EXCITATION COLLAPSE IN IN-CONTEXT CONTROL: SYMMETRIC FREEZING AND SYMMETRY-BROKEN IMITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context controllers are sequence models pretrained on families of dynamical systems so that they can control an unseen system from its interaction history. They are usually trained to imitate an oracle that knows the system, or to regress the oracle's privileged parameters. It is well known in adaptive control that controllers which hedge against parameter uncertainty can *turn off*: they stop exciting the system and so stop learning about it. We show that imitation pretraining produces this failure exactly whenever the family has a symmetry that the history cannot resolve before the controller acts; unknown actuator polarity is the standard example. The imitation optimum is equivariant, so from a cold start its closed loop stays on the symmetry-fixed actions for all time, which we call *symmetric freezing*. Because the frozen controller minimises the imitation loss over all history-measurable controllers, changing the architecture or adding capacity and data does not avoid it, and better optimisation moves the learner closer to it. The theory predicts degenerate latents in privileged latent distillation, a default direction of predictable sign under asymmetric priors, and exact freezing of equivariant architectures. We confirm each prediction for discrete and continuous symmetry groups and introduce two black-box diagnostics. Since the failure comes from the training target, we change the target. *Symmetry-broken imitation* (SBI) trains an unmodified network on a teacher that commits to one orbit element and revises this choice using exact privileged likelihoods. The teacher is provably consistent, and one round of on-policy relabelling transfers its revisions to the student. With one plain transformer and no knowledge of the symmetry at deployment, SBI removes all instability in a linear family (two seeds, 180 paired runs each), matching an explicit-copy quotient controller. In RMA-style latent distillation on a nonlinear pendulum with unknown motor polarity, changing only the regression target and adding one round of on-policy relabelling reduces falls from 32–43% to 10–13% across two seeds ($p < 10^{-12}$ in each), on par with classical multiple-model control. In a six-state family whose prior no model bank can cover, SBI beats multiple-model control with 1024 models on paired runs across three training seeds ($p = 8 \times 10^{-12}$ pooled).

1 INTRODUCTION

Sequence models can act as general-purpose controllers for families of dynamical systems: given the recent input–output history of an unknown system, they output the next input (Forgione et al., 2023; Bin Mohaya et al., 2026; Cole et al., 2025). In-context reinforcement learning (Laskin et al., 2023; Lee et al., 2023; Lin et al., 2024) and privileged teacher–student distillation in robot learning (Kumar et al., 2021; Chen et al., 2020) follow the same pattern. Training is almost always supervised: the model regresses the action, or the latent parameters, of a teacher that knows the system.

Adaptive control identified the main difficulty decades ago. A controller that minimises expected cost under parameter uncertainty shrinks its input when it is uncertain, so it gains less information, becomes more uncertain, and can eventually *turn off* (Aoki, 1967; Bar-Shalom, 1981; Wittenmark, 2002). Dual controllers were designed to prevent this (Mookerjee et al., 1985), and minimax adaptive controllers resolve a finite set of candidate systems optimally (Rantzer, 2020; 2021). Those

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.

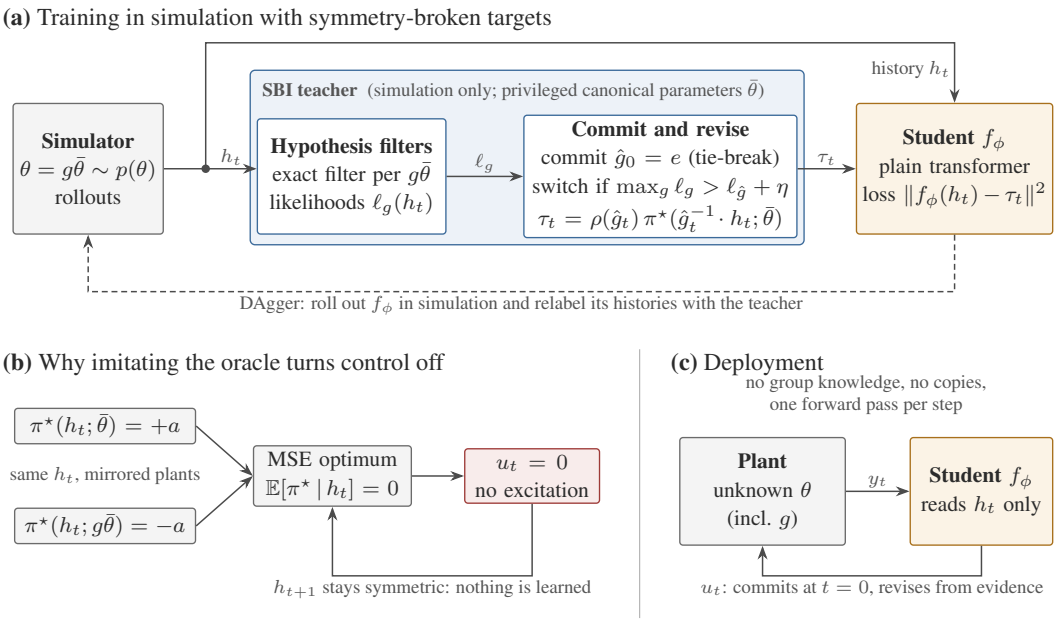

Figure 1: Symmetry-broken imitation (SBI). (a) In simulation, the teacher runs one exact filter per symmetry hypothesis $g\bar{\theta}$ using the privileged canonical parameters $\bar{\theta}$, commits to the identity at the cold start, switches only when another hypothesis leads by η nats, and emits the committed oracle's action τ_t (for latent distillation, the committed latent). The student is an unmodified network trained on τ_t; one DAgger round relabels its own rollouts. (b) The standard target $\pi^\star(h_t; \theta)$ takes opposite values on mirrored plants that share the history, so its MSE optimum is zero and the history stays symmetric (Theorems 1 and 2). (c) At deployment the student acts from the history alone, with no knowledge of the group and no copies.

controllers are designed by hand. In-context controllers are trained, and the controller they converge to is determined by the training objective.

We show that the standard objective makes turn-off exact. Suppose the sign of the input gain is unknown, a classical hard case (Nussbaum, 1983; Rantzer, 2020). Flipping the sign together with all past inputs leaves the distribution of the observations unchanged, while the oracle's action flips. The imitation optimum inherits this symmetry, so at a cold start the only action it can take is zero, and a zero action keeps the history symmetric. The optimal imitation controller therefore never acts. The imitation optimum is the closest history-measurable approximation of the oracle, so better optimisation, larger models or more data cannot change this outcome. The same argument applies to regressed latents, which collapse to degenerate values, and to equivariant architectures, which cannot break symmetry (Smidt et al., 2021; Kaba & Ravanbakhsh, 2023; Lawrence et al., 2025).

The analysis also suggests a remedy. The oracle's target depends on the unresolved group element, and a target that does not depend on it can be imitated without averaging. *Symmetry-broken imitation* (SBI) replaces the oracle by a teacher, available only in simulation, that commits to one orbit element by a fixed tie-break and switches when exact privileged likelihoods favour another element. Its targets are history-measurable up to identifiable parameters, so an unmodified network can learn them, and one round of DAgger (Ross et al., 2011) teaches the network when to revise its commitment. Figure 1 shows the pipeline.

Contributions. On the theory side, we prove that the imitation optimum is equivariant and that its closed loop from a cold start is confined to symmetry-fixed actions (Theorems 1 and 2), and that imitation selects this controller among all history-measurable controllers (Proposition 1). Further results cover asymmetric priors, approximate symmetry and equivariant architectures, and show that the loss of information perpetuates itself; all proofs are in Appendix B. On the method side, we propose symmetry-broken imitation, a change of training target with a consistency guarantee for its teacher (Proposition 5). SBI needs the symmetry only in simulation and leaves the architecture and

the deployment procedure unchanged, so it can replace the target in action imitation and in RMA-style latent distillation. For groups of up to 64 elements the teacher stays exact, and the student is then limited by its in-context control capacity rather than by the symmetry; a larger student reduces this limit. We also study two explicit quotient variants, quotient in-context control (QICC) and quotient latent distillation (QLD), which resolve the orbit with transformed copies of a canonical model. They work only when the in-context predictors are *frame-rigid*; otherwise a copy in the wrong frame adapts to the data and the orbit test loses its power, which we verify in two groups with two seeds each. Empirically, we confirm the predictions of the theory with mirror tests, symmetry-breaking times, a causal test with an asymmetric prior, exact freezing of equivariant controllers, degenerate latents, and a continuous rotation group. Finally, we report four evaluation choices (cold-start protocols, full-history classical baselines, calibrated filters, and hyperparameters selected on validation plants) that each changed our conclusions (Section 5.6).

Scope. We do not claim that learned in-context controllers should replace classical adaptive control. Where a bank of models covers the prior, multiple-model control works very well; we report it next to every result, and it is the stronger method in several of our low-dimensional families. Learned controllers are used for other reasons: they amortise identification across rich families, and they fit into learning pipelines such as privileged teacher–student distillation. For such pipelines the relevant comparison is the same pipeline trained with a different objective, and our claims about SBI concern that comparison.

2 SETTING

A system is indexed by $\theta \sim p(\theta)$. At time t the controller observes y_t and chooses u_t; the history is $h_t = (y_0, u_0, \dots, u_{t-1}, y_t)$. An in-context controller is a map $f : h_t \mapsto u_t$ shared across the family. The oracle $\pi^\star(h; \theta)$ knows θ. Imitation minimises $\mathbb{E}_{(\theta,h)\sim D}\|f(h) - \pi^\star(h;\theta)\|^2$ over histories generated by a behaviour policy. Evaluation uses a *cold start*: control begins at $t = 0$ without warm-up. Regret is the cost difference to the oracle on identical noise, and divergence (or, for the pendulum, a fall) is reported separately.

3 THEORY

A group G acts on parameters ($\theta \mapsto g\theta$), on actions through a linear representation ρ, and on observations through σ; it acts on histories componentwise.

Assumption 1. (i) $p(g\theta) = p(\theta)$. (ii) The dynamics are equivariant: $\sigma(g)y_{t+1}$ given $(g \cdot h_t, \rho(g)u_t)$ under $g\theta$ has the law of y_{t+1} given (h_t, u_t) under θ. (iii) $\pi^\star(g \cdot h; g\theta) = \rho(g)\pi^\star(h;\theta)$. (iv) The behaviour policy is equivariant in distribution.

For unknown polarity, $G = \{\pm 1\}$ acts by $B \mapsto -B$ and $\rho(-1)u = -u$, with σ trivial.

Theorem 1 (Equivariance of the imitation optimum)**.** *Under Assumption 1, $f^\star(h) = \mathbb{E}_D[\pi^\star(h;\theta) \mid h]$ satisfies $f^\star(g \cdot h) = \rho(g)f^\star(h)$ for all g and D-a.e. h. The same holds for the MSE-optimal estimate of any equivariant latent $z(\theta)$.*

Proof. By induction, (i), (ii) and (iv) make the law of (θ, h_t) G-invariant. With $(\theta', h') = g^{-1}(\theta, h)$, which has the same law, $f^\star(g \cdot h) = \mathbb{E}[\pi^\star(g \cdot h'; g\theta') \mid h' = h] = \rho(g)f^\star(h)$ by (iii). □

Theorem 2 (Symmetric confinement and freezing)**.** *If σ is trivial, the closed loop of $f^\star$ from a cold start satisfies $u_t \in \mathrm{Fix}(\rho)$ for all t almost surely. If $\mathrm{Fix}(\rho) = \{0\}$ (sign flips; rotations of inputs of dimension at least two), then $u_t \equiv 0$.*

Proof. The empty history is invariant. If h_t is invariant, Theorem 1 gives $f^\star(h_t) \in \mathrm{Fix}(\rho)$; then h_{t+1} is invariant because $\rho(g)u_t = u_t$ and σ is trivial. □

Proposition 1 (Imitation selects the frozen controller)**.** *For any history-measurable controller c, $\mathbb{E}\|c(h) - \pi^\star(h;\theta)\|^2 = \mathbb{E}\|f^\star(h) - \pi^\star\|^2 + \mathbb{E}\|c(h) - f^\star(h)\|^2$. Every controller that breaks the symmetry at the cold start therefore has strictly larger imitation loss than the frozen optimum, and improving optimisation, capacity, or data moves a learner toward freezing.*

Algorithm 1 Orbit resolution with explicit copies (QICC; QLD replaces predictors by filters)

commit $k \leftarrow e$; scores $S_g \leftarrow 0$
for $t = 0, 1, \dots$ **do**
  for $g \in G$ **do**
    $(a_g, \hat{y}^{(g)}_{t+1}(\cdot)) \leftarrow f_{\text{canon}}(g^{-1} \cdot h_t)$
  end for
  if $t > 0$ **then**
    $S_g \leftarrow \gamma S_g + \log p_g(y_t)$ for all g
  end if
  $g^\star \leftarrow \arg\max S_g$; **if** $S_{g^\star} > S_k + \eta$ **then** $k \leftarrow g^\star$
  apply $u_t = \rho(k) a_k$
end for

Corollary 1 (Equivariant architectures freeze)**.** *Any policy constrained to satisfy $f(g \cdot h) = \rho(g) f(h)$, whatever its parameters or training, obeys the conclusion of Theorem 2. This is the closed-loop, sequential counterpart of the static fact that equivariant networks cannot break the symmetries of their input (Smidt et al., 2021; Kaba & Ravanbakhsh, 2023; Lawrence et al., 2025).*

Proposition 2 (Asymmetric priors set the default)**.** *Let $p = \pi p_+ + (1 - \pi) p_-$ with $p_- = g_\# p_+$ for $G = \{\pm 1\}$. At any invariant history, $f^\star(h) = (2\pi - 1)\, m(h)$ with $m(h) = \mathbb{E}_{p_+}[\pi^\star(h; \theta) \mid h]$.*

Proposition 3 (Approximate symmetry, approximate freezing)**.** *If $p = (1 - \varepsilon) p_{\text{sym}} + \varepsilon q$ with p_{sym} satisfying Assumption 1 and $\text{Fix}(\rho) = \{0\}$, then at a symmetric history $\|f^\star(h)\| \le \varepsilon(h) \sup \|\pi^\star\|$, where $\varepsilon(h)$ is the posterior weight of q (ε at a cold start).*

Proposition 4 (Freezing perpetuates itself)**.** *For $G = \{\pm 1\}$ and $y = Cx + v$ with $v \sim \mathcal{N}(0, s^2)$, an input u_t yields at most $2(CBu_t)^2/s^2$ nats of information about g to first order. A frozen controller gains none; a controller committed to $|u_t| \ge \bar{u}$ gains at least $2(CB\bar{u})^2/s^2$ per step.*

Symmetry-broken imitation. Let G be finite and let $\bar{\theta}$ denote the canonical (orbit-representative) parameters, which are identifiable in context. A *symmetry-broken teacher* computes, from the history and the privileged $\bar{\theta}$, the log-likelihoods $\ell_g(h_t)$ of every hypothesis $g\bar{\theta}$. It commits to the identity at the cold start, switches to $\arg\max_g \ell_g$ when that hypothesis exceeds the committed one by a margin η, and acts as the oracle of the committed hypothesis: $\tau(h, \bar{\theta}) = \rho(\hat{g})\, \pi^\star(\hat{g}^{-1} \cdot h; \bar{\theta})$. For latent distillation its target is the committed latent $z(\hat{g}\bar{\theta})$.

Proposition 5 (SBI removes confinement; the teacher is consistent)**.** *(a) The teacher's target is a function of $(h, \bar{\theta})$ only, so the student's MSE optimum $\mathbb{E}[\tau \mid h]$ averages over the canonical parameters but not over the orbit. At the cold start it equals the canonical controller's action $\mathbb{E}[\pi^\star(h_0; \bar{\theta})]$, which is generically outside $\text{Fix}(\rho)$, so Theorem 2 does not apply. (b) With exact likelihoods, once the teacher is committed to the true element g_0 it ever switches away with probability at most $(|G| - 1)e^{-\eta}$, since each likelihood ratio against g_0 is a nonnegative martingale with unit mean (Ville's inequality). (c) While committed to a wrong element with inputs $|u_t| \ge \bar{u}$, the log-likelihood gap grows at the rate of Proposition 4, so the expected time to leave is at most of order $\eta s^2/(2(CB\bar{u})^2)$.*

Imitation from teacher-generated histories suffers covariate shift once the student acts. The standard remedy of rolling out the student and relabelling its histories with the teacher's targets (Ross et al., 2011) applies directly, because the teacher's selector runs on any history.

Explicit quotient variants. QICC trains a canonical controller with a one-step output predictor on the canonical family. At deployment it runs one copy per group element, scores the copies by prediction error, and commits with hysteresis (Algorithm 1). QLD does the same with a canonical latent regressor and one filter per hypothesis. For the orbit test to have power, a copy in the wrong frame must be unable to fit the data. This requires *frame-rigid* predictors, whose input gains are constrained to the canonical sign pattern. For continuous groups we resolve a discretised orbit. SBI can be read as distilling such a controller, with exact privileged likelihoods, into a single network.

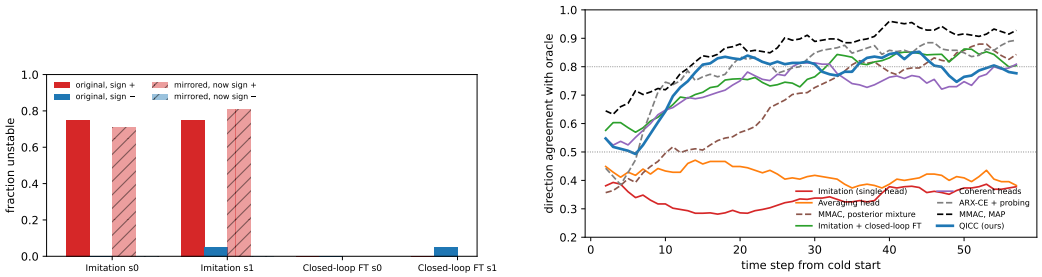

Figure 2: Left: mirror test. Failures of imitation-trained controllers follow the sign of the input gain. Right: direction agreement with the oracle along each controller's own cold-start trajectory. Averaging controllers never break the symmetry.

4 Diagnostics

Mirror test. Evaluate a controller on held-out systems and on their symmetry-transformed twins under identical noise. Failures that move with the transformation reveal a fixed default direction.

Symmetry-breaking time. Along the controller's own trajectory, record whether its action agrees in direction with the oracle's action on the same history. The symmetry-breaking time is the first step at which a 5-step average of this agreement reaches 0.8. For continuous groups we report the cosine between the two actions (chance level 0) and the action magnitude relative to the oracle's.

5 Experiments

All learned controllers are two-layer transformers (width 64) trained for 2000–3000 steps. Flagship comparisons use 60 new plants $\times$ 3 noise seeds (180 paired runs); other results use 15 held-out plants $\times$ 3 seeds. Classical hyperparameters were tuned on the evaluation plants, which favours the classical baselines; pendulum selection margins were chosen on disjoint validation plants. In paired comparisons we report the number of runs in which each controller does better. Appendix D gives 95% confidence intervals for every headline number, together with compute and hyperparameters.

5.1 Unknown polarity in linear systems

Plants are $x_{t+1} = Ax_t + Bu_t + w_t$, $y_t = x_{1,t} + v_t$, with an observed mode that is usually unstable and a random sign of B (Appendix A).

Imitation-trained transformers destabilise 40% and 42% of cold-start runs in two training seeds, and their early actions agree in direction with the oracle's only at chance level (0.49–0.50). The mirror test (Figure 2, left) shows that the failures concentrate on positive-gain plants (75% vs. 0–5%) and follow the sign when the dynamics are mirrored (71% and 81% vs. 0%). The trained networks thus realise a fixed default direction, a finite-sample departure from the exact frozen optimum. As a causal test we set $P(+) = 0.9$, for which Proposition 2 predicts a default toward $+$. In two seeds the trained controller destabilises 0% of positive-gain plants and 95–100% of negative-gain ones.

We then train a plain single-output transformer, the same architecture that turns off under oracle targets, on SBI targets. It commits immediately. Before DAgger it rarely learns *when* to switch, because phases of wrong commitment are short in teacher-generated data (seed 0: 0% unstable on positive-gain plants, 93% on negative-gain ones). One DAgger round removes all instability in both seeds (Table 1). On 60 new plants $\times$ 3 seeds, SBI+DAgger is 0% unstable in both seeds (median regret 5.53 and 5.14), as is frame-rigid QICC (5.79 and 4.29). In paired comparisons SBI is equal to or better than QICC (better in 101–112 of 180 runs, p between 0.12 and 0.001), and it uses one network and no knowledge of the group at deployment. In this low-dimensional family classical MMAC remains the best controller (median regret 2.11; better than SBI in 143 of 180 runs).

Table 1: Two-state linear family, cold start (15 plants × 3 seeds). SB: symmetry-breaking time.

Controller	median regret	unstable	SB
Imitation transformer (2 seeds)	3.0–4.5	40–42%	never
Imitation + closed-loop fine-tuning (2 seeds)	7.6–8.0	0–2%	30
QICC with free predictors (2 seeds)	10.5–38.6	0–49%	13
QICC, frame-rigid (2 seeds)	4.2–11.5	0%	–
SBI, teacher targets only (2 seeds)	2.0–3.5	22–31%	–
SBI + one DAgger round (2 seeds)	4.69–4.70	0%	–
ARX-CE with probing (classical)	14.4	0%	17
MMAC, MAP, 256–1024 models (classical)	1.71–1.90	0%	12

5.2 A HARD REGIME: WHEN A MODEL BANK CANNOT COVER THE PRIOR

In a six-state family with several unstable modes, a random sign and a 42-dimensional prior, a bank of 1024 models cannot cover the prior, whereas QICC needs only two hypotheses. On 60 new plants × 3 seeds, QICC destabilises 40–44% of runs across three training seeds, against 62% for MMAC (pooled paired comparison 217 vs. 121, $p = 2 \times 10^{-7}$). Against model-free ARX-CE the comparison is even (220 vs. 200, $p = 0.35$): ARX-CE diverges less often (28%), but on the runs that survive QICC's regret is one to two orders of magnitude lower. In this family the limiting factor is the quality of the canonical controller at this model size; orbit resolution is not the bottleneck.

With the plain single-output architecture and one DAgger round, SBI destabilises 37–44% of runs across three training seeds, against 62% for MMAC. Pooled over seeds it beats the 1024-model bank on the same runs (228 vs. 104, $p = 8 \times 10^{-12}$), and each seed does so individually. Against ARX-CE the pooled comparison is 234 vs. 181 ($p = 0.011$). When a bank cannot cover the prior, SBI therefore does better than bank-based control, and it needs no copies and no knowledge of the group at deployment.

5.3 OTHER GROUPS: PERMUTATIONS, ROTATIONS, AND FRAME RIGIDITY

Permutation. Two actuators with opposing effects are assigned to channels by an unknown permutation, so $\mathrm{Fix}(\rho)$ is the diagonal. Imitation destabilises 40% of runs. QICC's canonical controller transfers (0% unstable when the permutation is given), but with free predictors orbit resolution fails (96% unstable): in the wrong frame, the in-context predictors infer swapped actuator effects and predict as well as the correct copy. Frame-rigid predictors restore 0% instability in two seeds (median regret 5.63 and 8.59); classical MMAC reaches 1.75. In the sign family, free predictors were reliable in one seed and failed in another (33–49% unstable), and frame rigidity fixed both.

Rotation (continuous group). Two inputs act through $B_0 R(\varphi)$ with unknown $\varphi \sim U[0, 2\pi)$, and Theorem 2 predicts complete freezing. On 45 plants × 3 seeds, imitation destabilises 40% of runs. Its early actions are shrunk to 0.29 of the oracle's magnitude and carry no directional information (cosine 0.05, later -0.02), and its failure rate depends on the unknown angle (12–67% across quadrants). QICC over 8 rotated copies is 0% unstable in every quadrant, acts at full magnitude, and aligns with the oracle as the orbit is resolved (cosine 0.48). MMAC reaches a lower regret (4.2 vs. 13.4, both 0% unstable).

5.4 SCALING THE GROUP: WHERE THE STUDENT'S CAPACITY ENDS

We give m actuators independent unknown polarities, $G = \{\pm 1\}^m$ with $|G| = 2^m$, in coupled plants $x_{t+1} = Ax_t + B_0 \,\mathrm{diag}(s)\, u_t + w_t$ with some unstable modes and m noisy sensors, so that $2m^2$ parameters are unknown. Since $\mathrm{Fix}(\rho) = \{0\}$, imitation must freeze. The SBI teacher tests all 2^m hypotheses with exact Kalman filters, while the student runs one forward pass per step regardless of the size of the group.

The teacher is 0% unstable at every size, so its targets remain correct as the group grows, and SBI with DAgger improves on imitation at every size (Table 2). The student is not reliable in this multi-output family, however, and the all-positive column shows why: most of its failures occur on

Table 2: Independent unknown polarities of m actuators, 30 plants $\times$ 3 noise seeds: unstable runs (median regret in brackets). The all-positive column evaluates the SBI+DAgger student on plants whose polarities are all positive, where the teacher's first commitment is correct and it never switches.

m	$\lvert G\rvert$	unknowns	imitation	SBI	SBI+DAgger	all-positive	SBI teacher	MMAC (256)
2	4	8	42%	36%	19%	13%	0% (1.25)	0% (1.57)
4	16	32	92%	83%	74%	68%	0% (2.27)	0% (8.54)
6	64	72	97%	98%	93%	88%	0% (3.34)	0% (24.1)

Table 3: Pendulum with unknown motor polarity, cold start, 60 new plants $\times$ 3 noise seeds.

Controller	fell	median regret
Zero control	100%	–
Exactly equivariant imitation (Corollary 1)	100%	–
Action imitation (2 seeds)	78.9–84.4%	–
RMA-style latent regression (2 seeds)	32.2–42.8%	4.1–4.4
SBI-RMA, committed-latent targets, same model (2 seeds)	20.6–25.0%	0.94–4.33
SBI-RMA + one DAgger round (2 seeds)	10.0–13.3%	1.14–1.19
SBI-RMA + a second DAgger round (seed 0)	10.6%	1.02
QLD, explicit copies (2 seeds)	3.9–5.0%	0.84–2.26
MMAC, 256 models, calibrated EKFs (classical)	9.4%	0.99

all-positive plants, where no symmetry has to be broken. The share of failures attributable to the symmetry stays at about 5–6 percentage points, while the share that remains without any symmetry grows from 13% to 88% with the number of unknowns. At this model size the binding constraint is in-context identification and control of coupled multi-input systems. Additional capacity helps: a student with twice the depth, trained twice as long and without DAgger, is 57% unstable with four actuators (27% on all-positive plants), compared with 83% for the base model before DAgger and 68% on all-positive plants for the base model after DAgger. A bank of 256 models is stable at every size, and its regret grows with the number of unknowns.

5.5 Nonlinear plant and privileged latent distillation

The plant is an inverted pendulum $\ddot{\theta} = a \sin\theta - b\dot{\theta} + s\,k\,\mathrm{sat}(u)$ with unknown a, b, k and polarity s, a saturating input, and a noisy angle sensor. A run fails if the pendulum falls. Every filtering method uses the same calibrated time-varying EKF. RMA-style students regress the privileged latent (a, b, sk) and control with it (Kumar et al., 2021).

Table 3 confirms three predictions of the theory and shows the effect of SBI. First, exactly equivariant imitation drops every pendulum. Second, RMA-style students collapse their early estimate of the signed gain toward zero (median magnitude 1.40–1.96 against true values ≥ 3) and drop 32–43% of pendulums. Third, changing only the regression target to the SBI teacher's committed latent reduces falls to 20.6–25.0% (seed 0: runs lost only by SBI 11 vs. only by the RMA-style student 51, $p = 3 \times 10^{-7}$), and one DAgger round brings them to 10.0% and 13.3% (runs lost only by SBI vs. only by the RMA-style student: 5 vs. 64 and 4 vs. 57, $p < 10^{-12}$ in both seeds; paired regret 144 vs. 23 and 135 vs. 25). SBI-RMA is then statistically indistinguishable from classical MMAC in both seeds (falls 13 vs. 12 and 22 vs. 15, $p = 1$ and 0.32; regret 85 vs. 90 and 82 vs. 96, $p = 0.76$ and 0.33). A second DAgger round does not reduce falls further (10.6%); the remaining failures concentrate on plants whose polarity is opposite to the first commitment. QLD, which resolves the orbit with explicit copies, remains the safest learned controller (3.9–5.0%). Against SBI-RMA, runs lost only by SBI vs. only by QLD are 13 vs. 4 and 16 vs. 1 ($p = 0.049$ and 3×10^{-4}), and pooled against MMAC 14 vs. 32 ($p = 0.011$).

5.6 Evaluation pitfalls

Four protocol choices changed our conclusions. (1) Oracle warm-ups give full-history classical estimators informative data (ARX-CE regret 0.55 with warm-up, 8.4 from a cold start). (2) Baselines restricted to the learned controller's context window made a learned controller look better than classical control. (3) Steady-state filters over-weight the start-up transient and caused false switching. (4) Open-loop metrics such as the imitation loss and copy probes repeatedly failed to predict closed-loop behaviour; Proposition 1 explains why the imitation loss in particular is misleading.

6 Related work

Dual and adaptive control. Turn-off of cautious controllers is a classical topic (Aoki, 1967; Bar-Shalom, 1981; Wittenmark, 2002). Related work includes dual controllers that prevent it (Feldbaum, 1960; Mookerjee et al., 1985), including for an integrator with unknown gain (Åström & Helmersson, 1986) and with multiple models (Wenk & Bar-Shalom, 1980); the pitfalls of certainty equivalence (Polderman, 1986); Nussbaum gains (Nussbaum, 1983); multiple-model and supervisory switching control (Magill, 1965; Athans et al., 1977; Narendra & Balakrishnan, 1997; Morse, 1996; Hespanha et al., 2003); minimax adaptive control for unknown signs and finite model sets (Rantzer, 2020; 2021); and posterior sampling for LQ control (Abeille & Lazaric, 2018; Simchowitz & Foster, 2020). We show that learned controllers inherit turn-off exactly from their training objective, and we propose a remedy that changes the objective.

In-context control and RL. Transformers identify and control dynamical systems in context (Forgione et al., 2023; Bin Mohaya et al., 2026; Cole et al., 2025; Garg et al., 2022), and supervised pretraining yields in-context decision makers (Laskin et al., 2023; Lee et al., 2023; Lin et al., 2024; Raventós et al., 2023); exploration can use epistemic uncertainty in the predictive distribution (Dai et al., 2024). Meta-RL trained with reinforcement learning can learn Bayes-adaptive exploration (Duan et al., 2016; Zintgraf et al., 2020) or sample a task belief once per episode (Rakelly et al., 2019), which breaks the symmetry consistently over the episode. SBI obtains commit-and-revise behaviour from supervised targets alone. Sampling a new action at every step breaks the symmetry only at random from step to step, whereas control needs a commitment that persists and can be revised.

Privileged imitation. Imitating privileged experts averages away their information, which is known as the imitation gap (Weihs et al., 2021); related effects include latching and confounding (Swamy et al., 2022; Ortega et al., 2021; de Haan et al., 2019; Wen et al., 2020) and experts that no learner can reproduce (Walsman et al., 2023). Teacher–student distillation with privileged latents is standard in robot adaptation (Kumar et al., 2021; Chen et al., 2020). SBI is a closed-form way of making the teacher reproducible under symmetric ambiguity, with a consistency guarantee and without reinforcement learning.

Symmetry. Equivariant networks cannot break the symmetries of their input (Smidt et al., 2021; Kaba & Ravanbakhsh, 2023); randomised canonicalisation represents symmetry-broken outputs while retaining equivariance in distribution (Lawrence et al., 2025), and canonicalisation or frame averaging make arbitrary networks equivariant (Kaba et al., 2023; Puny et al., 2022). MDP homomorphic networks use symmetries of the state–action space (van der Pol et al., 2020). In our setting the symmetry acts on the unknown system, and building equivariance to it into the controller is what causes freezing. In control the broken symmetry must persist over time and be revised from closed-loop evidence; SBI's teacher and QICC's copies provide this.

7 Limitations

The models are small, and several results rest on one or two training seeds, including the pendulum SBI results (two seeds). SBI and the quotient variants need the symmetry group in simulation, and continuous groups require discretisation. In low-dimensional priors classical multiple-model control is competitive or better, and on the pendulum the explicit-copy QLD remains safer than SBI-RMA; a second DAgger round did not close this gap. In a coupled multi-output family with independent

polarities of 2–6 actuators, the SBI student is unreliable even where no symmetry has to be broken, while the teacher succeeds; at this model size, in-context control of such systems is the binding constraint. Trained imitation networks realise a fixed default direction instead of the exact frozen optimum, and the theory does not predict which direction they choose.

AI USE STATEMENT

AI-assisted tools were used during the preparation of this manuscript for language editing, grammar and style refinement, and assistance with software/code development and debugging. The authors independently designed the methodology, conducted and verified the experiments, analyzed and interpreted the results, and take full responsibility for the final manuscript and associated code. All AI-assisted material was reviewed and verified by the authors before use.

ETHICS STATEMENT

This work studies the safety of learned controllers. The failure mode we analyse does not show up in open-loop metrics and could cause physical damage if a controller with this failure were deployed; our diagnostics are intended to detect it before deployment.

REPRODUCIBILITY STATEMENT

All code, models, evaluation plants (with fixed random seeds), and raw per-run results are included in the supplementary material, with one script per experiment. The plant families and baselines are specified in Appendix A, complete proofs are in Appendix B, and statistical procedures, compute and hyperparameters are in Appendix D. All experiments run on a single CPU.

REFERENCES

Marc Abeille and Alessandro Lazaric. Improved regret bounds for Thompson sampling in linear quadratic control problems. In *ICML*, 2018.

Masanao Aoki. *Optimization of Stochastic Systems: Topics in Discrete-Time Systems*. Academic Press, 1967.

Karl Johan Åström and Anders Helmersson. Dual control of an integrator with unknown gain. *Computers & Mathematics with Applications*, 12(6, Part A):653–662, 1986.

Michael Athans, David Castanon, Keh-Ping Dunn, C. S. Greene, Wing H. Lee, Nils R. Sandell, Jr., and Alan S. Willsky. The stochastic control of the F-8C aircraft using a multiple model adaptive control (MMAC) method. *IEEE Transactions on Automatic Control*, 1977.

Yaakov Bar-Shalom. Stochastic dynamic programming: Caution and probing. *IEEE Transactions on Automatic Control*, 26(5):1184–1195, 1981.

Turki Bin Mohaya, Maitham F. AL-Sunni, John M. Dolan, and Peter Seiler. Generalizable optimal control with transformers: One policy across diverse systems. arXiv preprint arXiv:2603.14910, 2026.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, 2020.

Frank Cole, Yuxuan Zhao, Yulong Lu, and Tianhao Zhang. In-context learning of linear dynamical systems with transformers: Approximation bounds and depth-separation. In *Advances in Neural Information Processing Systems*, 2025.

Zhenwen Dai, Federico Tomasi, and Sina Ghiassian. In-context exploration-exploitation for reinforcement learning. *arXiv:2403.06826*, 2024.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *NeurIPS*, 2019.

Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779*, 2016.

A. A. Feldbaum. Dual control theory. *Automation and Remote Control*, 1960.

Marco Forgione, Filippo Pura, and Dario Piga. From system models to class models: An in-context learning paradigm. *IEEE Control Systems Letters*, 2023.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *NeurIPS*, 2022.

João P. Hespanha, Daniel Liberzon, and A. Stephen Morse. Hysteresis-based switching algorithms for supervisory control of uncertain systems. *Automatica*, 2003.

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

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *ICML*, 2023.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Robotics: Science and Systems*, 2021.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *ICLR*, 2023.

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

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. In *NeurIPS*, 2023.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. In *ICLR*, 2024.

D. T. Magill. Optimal adaptive estimation of sampled stochastic processes. *IEEE Transactions on Automatic Control*, 1965.

Purusottam Mookerjee, Yaakov Bar-Shalom, and John A. Molusis. Dual control and prevention of the turn-off phenomenon in a class of MIMO systems. In *Proceedings of the 24th IEEE Conference on Decision and Control*, 1985.

A. Stephen Morse. Supervisory control of families of linear set-point controllers, part I: Exact matching. *IEEE Transactions on Automatic Control*, 1996.

Kumpati S. Narendra and Jeyendran Balakrishnan. Adaptive control using multiple models. *IEEE Transactions on Automatic Control*, 1997.

Roger D. Nussbaum. Some remarks on a conjecture in parameter adaptive control. *Systems & Control Letters*, 1983.

Pedro A. Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, Tom Everitt, Corentin Tallec, Emilio Parisotto, Tom Erez, Yutian Chen, Scott Reed, Marcus Hutter, Nando de Freitas, and Shane Legg. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv:2110.10819*, 2021.

Jan Willem Polderman. On the necessity of identifying the true parameter in adaptive LQ control. *Systems & Control Letters*, 1986.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *ICLR*, 2022.

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

Anders Rantzer. Minimax adaptive control for state matrix with unknown sign. *IFAC-PapersOnLine*, 53(2):58–62, 2020.

Anders Rantzer. Minimax adaptive control for a finite set of linear systems. In *Learning for Dynamics and Control*, pp. 893–904. PMLR, 2021.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *NeurIPS*, 2023.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 627–635, 2011.

Max Simchowitz and Dylan J. Foster. Naive exploration is optimal for online LQR. In *ICML*, 2020.

Tess E. Smidt, Mario Geiger, and Benjamin Kurt Miller. Finding symmetry breaking order parameters with Euclidean neural networks. *Physical Review Research*, 2021.

Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Sequence model imitation learning with unobserved contexts. In *NeurIPS*, 2022.

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.

Aaron Walsman, Muru Zhang, Sanjiban Choudhury, Dieter Fox, and Ali Farhadi. Impossibly good experts and how to follow them. In *International Conference on Learning Representations*, 2023.

Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander Schwing. Bridging the imitation gap by adaptive insubordination. In *NeurIPS*, 2021.

Chuan Wen, Jierui Lin, Trevor Darrell, Dinesh Jayaraman, and Yang Gao. Fighting copycat agents in behavioral cloning from observation histories. In *NeurIPS*, 2020.

Carl J. Wenk and Yaakov Bar-Shalom. A multiple model adaptive dual control algorithm for stochastic systems with unknown parameters. *IEEE Transactions on Automatic Control*, 25(4):703–710, 1980.

Björn Wittenmark. Adaptive dual control. In Heinz Unbehauen (ed.), *Control Systems, Robotics and Automation, Encyclopedia of Life Support Systems (EOLSS)*. Eolss Publishers, 2002.

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 Families and baselines

Two-state family: $A_{11} \sim U[0.95, 1.2]$, $A_{12} \sim U[0.2, 0.6]$, $A_{21} \sim U[-0.3, 0.3]$, $A_{22} \sim U[0.3, 0.8]$, $B = s(U[0.2, 0.6], U[0.5, 1])^\top$, $\sigma_w = 0.1$, $\sigma_v = 0.2$, cost y^2+u^2. Hard family: six states, about half of the diagonal from $U[0.9, 1.15]$, couplings $U[-0.25, 0.25]$, $A_{1j} \sim U[0.2, 0.5]$, $B = s\,U[0.3, 1]^6$. Pendulum: $a \in [6, 14]$, $b \in [0.2, 1]$, $k \in [3, 8]$, $|u| \leq 2$, time step 0.05 s, angle noise 0.05, $|\theta_0| \leq 0.4$. SBI teachers use exact Kalman filters (linear families) or calibrated EKFs (pendulum) with the true canonical parameters, margin $\eta = 2$ nats, and a tie-break to the identity. The behaviour policy follows the teacher's committed action with dither 0.3, and half of the episodes start under the controller of a random plant from the full family. The DAgger round rolls out the student with dither on 1200–1500 fresh plants, relabels every visited history with the teacher, and fine-tunes on the union with the original data (2000 steps). MMAC uses models sampled from the prior, each with its own filter, and innovation likelihoods.

B PROOFS

Proposition 1. For a history-measurable c, write $c - \pi^\star = (c - f^\star) + (f^\star - \pi^\star)$. Because $f^\star(h) = \mathbb{E}[\pi^\star \mid h]$, the cross term $\mathbb{E}[(c(h) - f^\star(h))^\top \mathbb{E}[f^\star(h) - \pi^\star \mid h]]$ vanishes, which gives the decomposition. The cold-start history is invariant, so $f^\star(h_0) \in \mathrm{Fix}(\rho)$ by Theorem 1; if $c(h_0) \notin \mathrm{Fix}(\rho)$ with positive probability, then $\mathbb{E}\|c - f^\star\|^2 > 0$ and the imitation loss of c is strictly larger. □

Corollary 1. The induction in the proof of Theorem 2 uses only the equivariance of the policy, which holds by construction. □

Proposition 2. The posterior weight of the p_+ component at h is $\pi p_+(h)/(\pi p_+(h) + (1 - \pi)p_-(h))$, where $p_\pm(h)$ is the marginal likelihood of the history under each component. By Assumption 1, $p_-(h) = p_+(g^{-1} \cdot h)$, which equals $p_+(h)$ at an invariant history, so the weight is π. Within the p_- component, $\mathbb{E}_{p_-}[\pi^\star(h;\theta) \mid h] = \rho(g)\,\mathbb{E}_{p_+}[\pi^\star(g^{-1} \cdot h;\theta) \mid g^{-1} \cdot h] = -m(h)$ at an invariant history, by (iii). Hence $f^\star(h) = \pi m(h) - (1 - \pi)m(h) = (2\pi - 1)m(h)$. □

Proposition 3. By total expectation over the mixture components, $f^\star(h) = (1 - \varepsilon(h))\mathbb{E}_{p_{\mathrm{sym}}}[\pi^\star \mid h] + \varepsilon(h)\mathbb{E}_q[\pi^\star \mid h]$. At a symmetric history the first term lies in $\mathrm{Fix}(\rho) = \{0\}$ by Theorem 1 applied to p_{sym}, and the second is bounded by $\varepsilon(h) \sup \|\pi^\star\|$. When the first observation has the same law under both components, as in all our families, $\varepsilon(h_0) = \varepsilon$. □

Proposition 4. Along a symmetric history the filtered state estimates under θ and $-\theta$ coincide, so the two one-step predictive distributions of y_{t+1} given (h_t, u_t) are Gaussian with a common variance s^2 and means differing by $2CBu_t$ (to first order in the estimation error). The Kullback–Leibler divergence between $\mathcal{N}(\mu_1, s^2)$ and $\mathcal{N}(\mu_2, s^2)$ is $(\mu_1 - \mu_2)^2/(2s^2) = 2(CBu_t)^2/s^2$, which is zero for $u_t = 0$; a committed input $|u_t| \geq \bar{u}$ gives at least $2(CB\bar{u})^2/s^2$. □

Proposition 5. (a) The teacher's target is a measurable function of $(h, \bar{\theta})$, so $\mathbb{E}[\tau \mid h]$ integrates over the posterior of $\bar{\theta}$ only. At the cold start $\hat{g}_0 = e$ and $\mathbb{E}[\tau \mid h_0] = \mathbb{E}[\pi^\star(h_0;\bar{\theta}) \mid h_0]$. The canonical family is not closed under G, so Theorem 1 places no constraint on this action. (b) Let g_0 be the true element and suppose the teacher commits to g_0 at a stopping time σ; it switches to $\arg\max_g \ell_g$, so $\ell_{g_0}(h_\sigma) \geq \ell_g(h_\sigma)$ for all g. For each $g \neq g_0$, $\Lambda_t = \exp(\ell_g(h_t) - \ell_{g_0}(h_t))$ is a nonnegative martingale under the true law, because the inputs are functions of the history and the likelihoods are exact. Ville's inequality gives $P(\sup_{t \geq \sigma} \Lambda_t \geq e^\eta \mid \mathcal{F}_\sigma) \leq \Lambda_\sigma e^{-\eta} \leq e^{-\eta}$, and a union bound over the $|G| - 1$ alternatives completes the proof. (c) While the teacher is committed to a wrong element with $|u_t| \geq \bar{u}$, the expected increment of $\ell_{g_0} - \ell_{g'}$ is the one-step divergence, at least $\kappa = 2(CB\bar{u})^2/s^2$ to first order. By Wald's identity the expected time for the gap to rise by η plus its initial deficit is at most that total divided by κ, plus the expected overshoot. □

C ADDITIONAL RESULTS

Closed-loop fine-tuning of imitation controllers removes the default direction (two seeds); training the closed-loop objective from scratch was far worse. Selecting per-step multimodal heads by their prediction record fails (73% unstable) because the best head changes on 64% of steps; heads that stay fixed over an episode fix this. On the pendulum, multi-step forecasts (27–42% falls) and direction-neutral probing (49–91%) did not help. Adding QLD's hypotheses to MMAC's bank gave 11–13% falls, because the bank wins the likelihood comparison on 91–96% of steps. *Cautious commitment* showed the predicted U-shape on validation plants but did not reduce falls on test plants (5.6% and 4.4% vs. 5.0% and 3.9%). With a fixed global normalisation, action imitation still drops 87% of pendulums, so its failure is caused by the symmetry and not by the normalisation. A discrete-time Nussbaum controller failed (71% unstable).

Table 4: 95% intervals for the headline results (60 test plants $\times$ 3 noise seeds $=$ 180 runs each): Wilson intervals for the failure rate and percentile-bootstrap intervals (10,000 resamples) for the median regret of the runs that do not fail.

Family	Controller	failed [95% CI]	median regret [95% CI]
Linear	SBI + DAgger, seed 0	0.0% [0.0, 2.1]	5.53 [4.41, 6.74]
Linear	SBI + DAgger, seed 1	0.0% [0.0, 2.1]	5.14 [4.12, 7.16]
Linear	QICC frame-rigid, seed 0	0.0% [0.0, 2.1]	5.79 [4.42, 8.18]
Linear	QICC frame-rigid, seed 1	0.0% [0.0, 2.1]	4.29 [3.75, 6.18]
Linear	MMAC, 1024 models	0.0% [0.0, 2.1]	2.11 [1.92, 2.77]
Hard	SBI + DAgger, seed 0	37.2% [30.5, 44.5]	171 [78.6, 5.7e+03]
Hard	SBI + DAgger, seed 1	41.1% [34.2, 48.4]	136 [92.5, 664]
Hard	SBI + DAgger, seed 2	43.9% [36.8, 51.2]	142 [82.7, 367]
Hard	QICC, seed 0	42.2% [35.2, 49.5]	71.3 [49.7, 483]
Hard	QICC, seed 1	44.4% [37.4, 51.7]	699 [311, 9.41e+03]
Hard	QICC, seed 2	40.0% [33.1, 47.3]	1.43e+03 [98.7, 2.93e+04]
Hard	MMAC, 1024 models	61.7% [54.4, 68.5]	1.97e+03 [53.1, 6.39e+03]
Hard	ARX-CE with probing	28.3% [22.3, 35.3]	4.08e+04 [1.13e+04, 1.44e+05]
Pendulum	SBI-RMA + DAgger, seed 0	10.0% [6.4, 15.3]	1.14 [0.738, 1.71]
Pendulum	SBI-RMA + DAgger, seed 1	13.3% [9.1, 19.1]	1.19 [0.922, 1.76]
Pendulum	QLD, seed 0	5.0% [2.7, 9.2]	2.26 [1.54, 2.82]
Pendulum	QLD, seed 1	3.9% [1.9, 7.8]	0.84 [0.41, 1.06]
Pendulum	RMA-style, seed 0	42.8% [35.8, 50.1]	4.42 [3.12, 5.64]
Pendulum	RMA-style, seed 1	32.2% [25.8, 39.4]	4.13 [3.65, 4.82]
Pendulum	Action imitation, seed 0	78.9% [72.4, 84.2]	1.82 [0.87, 3.35]
Pendulum	Action imitation, seed 1	84.4% [78.4, 89.0]	2.15 [1.4, 3.53]
Pendulum	MMAC, 256 models	9.4% [6.0, 14.6]	0.989 [0.775, 1.31]

D STATISTICAL DETAILS, COMPUTE AND HYPERPARAMETERS

Tests. Paired comparisons use two-sided exact sign tests on per-run regret over identical plants, noise and initial states, counting a failure as the worst outcome, and exact McNemar tests on failures. The p-values are not adjusted for multiple comparisons.

Compute. All experiments ran on a single CPU core. Students are two-layer transformers of width 64 with four attention heads (116,417 parameters for the single-output model); the larger student in Section 5.4 has four layers. Training uses Adam (learning rate 10^{-3}, batch 128) for 3000 steps (6000 for the larger student), which takes 6–10 minutes. A DAgger round rolls out the student on 1000–1500 fresh plants and fine-tunes for 2000 steps at learning rate 3×10^{-4} (about 5 minutes). Evaluating 180 runs takes 1–3 minutes.

SBI hyperparameters. Switching margin $\eta = 2$ nats in every family; behaviour dither 0.3; half of the training episodes start under the controller of a random plant from the full family; context windows of the last 40–50 steps; on the hard family, a Huber loss on scale-normalised targets clipped to $[-50, 50]$, as for QICC on that family. None of these values was tuned on test plants.

E AUDIT OF IN-CONTEXT CONTROL EVALUATIONS

In a fully observed LQR setting, an audit of a study of "copycat" contamination found effects confounded with training budget, controllers worse than context-free baselines, and an expert-substitution condition that leaked the optimal gain. After these issues were fixed, expert-context pretraining produced controllers that copy the context policy and fail when generating their own context (regret 218–374 vs. 0.94 for least-squares certainty equivalence). Decorrelated behaviour, transition tokens, and per-window normalisation with a scale-invariant loss reduced self-deployment regret three- to five-fold in three of three paired seeds.