# WHAT DRIVES COUNTERFACTUAL-RESPONSE GAINS IN POLICY DISTILLATION?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can teacher interpretability improve policy distillation beyond ordinary off-policy sampling? We study two routes: selecting a decision tree's input vocabulary and augmenting its training data with matched state pairs drawn along response-validated features. Across 27 PPO and SAC teachers on four control tasks, counterfactual augmentation improves intervention consistency (IC), a descriptive statistic of teacher–student response agreement, by $0.097$ on fresh pairs at equal tree size and a 20,000-label budget. Sampling controls recover most of this gain; the increment over random pairs is $0.020$. A post-hoc expansion across 33 teachers evaluates IC and action error on identical endpoints: the scores are strongly associated, and only 58 of 2,376 method comparisons survive endpoint-error, return and competence matching, with a median absolute IC gap of $0.024$. This does not establish equivalence, but provides no demonstrated incremental diagnostic utility for IC. Restricting the vocabulary to validated features can omit information needed for control; a predictive sufficiency check repairs the Pendulum failure without establishing a general repair. The study combines a pre-registered core with explicitly post-hoc controls and sensitivities. Its contribution is an empirical decomposition of response gains and the limits of this gate and feature library, without an established interpretability-specific control or robustness advantage.

## 1 INTRODUCTION

When a learned model is compressed into a simpler one, the usual test of success is that the two produce the same outputs on the data. For neural control policies the simpler model is often a decision tree, a program or a symbolic expression, chosen because it can be inspected and verified (Bastani et al., 2018; Kohler et al., 2024; Landajuela et al., 2021), and the test is action agreement on the states the teacher visits, plus closed-loop return. Agreement on visited states is a weak constraint. A tree whose median action error against the teacher is a hundredth of the action bound may depend on different quantities: change one input while holding the others fixed, and the teacher's action moves while the tree's does not. The difference is invisible on visited states, where the inputs move together, and appears as soon as the closed loop leaves the training distribution. Whether this happens is an empirical question about what distillation preserves; it has been asked for language models (Haskins & Adams, 2025; Wadhwa et al., 2025) but not for control, where the student belongs to a different representation class and has no activations to align.

We measure selected state-response agreement rather than preservation of internal mechanisms. Our *validated handles* are scalar state features, either registered physical concepts or sparse-autoencoder latents (Cunningham et al., 2023; Gao et al., 2024), that pass an operational response gate. A matched pool state changes the target feature while approximately preserving a declared matching set; other coordinates can change. A pass establishes a response under this joint state change, not the isolated causal effect of the named scalar. Unlike representation-level causal abstraction (Geiger et al., 2021), this procedure does not align or intervene on a teacher's internal causal variables. Intervention consistency (IC) summarizes agreement between teacher and student action differences on the selected pairs.

The scientific question is how much response agreement comes from handle-directed sampling beyond ordinary off-policy data, and whether this gate selects a sufficient input vocabulary. DAgger already motivates collecting data beyond the teacher's visited distribution (Ross et al., 2011); our



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.

controls quantify the additional contribution of handle-directed endpoints. Twenty-seven PPO and SAC teachers on Pendulum, CartPole, Hopper and NavModes form the primary study. A fifth task was specified before running it. The pre-registered core was followed by multiple documented implementation corrections and post-hoc controls, including endpoint-matched and categorical sensitivities; affected stages were rerun as specified in Appendix C and Table 4.

Our contributions are:

- An **empirical decomposition of counterfactual-response gains**: at equal fitting budgets, ordinary sampling controls recover most of CF-aug's fresh IC improvement. Handle-directed endpoints add a small generator-specific increment, without an established control or robustness advantage (Sections 5.2 and 5.3).
- An **endpoint-matched diagnostic check**: IC and ordinary endpoint error largely agree in their method rankings. Large response gaps at matched visited-state error do not demonstrate a useful distinction among competent policies; the stricter post-hoc comparison has substantial attrition and does not establish incremental diagnostic utility (Section 5.1 and Appendix G).
- A **failure of this gate and feature library**: response validation can exclude information needed for control. A predictive sufficiency check repairs the Pendulum vocabulary loss, but its success is not general across tasks (Section 5.4).

IC remains a descriptive response statistic: it neither identifies internal mechanisms nor currently enables a demonstrated control decision beyond ordinary endpoint error.

## 2 Related work

**Distilling policies into verifiable surrogates.** VIPER distills a Q-function teacher into a decision tree with DAgger and criticality weights (Bastani et al., 2018); later work refines the tree (Roth et al., 2022; Vos & Verwer, 2024; Marton et al., 2024) or changes the class to programs (Kohler et al., 2024; Liu et al., 2024; Lin et al., 2023) or symbolic expressions (Landajuela et al., 2021; Sharan et al., 2022; Al-Tamari & Kadour, 2026), and all of it optimizes and reports behavioral agreement and return, as does the evaluation of interpretable RL by distillation (Kohler et al., 2025). Concept-bottleneck agents (Delfosse et al., 2024; Ye et al., 2024; Vincze et al., 2024; Grandien et al., 2024) are interpretable by construction and are the natural comparison for our vocabulary route; symbolic distillates can share their teachers' pathologies (Cyranka et al., 2023).

**Interpretability of neural policies and mechanism under distillation.** Linear probes (Alain & Bengio, 2016; Hewitt & Liang, 2019; Belinkov, 2021) decode what a representation contains but not what the network uses (Elazar et al., 2020); sparse autoencoders (Cunningham et al., 2023; Bricken et al., 2023; Shu et al., 2025) recover sparse directions, including in robot policies (Swann et al., 2026); and mechanistic studies of RL agents (Trim & Grayston, 2024; Ramanauskas & Şimşek, 2023; Langosco et al., 2021; Jiang et al., 2025) locate goal and navigation features. Our gate follows interchange-intervention and activation-patching practice (Geiger et al., 2021; Vig et al., 2020; Meng et al., 2022; Zhang & Nanda, 2023), applied in state space so that the same counterfactuals can be given to a tree. In transformers, distillation reorganizes circuits (Haskins & Adams, 2025) and a circuit-alignment objective can transfer them (Wadhwa et al., 2025); our study instead evaluates state-response agreement and does not test circuit transfer.

## 3 Setup: validated handles and a frozen harness

**Teachers.** Each teacher $\pi_T$ is a two-layer MLP trained with PPO (Schulman et al., 2017) or SAC (Haarnoja et al., 2018) and accepted only if its deterministic evaluation passes a pre-set target (Appendix A). The main study has 27 teachers on four tasks: Pendulum 12, CartPole 3, NavModes 6 (a navigation task whose wall gap gives the correct behavior discrete modes by construction) and Hopper 6. A fifth task, LunarLanderContinuous-v3, with six more teachers, was added later with every step pre-specified before any run on it (Section 5.5 and Appendix J).

**Validated handles.** A candidate handle is a scalar function $c(s)$ of the state with a pre-registered regime, response and predicted sign. To test it at a base state $s$, we search a pool of off-policy states

for a counterfactual $s'$ in which $c$ has moved by a dose $\delta\sigma_c$ while a declared matching set of other features stays within a residual, and compare the teacher's response at $s'$ and $s$. A handle passes if the realized sign matches the prediction, the scalar effect is monotone in $\delta \in \{\pm 0.5, \pm 1, \pm 2\}$ (categorical concepts instead use exact level swaps), and it exceeds a multiplicity-adjusted bootstrap reference threshold at enough doses (thresholds in Appendix B). Top-$k$ sparse-autoencoder latents enter the same gate two-sided. Of 81 pre-registered concept predictions on the original 21 teachers, 43 pass, 33 after excluding a post-hoc regime relabeling and a confounded concept; latents pass on all four tasks, and on Hopper they are the only handles. The gate intervenes on states, not probe directions, because decodable probe directions moved the action as predicted in 0 of 24 cases (Appendix B). The frozen handle sets are used unchanged by every later stage.

**Response statistic.** For a handle $i$ with matched pairs $\{(s_j, s'_j)\}$ pooled over doses, intervention consistency is

$$\mathrm{IC}_i(\pi_S) = \mathrm{corr}_j\big(\pi_T(s'_j) - \pi_T(s_j),\ \pi_S(s'_j) - \pi_S(s_j)\big), \tag{1}$$

the Pearson correlation over pairs and action dimensions of the *executed* actions (continuous outputs clipped to the environment's bound), and IC is the mean over a teacher's validated handles. A surrogate whose response is constant over a handle's pairs scores 0 on it. The correlation is scale-free by design, since fidelity already constrains action magnitude; a magnitude-sensitive companion gives the same results (Appendix I). Because the pairs are states, IC is defined for any surrogate class. IC measures counterfactual response agreement on the validated handles; two implementations with the same input-output function receive the same score, so it does not identify internal mechanism, and its incremental diagnostic utility beyond endpoint action error is unestablished.

**Harness.** Fidelity is $\mathrm{fid}(X) = \max(0, 1 - \mathrm{median}_{s\in X}\|\pi_T(s) - \pi_S(s)\|/a_{\max})$ for continuous executed actions, with $a_{\max}$ the action bound, and agreement rate for discrete ones; it is a median, so it says nothing about the error tail, which the harness records separately as the 95th-percentile error (Section 5.1). Fidelity is measured on teacher-visited states $D_1$, space-filling states $D_2$ and surrogate-visited states $D_3$. Out-of-distribution (OOD) behavior is retention $\mathrm{ret}(X) = \mathrm{fid}(X)/\mathrm{fid}(D_1)$, undefined when $\mathrm{fid}(D_1) < 0.2$, in three families: a *held-out region* of state space excluded from every surrogate's training data (Appendix A), the first 30 steps after shifted initial states at three severities, and dynamics parameters scaled by $\{0.667, 0.8, 1.25, 1.5\}$. The harness also records closed-loop return, switching-surface agreement and tree size. It was frozen before the baseline runs and subsequently corrected in several revisions; affected stages and later analyses are distinguished in Appendix C and Table 4.

**Validating the metric.** The metric is only useful if surrogates with similar fidelity on visited states but different counterfactual responses receive different scores; the following are sanity checks of that property, not evidence that IC identifies internal mechanism. The teacher scores $\mathrm{IC} = 1.000$ against itself; children fine-tuned from it score 0.73 to 1.00 and exceed independently seeded teachers in 29 of 31 cases over the five tasks (Appendix I). Adversarial impostors, students trained to match the teacher's actions ($D_1$ fidelity 0.69 to 1.00) while being penalized for responding to a given handle, score 0.01 to 0.69, the upper end on the metric-weak handles below; and IC decreases monotonically as a handle's effect is progressively corrupted. Two handles admit only partial impostors and are flagged as metric-weak (Appendix B).

## 4 Distillation methods

The surrogates of both routes and their controls are CART trees with at most $L \in \{8, 16, 32, 64\}$ leaves, trained on exactly $B = 20{,}000$ teacher-labeled states over five DAgger iterations; each batch is collected until it holds its quota of states outside the held-out region, so every method labels the same number of states, and the number is logged on every DAgger-trained record (B3 and B4 are fitted to the same DAgger data). The budget counts distillation labels only; the cached state pools, the gate, the latent extraction and the sufficiency gate are separate teacher access, accounted for in Appendix A, so the comparison is at equal fitting data, not at equal total teacher cost. B1, DAgger with CART (Ross et al., 2011), is the reference for every comparison. The other baselines populate the premise test: B2 is VIPER (Bastani et al., 2018) with two continuous criticality proxies, B3 a return-optimized tree with an imitation warm start (Marton et al., 2024), B4 genetic-programming symbolic regression standing in for deep symbolic policies (Landajuela et al., 2021), and B5 a hard-gated mixture of linear experts, the CART partition with a linear controller in each leaf, which is the

structure of sparse mixtures of shallow experts (Vincze et al., 2024) fitted by imitation. Appendix D gives every implementation.

**Vocabulary route.** The tree's inputs are the validated concepts $\phi_T(s)$ instead of the raw state (*concepts*); a second variant first distills the validated latents into short symbolic expressions and appends them (*distilled latents*). Controls are a random rotation of the state and a random subset of coordinates, both of the validated set's dimension, and the *unselected library* of every pre-registered concept regardless of gate outcome, which is larger.

**Data route: counterfactual augmentation (CF-aug).** The evaluation pair generator uses the declared matching sets and residual threshold to generate matched pairs $(s, s')$ for every validated handle from a pool restricted to outside the held-out region; both endpoints are teacher-labeled and replace up to half of every DAgger batch, as many as the pairs supply, so the query budget is unchanged (*replace*), or are appended (*add*). The realized fraction is 18 to 42 percent of each batch (Appendix D). The historical evaluation and augmentation generator applied the scalar dose grid to categorical handles too, with a 0.25-SD target tolerance, omitting some level transitions tested by the gate. A separate post-hoc NavModes analysis corrects this to exact level swaps (Appendix H); historical tables retain their original generator. The surrogate is still a raw-state tree of the same size. DAgger's later batches are the student's own rollouts, while the pairs sample the directions along which the teacher's decision depends on each handle; three controls separate the two, drawn from the same pool: *pool states*, unpaired states drawn uniformly; *random pairs*, the same base states each paired with a real pool state at the same displacement norm in a random direction; and *unvalidated pairs*, matched pairs for the library concepts that failed the gate. The first two always replace the same number of states as CF-aug; the third does so on every teacher except the three NavModes SAC teachers, where a CF-aug run at its lower dose provides the comparison, and it is absent where no concept failed the gate (CartPole, one LunarLander teacher; Appendix D).

**Protocol.** Every variant is paired with B1 at the same teacher, distillation seed (3 per teacher) and leaf budget. The pre-registered confirmatory analysis is a linear mixed model per OOD family, $\mathrm{ret} \sim \mathrm{variant} \times \mathrm{severity} + \mathrm{env} + \log_2 L$, with random intercepts for teacher and distillation seed nested within teacher, and Holm correction across variants; convergence is checked for every fit (Table 5). Per-task paired differences are reported with teacher-cluster bootstrap intervals and a sign test over per-teacher medians, since seeds and budgets within a teacher are correlated repeats; row-level tests are descriptive only. For the principal pooled contrasts, we foreground a post-hoc sensitivity analysis that averages paired differences within each teacher and reports the mean across teachers with a $t$-based 95 percent interval. These intervals describe variation across the sampled teachers, not uncertainty over new tasks; the pre-registered models remain in the appendix. Appendix C lists every deviation from the pre-registration, including which analyses were added after the pre-registered ones had been read.

# 5 Results

## 5.1 Matched aggregate action error underdetermines counterfactual response

We first ask whether matched visited-state action error constrains response agreement on the selected off-policy pairs. Behavior is matched as an aggregate, the median action error on visited states, which does not require two surrogates to agree on the same states. We group each teacher's baseline surrogates into behavioral-equivalence classes, sets whose $D_1$ fidelities differ by at most 0.05 (complete linkage, fidelity at least 0.2), and measure the spread of the response metrics within each class (Figure 1).

On Pendulum, 16 classes contain both a symbolic law and a DAgger or VIPER tree of the same teacher, at fidelities from 0.88 to 1.00; their median IC spread is 0.40, and the widest, a SAC teacher at fidelity 0.94 to 0.98, spans IC from $-0.22$ to 0.81 with held-out retention 0.00 to 0.72 and switching agreement 0.00 to 0.43. The extreme case is a constant program, torque $-0.54$ everywhere, with fidelity 0.995 (median error 0.005, 95th-percentile error 0.923): it never swings the pendulum up (return $-1296$ against $-226$ for a 64-leaf tree at fidelity 0.993) and has $\mathrm{IC} = 0.00$ against 0.85. These classes match the median error only. Requiring the 95th-percentile error to match as well, within 0.10 of the action bound, leaves eight such Pendulum classes, with a median IC spread of

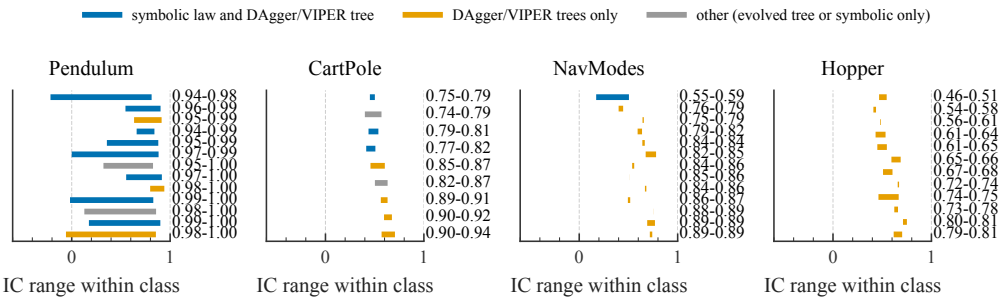



Figure 1: Spread of intervention consistency within behavioral-equivalence classes (median-error matching), one bar per class, with the class's $D_1$ fidelity range on the right. Shown are the 12 highest-fidelity classes per task plus the widest class mixing a symbolic law with a tree. Blue: a symbolic law and a DAgger or VIPER tree; orange: DAgger and VIPER trees only; grey: other combinations.

0.20 and a maximum of 0.70, so the gap survives matching on the error tail; requiring closed-loop return within 10 percent of the teacher's as well leaves none, because the symbolic laws never swing up; the classes surviving all three criteria are almost all tree-only (spreads 0.02 to 0.09), so the large gaps separate competent trees from incompetent programs, and the premise is a statement about matched aggregate open-loop error, not about matched closed-loop behavior (Appendix F). Within the DAgger and VIPER family the spread is small on NavModes, Hopper and LunarLander (median class spread 0.03, 0.07 and 0.02) and moderate on CartPole (0.11) and Pendulum (0.23, or 0.10 with the tail matched). Behavioral fidelity does not determine counterfactual response across representation classes, and among the DAgger and VIPER configurations tested, changing the fitting procedure moved IC comparatively little. We therefore examine changes to the input vocabulary and to the training distribution.

**Does IC add a diagnostic beyond endpoint error?** A separate post-hoc expansion fits four raw-state tree variants at all budgets and seeds across 33 teachers (1,584 fits) and scores both metrics on identical fresh endpoints. On the 27 primary teachers, CF-aug gains $+0.087$ IC while reducing mean endpoint error by $0.072$; the median within-teacher rank correlation between IC and negative endpoint error is $0.969$. Only 58 of 2,376 same-size, same-seed method comparisons survive endpoint-error, return and competence matching, across ten teachers, with median absolute IC gap $0.024$ and maximum $0.063$. None survive on Hopper or LunarLander. These selected comparisons do not establish equivalence or IC redundancy, but provide no demonstrated incremental diagnostic utility (Appendix G).

### 5.2 Counterfactual augmentation aligns responses, mostly through coverage

**Counterfactual responses.** At the same 20,000-label budget and tree size, CF-aug improves harness IC over the independent DAgger refit by $+0.100$ $[+0.077, +0.124]$ and fresh IC by $+0.097$ $[+0.075, +0.119]$ (mean paired differences averaged within each of 27 teachers, then across teachers; 95 percent $t$ intervals). Against the original DAgger baseline, the per-task median IC gains are 0.04 on Pendulum, 0.06 on CartPole, 0.08 on NavModes and 0.11 on Hopper. Teacher-level pooled $D_1$ fidelity differs by $-0.002$ $[-0.012, +0.007]$. Only Pendulum, with twelve teachers, reaches significance on the per-task sign test; three to six teachers on each other task limit task-specific inference. The pre-registered pooled models and full sensitivity analysis are in Appendix F.

**Is the gain an artifact of training on the metric?** CF-aug trains on the pair generator that IC evaluates, so we refit a DAgger tree at every cell with independent random draws and scored both on fresh pairs from an independently collected dataset that shares no state with the augmentation pool, and on pairs lying inside the held-out region, where no surrogate was trained (Table 1). On fresh pairs the gain is as large as on the harness pairs. On held-out-region pairs it vanishes on Pendulum and NavModes, is +0.05 with a wide interval on CartPole, and persists on Hopper (+0.11). The surrogate has learned the teacher's counterfactual responses on unseen states, not the specific pairs, but on most tasks only in the part of state space the pairs covered.

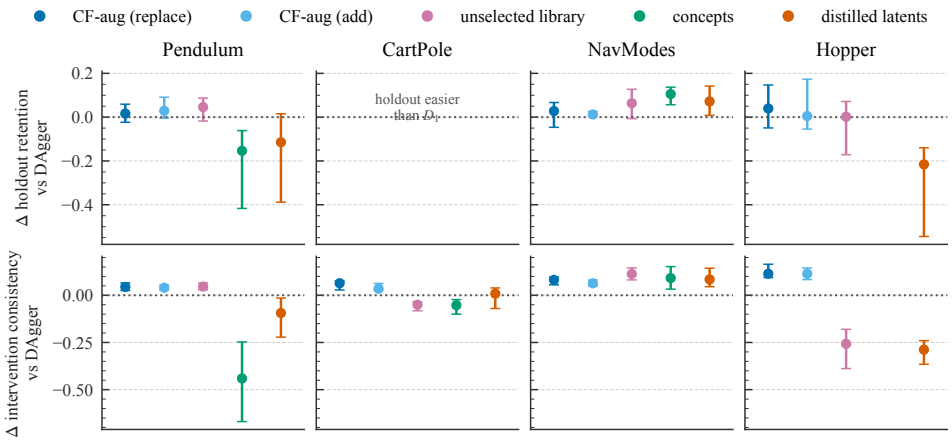



Figure 2: Paired differences from DAgger trees at matched teacher, distillation seed and leaf budget (median and teacher-cluster bootstrap 95% confidence interval over 36 to 144 pairs). Rows show held-out-region retention and intervention consistency; return is in Figure 5. CartPole's held-out region is easier than its visited distribution, so its retention panel is omitted; Hopper has no concept vocabulary.

Table 1: Median IC gain over a DAgger tree refit at the same cell with independent random draws. Left: CF-aug (replace) on the harness, fresh and held-out-region pair sets. Right: the three size-matched controls on fresh pairs. The pooled row gives mixed-model coefficients. * Holm-corrected $p < 0.05$: per task, a sign test over per-teacher medians across variants; pooled, the mixed model.

| | CF-aug | | | controls, fresh pairs | | |
|---|---|---|---|---|---|---|
| Task | harness | fresh | held-out | pool states | random pairs | unvalidated |
| Pendulum | $+0.04^*$ | $+0.05^*$ | $-0.01$ | $+0.03^*$ | $+0.03^*$ | $+0.02^*$ |
| CartPole | $+0.10$ | $+0.09$ | $+0.05$ | $+0.11$ | $+0.11$ | – |
| NavModes | $+0.06$ | $+0.07$ | 0.00 | $+0.06$ | $+0.03$ | $+0.05$ |
| Hopper | $+0.11$ | $+0.12$ | $+0.11$ | $+0.10$ | $+0.10$ | $+0.08$ |
| Pooled | $+0.100^*$ | $+0.097^*$ | $+0.033^*$ | $+0.084^*$ | $+0.077^*$ | $+0.061^*$ |

**Is the gain due to the handles?** The pairs differ from DAgger data in two ways: they come from an off-policy pool rather than from the surrogate's own rollouts, and they are arranged along the handles' directions. The controls of Section 4 keep the first and remove the second (Table 1, right). Unpaired pool states and random-direction pairs raise IC by 0.084 and 0.077 pooled on fresh pairs against 0.097 for CF-aug, on every task; pairs from gate-failed concepts raise it by 0.061, and at equal dose the validated pairs beat them by 0.02 to 0.03 on three tasks and not on NavModes (Appendix F). Across the 27 teacher means, CF-aug exceeds random pairs by $+0.021$ $[+0.009, +0.033]$ on harness pairs and $+0.020$ $[+0.007, +0.032]$ on fresh pairs; it exceeds pool states by $+0.016$ $[+0.007, +0.026]$ on harness pairs. The mixed-model increment over random pairs on held-out-region pairs is small and not significant ($+0.008$). Per task the increment over random pairs is 0.01 on Pendulum, CartPole and Hopper and 0.03 on NavModes (Appendix F). Excluding the two metric-weak handles changes the CF-aug gain by at most 0.01 (Appendix I). The controls also reproduce the robustness effects reported next. What aligns the tree's responses with the teacher's is mostly training on states off the surrogate's own trajectory; handle-directed placement adds a small pooled increment on this generator, without an established control advantage.

**Off-distribution behavior.** Retention gains are small and sensitive to the estimand; pool-state and random-pair controls reproduce the main effects (Appendix F and Table 7). Return is unchanged on Pendulum, CartPole and NavModes. Hopper gains 55 return, but all tree surrogates still fail closed-loop, so this does not establish improved usable control. The original augmentation run leaked held-out states; all reported comparisons use the corrected run (Appendix E).

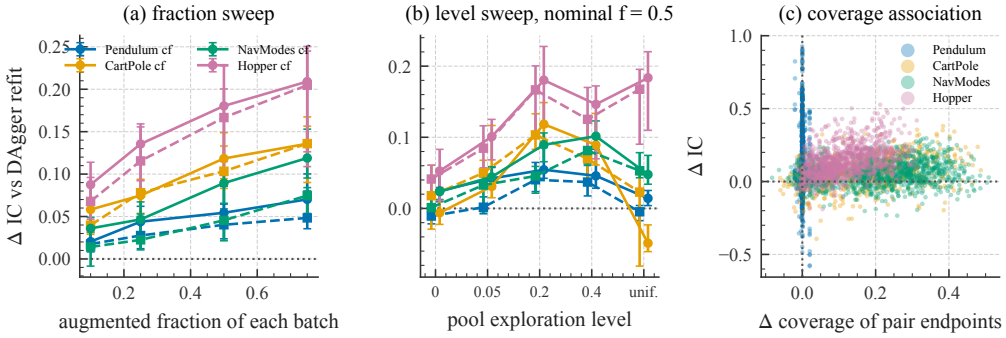



Figure 3: Coverage dose-response (5,508 surrogates; median and teacher-cluster bootstrap 95% confidence interval). Solid lines are CF-aug pairs, dashed lines unpaired states from the same pool. (a) IC gain over an independently drawn DAgger refit against the augmented fraction of each batch. (b) The same at a nominal half of each batch (three teachers fall short at some levels, Appendix I), against the pool's exploration level (on-policy, random-action probability 0.05, 0.2, 0.4, space-filling). (c) IC gain against coverage gain, one point per surrogate (an association; the measure is computed after training).

**Categorical matching sensitivity.** Exact-swap evaluation restores both missing extreme-to-extreme mode transitions for all four NavModes teachers with a validated mode handle. With corrected training and evaluation across all six teachers, CF-aug gains $+0.054$ IC $[+0.027, +0.081]$ over plain refitting, but its increments over pool states, random pairs and equal-dose unvalidated pairs have intervals spanning zero. This post-hoc check fixes transition coverage without establishing a NavModes-specific benefit of validated handles (Appendix H).

### 5.3 Augmentation dose and coverage predict IC gains

Coverage is the fraction of harness-pair endpoints within 0.25 standard deviations of a training state. Increasing the augmented fraction raises pooled IC gains from 0.04 to 0.11 for CF-aug and from 0.03 to 0.09 for pool states (Figure 3a). Across teacher means, CF-aug exceeds pool states by 0.017 to 0.024 at the four fractions, with positive interval lower bounds (Appendix F). More exploratory pools generally help, but space-filling sampling can hurt (Figure 3b). Coverage and IC gains are associated where coverage varies, although the teacher-level slope interval includes zero (Figure 3c; Appendix I). The tree never sees which states were paired: the residual benefit is from endpoint sampling. Coverage is measured after training, so these results support a sampling explanation without identifying causal mediation.

### 5.4 The selected vocabulary fails on Pendulum; a sufficiency check repairs it

Restricting a tree's inputs to validated concepts is the most direct way to make a surrogate "speak the teacher's language"; Figure 8 evaluates it in two steps. Against random rotations and random coordinate subsets of the same dimension, validated concepts win where a concept vocabulary exists: on CartPole the two-concept tree reaches the maximum return of 500 against 70 to 99 for two random coordinates or a random rotation, and on NavModes concept trees reach a median IC of 0.60 against 0.24 to 0.28, so selection is not just dimensionality reduction. Against the unselected library, which contains every pre-registered concept including those that failed the gate, the validated selection does not win on IC on any task; on held-out retention its point estimates lead on CartPole and NavModes, neither significant, and on return the library wins on Pendulum (Table 6).

On Pendulum the selection is harmful: concept-only trees lose 0.44 IC and 291 return relative to raw-state DAgger trees at the same budget, and random controls of the same dimension are equally poor. Two validation rulings caused it: angular velocity passed the gate only after a post-hoc regime relabeling and was excluded from the frozen set, and the pumping-phase concept was pre-registered with the wrong sign, so its corrected form counts as validated only for the three teachers trained after the correction. What remains for most teachers is energy and upright angle, with no velocity term, so no tree over these inputs can damp. The gate answered whether each feature moves the teacher's

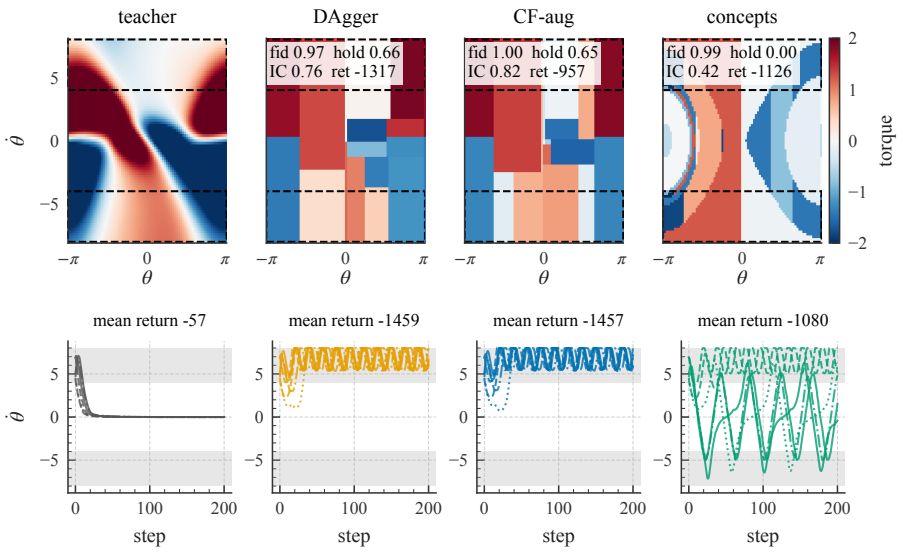



Figure 4: Pendulum, PPO teacher selected by a pre-committed rule (closest to the median CF-aug effect), trees at the 16-leaf budget. Top: torque over $(\theta, \dot{\theta})$ for the teacher, DAgger, CF-aug and the concept-only tree; dashed lines bound the held-out band $|\dot{\theta}| \in [4, 8]$; insets give $D_1$ fidelity, held-out retention, IC and return. Bottom: $\dot{\theta}$ over time from six shared starts inside the band.

torque in the predicted direction and never asked whether the selected features suffice. Raw and distilled latents are in Appendix F.

**A worked example.** Figure 4 shows a teacher chosen by rule, not by effect size. All three trees at the 16-leaf budget have $D_1$ fidelity between 0.97 and 1.00. The DAgger and CF-aug trees split on the same raw coordinates and differ only slightly (IC 0.76 against 0.82); neither swings up from a held-out start. The concept-only tree is a different controller: its torque depends on energy and angle alone, symmetric in the sign of $\dot{\theta}$, its IC is 0.42, and it cannot damp (Appendix L), while the library tree, which can use velocity, reaches IC 0.85. Two numbers in the figure cut against a single ranking: the concept tree has zero held-out retention (its torques in the held-out band are wrong by more than the bound at the median) yet the best mean return from the six held-out starts ($-1080$ against $-1459$ and $-1457$), because its wrong torques occasionally carry the pendulum through the balance band while the raw-state trees settle into a fast spin. Retention, return and IC rank these controllers differently because they measure different properties.

**A sufficiency gate.** A teacher-side check greedily extends the selected vocabulary until a tree reaches 0.9 of a raw-state reference's fidelity on fresh visited and space-filling validation states, both outside the held-out region (Appendix I). Eleven of twelve Pendulum sets initially fail. Extension recovers IC $+0.45$, return $+711$ and held-out retention $+0.30$ over the validated-only trees, bringing IC within 0.03 of the unselected library (Table 13). This is a Pendulum repair, not a general recipe: CartPole needs none, NavModes depends on the search starting point, and LunarLander's near-zero space-filling reference makes a relative pass uninformative. The result concerns this gate, library and search procedure, rather than causal feature selection generally.

### 5.5 Robustness and a fifth task

All effects have the same sign for PPO and SAC teachers, and the IC gain of CF-aug shrinks with tree size on Pendulum (Appendix F). A fifth task, LunarLanderContinuous-v3, was added with its concept registrations, predictions and held-out-region rule written before any run on it, and the main findings replicate there: 21 of 30 pre-registered concept tests pass, the metric validates, the premise holds, the augmentation raises IC by 0.11 with no held-out gain ($-0.04$) and its leaking first version reported a spurious +0.19, and concept trees lose to raw-state trees because no concept carries the contact flags. One nuance: on this task random-direction pairs recover 0.11 of the 0.14 gain but

Table 2: Scope across tasks. IC entries are median CF-aug gains over the independent DAgger refit (Tables 1 and 16); return is relative to the original DAgger baseline. Fresh and held-out-region pairs are separate tests: held-out pairs are not from the independent fresh dataset. Sufficiency outcomes refer to the vocabulary route.

| Task | Fresh IC | Held-out IC | Return | Sufficiency |
|---|---|---|---|---|
| Pendulum | $+0.05$ | $-0.01$ | Unchanged | Repairs vocabulary loss |
| CartPole | $+0.09$ | $+0.05$ | Unchanged | No repair needed |
| NavModes | $+0.07$ | $0.00$ | Unchanged | Search-dependent |
| Hopper | $+0.12$ | $+0.11$ | $+55$; trees fail | No concept vocabulary |
| LunarLander | $+0.11$ | $-0.03$ | $-8$; no gain | Inconclusive: weak reference |

unpaired states only 0.08, so handle-directed sampling contributes more there; the mean increment over random pairs across six teachers is $+0.030$ with a 95 percent interval of $[-0.006, +0.066]$, so it is suggestive rather than established (Appendix J).

Table 2 separates response agreement from control and vocabulary outcomes. Higher IC on fresh pairs is consistent across tasks; improved usable control and a general sufficiency repair are not established.

## 6 Limitations

The handles are features selected by an operational response gate, not identified internal mechanisms, and a surrogate faithful to them inherits the teacher's failures along with its competence. The IC gain of CF-aug is measured with pairs from the generator that produces its training data; fresh pairs from an independently collected dataset rule out memorization of training states but not a shared bias of the generator, and on three of the four tasks the gain does not reach held-out-region pairs. The principal pooled IC contrasts survive teacher-level inference, but three to twelve teachers per task limit task-specific conclusions. The post-hoc same-endpoint experiment finds strong IC–error association and small gaps among the few competent matched comparisons; severe attrition and matching on aggregate error prevent an equivalence conclusion (Appendix G). The categorical correction restores missing NavModes transitions, but only four teachers have a validated mode handle, and its six-teacher sensitivity analysis does not establish an IC advantage over the augmentation controls (Appendix H). Gate and evaluation still retain different pool and residual-acceptance conventions. On Hopper the handles are action-space latents, every tree fails closed-loop, and the effects are the largest, so the pooled coefficients overstate what the tasks with working surrogates show. CartPole's frozen held-out region is easier than its visited distribution; a harder region reproduces every CartPole conclusion (Appendix I). The baselines omit LLM-guided programmatic search (Liu et al., 2024), and B5 is the imitation-fitted form of sparse mixtures of shallow experts (Vincze et al., 2024). All teachers are small MLPs on low-dimensional tasks.

## 7 Conclusion

Counterfactual augmentation improves agreement with a teacher on selected off-policy response pairs, but ordinary pool-state and random-pair controls recover most of the gain. The small remaining generator-specific increment does not establish improved usable control, robustness or internal-mechanism preservation. The endpoint-matched expansion further narrows the diagnostic claim: IC gains accompany lower error on the same states, and the few competent comparisons surviving error and return matching have small response gaps. This is not an equivalence result; IC remains a descriptive response statistic whose incremental utility is unestablished. The vocabulary experiment identifies a failure of this operational gate and hand-designed library: selecting response-sensitive features can omit information needed for control. A predictive sufficiency check repairs the Pendulum loss, with no established general repair. The contribution is the controlled attribution of response gains to sampling and the explicit boundary of what these interventions and feature selections demonstrate.

## REPRODUCIBILITY STATEMENT

The versioned anonymous supplementary archive includes code, frozen configurations, historical raw records, teacher and child checkpoints, all 165 manifest-listed cached datasets, and the endpoint and categorical model/prediction artifacts. An artifact manifest hashes the shipped files; a table map links every manuscript table to record hashes, configuration hashes and analysis commands. Both new experiments include archived source snapshots matching their recorded source hashes. The README distinguishes reanalysis of saved records from retraining, and records the experiment environment. Historical runners explicitly select legacy categorical matching, new records carry the mode, and resume checks reject explicitly incompatible records; the corrected sensitivity evaluates both generators. Identity-bearing paths are redacted with an explicit redaction ledger. The archive is provided locally and is not a claim of a completed public release.

Historical per-cell RNG states and complete contemporaneous source manifests were not saved. Resumed runners skip cells without replaying their random draws, so even with the supplied inputs, bitwise replay of the historical training schedule is not guaranteed. Saved-record reanalysis is the reproducibility target for those tables; current source hashes are not retroactively attributed to historical fits. The newer experiments retain per-cell seeds, source manifests, saved models and raw predictions. Appendix C and Table 4 separates protocol amendments, historical results and post-hoc analyses.

## AI USE STATEMENT

We used Claude Fable 5.1 (Anthropic), a large language model coding assistant, under the authors' direction for four purposes: research execution, to implement the experiment and analysis code and to run the experiments; research ideation, to draft the stage protocols and to diagnose failures such as the held-out leak; retrieval and discovery, to search the literature and assemble the bibliography; and writing, to draft and edit the manuscript and produce the figures. OpenAI Codex also assisted with the subsequent code corrections, categorical matching implementation, sensitivity runs, analysis and manuscript integration. We did not use generative AI to generate data or to prove mathematical claims. The authors reviewed all generated code; the evaluation harness was frozen under version control before any method ran, the leak fix carries a runtime assertion, and the authors approved every gate amendment and protocol deviation before it was executed. The authors defined the research question, reviewed all results, checked every number in the paper against the analysis outputs, and take full responsibility for the content of this paper.

## REFERENCES

Shaker Al-Tamari and Waled Kadour. Explainable reinforcement learning via physics-aware policy distillation. *arXiv preprint arXiv:2607.24672*, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *arXiv preprint arXiv:1805.08328*, 2018.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *arXiv preprint arXiv:2102.12452*, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Jacek Cyranka, Kevin E M Church, and Jean-Philippe Lessard. Worrisome properties of neural network controllers and their symbolic representations. *arXiv preprint arXiv:2307.15456*, 2023.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. *arXiv preprint arXiv:2401.05821*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *arXiv preprint arXiv:2006.00995*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, 2021.

Nils Grandien, Quentin Delfosse, and Kristian Kersting. Interpretable end-to-end neurosymbolic reinforcement learning agents. *arXiv preprint arXiv:2410.14371*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

Reilly Haskins and Benjamin Adams. Distilled circuits: A mechanistic study of internal restructuring in knowledge distillation. *arXiv preprint arXiv:2505.10822*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Zeyu Jiang, Hai Huang, and Xingquan Zuo. Compositional concept-based neuron-level interpretability for deep reinforcement learning. *arXiv preprint arXiv:2502.00684*, 2025.

Hector Kohler, Quentin Delfosse, Riad Akrour, Kristian Kersting, and Philippe Preux. Interpretable and editable programmatic tree policies for reinforcement learning. *arXiv preprint arXiv:2405.14956*, 2024.

Hector Kohler, Quentin Delfosse, Waris Radji, Riad Akrour, and Philippe Preux. Evaluating interpretable reinforcement learning by distilling policies into programs. *arXiv preprint arXiv:2503.08322*, 2025.

Mikel Landajuela, Brenden K. Petersen, Sookyung Kim, Claudio P. Santiago, Ruben Glatt, Nathan Mundhenk, Jacob F. Pettit, and Daniel Faissol. Discovering symbolic policies with deep reinforcement learning. In *ICML*, 2021.

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. Goal misgeneralization in deep reinforcement learning. *arXiv preprint arXiv:2105.14111*, 2021.

Yu-An Lin, Chen-Tao Lee, Guan-Ting Liu, Pu-Jen Cheng, and Shao-Hua Sun. Program machine policy: Addressing long-horizon tasks by integrating program synthesis and state machines. *arXiv preprint arXiv:2311.15960*, 2023.

Max Liu, Chan-Hung Yu, Wei-Hsu Lee, Cheng-Wei Hung, Yen-Chun Chen, and Shao-Hua Sun. Synthesizing programmatic reinforcement learning policies with large language model guided search. *arXiv preprint arXiv:2405.16450*, 2024.

Sascha Marton, Tim Grams, Florian Vogt, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. Mitigating information loss in tree-based reinforcement learning via direct optimization. *arXiv preprint arXiv:2408.08761*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*, 2022.

Karolis Ramanauskas and Özgür Şimşek. Colour versus shape goal misgeneralization in reinforcement learning: A case study. *arXiv preprint arXiv:2312.03762*, 2023.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

Aaron M. Roth, Jing Liang, Ram Sriram, Elham Tabassi, and Dinesh Manocha. Msviper: Improved policy distillation for reinforcement-learning-based robot navigation. *arXiv preprint arXiv:2209.09079*, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

S P Sharan, Wenqing Zheng, Kuo-Feng Hsu, Jiarong Xing, Ang Chen, and Zhangyang Wang. Symbolic distillation for learned tcp congestion control. *arXiv preprint arXiv:2210.16987*, 2022.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

Aiden Swann, Lachlain McGranahan, Hugo Buurmeijer, Monroe Kennedy, and Mac Schwager. Sparse autoencoders reveal interpretable and steerable features in vla models. *arXiv preprint arXiv:2603.19183*, 2026.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Tristan Trim and Triston Grayston. Mechanistic interpretability of reinforcement learning agents. *arXiv preprint arXiv:2411.00867*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Mátyás Vincze, Laura Ferrarotti, Leonardo Lucio Custode, Bruno Lepri, and Giovanni Iacca. Smose: Sparse mixture of shallow experts for interpretable reinforcement learning in continuous control tasks. *arXiv preprint arXiv:2412.13053*, 2024.

Daniël Vos and Sicco Verwer. Optimizing interpretable decision tree policies for reinforcement learning. *arXiv preprint arXiv:2408.11632*, 2024.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Circuit distillation. *arXiv preprint arXiv:2509.25002*, 2025.

Zhuorui Ye, Stephanie Milani, Geoffrey J. Gordon, and Fei Fang. Licorice: Label-efficient concept-based interpretable reinforcement learning. *arXiv preprint arXiv:2407.15786*, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

# A ENVIRONMENTS, TEACHERS AND COMPUTE

**Tasks.** Pendulum-v1 (torque-actuated swing-up and balance, action scale 2), CartPole-v1 (discrete), Hopper-v4 (MuJoCo, 3 torques) from Gymnasium (Towers et al., 2024), and NavModes-v0. NavModes is a point mass in $[-1, 1]^2$ that starts near the bottom and must reach a goal at the top; a horizontal wall at $y = 0$ has one gap whose position (left, center or right) is sampled per episode and appears in the observation $[p_x, p_y, v_x, v_y, g_x, \Delta x_{\text{goal}}, \Delta y_{\text{goal}}, \Delta y_{\text{wall}}]$, where $g_x$ is the horizontal position of the gap and the $\Delta$ terms are offsets from the agent to the goal and to the wall. Actions are 2D accelerations; the horizon is 200 steps.

**A fifth task.** LunarLanderContinuous-v3 (Box2D) was added with every step pre-specified for the task before any run on it (Appendix J): observation $[x, y, v_x, v_y, \theta, \dot{\theta}, c_1, c_2]$ with two binary leg-contact flags, actions main-engine and lateral-engine commands in $[-1, 1]$, horizon 1,000 steps; three PPO teachers (1.5M steps) and three SAC teachers (400k steps) of width 256, all accepted at deterministic return $\geq 200$ (228 to 282).

**Teachers.** Two hidden layers, width 64 (256 on Hopper and LunarLander), tanh for PPO and ReLU for SAC. Acceptance targets on the mean deterministic evaluation return: CartPole $\geq 475$, Pendulum $\geq -180$, Hopper $\geq 2200$, LunarLander $\geq 200$, NavModes success rate $\geq 0.9$. One NavModes PPO seed that failed to learn was replaced by a new seed and one Hopper SAC teacher was retrained for longer. Three Pendulum seeds per algorithm were added after the pumping-phase replication (Appendix B).

**Held-out regions and perturbations.** Pendulum $|\dot{\theta}| \in [4, 8]$; CartPole pole angular velocity in $[0.25, 1]$ (the fast tail); NavModes $p_x \in [0.5, 1]$; Hopper forward speed in $[3.3, 6]$. Dynamics families: Pendulum mass, length, gravity, torque gain; CartPole pole mass, length, gravity, force gain; NavModes acceleration gain and damping; Hopper body mass, gravity, actuator gain, friction.

**Compute.** All experiments ran on one 20-core CPU workstation; the teachers are small enough that CPU training was faster than GPU. Median wall-clock per surrogate, including the full harness evaluation, was 3 to 5 s for DAgger trees and 6 to 18 s for CF-aug trees (dominated by the extra IC pair sets). The study comprises 1,503 baseline, 3,015 variant and 972 clean-rerun records, plus 936 data-route control records, 5,508 coverage-sweep, 576 sufficiency-gate (288 from the validated start and 288 greedy from the empty set) and 324 mixture-of-experts records on the four tasks (Appendix I), and 2,772 records for the fifth task (Appendix J), plus 36 dose-matched CF-aug records on NavModes: 15,642 records, all produced by the revision-2 rerun of every distillation stage (about six hours on the same machine) and the revision-3 and revision-4 reruns of the cells they corrected (Appendix C). The original run's records are kept in the supplement; the 714 harder-region CartPole records and the 3 Hopper children (Appendix I) come from the original harness.

**Teacher access beyond the distillation budget.** The 20,000-label budget counts the states each surrogate is fitted to. Table 3 lists the other teacher access, counted as distinct states the teacher acts on, read off the code under an idealized caching assumption (each state scored once); the code as run is less frugal, since latent matching recomputes activations over the pool once per latent, the sufficiency gate rescores its validation states for every candidate set, and DAgger's first iteration also queries the teacher for rollout actions at the held-out states it rejects. On Hopper a gate response is a 25-step rollout, so each gate state there costs 25 passes. Actual calls were not instrumented. Shared preparation is paid once per teacher and amortized over every surrogate of that teacher (12 per method); method-specific preparation is paid once per teacher and method. The augmentation and its controls draw their pools from the cached datasets and label their endpoints inside the 20,000 budget, so they add only the activation passes that latent matching needs; the sufficiency gate is the most expensive method-specific step. The comparisons in this paper are therefore at equal fitting data, and the vocabulary and data routes both presuppose the shared preparation, which a plain DAgger baseline does not need.

# B THE CAUSAL GATE

**Pre-registered concept library.** Pendulum: total energy (regime below the upright energy; response absolute torque; predicted sign $-$), pumping phase (same regime; signed torque; $+$), upright angle and angular velocity (balance band $|\theta| < \pi/6$; signed torque; $-$). CartPole: pole angle, pole

Table 3: Teacher access per teacher as distinct states scored, an idealized count that assumes caching (see text). Ranges span the five tasks. Fitting labels are the same 20,000 for every surrogate of every method.

| Step | Shared by | States per teacher |
|---|---|---|
| On-policy and broadened datasets ($2 \times$ 100,000 rollout steps) | every method | 200,000 |
| Activations of the broadened dataset for probes and the sparse autoencoder | vocabulary and data routes | 100,000 |
| Concept gate (512 base states, 6 doses, zero-dose null; 3 to 5 concepts) | both routes | 21,000 to 36,000 |
| Latent gate (same procedure per candidate latent) | both routes | per latent about 7,000 |
| Harness pairs (256 base states, 6 doses per handle; evaluation only) | every method | up to 3,100 per handle |
| CF-aug and controls: activations of the 20,000-state pool for latent matching (recomputed per latent as run) | data route | 20,000 |
| Sufficiency gate: 20,000 training labels, 5,000 validation rollout steps, 5,000 space-filling labels (validation sets rescored per candidate set as run) | gated vocabulary | 30,000 |
| DAgger fitting labels (4,000 of them from the teacher's own rollout in iteration 1) | every surrogate | 20,000 |

velocity, track margin (right-edge regime), all on the right-minus-left logit, $+$. NavModes: mode identity (gap position), gap error, wall side, goal distance and forward velocity on the acceleration channels. Hopper: height margin, forward speed and torso pitch were pre-registered; none passed at any declared horizon.

**Matching sets and regimes.** Holding all other concepts fixed is impossible when concepts are functions of the same coordinates (Pendulum energy is determined by angle and velocity), so each concept declares the set that is held fixed: for Pendulum energy, pumping phase and angular velocity, the upright angle; for upright angle, the angular velocity; for NavModes gap concepts, wall side, goal distance and velocity. Counterfactuals are drawn from the concept's regime, except for NavModes wall side, whose regime is defined by the concept itself.

**Thresholds.** For each concept, 512 base states are drawn from its regime and matched at each dose $\delta \in \{\pm 0.5, \pm 1, \pm 2\}$ standard deviations; a match must place the target within $0.25$ standard deviations of $c(s) + \delta\sigma_c$, a dose counts only if at least 30 base states are matched, and the frozen gate applies the residual cap of $0.75$ (z-units) to the *mean* matching residual within each dose. Evaluation pair construction instead filters individual pairs at the same threshold. We therefore reran the concept gate with individual-pair filtering as a sensitivity analysis: 126 of 135 decisions over the five tasks agree (Pendulum 45 of 48, CartPole 9 of 9, NavModes 26 of 30, Hopper 18 of 18, LunarLander 28 of 30), the per-pair rule passes 83 concepts against 76, and of the nine changed decisions only one concerns a handle in a frozen validated set (NavModes PPO seed 1, mode identity, which would be lost); the others are gains of concepts the rulings had excluded (Pendulum angular velocity on three teachers, NavModes velocity on three, LunarLander altitude on two). Dropping that one handle from the evaluation changes the NavModes CF-aug gain over the refit by 0.007 (+0.060 to +0.053) and the controls' gains by at most 0.005, so the augmentation comparison on that task does not depend on it; the check does not retrain vocabularies or add the eight newly passing concepts, so it speaks to that comparison only. The gate also scored the unclipped PPO mean and, on CartPole, the logit difference, whereas the harness scores executed actions; rerunning the frozen gate on executed actions (means clipped to the bound, the argmax action on CartPole) reproduces 131 of 135 concept decisions over the five tasks (all 18 Hopper decisions among them, since its responses are simulator quantities): three concepts excluded by rulings would pass (NavModes velocity on two teachers, LunarLander horizontal offset on one) and one validated handle would fail, NavModes PPO seed 1 mode identity, the same handle as under the per-pair rule, with the same small downstream change. Effects are mean response changes in units of the response's standard deviation over the dataset. The reference distribution bootstraps mean response changes under zero-dose matches (500 resamples; nearest same-concept-value neighbor, self excluded). The operational threshold $t_{\text{spec}}$ is its $1 - 0.05/m$ quantile in absolute value. For registered concepts, $m$ is the number of concepts tested

within one teacher. For latent candidates it is the candidate count times the number of response channels, separately within each teacher and layer. There is no joint correction across teachers, layers or concept and latent families. Shifted targets can change nuisance coordinates differently from zero-dose neighbors, and the finite bootstrap tail is not calibrated for those alternatives; the Bonferroni-style quantile therefore does not establish familywise false-positive control for causal use of the named feature. We treat it as a reproducible selection rule, not a causal significance guarantee. The effect threshold is $t_{\text{eff}} = \max(0.3, 2\,t_{\text{spec}})$. A scalar concept passes if the largest eligible effect exceeds $t_{\text{eff}}$ with the predicted sign, the Spearman rank correlation between dose and effect is at least $0.7$ in absolute value with the predicted sign, at least two doses exceed $t_{\text{eff}}$, and the concept's linear probe is decodable (held-out $R^2 \geq 0.8$ on the broadened dataset). Categorical concepts use level swaps and require at least one effect above $t_{\text{eff}}$; the dose-response test applies only when three or more distinct swap magnitudes exist. Latents pass two-sided (no predicted sign) under the same magnitude and dose tests, matched on the library minus the latent's best-correlated concept, from the full pool without a regime.

**Why interchange rather than probe patching.** A preliminary result motivated the design. Probe directions for the pre-registered concepts were highly decodable (held-out $R^2$ above 0.95), yet additive activation patches along them changed the teacher's action as predicted in 0 of 24 cases. Decodability did not imply use, so the gate intervenes on states rather than on probe directions.

**Metric-weak handles.** Two handles, Pendulum energy and the NavModes-PPO gap concepts, admit only partial adversarial impostors (IC about 0.5 at matched fidelity) and are flagged as metric-weak. The main tables include them unweighted; Appendix I reports the augmentation comparisons with them excluded.

**Amendments before freezing.** (i) Additive probe patches (0 of 24 passing) were replaced by interchange interventions. (ii) Categorical concepts use categorical swaps. (iii) A residual cap of 0.75, a multiplicity-adjusted reference threshold and a direction-over-full-swap ratio with a random-direction control were added, and claims were narrowed to behaviorally validated concepts. (iv) Declared matching sets and regime-restricted pools were introduced. (v) A simulator bug in the response functional (an environment wrapper that was never reset, which silently collapsed Hopper horizons to one step) voided the early Hopper results; the horizon was re-declared at 25 steps with gating per torque channel before rerunning. (vi) The pumping-phase concept failed with the wrong sign; the diagnosis (energy input for a torque-actuated pendulum is proportional to $u\,\dot{\theta}$, with no $\cos\theta$ lever term) was registered as a corrected concept before training three new seeds per algorithm, on which it passed in 4 of 6 cases.

## C Deviations from the pre-registered protocol

1. Additive activation patching was replaced by interchange interventions before any distillation (Appendix B).
2. The Hopper response horizon was re-declared after the simulator bug, before the rerun.
3. The corrected pumping-phase concept counts as pre-registered only for the teachers trained after its registration.
4. Retention is undefined when $D_1$ fidelity is below 0.2, a guard added after the freeze and applied retroactively to 336 baseline records written before it.
5. The first CF-aug run leaked held-out states (Section 5.2); all reported CF-aug numbers are from the clean rerun, with an independently drawn DAgger refit at the same cell for the paired controls.
6. The confirmatory mixed models were fitted after the per-task paired analysis had been read, so they are confirmatory in protocol but not blind.
7. The three data-route controls (pool states, random pairs, unvalidated pairs) were designed and run after the clean CF-aug results had been read. They are post hoc; their pool, pair counts, budgets, seeds and evaluation are those of the clean rerun.
8. The coverage dose-response sweep, the sufficiency gate and its distillation run, the concordance companion metric, and the Hopper SAC children were designed and run after the initial experiments (Appendix I). They are post hoc. The sufficiency gate's floor (0.9 of

the raw-state tree on visited and space-filling validation states) was fixed before its results were read; the space-filling requirement was added after a first version with visited states alone passed energy-only sets on balancing teachers, and that first version's results were discarded unread beyond that observation. The Hopper SAC fine-tuning recipe was chosen after the original recipe's collapse had been observed, and the collapsed children are kept.

9. The mixture-of-linear-experts surrogate (B5) was added after the initial experiments and run through the frozen harness under the Stage-3 seed scheme (independent draws per cell); it is post hoc, and the equivalence-class analysis that includes it is reported next to, not instead of, the pre-registered one.
10. The harder CartPole held-out region is an alternative evaluated in a separate results root; the frozen configuration is unchanged and remains the one every main-text number uses. The selection rule was fixed before the probe ran.
11. Revision 2 changed four things, and every distillation stage was then rerun under the revised harness. Apart from the subsequent corrections and separately identified post-hoc analyses below, the historical results come from that rerun, whose records sit in the supplement's revision-2 root next to the original ones, except the harder-region CartPole comparison (Table 14) and the Hopper children, which were run under the original harness and are marked as such. (a) DAgger batches are refilled until they hold their quota of allowed states, so budgets are exactly equal (the original run rejected held-out states without refilling, giving augmented variants up to 29 percent more labels on Pendulum); (b) fidelity, IC and training labels use executed actions, clipped to the environment's bound (the original run used the unclipped PPO mean, which is why NavModes PPO had zero space-filling fidelity); (c) the augmentation's latent pairs use each latent's saved matching set, as the harness always did (the original run used a fixed index); (d) the sufficiency gate scores candidate sets on its own validation rollouts and space-filling states, excluding the held-out region, instead of on the harness sets. Teachers, handles and the concept gate were not rerun; the gate's residual rule, a cap on the mean residual per dose, is unchanged, and Appendix B reports how many of its decisions change under the per-pair rule that the evaluation uses.
12. The fifth task was added after the initial four-task experiments. Its concept registrations, predicted signs, matching sets, harness configuration, region-selection rule and pipeline order were written before any run on the task and are timestamped in the supplement; VIPER was not run on it because Box2D state cannot be set from an observation. Its results are reported in their own section and are not pooled into the four-task estimates of Table 5.
13. Revision 3: the augmentation code selected contiguous slices of an array that held all base endpoints before all counterfactual endpoints, so with the large pair sets of the coverage sweep most sweep surrogates received base states only and no counterfactual endpoints; the main clean run's small pair sets were consumed in full, but base-heavy in early iterations. The selection now draws pair indices, enters both endpoints of every chosen pair in the same batch and spreads the pairs over iterations, and the realized fraction is recorded. The fresh evaluation pairs, which had resampled the same cached dataset as the augmentation pool, now come from an independently collected dataset of the same exploration level (rollout seed changed), and the exact endpoint overlap is logged. The clean augmentation run, the three data-route controls and the coverage sweep were rerun on all five tasks under this code; the historical CF-aug, control and sweep tables report that rerun; the later categorical sensitivity is reported separately in Appendix H. The leaked first run (Appendix E) and the stage-4 records were not rerun, so their augmentation used the earlier ordering; their final training sets are unaffected because those pair sets were consumed in full.
14. Revision 4 corrected three baseline-side implementation errors and reran the affected cells: (a) the symbolic baseline B4 had been fitted to the first 8,000 aggregated DAgger states and clipped to $[-1, 1]$ on every task, so on Pendulum it could apply at most half the torque bound; it is now fitted to all 20,000 states and clipped to the environment bound, and B4 was rerun on all five tasks; (b) the return-optimized baseline B3 scaled its leaf values by the action bound inside its fitness rollouts but not in the harness, so on Pendulum (the only task with a non-unit bound) the evaluated policy was not the optimized one; the scale now lives inside the tree and B3 was rerun on Pendulum; (c) the unvalidated-pair control

truncated a short pair set without replenishing it, so on 11 teachers it replaced fewer states than CF-aug; its matching now draws more base states, and a dose-matched CF-aug variant (the CF-aug pairs subsampled to the control's count) was added and both were rerun on all five tasks. The premise test and every number that depends on B3, B4 or the unvalidated control were regenerated. The same revision seeded the torch sampling stream and the action-space sampler in dataset collection, which the first version had left unseeded, so the cached datasets are reproducible only from the saved caches, which are preserved and listed in a manifest; the stage runners draw one random stream per teacher across cells and skip completed cells on resumption without replaying their draws, so a run resumed after an interruption gives different draws for its remaining cells than an uninterrupted run, and exact reproduction of the saved records requires the original execution schedule. The case studies of Figures 4 and 9 had been computed on a second machine whose copies of two caches were locally collected substitutes; they were recomputed on the reference caches, and the paper reports those values. Revision 4 also added a sensitivity check of the concept gate under the executed-action convention that the harness uses (Appendix B).

15. In a subsequent sensitivity analysis, categorical evaluation and augmentation were changed from scalar-grid targets to exact level swaps. The NavModes sensitivity protocol was written before examining its outcomes, but after the historical results; it is post hoc. It reuses frozen teachers and handles, rescores 288 saved models, and adds 384 fits at matched label budgets and augmentation doses. Historical tables are unchanged; Appendix H reports the separate results and remaining matching-rule differences.
16. The endpoint-matched expansion (Appendix G) was designed after the historical results and an endpoint pilot, with matching thresholds fixed before expansion outcomes were examined. The equal-task weighting check was added after the historical results. Both are descriptive post-hoc analyses, not extensions of the confirmatory registration.

Table 4: Result provenance. Historical estimates retain legacy categorical matching. Later experiments are reported separately; the archive distinguishes saved-record reanalysis from retraining.

| Result family | Protocol and corrections | Executable evidence |
|---|---|---|
| Historical baselines and augmentation | Registered core; successive budget, clipping, pairing and baseline fixes; post-hoc controls | Revision-root stage 3–6 records; explicit legacy runner mode |
| Endpoint expansion | Post-hoc fixed expansion after pilot; legacy categorical pairs | 1,584 saved models; predictions, cell seeds and source manifests |
| Categorical sensitivity | Post-hoc protocol fixed before outcomes; legacy/exact comparison | 384 new fits; 288 historical models rescored; matched doses |
| Equal-task sensitivity | Post-hoc reweighting of saved contrasts | Teacher means within each fixed task; input record hashes |

## D IMPLEMENTATION OF BASELINES, VARIANTS AND CONTROLS

**State pools.** Every teacher has two cached datasets of 100,000 states. The *on-policy* dataset holds states visited by the stochastic policy. The *broadened* dataset holds rollouts in which each action is, with probability 0.2, a uniformly random action, and otherwise the stochastic policy's action plus Gaussian noise with standard deviation 0.3 times half the action range (for discrete actions, a random action with a further probability 0.1). The gate, the harness pairs, the CF-aug pairs and every data-route control draw from the broadened dataset; the distillation pools additionally reject the held-out region before matching, with a runtime assertion that no augmented state lies in it.

**DAgger (B1).** Five iterations of 4,000 states each; the first iteration rolls out the teacher, later iterations roll out the current tree, and the teacher labels every state. Held-out-region states are rejected from each batch before labeling. Trees are CART regressors (classifiers on CartPole) with at most $L$ leaves, fitted on the aggregate of all batches.

**VIPER (B2).** DAgger with per-state sample weights from a criticality proxy. States are recorded with their simulator state, and $Q(s,a) = r + \gamma V(s')$ is computed by one simulator step from $s$ under $a$, with $V$ the teacher's critic (the PPO value head, or the SAC $Q_1$ at the policy action). On CartPole

the weight is $\max_a Q(s,a) - \min_a Q(s,a)$ over both actions, as in the original. For continuous actions two proxies replace the enumeration: *drop*, $\max(0,\, Q(s,\pi(s)) - \frac{1}{4}\sum_{k=1}^{4} Q(s,\pi(s)+\epsilon_k))$ with $\epsilon_k \sim \mathcal{N}(0, 0.1^2 I)$, and *grad*, the norm of the finite-difference gradient of $Q$ in the action with step 0.1. Weights are shifted to be positive and normalized to mean 1.

**Return-optimized trees (B3).** A $(1+\lambda)$ evolution strategy over axis-aligned trees of fixed depth $\log_2 L$, whose internal nodes carry a coordinate and a threshold and whose leaves carry an action. Leaf values lie in $[-1,1]$ and are scaled by the action bound inside the tree, so the imitation warm start, the fitness rollouts and the harness score the same physical policy (Appendix C, item 14). The tree is warm-started for 60 generations on imitation loss to the DAgger data (mutation scale 0.2, 48 offspring), then evolved for 30 generations on the mean return of 3 seeded episodes (population 24, mutation scale 0.1). This baseline converged on Pendulum and CartPole only.

**Symbolic regression (B4).** gplearn genetic programming with population 500, 20 generations, primitives $\{+, -, \times, \div, \sin, \cos\}$ and parsimony coefficient 0.01, fitted to the teacher's actions on all 20,000 DAgger states with uniform weights, one program per action dimension, its output clipped to the environment's action bound (on CartPole, one program for the signed right-minus-left margin, thresholded at 0). Complexity is the program length in tokens; the program with the lowest raw fitness (imitation loss) among those within the token budget is selected from the final population, and budgets with no such program are reported as unrealizable.

**Vocabulary variants.** The concepts variant feeds $\phi_T(s)$, the validated non-post-hoc, non-confounded concepts, to the same DAgger procedure. The distilled-latent variant first fits, for each of the top three validated latents by effect size, a gplearn program (population 500, 20 generations, the same primitives, parsimony 0.01, at most 16 tokens) to the latent's value on the distillation states, drops latents with no program within budget, and then trains the tree over concepts plus distilled latents. The random-rotation control multiplies the state by the first $k$ columns of a random orthogonal matrix; the random-coordinate control keeps $k$ random coordinates; $k$ is the dimension of the validated set.

**Counterfactual augmentation and its controls.** CF-aug generates up to 600 pairs per validated handle from a 20,000-state pool sampled from the allowed broadened states with the evaluation pair generator's matching rules, then replaces part of each DAgger batch with pair endpoints, or appends them in the add variant. Pair indices are drawn at random, both endpoints of every chosen pair enter the same batch, and the pairs are spread evenly over the five iterations (Appendix C, item 13). The replaced share is the smaller of half the batch (2,000 of 4,000 states) and a fifth of the available pairs times two, so it is limited by how many pairs the handles supply: recorded per surrogate, it is 18 to 30 percent on Pendulum (median 24), 30 on CartPole, 30 to 42 on NavModes (median 39), 18 on Hopper and 18 to 48 on LunarLander (median 36); no main-study surrogate reached 50 percent. The coverage sweep (Appendix I) builds larger pair sets so that its nominal fractions are realized exactly, except in 48 of its 6,732 cells (36 Hopper and 12 Pendulum cells at the nominal 0.5 in the exploration sweep, realized 0.18 to 0.47), which the exploration-level summaries include at their nominal level; Appendix I reports the sweep with the three affected teachers excluded. The pool-states control replaces the same number of states with unpaired draws from the same pool. The random-pairs control keeps every CF-aug base state and replaces its partner by a pool state whose standardized distance from the base lies within 20 percent of the true counterfactual's displacement norm, chosen at random among such states or, if none exists, the nearest in norm; the median displacement norms of true and random pairs agree within 3 percent on every teacher. The unvalidated-pairs control runs the matching procedure for every library concept that failed the gate or was excluded by a ruling (including post-hoc and confounded rulings, and the corrected pumping concept only on the teachers it was registered for), with 2,048 base states per concept (revision 4; the first version used 256 and on 11 teachers fell short of the CF-aug count without replenishing), then subsamples to the CF-aug pair count. The control is absent on CartPole and on LunarLander PPO seed 2, which have no gate-failed concept, and on the three NavModes SAC teachers, whose only gate-failed concept is wall side, it reaches about 1,600 of 4,200 pairs; for those teachers a dose-matched variant subsamples the CF-aug pairs to the same count (Appendix F). Realized fractions are recorded on every control record.

**Adversarial impostors.** An MLP with two hidden layers of width 128 (ReLU) trained for 300 epochs with Adam (learning rate $10^{-3}$, batch 1,024) on the loss $\mathcal{L}_{\text{fit}} + \lambda\, \mathbb{E}_{(s,s')} \|f(s') - f(s)\|^2$, where $\mathcal{L}_{\text{fit}}$ is the mean squared error to the teacher's actions on $D_1$ (cross-entropy on CartPole, with

the sensitivity term on softmax outputs), the expectation runs over 512 pairs per step drawn from the target handle's own matched pairs, and $\lambda = 1$. Fine-tuned children, the positive control, continue the teacher's training from its final checkpoint with a new seed at learning rate $10^{-4}$.

## E THE LEAKED AUGMENTATION RUN

Our first full run of CF-aug reported held-out retention gains of 0.03 to 0.13 per task (+0.108 pooled). That run sampled the counterfactual pool from a broadened state dataset without rejecting the held-out region, and the pairs were appended to each DAgger batch after the batch's own rejection step. As a result, 1 to 6 percent of CF-aug's training states lay inside the region that every other surrogate was denied (7 to 24 percent of pair endpoints, depending on the task). Restricting the pool before matching reduces the pooled held-out effect from +0.108 to +0.037, and all results here come from that clean run, with a runtime assertion that no training state falls in the region. IC was little affected (per-task gains of 0.05 to 0.12 leaked against 0.04 to 0.11 clean). The leak fractions were measured on the original run; the leaking variant was rerun under the revised harness, but not under the pair-aware selection of revision 3, for the comparison above. The failure is easy to reproduce whenever augmentation draws from a different state source than the main data pipeline, and a held-out region that is filtered in only one of the two sources gives no warning.

## F ADDITIONAL RESULTS

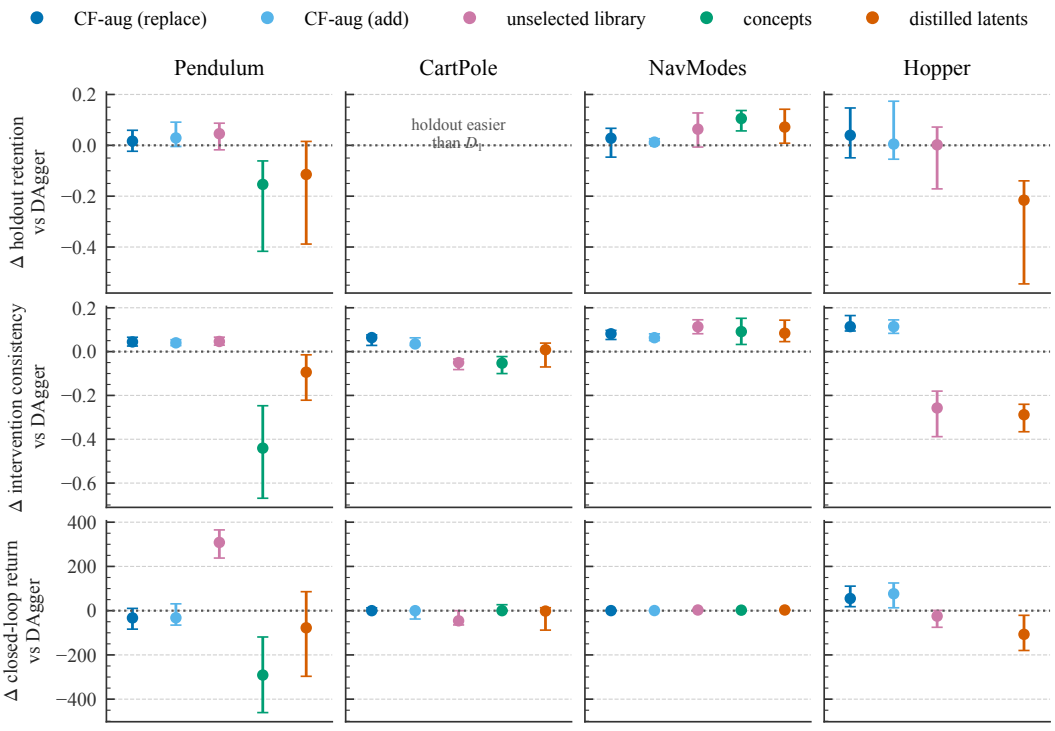



Figure 5: The overview of Figure 2 with the closed-loop return row that the main text omits: paired differences from DAgger trees at matched teacher, distillation seed and leaf budget (median and 95% bootstrap confidence interval).

**Off-distribution behavior.** Teacher-level mean gains over DAgger are small: held-out retention $+0.042$ $[+0.001, +0.083]$ and init-shift retention at severity 1.0 $+0.031$ $[+0.004, +0.059]$. The pre-registered models give weaker evidence: the pooled effect is +0.037 for the replace variant (Holm $p = 0.085$) and +0.044 for the add variant ($p = 0.042$), and against the independently drawn refit the median difference is $-0.01$ on Pendulum, 0.00 on CartPole, +0.03 on NavModes and +0.11 with

Table 5: Pooled mixed-model estimates relative to DAgger trees (27 teachers, 3 seeds, 4 budgets; random intercepts for teacher and for seed within teacher; the concepts column uses the 21 teachers with a concept vocabulary, since Hopper has none). CF-aug is the replace variant. Return is reported raw and standardized within each task (SD units over all records of that task), because the four tasks have reward scales from tens to thousands. Holm correction is across variants within each row. Every fit reported in this paper converged; the held-out retention, raw return, standardized return and leaked held-out models, and the controls' held-out-region IC model, needed a derivative-free optimizer after L-BFGS stopped short, with estimates unchanged to the precision shown. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

| | CF-aug | concepts | unselected library | distilled latents |
|---|---|---|---|---|
| Intervention consistency | $+0.091^{***}$ | $-0.269^{***}$ | $-0.002$ | $-0.108^{***}$ |
| Held-out retention | $+0.037$ | $-0.128^{***}$ | $+0.032$ | $-0.126^{***}$ |
| Init-shift retention | $+0.028$ | $-0.104^{***}$ | $-0.042$ | $-0.011$ |
| Dynamics retention $\times$ severity | $+0.020^{*}$ | $+0.014$ | $-0.013$ | $-0.004$ |
| Closed-loop return, raw | $+13$ | $-161^{***}$ | $+136^{***}$ | $-62^{**}$ |
| Closed-loop return, task SD units | $+0.12$ | $-0.25^{***}$ | $+0.47^{***}$ | $-0.05$ |
| $D_1$ fidelity | $-0.002$ | $-0.015^{*}$ | $-0.012$ | $-0.022^{***}$ |

a wide interval on Hopper (Table 7). Retention under shifted initial states is unchanged in the pooled estimate (+0.028, not significant), with a gain on Hopper only (+0.09). Under scaled dynamics the level effect is null but the interaction with severity is positive (+0.020, Holm $p = 0.013$), most visibly on Hopper (Appendix F). Return is unchanged on Pendulum, CartPole and NavModes and rises by 55 on Hopper, where every tree surrogate otherwise fails closed-loop. Pool states and random pairs give the same Hopper gains, so these effects are also coverage effects (Table 7). A first run of CF-aug leaked held-out states into its training data and reported a held-out gain of 0.11 pooled where the clean run gives 0.04 (Appendix E).

**Robustness.** All effects have the same sign for PPO and SAC teachers; SAC teachers gain more IC from CF-aug (+0.17 against +0.10 on Hopper, +0.07 against +0.03 on Pendulum), and the concept-only failure on Pendulum is larger for SAC (IC $-0.58$, return $-522$) than for PPO ($-0.41$, $-117$). The IC gain of CF-aug shrinks with tree size on Pendulum (+0.09 at 8 leaves, +0.02 at 64), consistent with larger trees covering the counterfactual directions on their own, and is flat on Hopper (+0.13 to +0.14 from 16 leaves). The loosest gate, which admits every pre-registered concept, is exactly the unselected library, so the two tested vocabularies bracket a threshold sweep at its ends; intermediate subsets were not tested. On Hopper every tree fails closed-loop, and the population-level gains in init-shift retention (+0.09) and return (+55) are visible on the case-study teacher at 32 leaves, where the CF-aug tree stays upright twice as long from shifted starts, and not at 16, where it falls sooner and loses fidelity despite a higher IC (Appendix K). Appendix F gives per-task tables, severity curves and budget dependence.

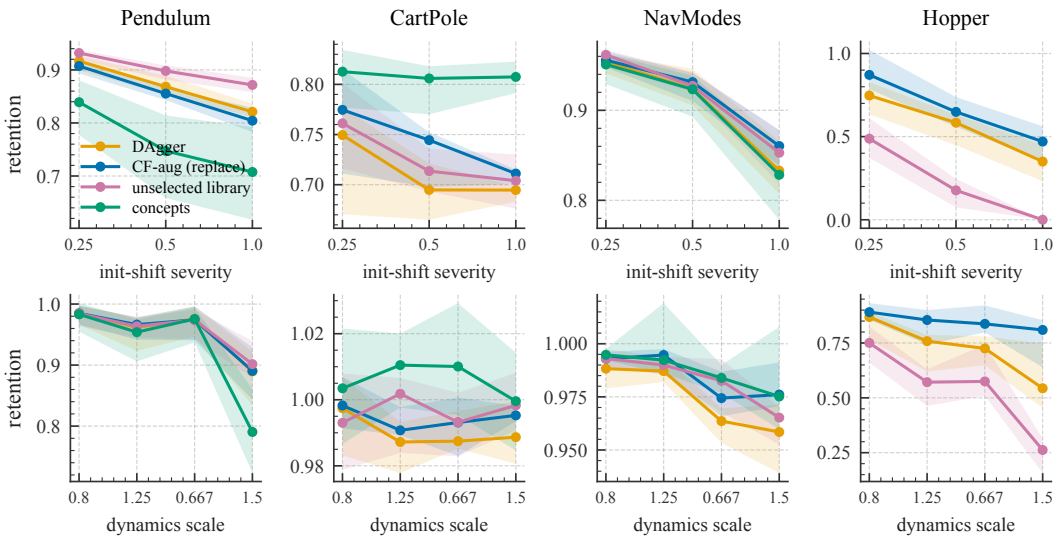



Figure 6: Retention against severity for shifted initial states (top) and scaled dynamics (bottom); median and 95% bootstrap confidence interval over teachers, seeds and budgets.

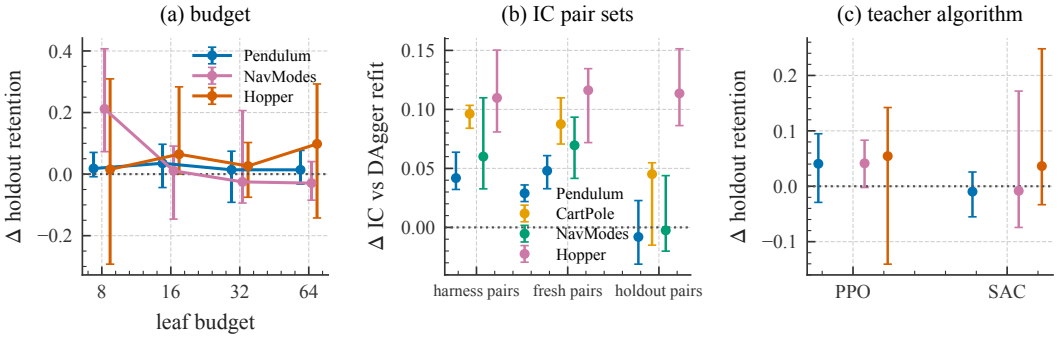



Figure 7: (a) Held-out retention gain of CF-aug over DAgger by leaf budget. (b) IC gain over an independently drawn DAgger refit on harness, fresh and held-out-region pairs. (c) Held-out retention gain by teacher algorithm.

**CF-aug against the random-pair control.** Pooled mixed-model coefficients with random pairs as the reference: IC on harness pairs $+0.021 \pm 0.005$ ($p = 10^{-4}$), on fresh pairs $+0.020 \pm 0.005$ ($p = 4 \times 10^{-4}$), on held-out-region pairs $+0.008 \pm 0.009$ ($p = 0.83$); pool states $+0.004$ and unvalidated pairs $-0.019$ on harness pairs. Per task, CF-aug minus random pairs on harness pairs: Pendulum $+0.01$, CartPole $+0.01$, NavModes $+0.03$, Hopper $+0.01$; minus pool states: $+0.01$, $-0.01$, $+0.02$, $+0.01$.

**Validated against gate-failed pairs at equal dose.** On every teacher but the three NavModes SAC teachers the unvalidated control now supplies as many pairs as CF-aug, so the two differ only in which concepts generated them. At those equal-dose cells CF-aug minus the unvalidated control on harness IC is $+0.025$ $[+0.014, +0.039]$ on Pendulum, $+0.005$ $[-0.020, +0.017]$ on the three NavModes PPO teachers, $+0.027$ $[+0.021, +0.045]$ on Hopper (sign test $p < 0.05$), $+0.026$ $[0.000, +0.052]$ on LunarLander and $+0.024$ $[+0.016, +0.034]$ pooled over the four tasks (252 cells). The NavModes SAC teachers have one gate-failed concept, wall side, which yields about 1,600 pairs against 4,200 for CF-aug, so their control replaces 16 percent of each batch; a CF-aug variant subsampled to the same 16 percent scores $-0.013$ $[-0.023, -0.007]$ against it on harness pairs and $-0.023$ on fresh pairs, while full-dose CF-aug scores $+0.028$ $[-0.016, +0.061]$. On those

Table 6: Median paired differences from the DAgger baseline trees (B1) per task, pooled over budgets. * Holm-corrected teacher-level sign test $p < 0.05$ across variants within task and metric; with three to six teachers only Pendulum (twelve) can reach significance.

| Task | Variant | held-out ret. | init-shift ret. | IC | return | $D_1$ fid. |
|---|---|---|---|---|---|---|
| Pendulum | CF-aug (replace) | $+0.02$ | $-0.01$ | $+0.04^*$ | $-33$ | 0.00 |
| | concepts | $-0.15$ | $-0.10$ | $-0.44^*$ | $-291$ | 0.00 |
| | unselected library | $+0.05$ | $+0.04^*$ | $+0.05^*$ | $+308^*$ | 0.00 |
| | distilled latents | $-0.11$ | $-0.03$ | $-0.09$ | $-77$ | 0.00 |
| CartPole | CF-aug (replace) | 0.00 | $+0.03$ | $+0.06$ | 0 | 0.00 |
| | concepts | $+0.10$ | $+0.12$ | $-0.05$ | 0 | $-0.07$ |
| | unselected library | 0.00 | 0.00 | $-0.05$ | $-46$ | $-0.01$ |
| | distilled latents | $+0.02$ | $-0.01$ | $+0.01$ | $-2$ | $-0.01$ |
| NavModes | CF-aug (replace) | $+0.03$ | $+0.03$ | $+0.08$ | $+0.2$ | $+0.02$ |
| | concepts | $+0.11$ | $-0.01$ | $+0.09$ | $+2.3$ | $+0.01$ |
| | unselected library | $+0.06$ | 0.00 | $+0.11$ | $+3.0$ | $+0.02$ |
| | distilled latents | $+0.07$ | $-0.02$ | $+0.08$ | $+3.1$ | $+0.02$ |
| Hopper | CF-aug (replace) | $+0.04$ | $+0.09$ | $+0.11$ | $+55$ | $-0.01$ |
| | unselected library | 0.00 | $-0.30$ | $-0.26$ | $-25$ | $-0.09$ |
| | distilled latents | $-0.22$ | $+0.01$ | $-0.29$ | $-107$ | $-0.27$ |

Table 7: Data-route controls: median paired differences from an independently drawn DAgger refit at the same cell (IC, held-out retention, $D_1$ fidelity) or from the DAgger baseline trees B1 (init-shift retention at severity 1.0, return), per task and pooled over budgets. * Holm-corrected teacher-level sign test $p < 0.05$ across the four augmentations within task and metric (only Pendulum, with twelve teachers, can reach it). The unvalidated control is absent on CartPole, and on the three NavModes SAC teachers it replaces 16 percent of each batch against 42 for CF-aug (Appendix D).

| Task | Augmentation | IC harn. | IC fresh | IC held-out | held-out ret. | init-shift | return | $D_1$ fid. |
|---|---|---|---|---|---|---|---|---|
| Pendulum | CF-aug | $+0.04^*$ | $+0.05^*$ | $-0.01$ | $-0.01$ | $-0.01$ | $-33$ | 0.00 |
| | pool states | $+0.03^*$ | $+0.03^*$ | 0.00 | 0.00 | 0.00 | $-29$ | 0.00 |
| | random pairs | $+0.03^*$ | $+0.03^*$ | 0.00 | $+0.01$ | 0.00 | $-12$ | 0.00 |
| | unvalidated pairs | $+0.02^*$ | $+0.02^*$ | $+0.01$ | 0.00 | $-0.02$ | $+41$ | 0.00 |
| CartPole | CF-aug | $+0.10$ | $+0.09$ | $+0.05$ | 0.00 | $+0.03$ | 0 | 0.00 |
| | pool states | $+0.13$ | $+0.11$ | $+0.03$ | $+0.03$ | $+0.04$ | 0 | $-0.02$ |
| | random pairs | $+0.12$ | $+0.11$ | $+0.05$ | $+0.01$ | $+0.04$ | 0 | 0.00 |
| NavModes | CF-aug | $+0.06$ | $+0.07$ | 0.00 | $+0.03$ | $+0.03$ | 0 | $+0.01$ |
| | pool states | $+0.05$ | $+0.06$ | 0.00 | 0.00 | $+0.01$ | 0 | $+0.02$ |
| | random pairs | $+0.03$ | $+0.03$ | $-0.03$ | $+0.03$ | $+0.01$ | 0 | $+0.02$ |
| | unvalidated pairs | $+0.05$ | $+0.05$ | $+0.01$ | $+0.03$ | 0.00 | 0 | $+0.01$ |
| Hopper | CF-aug | $+0.11$ | $+0.12$ | $+0.11$ | $+0.11$ | $+0.09$ | $+55$ | $-0.02$ |
| | pool states | $+0.09$ | $+0.10$ | $+0.10$ | $+0.05$ | $+0.08$ | $+94$ | $-0.02$ |
| | random pairs | $+0.08$ | $+0.10$ | $+0.08$ | $+0.09$ | $+0.08$ | $+95$ | $-0.02$ |
| | unvalidated pairs | $+0.08$ | $+0.08$ | $+0.11$ | $+0.06$ | $+0.08$ | $+74$ | $-0.04$ |

three teachers the gate-failed pairs are therefore at least as useful as the validated ones at equal dose, and CF-aug's advantage there is a dose effect; elsewhere the validated pairs carry a small advantage at equal dose.

**Equivalence.** With an equivalence margin of $\pm 0.02$ IC, the teacher-cluster interval of CF-aug minus the control on harness pairs lies inside the margin against random pairs on Pendulum ($+0.01$ $[0.00, +0.02]$) and CartPole ($+0.01$ $[0.00, +0.02]$) and against pool states on Hopper ($+0.01$ $[+0.01, +0.02]$); it crosses the margin against pool states on Pendulum ($+0.01$ $[0.00, +0.02]$) and CartPole ($-0.01$ $[-0.04, 0.00]$) and against random pairs on Hopper ($+0.01$ $[+0.01, +0.02]$), and lies outside it on NavModes ($+0.02$ $[0.00, +0.04]$ and $+0.03$ $[+0.01, +0.06]$), where CF-aug beats both controls. Restricting each comparison to cells whose $D_1$ fidelity differs by at most 0.02 leaves the point estimates small ($+0.004$ to $+0.011$ on Pendulum, $-0.003$ and $-0.006$ on CartPole, $+0.018$

and $+0.028$ on NavModes, $+0.017$ and $+0.021$ on Hopper), with intervals that stay inside the margin on Pendulum against random pairs only. On Pendulum CF-aug beats both controls on harness IC under the teacher-level test (by 0.01, Holm $p < 0.05$): a small advantage can be statistically supported and practically negligible at once.

**Teacher-level sensitivity check.** The pooled models carry random intercepts for teacher and seed but no teacher-specific treatment slopes. As a check that treats the 27 teachers as the only independent units, every contrast was averaged within teacher over seeds and budgets and summarized by the mean of the teacher means with a $t$-based 95 percent interval: CF-aug minus the refit $+0.100$ $[+0.077, +0.124]$ on harness IC, $+0.097$ $[+0.075, +0.119]$ on fresh pairs and $+0.033$ $[+0.009, +0.056]$ on held-out-region pairs (26 teachers); minus random pairs $+0.021$ $[+0.009, +0.033]$; minus pool states $+0.016$ $[+0.007, +0.026]$; minus unvalidated pairs $+0.041$ $[+0.019, +0.063]$ (24 teachers, three of them at unequal dose); held-out retention $+0.042$ $[+0.001, +0.083]$ and init-shift retention $+0.031$ $[+0.004, +0.059]$ against DAgger; $D_1$ fidelity $-0.002$ $[-0.012, +0.007]$. The principal IC contrasts therefore survive teacher-level inference. The coverage slope does not: the mean of per-teacher slopes over the 19 teachers whose coverage varies is $+1.1$ $[-0.3, +2.6]$ IC per unit coverage, so the precision of the pooled slope is optimistic, and the coverage association rests on the within-task rank correlations and the dose-response curves rather than on that coefficient. CF-aug minus pool states at each sweep fraction is $+0.017$ to $+0.024$ with lower bounds of $+0.006$ to $+0.016$.

**Equal-task weighting.** Teacher-weighted averages give Pendulum twelve of the 27 units, versus three CartPole and six each NavModes and Hopper. A post-hoc sensitivity gives each of these four fixed tasks weight $1/4$, averaging teacher contrasts within task first. Fresh IC changes by $+0.0961$ $[+0.0801, +0.1120]$ against refitting, $+0.0097$ $[+0.0021, +0.0174]$ against pool states and $+0.0154$ $[+0.0061, +0.0246]$ against random pairs. The qualitative conclusion survives this weighting. Intervals use variance $\sum_e s_e^2/(16n_e)$ and Welch–Satterthwaite degrees of freedom; they describe teacher variation within the four fixed tasks, not uncertainty over new tasks. The primary teacher-weighted estimand and task-specific results are retained. `scripts/analyze_equal_task.py` records all six harness/fresh contrasts and input hashes.

**IC and action error on different off-policy distributions.** Fidelity $D_2$ uses Sobol states in a box defined by marginal percentiles of the broadened data, with task-specific consistency repairs; Hopper instead uses sampled broadened rollout states. IC uses matched endpoints from broadened rollouts, so this comparison changes both the scoring functional and the evaluation distribution. Against DAgger, the augmentation changes $D_2$ by $-0.01$ on Pendulum, $+0.01$ on CartPole, $+0.02$ on NavModes and $-0.03$ on LunarLander while raising IC by 0.04, 0.06, 0.08 and 0.11, and the two controls behave alike; only on Hopper do both move together ($D_2$ $+0.08$, IC $+0.11$). The concept vocabulary on NavModes raises IC by 0.09 while leaving $D_2$ unchanged ($-0.06$ $[-0.14, +0.06]$), and on Pendulum lowers both ($-0.35$ and $-0.44$). These scores rank methods differently on four of five tasks, but the distribution mismatch prevents attributing that difference to pairwise scoring. The completed same-endpoint comparison in Appendix G finds strong agreement with ordinary endpoint error and no established incremental diagnostic utility.

**Raw and distilled latents.** Raw latents are a different case. As tree inputs they are highly effective (CartPole: space-filling fidelity 0.99 and IC 0.97, against 0.75 to 0.88 and 0.53 to 0.59 for the same-dimension controls), but they are not interpretable and not size-comparable, and distilling each latent into a short symbolic expression loses the advantage on every task (IC $-0.108$ pooled relative to DAgger; $-0.09$ on Pendulum and $-0.29$ on Hopper, and no better than the unselected library on CartPole and NavModes). For the latent route, abstraction rather than selection is the bottleneck.

**Budget dependence.** Median IC gain of CF-aug (replace) over DAgger at 8, 16, 32 and 64 leaves: Pendulum $+0.09, +0.08, +0.04, +0.02$; CartPole $+0.01, +0.07, +0.06, +0.07$; NavModes $+0.01, +0.09, +0.09, +0.09$; Hopper $+0.08, +0.13, +0.14, +0.14$.

## G ENDPOINT-MATCHED DIAGNOSTIC ANALYSIS

**Design.** After a pilot, we fixed the expansion protocol before examining its results (`docs/endpoint-matched-expansion.md`). All 33 teachers, four leaf budgets (8, 16, 32, 64), three distillation seeds and four methods (refit, CF-aug, pool states, random pairs) yield 1,584 new

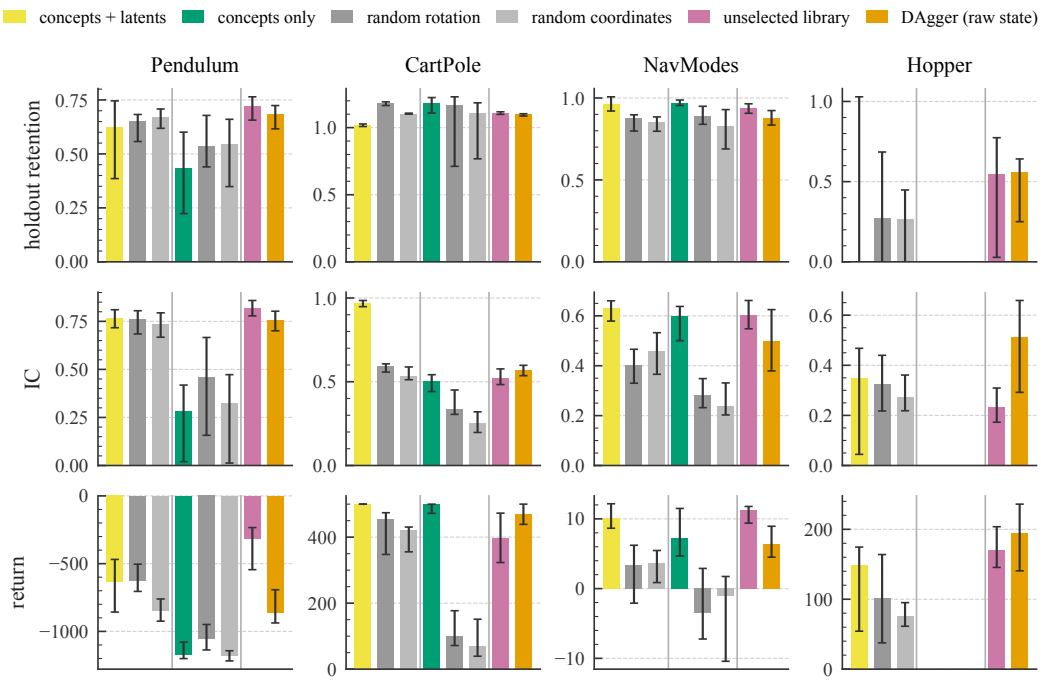



Figure 8: The vocabulary route and its controls (median and 95% bootstrap confidence interval over teachers, seeds and budgets). Groups, separated by vertical lines: validated concepts plus top-3 raw latents (hatched; not size-comparable) with a random rotation and random coordinates of the same dimension; validated concepts only, with matched controls; the unselected concept library and DAgger on the raw state.

fits. Every fit uses 20,000 labels and a stable per-cell seed. All policies of a teacher share the fresh endpoints and 20 return episode seeds. IC and normalized action error use identical clipped actions; mean error and IC use equal-handle weighting. The expansion retains historical scalar-grid categorical matching; the separate exact-swap sensitivity is in Appendix H. Models, predictions, inputs, source hashes and versions were saved; this is a new refit experiment, not an exact replay of historical policies.

**Unconditional comparisons.** Each paired difference is averaged within teacher over the twelve budget/seed cells, then summarized across teachers with a descriptive $t$-based 95% interval. The four-task CF-minus-refit IC gain of $+0.087$ $[+0.064, +0.110]$ accompanies lower mean endpoint error, $-0.072$ $[-0.090, -0.054]$. Against pool states the corresponding differences are $+0.024$ $[+0.012, +0.035]$ and $-0.017$ $[-0.024, -0.011]$; against random pairs, $+0.019$ $[+0.009, +0.029]$ and $-0.012$ $[-0.017, -0.006]$. Across 33 teachers, the median within-teacher rank correlation between IC and negative mean endpoint error is 0.969. These estimates concern the new fits; they do not replace the historical estimates in Table 1.

**Matching and attrition.** At each teacher, budget and distillation seed, compare all six pairs of methods. Endpoint matching requires mean-error difference at most 0.02 and 95th-percentile error difference at most 0.10. Return matching requires a difference at most 10% of absolute teacher return (scale at least 1). Both policies must achieve at least 90% of teacher return above the fixed task floor. These operational filters were fixed before outcomes, but use noisy 20-episode estimates and are not equivalence tests. Only 58 of 2,376 comparisons survive all filters (ten teachers, 77 distinct policies; Table 8). Median absolute IC gap is 0.0237, 95th percentile 0.0502 and maximum 0.0632. Policy pairs are dependent; these quantiles are descriptive, and selection changes the estimand. No systematic CF advantage is established among retained comparisons; several method contrasts have fewer than three teachers and no reported interval.

Table 8: Endpoint-matched expansion: sequential attrition of same-budget, same-seed method comparisons. Last two columns concern comparisons surviving all three filters.

| Task | All | Endpoint | + Return | + Competence | Teachers | Median $\lvert\Delta\mathrm{IC}\rvert$ |
|---|---|---|---|---|---|---|
| Pendulum | 864 | 330 | 43 | 5 | 2 | 0.019 |
| CartPole | 216 | 81 | 42 | 37 | 3 | 0.028 |
| NavModes | 432 | 126 | 39 | 16 | 5 | 0.014 |
| Hopper | 432 | 85 | 77 | 0 | 0 | – |
| LunarLander | 432 | 55 | 14 | 0 | 0 | – |

**Sensitivity and limits.** Matching aggregate endpoint errors does not match statewise errors. Conditioning additionally on both policies having small error at both ends of the same pairs generally yields small IC gaps, but a NavModes pool-versus-random comparison reaches 0.154 on only 82 pairs across two handles at an error threshold of 0.10. It is neither a full-distribution result nor a CF advantage. On CartPole, endpoint error below 1 forces exact binary action agreement; its zero conditional gaps are tautological. The pre-specified 54-setting sensitivity grid and cross-budget checks are retained in the companion JSON rather than selecting favorable thresholds. Severe attrition, fixed tasks and the restriction to four raw-state tree variants limit generalization. Strong score association and small selected gaps do not prove IC redundant or equivalent to error. No reproducible competent-policy failure uniquely diagnosed by IC has been established.

**Reanalysis.** Run `scripts/analyze_endpoint_expansion.py` from the revision root with `-root results/endpoint_expansion_v1` and an output report path. The report `paper/endpoint_matched_expansion_results.md` and companion JSON preserve task-specific contrasts, the full matching grid, teacher-level values and hashes. The original expansion completed in 19 minutes on the 20-core workstation. The reproduction bundle includes the saved models and arrays; source snapshots are checked against recorded hashes where available.

## H Categorical matching sensitivity on NavModes

**Correction and scope.** The categorical gate enumerates exact target levels, whereas the historical evaluation and augmentation code applied the scalar SD grid. We share the gate's target-level

enumeration and nearest-neighbor logic with evaluation, preserving evaluation's declared pool restrictions, per-pair residual cap of 0.75 and minimum of 30 retained pairs per transition. The frozen gate retains its full categorical candidate pool and mean-residual convention. Scalar and latent construction are unchanged; the historical generator remains available as an explicit legacy mode. This corrects target coverage, rather than making all gate and evaluation rules identical.

**Design and provenance.** The post-hoc protocol uses all six NavModes teachers, four leaf budgets (8, 16, 32, 64) and three distillation seeds, with 20,000 labels per fit. Fresh pairs use the independently collected broadened dataset; augmentation excludes the held-out region. Historical fresh pairs, teacher actions and augmentation arrays were reconstructed exactly against saved artifacts. All 288 historical models were rescored under both pair definitions. Corrected CF and random-pair fits use the same per-cell seeds and common augmentation count as their legacy comparators; unchanged refit and pool-state models are reused. A separate family fits old/new CF and unvalidated controls at a common dose across all four conditions, with its own fixed seed namespace. There are 864 condition records, comprising 384 new fits and 480 verified model reuses; these are repeated measurements, not 864 independent units. Return uses the same 20 episode seeds per teacher. The six workers completed in 4.2 minutes on the experiment workstation.

**Realized transitions.** Mode identity passes the frozen gate for PPO seed 1 and SAC seeds 0, 1 and 2. Legacy fresh evaluation includes the four adjacent directed transitions but neither extreme-to-extreme transition. Exact swaps add both missing transitions for every one of these teachers (Table 9); adjacent-transition counts are unchanged. PPO seeds 2 and 3 have no validated mode handle. Counts concern the mode handle itself: other handles can also supply cross-mode augmentation endpoints. In the allowed training pool, legacy mode-handle construction yields zero pairs for SAC seed 2, versus 454 before the cap under exact swaps. The full per-transition and training-pool audits accompany the analysis.

Table 9: Fresh mode-handle pairs for the two previously absent directed transitions. Legacy counts are zero throughout; exact counts are also the counts of unique state pairs.

| Teacher | $-0.6 \to +0.6$ | $+0.6 \to -0.6$ |
|---|---|---|
| PPO seed 1 | 89 | 69 |
| SAC seed 0 | 77 | 74 |
| SAC seed 1 | 75 | 74 |
| SAC seed 2 | 94 | 65 |

**Inference and results.** Primary IC averages only handles present under both evaluation definitions. Each paired contrast is averaged over the 12 budget/seed cells within teacher, then summarized across the six teachers with a descriptive $t$-based 95 percent interval; intervals are post hoc and not multiplicity-adjusted. Under corrected training and evaluation, CF improves IC and mean endpoint error over refitting (Table 10), while its IC intervals against all three augmentation controls span zero. In the separate equal-dose family, CF has slightly higher mean endpoint error than unvalidated pairs. IC gains therefore do not establish a benefit specific to validated handles.

Table 10: CF minus comparator under exact-swap training and evaluation. Mean paired differences and teacher-level 95% intervals ($n = 6$); lower endpoint error is better. The unvalidated comparison uses its own common-dose CF fits.

| Comparator | IC difference | Mean endpoint-error difference |
|---|---|---|
| Refit | $+0.054\ [+0.027, +0.081]$ | $-0.056\ [-0.074, -0.037]$ |
| Pool states | $+0.014\ [-0.011, +0.038]$ | $-0.007\ [-0.020, +0.007]$ |
| Random pairs | $+0.024\ [-0.008, +0.056]$ | $-0.005\ [-0.017, +0.008]$ |
| Unvalidated pairs | $-0.010\ [-0.021, +0.002]$ | $+0.005\ [+0.002, +0.009]$ |

**Rescoring versus retraining.** Rescoring fixed legacy CF models changes mean IC by $+0.0041$ $[-0.0020, +0.0102]$. Holding exact evaluation fixed, corrected training changes the CF-minus-pool increment by $+0.0052$ $[-0.0039, +0.0144]$ and the CF-minus-random increment by $+0.0114$ $[-0.0238, +0.0465]$. Corrected-training return differences are $-0.64$ $[-1.89, +0.60]$ against refitting, $+0.64$ $[-0.67, +1.95]$ against pool states and $-0.13$ $[-1.17, +0.91]$ against random pairs.

These checks restore categorical coverage without establishing a control advantage or eliminating shared-generator bias.

**Artifacts.** The protocol is `docs/categorical-sensitivity.md`; the analysis entry point is `scripts/analyze_categorical_sensitivity.py`, and the report and companion JSON are `paper/categorical_sensitivity_results.md` and `paper/categorical_sensitivity_results.json`. Records, inputs, predictions and model hashes are under `alt/v2/results/categorical_sensitivity_v1`. Local reanalysis reproduces remote statistics to numerical precision. These artifacts remain separate from the historical result roots and are included in the versioned supplementary archive.

# I COVERAGE SWEEP, SUFFICIENCY GATE, POSITIVE CONTROL, FURTHER CLASSES AND REGIONS

**Exploration levels.** We collect a pool at exploration level $e$ from rollouts of the stochastic teacher in which each action is uniformly random with probability $e$ and otherwise carries Gaussian noise with standard deviation $1.5e$ times half the action range (for discrete actions, a further random action with probability $e/2$); $e = 0$ is the on-policy dataset and $e = 0.2$ the broadened dataset of the main study. Levels 0.05 and 0.4 use 30,000 newly collected states per teacher; space-filling pools are Sobol samples in the box spanned by the 1st to 99th percentiles of the broadened states (on Hopper, where such states are not valid, a broadened sample). Every pool rejects the held-out region before matching. CF-aug pairs at each level come from the evaluation pair generator described in Section 4 with up to 1,024 base states per dose, so that fractions up to 0.75 are realizable.

**Coverage.** For a surrogate with training set $X$, coverage is the fraction of the harness pair endpoints, standardized by the broadened dataset's mean and standard deviation, whose nearest neighbor in $X$ lies within 0.25 standard deviations. The coverage association regresses the paired IC gain on the paired coverage gain over all 5,184 augmented surrogates, with random intercepts for teacher and seed nested within teacher, and then adds an indicator for CF-aug against unpaired states.

Table 11: Coverage sweep: median IC gain over an independently drawn DAgger refit at the same cell (pooled over budgets and seeds; $n$ = 144 Pendulum, 36 CartPole, 72 NavModes and Hopper, 324 pooled). Top: augmented fraction at the standard pool. Bottom: exploration level at a nominal half of each batch; the on-policy and 0.05 pools of two Hopper SAC teachers and the space-filling pool of one Pendulum SAC teacher supply fewer pairs (realized 0.18 to 0.47), and excluding those three teachers changes the pooled CF-aug medians to $+0.02$, $+0.05$, $+0.09$, $+0.07$ and $+0.03$ and the pool-state medians to $0.00$, $+0.03$, $+0.06$, $+0.06$ and $+0.03$ (on-policy, 0.05, 0.2, 0.4, uniform), with the same ordering.

| | | augmented fraction | | | |
|---|---|---|---|---|---|
| Task | augmentation | 0.10 | 0.25 | 0.50 | 0.75 |
| Pendulum | CF-aug / pool | $+0.02$ / $+0.02$ | $+0.04$ / $+0.03$ | $+0.05$ / $+0.04$ | $+0.07$ / $+0.05$ |
| CartPole | CF-aug / pool | $+0.06$ / $+0.04$ | $+0.08$ / $+0.08$ | $+0.12$ / $+0.10$ | $+0.14$ / $+0.14$ |
| NavModes | CF-aug / pool | $+0.04$ / $+0.01$ | $+0.05$ / $+0.02$ | $+0.09$ / $+0.05$ | $+0.12$ / $+0.08$ |
| Hopper | CF-aug / pool | $+0.09$ / $+0.07$ | $+0.14$ / $+0.12$ | $+0.18$ / $+0.17$ | $+0.21$ / $+0.21$ |
| Pooled | CF-aug / pool | $+0.04$ / $+0.03$ | $+0.06$ / $+0.05$ | $+0.10$ / $+0.07$ | $+0.11$ / $+0.09$ |
| | | pool exploration level (and space-filling) | | | |
| Task | augmentation | 0 (on-policy) | 0.05 | 0.4 | uniform |
| Pendulum | CF-aug / pool | $+0.03$ / $-0.01$ | $+0.04$ / $0.00$ | $+0.05$ / $+0.04$ | $+0.01$ / $-0.01$ |
| CartPole | CF-aug / pool | $-0.01$ / $+0.02$ | $+0.03$ / $+0.05$ | $+0.09$ / $+0.07$ | $-0.05$ / $+0.02$ |
| NavModes | CF-aug / pool | $+0.02$ / $0.00$ | $+0.04$ / $+0.03$ | $+0.10$ / $+0.08$ | $+0.05$ / $+0.05$ |
| Hopper | CF-aug / pool | $+0.05$ / $+0.04$ | $+0.10$ / $+0.09$ | $+0.15$ / $+0.13$ | $+0.18$ / $+0.17$ |
| Pooled | CF-aug / pool | $+0.03$ / $0.00$ | $+0.06$ / $+0.03$ | $+0.08$ / $+0.07$ | $+0.04$ / $+0.04$ |

**Pooled exploration effects.** Unpaired on-policy states give $+0.003$ IC against the refit, while matched pairs from the same pool give $+0.025$, at similar measured coverage. A random-action probability of 0.05 gives gains of 0.03 to 0.06; levels 0.2 and 0.4 give 0.07 to 0.10. Space-filling

states give about 0.04 pooled but hurt CartPole ($-0.05$), where uniform draws include states the teacher never reaches. These comparisons concern sampling rules, not a pairwise training objective.

**Coverage association.** Spearman correlation between the IC gain and the coverage gain: Pendulum $+0.02$ ($n = 2{,}304$; coverage is 0.97 for every Pendulum training set, so its slope is not interpretable), CartPole $+0.44$ (576), NavModes $+0.19$ (1,152), Hopper $+0.46$ (1,152), pooled $+0.22$ (5,184). Mixed-model slopes of IC gain per unit coverage gain: CartPole $+0.47 \pm 0.04$, NavModes $+0.14 \pm 0.02$, Hopper $+0.63 \pm 0.03$, pooled $+0.26 \pm 0.02$; the teacher-level check in Appendix F does not confirm the pooled slope's precision, so the association is read from the within-task correlations and the dose-response curves. With coverage in the model, the CF-aug indicator against unpaired states is $+0.028$ ($p < 10^{-18}$): the pairs' increment is not explained by the coverage they add. CF-aug minus pool states at the same cell is $+0.012$, $+0.014$, $+0.016$ and $+0.019$ pooled at the four fractions (teacher-level sign tests $p \leq 0.002$), and at half of each batch $+0.015$ on Pendulum, $+0.009$ on CartPole, $+0.041$ on NavModes and $+0.015$ on Hopper.

**Concordance.** The concordance correlation coefficient $2\,\mathrm{cov}(x, y)/(\mathrm{var}\, x + \mathrm{var}\, y + (\bar{x} - \bar{y})^2)$ of the same response differences penalizes magnitude error as well as ordering error. Over all 5,508 surrogates its rank correlation with IC is 0.991 (Pendulum), 0.999 (CartPole), 0.998 (NavModes) and 0.996 (Hopper). At half of each batch, the augmentation gains under concordance are $+0.07$, $+0.13$, $+0.08$ and $+0.17$ for CF-aug against $+0.05$, $+0.12$, $+0.09$ and $+0.18$ under IC.

**Sufficiency gate.** A 64-leaf CART tree fitted on 20,000 teacher-labeled allowed states from the broadened dataset scores each candidate set; the reference is the same tree over the raw state. Both are scored on two fresh validation sets that exclude the held-out region and are disjoint from the harness sets: $V_1$, states from new rollouts of the teacher, and $V_2$, new space-filling states. The floor is 0.9 of the reference fidelity on $V_1$ and on $V_2$; both sets are needed, because upright angle alone scores 0.99 on the visited states of a balancing Pendulum teacher and 0.12 on space-filling states. Greedy extension adds, at each step, the library concept that most reduces the binding deficit (the smaller of the two fidelity margins). The library is the pre-registered set, including gate failures and concepts excluded by a ruling, and the corrected pumping concept on the teachers it was registered for. Table 12 lists every decision. The gated set is then distilled as in the concepts variant, with the Stage-4 random streams.

Table 12: Sufficiency-gate decisions per teacher. Fidelities are $V_1$/$V_2$; the floor is 0.9 of the raw tree's. Concept labels: P1 energy, P2 pumping phase, P2′ corrected pumping, P3 upright angle, P4 angular velocity; C1 pole angle, C2 pole velocity; N1 mode, N2 gap error, N3 wall side, N4 goal distance, N5 velocity.

| Task | teacher | raw tree | validated set (fid.) | pass | gated set (fid.) | library (fid.) |
|---|---|---|---|---|---|---|
| Pendulum | PPO 0 | 0.97/0.88 | P3 (0.95/0.54) | no | P3 P2 P1 (0.99/0.86) | 0.89/0.93 |
| | PPO 1 | 0.97/0.89 | P1 P3 (0.95/0.57) | no | P1 P3 P4 (0.90/0.89) | 0.91/0.92 |
| | PPO 2 | 0.90/0.90 | P3 (0.98/0.20) | no | P3 P4 P1 (1.00/0.92) | 1.00/0.92 |
| | PPO 10 | 0.99/0.86 | P3 P2′ (0.99/0.67) | no | P3 P2′ P1 (0.91/0.90) | 0.91/0.91 |
| | PPO 11 | 0.99/0.88 | P3 P2′ (0.99/0.71) | no | P3 P2′ P2 (0.93/0.88) | 0.91/0.90 |
| | PPO 12 | 0.93/0.89 | P1 P3 P2′ (0.95/0.89) | yes | same | 0.90/0.92 |
| | SAC 0 | 0.78/0.85 | none (0/0) | no | P2 P3 (0.97/0.77) | 0.86/0.89 |
| | SAC 1 | 0.91/0.86 | P1 P3 (0.98/0.40) | no | P1 P3 P4 (0.92/0.92) | 0.89/0.92 |
| | SAC 2 | 0.90/0.88 | P1 P3 (0.94/0.44) | no | P1 P3 P4 (0.94/0.92) | 0.94/0.91 |
| | SAC 10 | 0.87/0.82 | P3 (0.96/0.11) | no | P3 P2 (0.87/0.77) | 0.95/0.92 |
| | SAC 11 | 0.84/0.89 | P3 (0.99/0.12) | no | P3 P2 (0.93/0.89) | 0.93/0.93 |
| | SAC 12 | 0.94/0.88 | P3 P2′ (0.94/0.39) | no | P3 P2′ P1 (0.95/0.89) | 0.95/0.94 |
| CartPole | PPO 0–2 | 0.78–0.85/0.86–0.88 | C1 C2 (0.79–0.83/0.80–0.82) | yes | same | 0.84–0.86/0.80–0.82 |
| NavModes | PPO 1 | 0.91/0.62 | N1 N2 N4 (0.91/0.62) | yes | same | 0.90/0.54 |
| | PPO 2 | 0.90/0.57 | N2 N4 (0.90/0.62) | yes | same | 0.91/0.50 |
| | PPO 3 | 0.90/0.51 | N2 N4 (0.90/0.45) | no | N2 N4 N1 (0.91/0.52) | 0.90/0.40 |
| | SAC 0–2 | 0.87–0.89/0.44–0.52 | N1 N2 N4 N5 (0.88–0.89/0.33–0.37) | no | full library (0.88–0.89/0.25–0.37) | same |

**Greedy selection from the empty set.** As a comparison for the validated start, the same greedy search was run from the empty set (v1g) with the same floor, and the selected sets were distilled under the same seeds. On Pendulum it selects two or three concepts (against two or three from the validated start), on CartPole one or two (against two) and on NavModes two to five (against two, three or five), and it passes the floor for two NavModes SAC teachers that the validated start could not extend to a passing set: seed 0 with mode, gap error and goal distance (space-filling fidelity 0.407 against a floor of 0.400) and seed 2 with mode, gap error, wall side and velocity (0.397 against 0.394), while seed 1 fails from both starts. At matched cells the distilled v1g trees differ from the

Table 13: The sufficiency gate. A validated concept set passes if a 64-leaf tree over it, fitted on teacher-labeled states, reaches 0.9 of the raw-state tree's fidelity on fresh visited and fresh space-filling validation states; the gate extends failing sets greedily from the library. Right: median paired differences of the gated vocabulary from the validated-only and the unselected-library trees at matched cells ($^*$ teacher-level sign test $p < 0.05$). On CartPole the gated set equals the validated set, so the differences are sampling variation between independent DAgger runs.

| | | | gated minus validated | | | gated minus library | | |
|---|---|---|---|---|---|---|---|---|
| Task | pass as is | gated size (library) | IC | return | held-out | IC | return | held-out |
| Pendulum | 1 of 12 | 2–3 (4–5) | $+0.45^*$ | $+711^*$ | $+0.30^*$ | $-0.03^*$ | $-31$ | $+0.04$ |
| CartPole | 3 of 3 | 2 (3) | $-0.02$ | 0 | $+0.01$ | $-0.06$ | $+43$ | $+0.10$ |
| NavModes | 2 of 6 | 2–5 (5) | $+0.01$ | 0 | $-0.03^*$ | $-0.01$ | 0 | $+0.01$ |

validated-start trees by 0.00 IC on CartPole and NavModes and by $-0.08$ IC $[-0.29, 0.00]$ and $-140$ return (sign test $p < 0.05$) on Pendulum, where the validated start keeps upright angle in every set while the empty start drops it for the PPO teachers; against the validated-only trees v1g still gains $+0.23$ IC and $+331$ return on Pendulum.

**A harder CartPole held-out region.** The frozen region (pole angular velocity in $[0.25, 1]$) gives every surrogate retention above 1. A probe fixed the selection rule before running: among one-sided tails at the pooled visited 90th percentile of each coordinate's magnitude, take the one with the lowest median DAgger retention over the three teachers at 16 and 64 leaves, subject to at least 50 visited states per teacher. It selected the positive pole-angle tail (angle above 0.0079 rad), with median retention 0.81 against 1.09 for the frozen region. Stages 3 to 6 were run for CartPole under this region in a separate results root that shares teachers, handles and code with the main study (714 records). This comparison was made under the original harness, before revision 2, and was not repeated under the revised one; both halves of Table 14 are original-harness numbers, so its frozen-region rows differ from Table 6. Every CartPole conclusion survives: the augmentation's IC gain is +0.07 and its controls give +0.06 to +0.08; its held-out retention gain over the refit at the same cell is +0.04 and the pool-states control gives the same; the concept vocabulary has the best held-out retention and the largest init-shift gain, as it had for init-shift under the frozen region; and the mixture of linear experts gains IC and return.

Table 14: CartPole under the frozen and the harder held-out region, both under the original harness (before revision 2): median paired differences from DAgger trees ($n = 36$; $^*$ row-level Wilcoxon, Holm-corrected across variants, descriptive only). Held-out retention is the informative column that the frozen region lacked.

| Region | Variant | held-out ret. | init-shift ret. | IC | return | $D_1$ fid. |
|---|---|---|---|---|---|---|
| frozen | CF-aug | 0.00 | $+0.06^*$ | $+0.05^*$ | $+5$ | $-0.02$ |
| | concepts | $+0.05$ | $+0.14^*$ | $-0.03$ | $+50$ | $-0.04^*$ |
| | unselected library | $+0.01$ | $+0.07^*$ | $-0.04$ | 0 | $-0.01$ |
| | distilled latents | 0.00 | $+0.04$ | $+0.01$ | $+38^*$ | $+0.01$ |
| harder | CF-aug | $+0.01$ | $+0.08^*$ | $+0.07^*$ | $+52^*$ | $-0.02^*$ |
| | concepts | $+0.11^*$ | $+0.16^*$ | $-0.01$ | $+82^*$ | $-0.08^*$ |
| | unselected library | 0.00 | $+0.05$ | $-0.01$ | $-8$ | $-0.01$ |
| | distilled latents | $-0.04^*$ | $-0.01$ | $-0.04^*$ | $-7$ | 0.00 |

**A fourth surrogate class.** B5 fits the CART partition of B1 on the same DAgger data, then a ridge regression (regularization 1) on the states of each leaf for continuous actions, or a logistic regression for discrete ones, falling back to the leaf constant when a leaf has fewer states than coordinates plus two or a single class. It runs under the Stage-3 seed scheme at every teacher, seed and budget (324 records). Table 15 gives the paired differences from B1. In the premise test with B5 added, classes that mix it with DAgger or VIPER trees have a median IC spread of 0.39 on Pendulum (21 classes), 0.14 on CartPole (5), 0.10 on NavModes (9) and 0.13 on Hopper (7), against 0.28, none, 0.05 and 0.07 for tree-only classes.

Table 15: Mixture of linear experts (B5) minus DAgger tree (B1) at matched teacher, seed and budget: median paired differences, * teacher-level sign test $p < 0.05$. $n$ = 144, 36, 72, 72.

| Task | $D_1$ fid. | $D_2$ fid. | held-out ret. | init-shift ret. | IC | switching | return |
|---|---|---|---|---|---|---|---|
| Pendulum | 0.00 | $+0.06^{*}$ | $-0.01$ | $+0.06^{*}$ | $+0.09^{*}$ | $+0.15^{*}$ | $+209^{*}$ |
| CartPole | 0.00 | $+0.07$ | 0.00 | $-0.01$ | $+0.02$ | $+0.02$ | 0 |
| NavModes | $+0.14^{*}$ | $+0.11^{*}$ | $+0.08^{*}$ | $+0.01$ | $+0.24^{*}$ | $+0.15^{*}$ | $+5^{*}$ |
| Hopper | $+0.17^{*}$ | $+0.22^{*}$ | $+0.20$ | 0.00 | $+0.17^{*}$ | $+0.20^{*}$ | $+792^{*}$ |

**Hopper SAC positive control.** The Stage-2 child recipe fine-tuned each teacher from its final checkpoint with a fresh replay buffer, 512 random-action warm-up steps and learning rate $10^{-4}$; the three Hopper SAC children collapsed to returns of 317, 298 and 746 against a target of 2,200 and were unusable. The revised recipe fills the buffer with the parent's own policy for 5,000 steps before any update, then trains for 10,000 steps with policy learning rate $3 \times 10^{-5}$, critic learning rate $10^{-4}$ and $\log \alpha = -6$. The children reach returns of 2,928, 2,271 and 2,910 and score IC 0.88, 0.89 and 0.90 against their parents on the parents' harness pairs, against 0.46, 0.35 and 0.41 for an independently seeded teacher.

## J THE FIFTH TASK: LUNARLANDER

**Summary.** We added LunarLanderContinuous-v3 with the full pipeline, from teacher training through the gate, the frozen-harness configuration, the baselines, both routes, the controls and the later additions, and we wrote its concept registrations, predictions and held-out-region rule before any run on the task. It differs from the four tasks in kind: contact-driven dynamics with a regime change at touchdown, two coupled engines with dead zones, and binary contact flags in the observation. The main findings replicate on it, with the exceptions stated below.

**Gate and metric.** Of 30 pre-registered concept tests, 21 pass: vertical velocity and tilt on every teacher, altitude on four of six, angular velocity on two and horizontal offset on three, which matches the pre-specified predictions except for angular velocity, predicted to pass and passing on two teachers (the offset was registered as least certain). The metric validates as before: fine-tuned children score IC 1.00 (PPO) and 0.74 to 0.77 (SAC) against 0.84 to 0.90 and 0.30 to 0.52 for independently seeded teachers, and adversarial impostors score 0.02 to 0.54 at fidelity 0.78 to 0.96. The rule selected the high-altitude band ($y > 1.25$, 4 to 14 percent of visited states) as the held-out region; DAgger trees retain 0.56 there at the median.

**Premise.** Trees and symbolic laws at matched fidelity again differ in mechanism (five such classes, median IC spread 0.17 and up to 0.19), and the mixture of linear experts widens it (classes mixing it with trees have a median IC spread of 0.29 against 0.02 for trees alone) while beating DAgger trees by 0.17 in fidelity, 0.27 in IC and 195 in return.

**Data route.** Counterfactual augmentation raises IC by 0.11 on harness and fresh pairs at equal budget, improves neither held-out retention ($-0.04$) nor init-shift retention (0.00), and the first, leaking version of the augmentation reported a held-out gain of $+0.19$ that the clean run turned into $-0.04$, so the leak of Section 5.2 reproduces on a task built after the fix. The controls place this task at the endpoint-selection end of the five: random-direction pairs recover 0.11 of the 0.14 pooled gain, so the handle-specific increment is 0.03 ($p = 0.03$ on harness pairs, 0.04 on fresh pairs; teacher-level interval $[-0.01, +0.07]$ over six teachers), and unpaired pool states recover only 0.08, so handle-directed sampling contributes 0.05 beyond unpaired coverage here (Table 16).

**Vocabulary route.** As predicted, no concept carries the contact flags, so trees over the validated concepts lose fidelity ($-0.03$), IC ($-0.05$) and held-out retention ($-0.10$) against raw-state trees, and the unselected library does the same ($-0.04$ IC, $-0.10$ held-out). The sufficiency gate passes one of six validated sets as it is; one more passes after adding altitude, and for the other four the validated set cannot be extended to the floor on space-filling states, where the raw-state reference tree itself scores only 0.05 to 0.22. A greedy search from the empty set passes all three SAC teachers and none of the PPO teachers. Passing a space-filling floor this close to zero is not meaningful sufficiency, so on this task the gate's verdicts say little either way. Over the six teachers the gated

trees differ from the validated ones by $+0.01$ IC, $+0.02$ fidelity and $+30$ return, none significant, and the trees from the empty-start sets score the same.

**Registration.** The task document (in the supplement, with its version timestamp) fixed, before any run on the task: five scalar concepts with regimes, response channels, predicted signs and matching sets (altitude $y$ in the airborne descending regime on the main engine, sign $-$; vertical velocity $v_y$ on the main engine, $-$; tilt $\theta$ on the lateral engine, $+$; angular velocity on the lateral engine, $+$; horizontal offset $x$ on the lateral engine, $-$; matching sets follow the coordinate-sharing rule of Appendix B); the harness configuration (action bound 1; init-shift by adding up to 4 and 2 units of horizontal and downward velocity, 0.5 rad of tilt and 1 rad/s of spin times severity; dynamics families gravity, engine gain and lander mass at the four scales; switching slice on $v_y$; safety as no crash; invariance as staying within 0.4 rad of upright); the region-selection rule (the pre-declared probe of Appendix I with a visited fraction between 2 and 25 percent); and the pipeline order. The predictions written in that document were: vertical velocity and tilt pass everywhere, altitude on most teachers, angular velocity passes, offset is least certain; the premise holds; counterfactual augmentation raises IC and unpaired states reproduce it; the validated concepts suffice on the gate but both concept trees fail at touchdown and the raw-state tree wins on return. All held except the angular-velocity prediction (two of six passes) and the sufficiency-gate prediction: the validated sets fail the gate on space-filling states for five of six teachers.

**Gate outcome.** Passes per concept over the six teachers: altitude 4, vertical velocity 6, tilt 6, angular velocity 2, horizontal offset 3 (21 of 30). Latents pass on five of six teachers (0 to 14 per teacher); SAC seed 2 has no validated latent handles. The SAC teachers are far from linear in the handles (a linear surrogate scores IC 0.50 to 0.59 against 0.85 to 0.98 for PPO), which is why cross-seed IC is high for PPO and low for SAC.

Table 16: LunarLander (pre-specified before any run on it): median paired differences from DAgger trees (B1) at matched teacher, seed and budget ($n = 72$; 70 where retention is undefined). Top: variants against Stage-3 B1. Bottom: the clean augmentation and its controls against an independently drawn refit at the same cell on the three pair sets. $^*$ teacher-level sign test $p < 0.05$ (Holm-corrected across variants within metric, which with six teachers no Stage-3 variant or control passes; the mixture-of-experts row is uncorrected).

| Variant | held-out ret. | init-shift ret. | IC | return | $D_1$ fid. |
|---|---|---|---|---|---|
| CF-aug (clean, replace) | $-0.04$ | 0.00 | $+0.11$ | $-8$ | $+0.03$ |
| CF-aug (first run, leaked) | $+0.19$ | $+0.05$ | $+0.15$ | $+24$ | $+0.05$ |
| concepts | $-0.10$ | $-0.10$ | $-0.05$ | $-16$ | $-0.03$ |
| unselected library | $-0.10$ | 0.00 | $-0.04$ | $-8$ | 0.00 |
| distilled latents | $-0.08$ | $-0.07$ | 0.00 | $-13$ | $+0.02$ |
| mixture of linear experts | $+0.21$ | $+0.20^*$ | $+0.27^*$ | $+195^*$ | $+0.17^*$ |
| | IC harness | IC fresh | IC held-out | $D_1$ fid. | held-out ret. |
| CF-aug | $+0.11$ | $+0.11$ | $-0.03$ | $+0.01$ | $-0.09$ |
| pool states | $+0.07$ | $+0.07$ | $-0.01$ | $+0.03$ | $-0.06$ |
| random pairs | $+0.10$ | $+0.11$ | $-0.03$ | $+0.02$ | $-0.06$ |
| unvalidated pairs | $+0.10$ | $+0.12$ | $+0.01$ | $-0.01$ | $+0.04$ |

**Pooled controls.** Mixed-model coefficients against the refit: CF-aug $+0.137$, random pairs $+0.107$, unvalidated pairs $+0.108$, pool states $+0.084$ on harness pairs, and $+0.135$, $+0.103$, $+0.115$, $+0.082$ on fresh pairs from the independent dataset. With random pairs as the reference, CF-aug is $+0.030 \pm 0.012$ ($p = 0.03$; $+0.031$, $p = 0.04$, on fresh pairs), pool states $-0.023 \pm 0.012$ ($p = 0.05$) and unvalidated pairs $-0.001$; at equal dose CF-aug minus the unvalidated control is $+0.026$ $[0.000, +0.052]$ over the five teachers that have a gate-failed concept. On held-out-region pairs every augmentation is at or below zero ($-0.01$ to $-0.05$, none significant): the high-altitude band is the start of every episode, the pool excludes it, and the teacher's counterfactual responses there are unlike those below it. Every augmentation also lowers held-out retention against the refit ($-0.04$ to $-0.10$, significant for CF-aug only), the one task where the coverage effect has a cost. In the coverage sweep the pairs stay ahead of unpaired states by 0.03 to 0.07 at every fraction and the CF-aug indicator adjusted for coverage is $+0.023$ ($p = 4 \times 10^{-5}$).

**Premise and the fourth class.** Sixteen behavioral-equivalence classes among the baselines; the five that mix trees with symbolic laws span IC by 0.17 at the median and up to 0.19 (PPO seed 2, fidelity 0.53 to 0.57, IC 0.47 to 0.67). With the error tail matched, four such classes remain (median 0.12, up to 0.19), and with return matched as well, none. With the mixture of linear experts added, classes mixing it with trees have a median IC spread of 0.29 (two classes) against 0.02 for tree-only classes (four), the largest being 0.50 (SAC seed 0, mixtures of 8 to 64 leaves against 8- and 64-leaf trees at fidelity 0.81 to 0.86).

**Sufficiency gate.** One of six validated sets passes as it is (SAC seed 0, whose raw-state tree has space-filling fidelity 0.00, so only the visited floor binds). The other five fail on space-filling states: the raw-state reference scores 0.05 to 0.22 there, the floor is 0.04 to 0.19, and the validated sets score 0.00 to 0.11. Adding altitude lifts SAC seed 2 over its floor; for the other four the full library of five concepts stays below it. The greedy search from the empty set passes SAC seed 1 as well (altitude, vertical velocity, tilt and angular velocity, 0.089 against a floor of 0.080) and SAC seed 0 with altitude alone, and fails the three PPO teachers with the full library, so what the validated start cannot extend to the floor is not always unreachable. Over the six teachers the gated trees differ from the validated ones by $+0.01$ IC, $+0.02$ fidelity, $+0.07$ held-out retention and $+30$ return, none significant, and the trees from the empty-start sets score the same.

## K HOPPER CASE STUDY

At 16 leaves the DAgger tree has $D_1$ fidelity 0.64, IC 0.46 and return 280; the CF-aug tree has fidelity 0.61, IC 0.67 and return 223. From six shared shifted starts the DAgger tree falls after 69 steps on average (return 184) and the CF-aug tree after 52 (return 136). At 32 leaves the CF-aug tree is ahead on IC (0.73 against 0.67) and return (401 against 246) at lower fidelity (0.67 against 0.71), and from shifted starts it stays upright for 96 steps against 50. The Hopper handles are torque-channel latents, so the IC gain reflects how the teacher's torques respond to changes in those latents; on this teacher it keeps the larger tree upright longer without teaching it to hop, costs visited-state fidelity at both sizes, and at 16 leaves buys no closed-loop benefit at all.

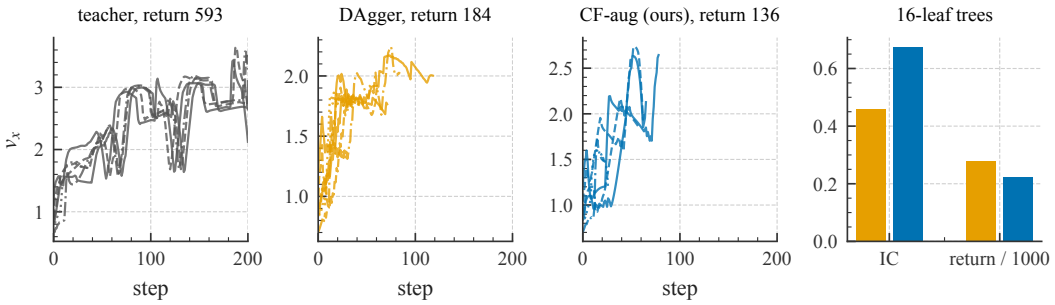



Figure 9: Hopper, PPO teacher selected by the same rule, 16-leaf trees, six shifted-start rollouts each: forward velocity until the fall, and IC and return of the two trees. On this teacher the 16-leaf CF-aug tree has the higher IC and the lower fidelity and return.

## L PRINTED TREES

The concept-only tree at the 8-leaf budget for the Pendulum teacher of Figure 4, a separate fit for the same teacher (the figure shows the 16-leaf budget), which has $D_1$ fidelity 0.99 and IC 0.41:

```
angle <= 0.05:
  energy <= 0.12 -> -0.21
  energy > 0.12:
    angle <= -0.88 -> 0.80
    angle > -0.88:
      energy <= 2.53 -> -1.82
      energy > 2.53:
        angle <= -0.08 -> 0.75
        angle > -0.08 -> 0.11
angle > 0.05:
  energy <= 5.01:
    energy <= 0.02 -> -0.26
    energy > 0.02 -> -1.05
  energy > 5.01 -> 0.46
```

No branch depends on the sign of $\dot{\theta}$, so the tree applies the same torque when the pendulum passes a given angle moving up or down and cannot apply the appropriate velocity-dependent damping in both directions near the top.