6 is a worst-case statement about approximation over $\mathcal{S} \times \mathcal{A}$, and it holds for every strictly equivariant kernel and invariant reward, whatever architecture or training procedure produced them. No strictly equivariant approximation is uniformly closer to the observed environment than half of the largest deviation that environment exhibits.

Corollary 7 (Class separation)**.** *Assume A1–A6 and write $\Delta = \sup_{x,g} \epsilon_P(x; g)$. Every strictly G-equivariant kernel Q satisfies $\sup_x d_P(P(\cdot \mid x), Q(\cdot \mid x)) \ge \Delta/2$, while the member $K_{\theta^\dagger}(\cdot \mid \cdot, \eta^\dagger)$ supplied by A6 satisfies $\sup_x d_P(P(\cdot \mid x), K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger)) \le \delta_P$. Whenever $\delta_P < \Delta/2$, the family approximates the observed environment strictly better than any strictly equivariant kernel, by a margin of at least $\Delta/2 - \delta_P$. The reward statement is identical with ϵ_r and δ_r in place of ϵ_P and δ_P.*

Proof. The first inequality is Proposition 6 with the supremum written as Δ. The second is A6, which supplies $\theta^\dagger$ and $\eta^\dagger$ with that uniform error. Subtracting the second from the first gives the margin. □

C FIXED DIRECTIONS, SUFFICIENCY, AND MINIMALITY OF THE EXTENSION

Let $\rho : G \to O(W)$ be the declared context representation of the task on a finite-dimensional Euclidean space W, let $P_{\text{fix}} = |G|^{-1} \sum_{g \in G} \rho(g)$ be the Reynolds projector, and set $\text{Fix}(G) = \text{range}\, P_{\text{fix}}$ and $\mathcal{V} = \text{Fix}(G)^{\perp}$, whose dimension n is the rank of the breaking representation. This section treats the environment itself as a member of a slice family, that is, a map $w \mapsto P_w$ from W to transition kernels with

$$P_{\rho(g)w}(g \cdot s' \mid g \cdot s,\, g \cdot a) = P_w(s' \mid s, a) \qquad \text{for all } g \in G,\ w \in W,\ (s, a, s'), \tag{23}$$

the environment-side counterpart of A4. The observed environment is P_w for one unknown w.

Lemma 8 (Fixed directions carry no breaking). *Write $w = w_{\text{fix}} + w_{\perp}$ with $w_{\text{fix}} = P_{\text{fix}} w$ and $w_{\perp} = (I - P_{\text{fix}})w \in \mathcal{V}$. Then $w - \rho(g)w = w_{\perp} - \rho(g)w_{\perp}$ for every $g \in G$. If $w \in \text{Fix}(G)$, the kernel P_w is strictly G-invariant and its deviation fields vanish identically.*

Proof. Each $\rho(g)$ is orthogonal and satisfies $\rho(g)P_{\text{fix}} = P_{\text{fix}}$, so $\rho(g)w_{\text{fix}} = w_{\text{fix}}$ and the fixed component cancels from the difference, which gives the first claim. If $w \in \text{Fix}(G)$ then $\rho(g)^{-1}w = w$, and equation 23 applied at $\rho(g)^{-1}w$ gives $P_w(g \cdot s' \mid g \cdot s, g \cdot a) = P_{\rho(g)^{-1}w}(s' \mid s, a) = P_w(s' \mid s, a)$. The deviation fields of Equation 2 are then zero for every x and every g. □

Lemma 8 settles what the extension may discard. The orbit displacement that drives the band of Proposition 3 depends on w only through $w_{\perp}$, and the next lemma reproduces the observed environment from a context in $\mathcal{V}$ alone, so n coordinates suffice and none of them is spent on $\text{Fix}(G)$.

Lemma 9 (Group-moved causes are realizable). *Let the observed environment be $P_{w^\dagger}$, set $\eta^\dagger = (I - P_{\text{fix}})w^\dagger$ with $\|\eta^\dagger\| \le B_\eta$, and define $K^\star(\cdot \mid x, \eta) = P_{P_{\text{fix}}w^\dagger + \eta}(\cdot \mid x)$ for $\eta \in \mathcal{V}$. Then $K^\star$ is jointly equivariant and $K^\star(\cdot \mid x, \eta^\dagger) = P_{w^\dagger}(\cdot \mid x)$ for every x. Hence A6 holds for every δ_P above the approximation error $\inf_{\theta \in \Theta} \sup_x d_P\big(K^\star(\cdot \mid x, \eta^\dagger), K_\theta(\cdot \mid x, \eta^\dagger)\big)$ of the model class at the observed slice.*

Proof. Because $\rho(g)P_{\text{fix}} = P_{\text{fix}}$, the group changes only the breaking component, and equation 23 gives, for every $g \in G$,

$$K^\star(g \cdot s' \mid g \cdot x, \rho(g)\eta) = P_{\rho(g)(P_{\text{fix}}w^\dagger + \eta)}(g \cdot s' \mid g \cdot x) = P_{P_{\text{fix}}w^\dagger + \eta}(s' \mid x) = K^\star(s' \mid x, \eta), \tag{24}$$

which is joint equivariance. At $\eta = \eta^\dagger$ the subscript is $P_{\text{fix}}w^\dagger + (I - P_{\text{fix}})w^\dagger = w^\dagger$. □

A load on one limb is such a cause, because the declared group carries it to the corresponding limb. The next statement settles what the extension may not discard.

Proposition 10 (Minimality of the extension). *Let $U \subseteq \mathcal{V}$ be non-empty and open, and suppose the environments the family must cover include $\{P_w : w \in U\}$ with $P_{w_1} \neq P_{w_2}$ whenever $w_1 \neq w_2$. Let Z be a finite-dimensional G-representation with $\dim Z = k$, let $\varphi : U \to Z$ be continuous, and suppose one member K_θ of a jointly equivariant family over $\mathcal{S} \times \mathcal{A} \times Z$ satisfies $K_\theta(\cdot \mid x, \varphi(w)) = P_w(\cdot \mid x)$ for every $w \in U$ and every x. Then $k \ge n$.*

Proof. The reproduction hypothesis makes φ injective on U: if $\varphi(w_1) = \varphi(w_2)$ then $P_{w_1}(\cdot \mid x) = K_\theta(\cdot \mid x, \varphi(w_1)) = K_\theta(\cdot \mid x, \varphi(w_2)) = P_{w_2}(\cdot \mid x)$ for every x, and the separation hypothesis forces $w_1 = w_2$. Suppose $k < n$. Identify $\mathcal{V}$ with $\mathbb{R}^n$ and Z with $\mathbb{R}^k$, and compose φ with the inclusion $\mathbb{R}^k \hookrightarrow \mathbb{R}^n$ onto the first k coordinates. The result is a continuous injection of the open set $U \subseteq \mathbb{R}^n$ into $\mathbb{R}^n$ whose image lies in a proper linear subspace and therefore has empty interior, contradicting invariance of domain, which makes the image of an open set under a continuous injection open. Hence $k \ge n$. For linear φ the conclusion follows from injectivity alone, since an open subset of $\mathcal{V}$ spans $\mathcal{V}$. □

D ONE-STEP ERROR OF A SYMMETRY-EXPANDED ROLLOUT

The band of Proposition 3 concerns the fixed section $(\theta^\dagger, \eta^\dagger)$, which is the best member of the class rather than the one training returns. The statement below is the corresponding bound for the learned pair $(\hat{\theta}, \hat{\eta})$ evaluated at a transformed state-action pair, which is what a symmetry-expanded rollout generates. Write

$$e_{\hat{\theta},\hat{\eta}}(x) = d_P\big(K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger),\, K_{\hat{\theta}}(\cdot \mid x, \hat{\eta})\big) \tag{25}$$

for the learned section error at x.

Proposition 11 (Symmetry-expanded one-step error)**.** *Under A1–A6, for every x and every $g \in G$,*

$$d_P\big(P(\cdot \mid g \cdot x),\, K_{\hat{\theta}}(\cdot \mid g \cdot x, \hat{\eta})\big) \le \delta_P + e_{\hat{\theta},\hat{\eta}}(x) + \omega_P\big(\|\eta^\dagger - g^{-1} \cdot \eta^\dagger\|\big) + \omega_P\big(\|\hat{\eta} - g^{-1} \cdot \hat{\eta}\|\big)\,, \tag{26}$$

and the same bound holds for the reward with d_P, δ_P, ω_P replaced by absolute reward error, δ_r, and ω_r.

Proof. Fix x and g. By the triangle inequality and A6 applied at $g \cdot x$,

$$d_P\big(P(\cdot \mid g \cdot x), K_{\hat{\theta}}(\cdot \mid g \cdot x, \hat{\eta})\big) \le \delta_P + d_P\big(K_{\theta^\dagger}(\cdot \mid g \cdot x, \eta^\dagger),\, K_{\hat{\theta}}(\cdot \mid g \cdot x, \hat{\eta})\big)\,. \tag{27}$$

Both contexts are admissible after transport, by the isometry argument of Proposition 3, so A4 applies to each and gives $K_{\theta^\dagger}(\cdot \mid g \cdot x, \eta^\dagger) = g_\# K_{\theta^\dagger}(\cdot \mid x, g^{-1} \cdot \eta^\dagger)$ and $K_{\hat{\theta}}(\cdot \mid g \cdot x, \hat{\eta}) = g_\# K_{\hat{\theta}}(\cdot \mid x, g^{-1} \cdot \hat{\eta})$. Invariance of d_P under a common push-forward removes $g_\#$ from both arguments, so the second term equals

$$d_P\big(K_{\theta^\dagger}(\cdot \mid x, g^{-1} \cdot \eta^\dagger),\, K_{\hat{\theta}}(\cdot \mid x, g^{-1} \cdot \hat{\eta})\big)\,. \tag{28}$$

Inserting $K_{\theta^\dagger}(\cdot \mid x, \eta^\dagger)$ and $K_{\hat{\theta}}(\cdot \mid x, \hat{\eta})$ and applying the triangle inequality twice more bounds this by

$$\omega_P\big(\|g^{-1} \cdot \eta^\dagger - \eta^\dagger\|\big) + e_{\hat{\theta},\hat{\eta}}(x) + \omega_P\big(\|\hat{\eta} - g^{-1} \cdot \hat{\eta}\|\big)\,, \tag{29}$$

where the two outer terms use A5 at the fixed and the learned context respectively. Collecting terms gives equation 26. The reward case replaces each kernel by the corresponding reward prediction and each application of A5 and A6 by its reward counterpart; no push-forward is involved. □

The bound separates four contributions: realizability of the class, the error of the learned section relative to the best section, and two context-transport terms. The last two are the price of evaluating a model at a transformed pair while holding its context fixed, one for the fixed section and one for the learned one, and they are what the expanded rollout pays for reusing the symmetry. Both vanish when g stabilises the corresponding context.

For rollouts the accounting is at the level of state-action pairs rather than states. If the policy selects a' at the transformed state $g \cdot s$, the source pair to which equation 26 applies is $x = (s,\, g^{-1} \cdot a')$, since the policy is not required to be equivariant and need not select $g \cdot a$ there.

E FINITE-GROUP AVERAGING

The family is built to be exactly equivariant rather than trained towards equivariance, by averaging each head over the group.

Lemma 12 (Output-level averaging)**.** *Let G be finite and let ρ_{in} and ρ_{out} be representations of G on the input and output spaces. For any map h define*

$$f(x) \;=\; \frac{1}{|G|} \sum_{g \in G} \rho_{\mathrm{out}}(g)^{-1}\, h(\rho_{\mathrm{in}}(g)\, x)\,. \tag{30}$$

Then f is equivariant, so $f(\rho_{\mathrm{in}}(u)x) = \rho_{\mathrm{out}}(u) f(x)$ for every $u \in G$.

Proof. Substituting $\rho_{\mathrm{in}}(u)x$ into equation 30 gives $\frac{1}{|G|}\sum_g \rho_{\mathrm{out}}(g)^{-1} h(\rho_{\mathrm{in}}(gu)x)$, because ρ_{in} is a homomorphism. Writing $k = gu$ and using $\rho_{\mathrm{out}}(g)^{-1} = \rho_{\mathrm{out}}(u)\rho_{\mathrm{out}}(k)^{-1}$ turns the sum into $\rho_{\mathrm{out}}(u)\frac{1}{|G|}\sum_k \rho_{\mathrm{out}}(k)^{-1} h(\rho_{\mathrm{in}}(k)x)$, since k ranges over G as g does. The remaining factor is $f(x)$. □

The deterministic kernel used in our experiments, the velocity field, and the scalar heads are built by equation 30, so Lemma 12 makes each of them exactly equivariant. The construction needs ρ_{out} to be a linear representation and the averaging to commute with the vector-space structure of the output.

F CONTROLLED LOCOMOTION TASKS

Table 2: Controlled task settings, with budgets in real environment steps. Forces act in world coordinates at the target body's centre of mass. $\mathbb{Z}_n$ denotes the cyclic group of order n; Hexapod and Ant use state-space lifts of six-leg and four-leg rotations. Humanoid uses left–right reflection, Walker2d exchanges its two planar legs, and manipulation reflects the task coordinates.

Base task	Declared group	Added force and target	Force (N)	Steps
Locomotion comparison				
Hexapod	$\mathbb{Z}_{12}$	Downward, one leg mount	6.0622	1M
Ant	$\mathbb{Z}_8$	Downward, right rear leg	6.0179	1M
Humanoid	$\mathbb{Z}_2$	Downward, right thigh	128.3648	1M
Walker2d	$\mathbb{Z}_2$	Downward, one thigh	220.4408	1M
Manipulation comparison				
FetchPushDense-v4	$\mathbb{Z}_2$	Lateral ($+y$), object	3.2	500k
FetchSlideDense-v4	$\mathbb{Z}_2$	Lateral ($+y$), object	0.4	500k

Hexapod is our own six-legged MuJoCo robot. Its legs are mounted at 60° intervals around the torso, each with a hip and an ankle motor, so its action space has twelve dimensions. Its total mass is 1.15 kg. Ant, Humanoid, and Walker2d are the `Ant-v2`, `Humanoid-v2`, and `Walker2d-v2` environments of OpenAI Gym (Brockman et al., 2016). We control the added asymmetry by applying a constant downward force to one rigid body throughout interaction. The force is expressed in world coordinates and applied at that body's centre of mass. Its magnitude and location are absent from the observation. The action space and reward function retain their original definitions. The four settings use forces of 6.0622 N on Hexapod's `leg_0_mount`, 6.0179 N on Ant's `right_back_leg`, 128.3648 N on Humanoid's `right_thigh`, and 220.4408 N on Walker2d's `thigh`. Figure 7 shows the loaded bodies and how the declared group moves the load around its orbit.

The group acts on both the robot coordinates and the possible locations of the load. Hexapod and Ant use cyclic groups of orders 12 and 8, respectively, which give linear lifts of the six-leg and four-leg rotations in the state representation. The order doubles because one full rotation maps the observed torso quaternion q to $-q$, the same orientation written as a different vector. Humanoid uses a left–right reflection, and Walker2d uses exchange of its two planar legs; both groups have order two.

G REPLAY-FRACTION SCHEDULES AND RESULT AGGREGATION

GSB uses one synthetic-replay schedule per task, shared by all three seeds. Hexapod uses the fixed schedule, and Ant, Humanoid, and Walker2d use the decaying schedule. Both begin standard model rollouts at 50,000 real environment steps and increase the synthetic fraction linearly to 25% at 150,000 steps, when symmetry expansion begins. The fixed schedule keeps this fraction through the remaining training budget. The decaying schedule holds it until 250,000 steps and reduces it linearly to zero at 500,000 steps, after which model training also stops. Figure 8 contrasts these schedules. Manipulation uses zero synthetic replay throughout.

All methods use training seeds 42, 43, and 44. For a run i, let $\mathcal{T}_i$ contain the available evaluation steps in the inclusive window $[900{,}000, 1{,}000{,}000]$ for locomotion and $[400{,}000, 500{,}000]$ for manipulation, and let $J_i(t)$ be its logged mean evaluation return at step t. We report

$$\bar{J}_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} J_i(t), \qquad \bar{J} = \frac{1}{3} \sum_{i=1}^{3} \bar{J}_i, \qquad s_J = \sqrt{\frac{1}{2} \sum_{i=1}^{3} (\bar{J}_i - \bar{J})^2}. \tag{31}$$

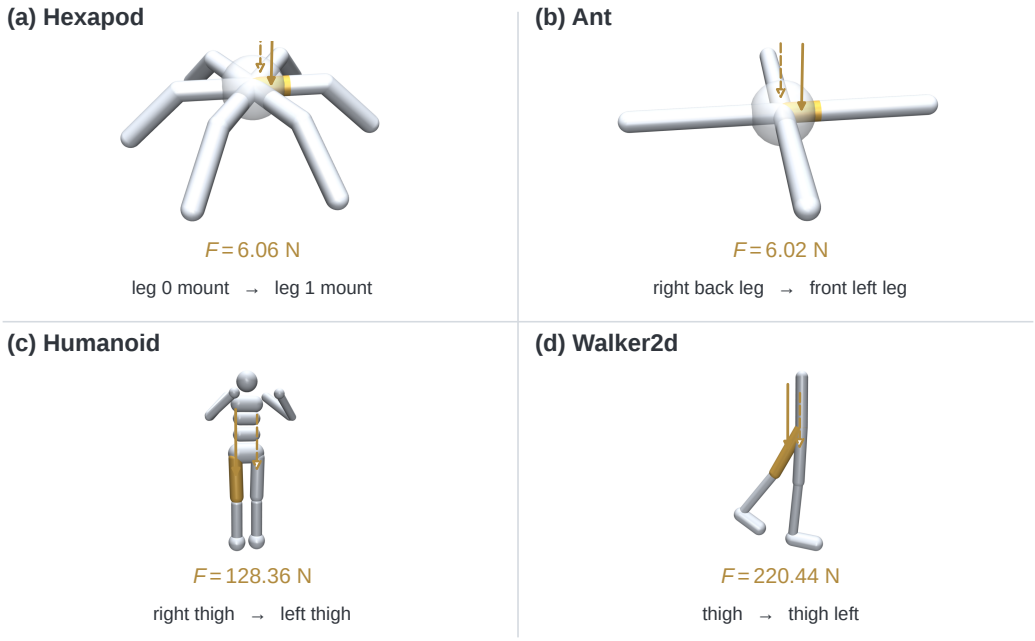

Figure 7: The declared group moves each load to a corresponding body. Solid ochre arrows show the applied force at the loaded body's centre of mass, and dashed arrows show its image under one group element.

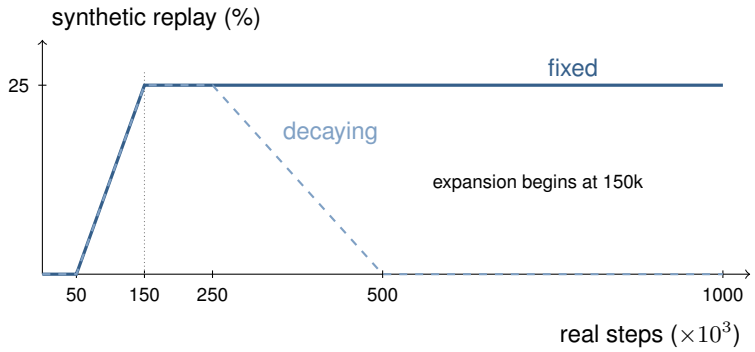

Figure 8: The fixed and decaying synthetic-replay schedules. Model and context updates begin at 1,000 real steps and SAC updates at 10,000 steps; both schedules use the same 50k–150k ramp. Synthetic transitions expire after 10,000 real steps.

Thus each training run receives equal weight, and s_J is the sample standard deviation of run-level final-window returns. The probability of improvement in Section 4 takes, on each task, the baseline with the highest $\bar{J}$, counts the share of the nine pairs of run-level returns in which the GSB run is higher, and averages these shares over the six tasks; its interval resamples runs with replacement within each task and method, 20,000 times (Agarwal et al., 2021). The Humanoid PE-SAC run with seed 42 ends at 980,000 steps and contributes nine evaluations; the table averages its available window, and the curves use the steps all three seeds share.

H Baselines and Settings

All methods use the SAC hyperparameters of Table 3 on every task, with actor and critic networks of two hidden layers of width 256. Vanilla SAC, EMLP, RPP, and PE-SAC run in one JAX implemen-

tation, and GSB runs in PyTorch. EMLP makes the actor and critic equivariant. For its equivariant layers and those of RPP and PE-SAC, the width is a channel budget divided among the group representations. RPP pairs a residual-pathway actor with a plain critic. PE-SAC uses partially equivariant actor and critic networks, a gating network with two hidden layers of width 128, and an auxiliary dynamics network with two hidden layers of width 256. RPP, PE-SAC, and GSB clip gradient norms at 0.5. RPP, EMLP, and PE-SAC add an actor weight decay of 10^{-6}. Replay buffers retain every real transition of a run. Locomotion evaluates ten episodes every 10,000 steps, and manipulation evaluates twenty held-out episodes every 10,000 steps.

Table 3: SAC hyperparameters shared by all methods on all tasks.

Setting	Value
Actor, critic, and temperature learning rates	3×10^{-4}
Discount γ	0.99
Target smoothing τ	0.005
Temperature	learned, initialised at 1.0
Target entropy	$-\lvert\mathcal{A}\rvert/2$
Batch size	256
Updates per environment step	1
Initial steps with uniformly random actions	10,000

GSB updates its model parameters and context with one Adam optimiser at learning rate 3×10^{-4}, gradient-norm clipping at 1.0, and two gradient steps of batch 256 per environment step, starting at 1,000 steps. In locomotion the dynamics model is a deterministic network with two hidden layers of width 256 and tanh activations on (s, a, η), averaged over the group (Appendix E). It predicts the state change in group-whitened coordinates under a squared-error loss, and an invariant reward head of the same size sits beside it. The context carries the regular representation of the physical symmetry, which permutes the six or four leg positions of Hexapod and Ant and the two sides of Humanoid and Walker2d. Restricted to its breaking subspace, it has one dimension fewer than there are positions, whichever body the force loads. It satisfies $\|\eta\| \leq 1$ and starts at zero, the group-fixed point. Model rollouts have horizon one and start every 250 steps from 64 real states, and once symmetry expansion begins each start yields $|G|$ transitions. Table 4 lists the group order, context dimension, and synthetic-replay schedule of each locomotion task. In manipulation the context is fitted by the two-step prediction error of the relative object–gripper velocity, while the model parameters minimise the model and reward losses with $\hat{\eta}$ detached. The actor reads the detached context projected onto the breaking subspace. From 10,000 steps, 128 of the 256 rows in each SAC batch are mirrored real transitions whose relative-velocity slots are corrected, with reward and termination unchanged.

Table 4: GSB settings per locomotion task, identical across seeds. Appendix G defines the two schedules.

	Hexapod	Ant	Humanoid	Walker2d
Group order $\lvert G\rvert$	12	8	2	2
Context dimension	5	3	1	1
Synthetic-replay schedule	fixed	decaying	decaying	decaying

I MANIPULATION CONTROL COMPONENTS

Table 5 specifies the matched controls behind the component gains of Figure 6(d) and behind the comparison with training without transported transitions. Against that last row, the plain mirror changes held-out return by -2.15, -1.48, and -2.10 on FetchSlide and by -0.41, -1.09, and -0.81 on FetchPush for seeds 42, 43, and 44. GSB changes it by $+1.40$, $+3.38$, and $+2.55$ on FetchSlide and by $+1.11$, $+1.38$, and $+0.50$ on FetchPush.

Table 5: Matched controls behind Figure 6(d). Every row uses standard twin critics that do not read the context, and no row generates model rollouts. The correction gain pairs the first two rows, which differ only in whether the relative-velocity slots of the transported transitions are corrected. The last row transports no transitions and is otherwise identical to the first. It is the reference for the losses and gains against training without transported transitions. The actor gain pairs the first and third rows on FetchSlide.

Configuration	Context	Actor	Transported data
GSB, both tasks	Learned	Jointly equivariant; detached $\hat{\eta}$	Half of each batch, corrected slots
Uncorrected mirror	Learned	Same actor	Same share, plain mirror
Plain actor, Slide	Learned, unused by the actor	Unconstrained; state only	Same share, corrected slots
No transported data	Learned	Same actor	None

The correction carries the learned slice's own error in place of the deviation. For deterministic dynamics $s' = F(x)$ and the deterministic model f, joint equivariance gives $f(g \cdot x, g \cdot \hat{\eta}) = g \cdot f(x, \hat{\eta})$, so the corrected slots equal $[f(g{\cdot}x, \hat{\eta}) + g{\cdot}(F(x) - f(x, \hat{\eta}))]_K$. The group acts isometrically and maps the slots K to themselves, as the reflection does for the relative velocity, so their distance to the true outcome $[F(g \cdot x)]_K$ is at most $\|[F(x) - f(x, \hat{\eta})]_K\| + \|[F(g \cdot x) - f(g \cdot x, \hat{\eta})]_K\|$, the learned slice's errors at the observed and at the transformed pair, the second of which equation 7 bounds. The plain mirror differs from the true outcome by $\|[F(g \cdot x) - g \cdot F(x)]_K\|$, the deviation of equation 2 on those slots, whatever the model learns.

J Deterministic kernel

The experiments in this paper use the deterministic kernel $K_\eta = \delta_{f_\eta}$, for which the discrepancy between two context slices is the distance between the predictions themselves,

$$W_2(K_{\eta_1}, K_{\eta_2}) = \|f_{\eta_1} - f_{\eta_2}\|, \tag{32}$$

taken in the ground norm the run declares. The modulus condition equation 4 therefore holds with $\omega_P(t) = L_f t$ whenever every prediction in the family is L_f-Lipschitz in the context, as networks with Lipschitz activations and bounded weights are, so the bounds of Section 3.3 apply to this kernel directly. Because the model predicts in whitened coordinates $y = Ws$, a context modulus L in the whitened metric becomes $\|W^{-1}\|_{\mathrm{op}} L$ in the physical Euclidean metric.

K Locomotion control components

Table 6 compares GSB with a control that runs its Hexapod pipeline with the context pinned. The control shares GSB's code, seeds, SAC settings, model-training start, rollout schedule, synthetic-replay cap, and budget. It keeps the jointly equivariant model and symmetry expansion and pins the context at the group-fixed point 0, the strictly equivariant case of Proposition 4. The frozen-context runs match the GSB runs evaluation for evaluation until model replay begins at 50,000 steps, so their final returns differ from GSB's through the learned context alone. GSB exceeds the frozen-context control in every seed, by 520.4 to 908.1.

Table 6: Hexapod final-window return of GSB and its frozen-context control, seeds 42–44, with the mean and sample standard deviation across seeds. PE-SAC, the strongest baseline in Table 1, is shown for reference.

Configuration	Context	Model	Expansion	Seed 42	Seed 43	Seed 44	Mean $\pm$ SD
GSB	Learned	Jointly equivariant	Yes	4973.2	4688.2	3586.2	4415.8 ± 732.5
Frozen context	Fixed at 0	Jointly equivariant	Yes	4235.8	3780.1	3065.8	3693.9 ± 589.7
PE-SAC	–	–	–	2932.0	3606.4	3418.5	3319.0 ± 348.1