CONDITIONAL POLICY ADVANTAGE: CHANGING THE EXPERT OR TEST PHYSICS CAN REVERSE POLICY RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does telling a robot controller the physics it faces, such as gravity, really help it handle changing physics? Across four simulated locomotion tasks, we show that the answer, and which architecture uses physics best, depends on two rarely reported choices: the expert whose actions the controller learns to copy, and the physical conditions used for testing. With training states fixed, changing only the expert reverses architecture rankings. With trained controllers fixed, changing only the test physics reverses whether physics inputs help. Even test sets drawn from one distribution rank architectures in opposite orders in 7 of 24 comparisons. These reversals are predictable from separate data, and inside a single controller the expert decides where accurate physics hurts. We call this dependence *Conditional Policy Advantage* and propose an Expert-by-Test Grid to report it.

1 INTRODUCTION

Does telling a controller the physics it faces help it handle changing physics? Physics-conditioned controllers use known properties, such as mass and joint damping, to choose actions (Yu et al., 2017). One training approach learns an expert policy at each physical setting and combines their actions through behavior cloning (BC). Testing whether the physics inputs help, and which architecture uses them best, then involves two choices that papers rarely report: which expert supplies actions at each training setting, and which physical settings are used for testing.

Our central claim is that a policy's advantage depends jointly on these two choices, so a ranking from one expert assignment and one test set need not generalize. We call a return gap under a stated expert and test its *Conditional Policy Advantage*. The expert trained at a setting is often not the best expert there: another expert scores higher at 33 of 44 training settings. We therefore compare the expert trained at each setting (OWN) with the expert achieving the highest return there (BEST), with states and training fixed (Figure 1, top). This alone reverses which architecture wins. Holding the trained policies fixed and changing only the test physics does the same, even between test sets drawn from one distribution (bottom). More episodes cannot fix this, because they reduce only reset noise.

Prior work selects or combines experts to improve learning (Cheng et al., 2020; Liu et al., 2023). We instead ask when a policy's advantage survives these two choices, and turn the answer into a reporting protocol, the Expert-by-Test Grid (Section 6). Our contributions are:

1. **The expert changes the answer.** With training states fixed, changing only the expert reverses architecture rankings, including comparisons with matched backbones and training schedules (Section 3).
2. **The test changes the answer.** With trained policies fixed, changing only the test physics reverses whether physics inputs help. Even test sets drawn from one distribution rank Concat and FiLM in opposite orders in 7 of 24 comparisons (Section 4).
3. **Explaining sensitivity to expert and test choices.** Reversals across test sets follow from how much a gap varies across physics and can be predicted from separate data. Inside a single policy, the expert decides where accurate physics hurts, and relabeling alone repairs or creates these failures (Sections 4.3 and 5).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

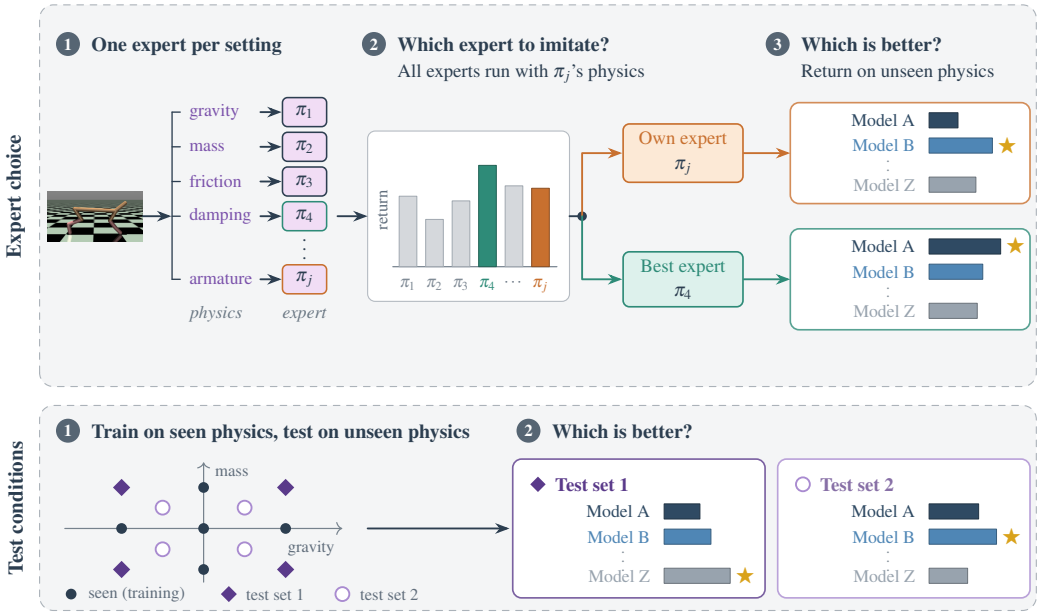

Figure 1: Two choices can change which policy wins. Top: each physical setting has its own expert. At setting j, students imitate either the expert trained there (Own) or the expert with the highest return there (Best), and the best student depends on this choice. Training states are held fixed. Only the expert providing the action labels changes. Bottom: students trained on seen physics can rank differently on two unseen test sets. Models A to Z stand for the compared students, such as Concat, FiLM, Gated-LoRA, and Physics-blind. A star marks the best student. Bar lengths are schematic.

2 Study Design

2.1 Fixed States and Alternative Experts

Physical settings. The conditioned policy receives known physical parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. Changing these values leaves the state and action spaces unchanged. Training uses the default (nominal) setting and two endpoints for each parameter:

$$\Theta_{\text{train}} = \{\boldsymbol{\theta}_0\} \cup \{\boldsymbol{\theta}_0 + \delta_i^- e_i, \boldsymbol{\theta}_0 + \delta_i^+ e_i : i = 1, \ldots, d\}. \tag{1}$$

Here e_i changes only parameter i, and δ_i^- and δ_i^+ give its lower and upper offsets. We use $d = 5$, giving $J = 11$ configurations. The parameters are gravity, mass, friction, damping, and armature. The environments are HalfCheetah, Ant, Hopper, and Walker2d, using Gymnasium's v4 variants. The original test set has thirteen unseen configurations that combine changes in two or three parameters. The ranges and exact targets are specified in Appendix A.

Expert bank and states. Each environment has a fixed bank of PPO experts (Schulman et al., 2017), one trained at each configuration. Let D_j be the states collected from expert π_j at its training configuration. Both assignments use the same pooled states and sample weights. Only the expert providing the action labels changes.

Training-only expert selection. Let G_{ij}^{train} be the mean return of expert i at training configuration j. We compare two assignment rules:

$$b_{\text{OWN}}(j) = j, \qquad b_{\text{BEST}}(j) = \arg\max_i G_{ij}^{\text{train}}. \tag{2}$$

BEST selects the largest observed return among the available experts. Students without a retrieved expert action (unanchored policies) fit

$$L_m(b) = \frac{1}{N} \sum_{j=1}^{J} \sum_{s \in D_j} \left\| f_m(s, \boldsymbol{\theta}_j) - \pi_{b(j)}(s) \right\|_2^2, \qquad N = \sum_j |D_j|. \tag{3}$$

Anchored students learn a residual on top of an expert action, using the same labels and states.

2.2 POLICY FAMILIES AND TRAINING CONTROLS

Table 1 shows how each policy uses physics. Concat joins the state s and physical parameters $\boldsymbol{\theta}$ into one input vector. FiLM takes state features and scales and shifts them using functions of $\boldsymbol{\theta}$. Physics-blind receives only s. It is a separately trained student using the same imitation data, and tests whether explicit physics inputs help beyond the state information. These three models share hidden blocks and a training schedule, and Concat versus FiLM is our main comparison. An initial study also compares Concat with Gated-LoRA and three anchored methods.

Table 1: Policy designs and training phases. Each method uses the same training setup under both expert assignments.

Policy	How physics affects the action	Phases
Concat	Appended to the state input	One
FiLM	Scale and shift of hidden features	One
Physics-blind	No physics input	One
Gated-LoRA	Gated low-rank updates to a frozen trunk	Two
Anchor-MLP	Retrieved expert plus an MLP residual	One
Anchor-Joint	Retrieved expert plus a joint-gated residual	Two
Anchor-Factored	Retrieved expert plus a factored-gate residual	Two

Gated-LoRA trains a state-only base network, freezes it, and then learns low-rank weight updates scaled by gates computed from $\boldsymbol{\theta}$. Its block order and skip connections differ from Concat, so this comparison covers complete training procedures. The anchored policies start from a retrieved expert's action and add a learned correction (Appendix B).

2.3 RETURN GAPS AND THEIR UNCERTAINTY

Evaluation. We compare raw return averaged over the test configurations, separately for each task. The initial comparison uses 13 configurations, 20 episodes per configuration, and three BC fits that share one expert bank and state pool. Policies use matched reset seeds, which fix the random initial state of each episode, and take no test-time gradient steps. Expert selection and student fitting use training settings only.

For policies A and B, expert assignment b, and a test set T, define

$$D_b(T) = \frac{1}{|T|} \sum_{\theta \in T} \left[\bar{R}_{A,b}(\theta) - \bar{R}_{B,b}(\theta) \right]. \tag{4}$$

This is *Conditional Policy Advantage*: a return gap tied to a specified assignment and test set. Here, $\bar{R}$ averages episodes. Positive values favor A. We call a reversal supported when both 95% gap intervals exclude zero and have opposite signs under two assignments or test sets. We separately report cases where only the mean ordering changes.

Two sources of uncertainty. Let a test set contain C physics configurations with R reset episodes each. For the paired return difference between two policies, with independently sampled configurations and independent reset errors within a configuration,

$$\mathrm{Var}\left[D_b(T)\right] = \frac{\sigma^2_{\mathrm{phys}}}{C} + \frac{\sigma^2_{\mathrm{reset}}}{CR}. \tag{5}$$

Here σ^2_{phys} is the variance of the gap across physical configurations and $\sigma^2_{\mathrm{reset}}$ its variance across resets within a configuration. Intervals for a fixed test set contain only the second term, which more resets can shrink. Claims about the physical distribution need both terms. Section 4.3 uses this decomposition to predict which comparisons reverse.

3 THE EXPERT CHANGES THE ANSWER

We first hold training states fixed, change only the expert that labels them, and ask which architecture uses physics best. Our main pair is Concat and FiLM, which share hidden blocks, optimizer, and training schedule. We follow the same pair when we change the test in Section 4.

3.1 A MATCHED PAIR CHANGES ITS WINNER

We train Concat and FiLM with three BC seeds per task. Within each assignment, paired models share training states and expert labels. Figure 2 shows all nine comparisons. In HalfCheetah, OWN favors FiLM in all three fits, while BEST favors Concat. The preferred architecture therefore changes even with matched backbones and training schedules. Ant has one supported reversal, while Walker2d has none. A separate Hopper study adds six FiLM comparisons with four supported reversals (Appendix F).

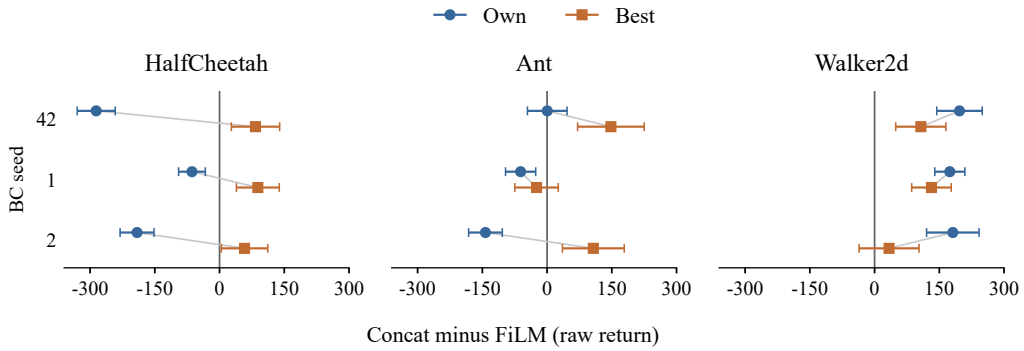

Figure 2: Concat minus FiLM with matched backbones and training schedules. Each row is one BC seed. Bars: paired pointwise 95% intervals on fixed test physics. Positive gaps favor Concat. All nine planned comparisons are shown.

3.2 THE SAME EFFECT ACROSS FIVE POLICIES

The initial study compares Concat, Gated-LoRA, and three anchored policies (Table 2). The HalfCheetah gap between Concat and Gated-LoRA changes from -431.5 to $+200.5$ raw return, in the same direction for all three training seeds. The policy with the highest mean return changes in 2 of 4 tasks, although some margins are small: on Walker2d under BEST, Gated-LoRA leads Concat by only 2.4 raw return.

Table 2: Mean raw return over 13 test settings and three BC seeds. Cells show OWN $\rightarrow$ BEST. Appendix D gives training variation.

Method	HalfCheetah	Ant	Hopper	Walker2d
Concat	414 $\rightarrow$ 1044	702 $\rightarrow$ 1237	1171 $\rightarrow$ 1116	1110 $\rightarrow$ 1614
Gated-LoRA	845 $\rightarrow$ 844	1017 $\rightarrow$ 1461	740 $\rightarrow$ 877	941 $\rightarrow$ 1617
Anchor-MLP	798 $\rightarrow$ 669	863 $\rightarrow$ 1041	923 $\rightarrow$ 1108	1036 $\rightarrow$ 1254
Anchor-Joint	677 $\rightarrow$ 749	866 $\rightarrow$ 1140	443 $\rightarrow$ 927	962 $\rightarrow$ 1164
Anchor-Factored	842 $\rightarrow$ 769	841 $\rightarrow$ 1187	819 $\rightarrow$ 1074	819 $\rightarrow$ 1165

Supported reversals. With episode-level intervals and two additional training fits for HalfCheetah and Ant, Concat/Gated-LoRA reverses in 6 of 16 comparisons: five favor Concat in HalfCheetah and one favors Gated-LoRA in Walker2d (Table 3). Including the anchored policies gives 49 supported reversals among 160 comparisons, with at least one in every task and training-seed group. We report anchored pairs separately because reassignment also changes the match between the retrieved expert and the action labels (Appendix F).

Table 3: Assignment reversals with training states fixed, on the original test set. Supported reversals have opposite signs and both pointwise 95% intervals exclude zero. The unanchored column reports Concat/Gated-LoRA. The other columns add the anchored methods. Comparisons share fitted policies.

	Unanchored pair	All five methods	
Task	Supported	Supported	Mean reversal
HalfCheetah	5/5	24/50	30/50
Ant	0/5	13/50	22/50
Walker2d	1/3	7/30	11/30
Hopper	0/3	5/30	8/30
Total	6/16	49/160	71/160

3.3 WHY THE EXPERT MATTERS, AND REPLICATION

Expert selection. The expert trained at a setting is often not the best there (Figure 3, Appendix C), and several settings share one best expert. Reassignment therefore changes both the source of action labels and the number of experts supplying them. The selected experts keep positive average gains over OWN on new resets in all four HalfCheetah banks: the original 1M bank, a 3M bank, and two independently trained 1M banks, A and B.

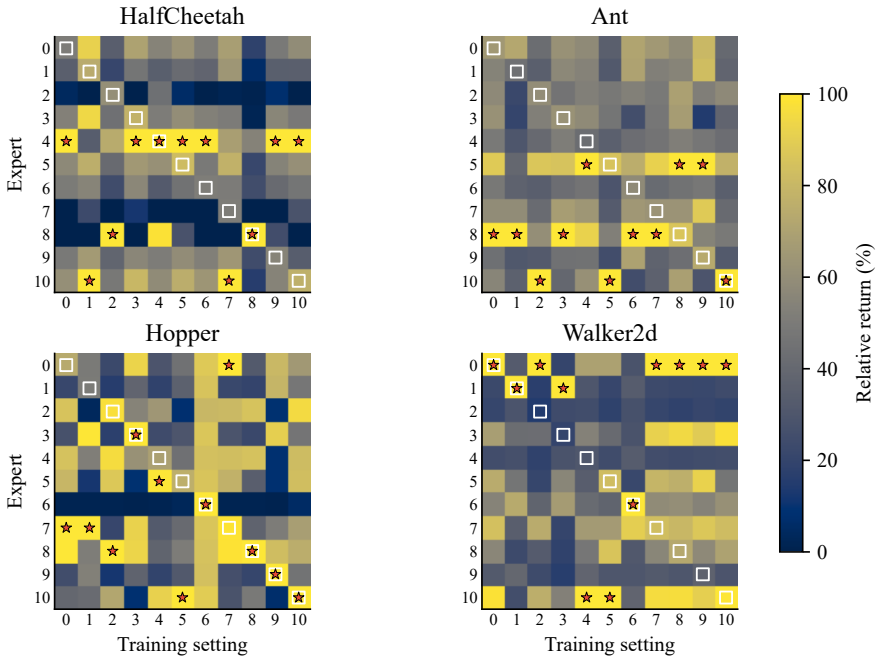

Figure 3: Training returns normalized by each column's maximum. Rows are experts and columns are training settings. Squares: Own. Stars: Best. Best entries often lie off the diagonal, and several columns share a winner. Negative values are clipped to zero. Indices follow Appendix A.

Other banks and tasks. In banks A and B, Concat gains more from reassignment than Gated-LoRA at every BC seed on both the original and interior test sets (Section 4.1), with two mean reversals on the interior set in each bank (Appendix E). In the 3M bank, Gated-LoRA instead gains more on the interior set, so another bank can change the response (Appendix E). In a registered Reacher extension, Best improves all three students without changing their order, so reassignment does not always reverse a ranking (Appendix K).

4 THE TEST CHANGES THE ANSWER

Having shown that the expert changes the preferred architecture, we now fix the trained policies and vary only the test physics. This changes both the benefit of physics inputs (Section 4.1) and the ranking between conditioned architectures (Section 4.2), and Section 4.3 explains why.

4.1 SAME POLICIES, DIFFERENT TEST PHYSICS

Both expert assignments are tested on combinations of endpoint values (original) and interior values (labeled fresh in Figure 4). Both sets change several parameters at once. Eight of the thirteen interior configurations lie outside the training settings' convex hull.

Do physics inputs help? It depends on the test. In bank A under Best, Concat and FiLM lose to Physics-blind on original physics and win on the interior set in all three BC seeds and both reset sets. All 24 gap intervals exclude zero (Figure 4). FiLM's interior-set advantage is unresolved in the original 1M bank, and the 3M controls are mixed (Appendix E).

Physics, not reset noise, drives the change. In a crossed Concat/Gated-LoRA evaluation on independent HalfCheetah banks A and B, changing physics at fixed resets gives 2 supported reversals in 20 comparisons. Changing resets at fixed physics gives 0 in 20 (Appendix G).

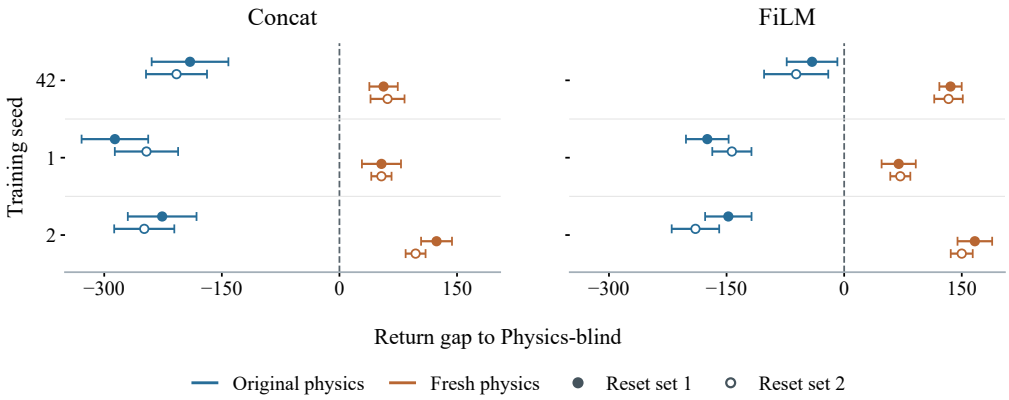

Figure 4: Return gaps to Physics-blind in bank A under Best. Positive values favor the conditioned policy. Bars: paired pointwise 95% intervals. Changing test physics reverses all six comparisons under either reset set.

4.2 INDEPENDENT TEST SETS FROM ONE DISTRIBUTION

Can rankings change when test sets come from the same distribution? We compare 24 fixed Concat/FiLM pairs across four tasks, three training seeds, and two assignments. Each pair is evaluated on eight independent test sets, giving 192 return-gap intervals. Each set contains 32 physics configurations with four episodes each. Each configuration moves two or three random axes 25% to 100% of the way from nominal toward a random endpoint, by the same rule for all sets.

In 12 of 24 policy comparisons, different sets support opposite rankings (Figure 5): 3, 2, 1, and 6 of six in HalfCheetah, Ant, Walker2d, and Hopper. A Bonferroni correction across all 192 intervals, added after the study, retains seven, with at least one in every task. Each interval describes its own test set, so a clear result on one sampled set need not hold on another. Appendix I.1 gives the sampling and interval rules.

Reversals robust to single-configuration removal. Removing each test configuration in turn, five comparisons keep both a positive and a negative supported gap after every deletion (5/24), so these reversals do not depend on one configuration (Table 10 in Appendix I.1).

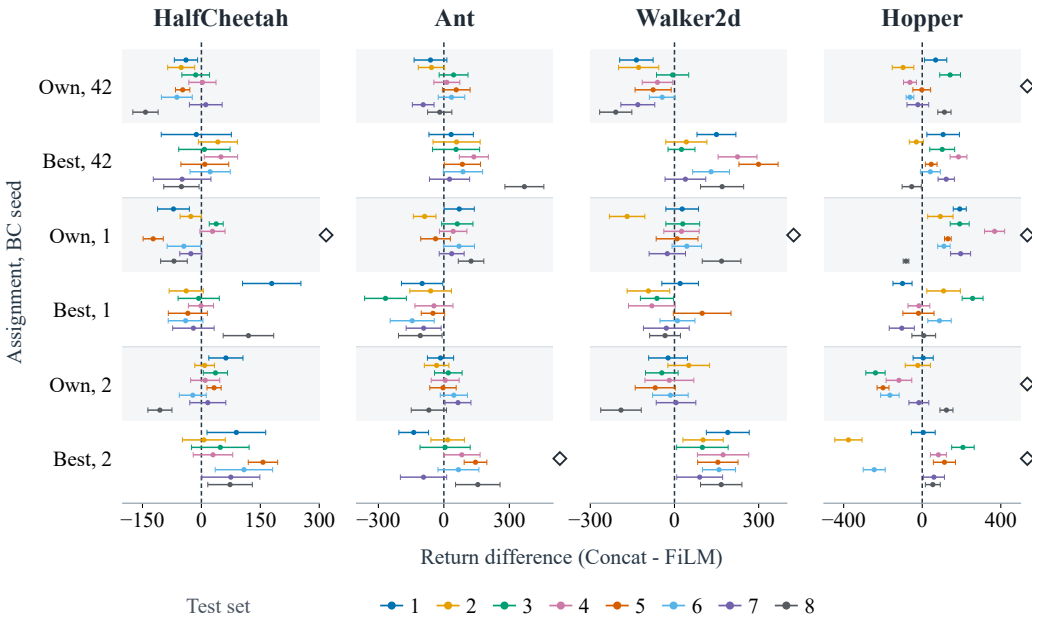

Figure 5: Concat minus FiLM on eight independent test sets, each with 32 physics settings and four episodes per setting. Bars: pointwise 95% intervals conditional on test physics. Each row fixes two trained policies, and colors are test sets, so bars on opposite sides of zero show a reversal without retraining. Positive gaps favor Concat. Diamonds mark reversals that survive correction across all 192 intervals. Horizontal scales differ by task.

4.3 Why some comparisons reverse

Equation 5 explains why a precise result on one test set can mislead: a new test set also redraws the physics. Adding the physics term to the same 192 comparisons keeps every estimate but cuts supported rankings from 100 to 9 (Appendix I.2).

We turn this into a risk score. For each comparison, we estimate the mean gap and both variance terms on a separate reference set. Treating a test-set mean as normal, let p_+ and p_- be the chances that one test set supports a positive or a negative gap. By inclusion and exclusion, the chance that K independent sets show both signs is

$$\text{risk}_K = 1 - (1 - p_+)^K - (1 - p_-)^K + (1 - p_+ - p_-)^K, \tag{6}$$

and we use $K = 8$ to match our study. Computed without the test sets, this score separates the comparisons that reversed from those that did not (AUC 0.81 after correction, Figure 6). A comparison is fragile when its gap varies strongly across physics relative to its average.

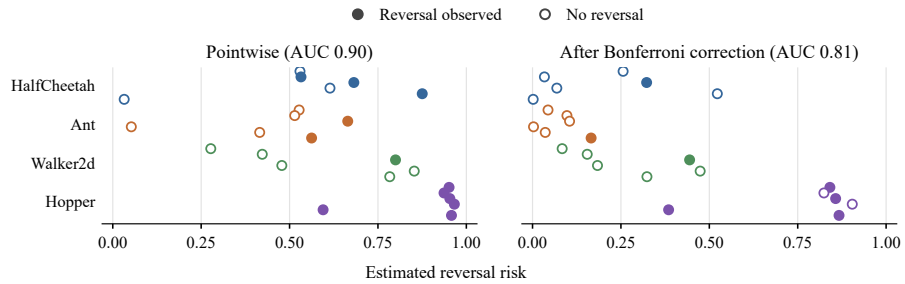

Figure 6: Reversal risk (Equation 6) from separate reference data. Each point is one fixed Concat/FiLM comparison. Filled markers: supported gaps of both signs were observed. Left: pointwise intervals. Right: Bonferroni correction across 192 intervals.

5 THE EXPERT DECIDES WHERE ACCURATE PHYSICS HURTS

The comparisons above leave open whether accurate physics inputs can hurt an individual policy, and whether the expert labels contribute to that failure. Unlike the Physics-blind comparison of Section 4.1, which trains a separate student, we give each saved Concat and FiLM policy either the true physics or the nominal setting, while the simulator runs the true physics. The true input helps most at the training settings. Yet in HalfCheetah and Walker2d it already hurts halfway between the nominal setting and a training endpoint (Appendix J).

Losses follow the label gap. These losses appear only along directions where the label expert changes. Here a loss means a median true-input benefit below -10% across the six policies of a task and assignment. For each direction, we measure the action gap J between the two label experts as the root mean square of their action difference over visited states. Across four tasks and both assignments, J predicts where losses occur with an AUC of 0.89: losses appear in 11 of 12 directions with $J > 1.5$ and in none of 16 with $J \leq 0.5$.

Changing only the labels moves the failures. We test this directly in HalfCheetah by changing only the labels, with states and training fixed (Figure 7). Giving the gravity and damping endpoints the nominal expert's labels largely restores the halfway settings in every seed, at the cost of lower endpoint returns. Giving one mass endpoint the labels of a different but equally strong expert creates a new failure at the predicted halfway setting, while the average change elsewhere is small.

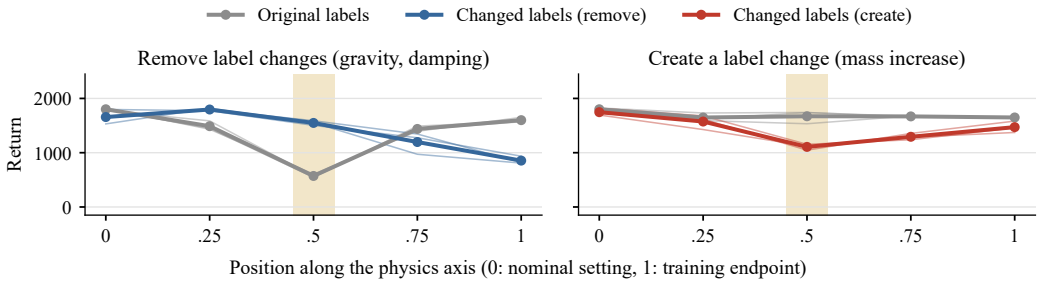

Figure 7: Changing only the expert labels moves failures of HalfCheetah Concat students. Return with the true physics input along relabeled directions. Thin lines: BC seeds. Thick lines: means. Left: average of the four gravity and damping directions. Shading marks the halfway setting where the change was predicted.

Why label gaps matter. At a state s, let v be the difference between the target action at $\boldsymbol{\theta}$ and the policy's action with the nominal input, and u the change that the true input produces. Under a squared action error, the nominal input costs $\|v\|^2$ and the true input costs $\|u - v\|^2$, so the true input improves by

$$\|v\|^2 - \|u - v\|^2 = 2u^\top v - \|u\|^2, \tag{7}$$

which equals $k(2-k)\|v\|^2$ for $u = kv$. A change helps only if it points the right way without overshooting ($0 < k < 2$). Where no change is required ($v \approx 0$), any change is a loss.

Proposition 1 (label conflict). Let two training settings at distance w take labels whose actions differ by J at a state s, and let the student be L-Lipschitz in its physics input. Then its Euclidean action errors e_- and e_+ at the two settings satisfy

$$e_- + e_+ \;\geq\; J - L\,w, \tag{8}$$

by the triangle inequality. Fitting both labels within ϵ requires $L \geq (J - 2\epsilon)/w$, so a large label gap demands a steep response to physics between the two settings. With the results above, this explains why large label gaps mark vulnerable intermediate settings. Expert assignment thus shapes where physics-conditioned policies fail.

6 EXPERT-BY-TEST GRID: A REPORTING PROTOCOL

Each finding above implies something a comparison must report. The *Expert-by-Test Grid* places expert assignments in rows and test sets in columns. Each cell reports the return gap for one method pair and its uncertainty, so rows test robustness to test sets and columns test robustness to expert assignment. We recommend four steps:

1. **Fix the comparison grid.** Set assignments, test sets, and analysis rules before seeing the results. Use training-only expert cross-evaluation to define a second assignment.
2. **Pair the training comparisons.** Fix states, sample weights, and each method's training setup across assignments. Include a physics-blind student and a nominal-input test.
3. **Cross physics and resets.** Pair episode returns on at least two physics sets chosen in advance, with matched resets. Claims about a physical distribution require independently sampled configurations. Report the mean gap and both variance components in Equation 5, and use the risk score of Section 4.3 to flag sensitive comparisons.
4. **Report the whole grid.** Show raw-return gaps, unresolved orderings, and all planned seeds. Distinguish independent expert banks from repeated BC fits within a bank.

Report an advantage across the tested grid only when every cell supports the same sign, and name the assignments and physics sets covered by that claim.

7 RELATED WORK

Conditioned policies. UP-OSI combines a universal policy with online estimation of dynamics parameters (Yu et al., 2017). We supply the parameters directly and test how policies use known physics. Our FiLM and physics-gated LoRA students adapt feature modulation (Perez et al., 2018) and low-rank updates (Hu et al., 2021) to imitation.

Distillation and imperfect experts. Policy distillation combines task-specific experts into one student (Rusu et al., 2015), and Distral jointly learns task policies and a shared policy (Teh et al., 2017). Distillation objectives (Czarnecki et al., 2019) and demonstration data (Mandlekar et al., 2021) change learning outcomes, and a more accurate teacher need not produce a better student (Cho & Hariharan, 2019). MAMBA benchmarks against the state-wise best of several experts (Cheng et al., 2020), and MAPS actively chooses which expert to imitate (Liu et al., 2023). We instead change only the source of action labels, with states and objective fixed.

Evaluation methodology. Random seeds, environment variability, and implementation choices affect RL comparisons (Henderson et al., 2018; Engstrom et al., 2019). Agarwal et al. (2021) advocate interval estimates for few training runs, Packer et al. (2018) assess generalization beyond the training distribution, and DomainBed shows why explicit model selection matters (Gulrajani & Lopez-Paz, 2020). None of these studies varies the expert that supplies imitation labels. We vary it together with the test physics and predict which comparisons reverse.

8 CONCLUSION

Whether physics inputs help, and which architecture benefits most, depends on the expert assignment and the test physics. Changing only the expert reverses architecture rankings, and changing only the test physics reverses whether physics inputs help. These reversals are predictable from how much a gap varies across physics, and inside a single policy the expert decides where accurate physics hurts. A ranking from one expert and one test set therefore need not generalize, so we recommend the Expert-by-Test Grid with physics-blind and nominal-input controls.

Reproducibility statement

The supplement provides code, recorded results, and a script that checks the saved numbers offline. Appendices A and B give the training details. Reproducing the exact runs also requires the archived model checkpoints and demonstration data. These are not included in this code-and-results package.

AI use statement

We used generative AI to polish writing, identify related work, and improve figure presentation. We did not use it for research ideas, experiment design, method development, mathematical proofs, or result interpretation. All AI-assisted content was reviewed by the authors.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320, 2021.

Ching-An Cheng, Andrey Kolobov, and Alekh Agarwal. Policy improvement via imitation of multiple oracles. *Advances in Neural Information Processing Systems*, 33:5587–5598, 2020.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4793–4801. IEEE, 2019.

Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1331–1340. PMLR, 2019.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International conference on learning representations*, 2019.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Xuefeng Liu, Takuma Yoneda, Chaoqi Wang, Matthew Walter, and Yuxin Chen. Active policy improvement from multiple black-box oracles. In *International Conference on Machine Learning*, pp. 22320–22337. PMLR, 2023.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

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

Yee Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust multitask reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

Wenhao Yu, Jie Tan, C Karen Liu, and Greg Turk. Preparing for the unknown: Learning a universal policy with online system identification. *arXiv preprint arXiv:1702.02453*, 2017.

A CONFIGURATIONS AND DATA COLLECTION

The parameter order is gravity, mass scale, friction scale, damping scale, and armature scale. Table 4 lists the values. Configuration index 0 is the nominal setting. Indices 1 and 2 change gravity to endpoint 1 and endpoint 2. Indices 3 and 4 change mass, followed by friction (5, 6), damping (7, 8), and armature (9, 10). All other parameters keep their nominal values. These indices also identify the experts.

Table 4: Physics values used in the controlled study. Scale parameters multiply the corresponding nominal simulator values. Gravity is signed.

Parameter	Endpoint 1	Nominal	Endpoint 2
Gravity	-6.0	-9.81	-14.0
Mass scale	0.75	1.0	1.3
Friction scale	0.5	1.0	2.0
Damping scale	0.5	1.0	2.0
Armature scale	0.5	1.0	2.0

For every coordinate pair (i, j) with $i < j$, the two-coordinate test configuration uses endpoint 1 for i and endpoint 2 for j. This gives ten targets. The remaining three targets are

$$(-6.0, 1.3, 0.5, 1, 1), \qquad (-14.0, 1, 1, 2, 0.5), \qquad (-9.81, 0.75, 2, 0.5, 1).$$

These targets are selected combinations of parameter values. They do not cover all possible combinations. The exact list is implemented in `src/bench_d5.py` and stored in the bank-reference files.

Demonstration and evaluation episodes. Each expert collects 50 demonstration episodes at its own training configuration. Episode e, indexed from zero, resets with seed $7000 + e$. Evaluation uses 20 episodes with seed $2000 + e$. Each episode runs for at most 1,000 steps and stops earlier if the environment terminates or truncates. Actions are deterministic. Expert labels and deployed actions are clipped to the environment action bounds. The state pool is collected again with the same policies and reset seeds for each run. Both assignments use this same pool.

B TRAINING DETAILS

All student networks use three hidden layers of width 256 and a linear output layer. Physics inputs are standardized. The layer operations and training schedules are given below.

Unanchored policies. Concat appends standardized physics to the state. Each hidden block applies a linear map, ELU, and LayerNorm, without skip connections. FiLM uses the same state-only blocks, with physics-dependent scale and shift after normalization. These maps start at the identity. Physics-blind uses the same blocks without physics input.

Gated-LoRA first fits a physics-blind trunk. Its hidden blocks apply a linear map, LayerNorm, and ELU, with skip connections at layers two and three. It then freezes the trunk and adds rank-8 updates to the three hidden maps and the output map. For layer input x,

$$\Delta h = (xA \odot g(\boldsymbol{\theta}))B, \qquad g_k(\boldsymbol{\theta}) = \prod_{i=1}^{5} \sigma(w_{ik}\theta_i + b_{ik}). \tag{9}$$

The B factors start at zero.

Retrieval-anchored policies. At test time, let u be the nearest training configuration in standardized coordinates. The predicted action before clipping is

$$a(s, \boldsymbol{\theta}) = \pi_u(s) + \Delta_\psi(s, z(\boldsymbol{\theta}) - z(\boldsymbol{\theta}_u)),$$

Table 5: Trainable parameters in the conditioned fitting stage. Adapter policies also use a previously trained trunk. These counts do not measure total model capacity or training cost.

Method	HalfCheetah	Ant	Hopper	Walker2d
Concat	140,550	143,624	138,243	140,550
Gated-LoRA	12,792	12,888	12,720	12,792
Anchor-MLP	140,550	143,624	138,243	140,550
Anchor-Joint	16,206	16,800	15,387	16,206
Anchor-Factored	16,334	16,928	15,515	16,334

where z standardizes each coordinate using its mean and standard deviation across training settings. At training time, each sample from configuration j draws anchor u uniformly from the 11 bank policies. With probability 0.15 this draw is replaced by $u = j$. The residual fits

$$\left\|\pi_u(s) + \Delta_\psi(s, z(\boldsymbol{\theta}_j) - z(\boldsymbol{\theta}_u)) - \pi_{b(j)}(s)\right\|_2^2.$$

Even when the anchor matches the training configuration, the target residual under BEST need not be zero if $b(j) \neq j$.

Anchor-MLP fits a separate three-layer residual MLP with the same block order as Concat. Its input is the state and standardized parameter difference. Anchor-Joint and Anchor-Factored use a frozen pretrained residual trunk with gated rank-8 updates. The joint gate is $g(\delta) = \sigma(W\delta + b)$. The factored gate uses the product above. Within each assignment and seed, these two policies share the same pretrained trunk weights. The shared trunk is not used by Anchor-MLP's separate residual network. All three residual output heads start at zero, so their initial deployed policy equals the retrieved expert. The gated anchored models initialize B from a normal distribution with standard deviation $0.1/\sqrt{8}$.

BC optimization. Training uses AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, a cosine learning-rate schedule, and gradient-norm clipping at 1.0. Concat, FiLM, and Physics-blind train for 200 epochs. Gated-LoRA uses 200 pretraining epochs and 200 adaptation epochs. The gated anchored models use 200 shared trunk-pretraining epochs and 200 residual-training epochs. Anchor-MLP trains its separate residual network for 200 epochs. Each phase restores the weights with the lowest observed mean training MSE. No test returns are used for checkpoint selection. These schedules differ between methods and stay fixed across assignments. The reported study uses zero-shot evaluation only.

PPO training. The experts use Stable Baselines3 PPO with learning rate 3×10^{-4}, 2,048 steps per update per environment, eight parallel environments, batch size 256, ten optimization epochs per update, discount 0.99, GAE parameter 0.95, clipping 0.2, entropy coefficient 0.005, and maximum gradient norm 0.5. Budgets are one million total steps for HalfCheetah, Hopper, and Walker2d, and two million for Ant. Original expert checkpoints do not record training seeds. The checkpoint files are therefore needed to reproduce these exact banks. Different BC seeds reuse the same bank. They do not represent independently trained expert banks.

Compute and implementation. The experiments ran on a workstation with an Intel i7-14700KF CPU and an RTX 4070 SUPER GPU. The reported paired runs use CPU computation. Implementation is in `src/models.py`, `src/anchored.py`, `src/training.py`, and `run_anchored_d5.py`.

C TRAINING-ONLY EXPERT CROSS-EVALUATION

For each environment, cross-evaluation scores all 11^2 expert and training-configuration pairs over twenty episodes each, or 2,420 episodes. Best takes the largest column mean, breaking ties by index. Selection uses no held-out scores. Table 6 summarizes the comparisons.

Independent expert checks. Twenty new resets per training setting give selected-minus-Own mean gains of 471.0, 855.8, 169.2, and 319.3 for the original 1M, original 3M, A, and B HalfCheetah banks.

Gains are positive at 9/11, 7/11, 7/11, and 7/11 settings. The selected experts remain fixed during this check.

Table 6: Experts beaten by another bank policy at their own training or held-out settings. Median raw-return gaps use only positive cases. Counts use 20-episode means.

Environment	Own expert beaten	Median gap	Target PPO beaten	Median gap
HalfCheetah	9/11	555	10/13	250
Ant	10/11	495	11/13	489
Hopper	6/11	549	3/13	383
Walker2d	8/11	1000	12/13	481

D RAW SCORES AND TRAINING VARIATION

Raw return averages thirteen configuration means, each from twenty episodes. Standard deviations use the three primary BC seeds and denominator two. They describe training variation within one bank. They are not the episode-level intervals used to identify supported reversals. Extra seeds 789 and 2024 enter the census, not these three-seed means.

Table 7: Raw-return means and sample standard deviations across the three primary training seeds. Full configuration and seed scores are in the supplement.

Environment	Student	OWN	BEST
HalfCheetah	Concat	414.0 ± 112.3	1044.1 ± 37.0
	Gated-LoRA	845.5 ± 20.9	843.6 ± 38.7
	Anchor-MLP	798.4 ± 46.0	669.5 ± 24.4
	Anchor-Joint	677.5 ± 111.7	748.5 ± 93.6
	Anchor-Factored	841.7 ± 38.5	768.8 ± 97.9
Ant	Concat	701.6 ± 59.4	1237.0 ± 36.2
	Gated-LoRA	1016.8 ± 30.1	1461.2 ± 130.2
	Anchor-MLP	863.5 ± 52.0	1041.3 ± 20.6
	Anchor-Joint	866.1 ± 163.8	1139.7 ± 8.7
	Anchor-Factored	840.7 ± 78.5	1187.3 ± 10.1
Hopper	Concat	1170.9 ± 124.2	1116.4 ± 81.5
	Gated-LoRA	740.2 ± 59.0	876.9 ± 131.9
	Anchor-MLP	923.4 ± 158.8	1108.2 ± 10.1
	Anchor-Joint	442.7 ± 144.6	927.2 ± 96.2
	Anchor-Factored	818.9 ± 210.8	1074.5 ± 252.2
Walker2d	Concat	1109.8 ± 25.0	1614.2 ± 84.8
	Gated-LoRA	941.0 ± 47.2	1616.6 ± 50.7
	Anchor-MLP	1035.8 ± 54.8	1253.5 ± 61.0
	Anchor-Joint	962.4 ± 109.5	1163.9 ± 14.4
	Anchor-Factored	818.9 ± 57.8	1165.3 ± 134.9

Secondary normalized scores. Let $O(k)$ be target-trained PPO return and G_{ik}^{test} expert i's return at target k. The supplemental normalized score is

$$\text{Ref}(k) = \max\{O(k), \max_i G_{ik}^{\text{test}}\}, \qquad Q = \frac{100}{K} \sum_{k=1}^{K} \frac{R(k)}{\text{Ref}(k)}. \tag{10}$$

We call Q %REF. The denominator is fixed across methods and assignments. It is a finite-bank reference, not an upper bound. Normalization weights configurations differently from the raw mean. For example, HalfCheetah Gated-LoRA changes from 845.5 to 843.6 in raw return, while %REF changes from 67.6 to 54.5. The ranking census uses raw return.

Assignment interaction. For Concat minus Gated-LoRA, the interaction is

$$I = D_{\text{BEST}} - D_{\text{OWN}} = \Delta_{\text{Concat}} - \Delta_{\text{Gated-LoRA}}, \tag{11}$$

where Δ_m is method m's change under reassignment. A positive interaction alone does not imply a reversal. The gaps under the two assignments must have opposite signs.

E INDEPENDENT BANKS AND THE HIGHER-BUDGET FOLLOW-UP

We fixed the follow-up plan before training two new HalfCheetah banks. A uses expert seeds 17000–17010 and B uses 29000–29010, both at 1M PPO steps. The existing 3M bank uses seeds $42 + i$ for configuration i. Students use BC seeds 42, 1, and 2. Assignments share the bank-specific state pool. Each test set has thirteen configurations. The interior configurations were fixed in the follow-up plan. Original and interior evaluations use resets 2000–2019 and 12000–12019, respectively.

All core seed comparisons. Here $D = R_{\text{Concat}} - R_{\text{Gated-LoRA}}$ and $I = D_{\text{BEST}} - D_{\text{OWN}}$. Neither new bank repeats the original mean reversal on original targets. Each has two mean reversals on interior targets. Some gaps are too small for both intervals to exclude zero.

Bank	Targets	BC seed	D_{Own}	D_{Best}	I
3M	Original	42	-250.6	-588.0	-337.4
3M	Interior	42	245.7	-405.9	-651.6
3M	Original	1	-727.5	-799.0	-71.5
3M	Interior	1	206.3	-279.4	-485.7
3M	Original	2	-832.8	-366.3	466.5
3M	Interior	2	-70.2	-286.2	-216.0
A (1M)	Original	42	-584.4	-140.4	444.0
A (1M)	Interior	42	-136.2	-63.6	72.6
A (1M)	Original	1	-398.6	-188.7	209.8
A (1M)	Interior	1	-64.7	11.1	75.8
A (1M)	Original	2	-470.3	-340.6	129.7
A (1M)	Interior	2	-85.2	10.1	95.3
B (1M)	Original	42	-447.9	-61.1	386.8
B (1M)	Interior	42	-227.7	-71.2	156.5
B (1M)	Original	1	-432.2	-100.0	332.1
B (1M)	Interior	1	-437.0	150.8	587.8
B (1M)	Original	2	-622.1	-12.4	609.7
B (1M)	Interior	2	-340.7	69.3	410.0

The 3M result. All three interior-target interactions are negative. Seeds 42 and 1 reverse toward Gated-LoRA with both endpoint intervals excluding zero. Original-target interactions have mixed signs:

Seed	Original I [95% interval]	Interior I [95% interval]
42	-337.4 $[-445.3, -229.5]$	-651.6 $[-756.6, -546.7]$
1	-71.5 $[-197.0, 54.0]$	-485.7 $[-638.4, -332.9]$
2	466.5 $[363.7, 569.3]$	-216.0 $[-349.4, -82.6]$

Training budget, trained experts, and demonstration states differ between the original 1M and 3M banks, so this comparison does not isolate a budget effect.

Physics-blind controls outside bank A. The table includes every planned Best-assignment control in the original 1M bank and 3M. These evaluations change physics and resets together.

Table 8: Best-assignment return minus Physics-blind. Brackets give pointwise conditional 95% intervals for fixed policies and targets. The last upper bound (0.04) is positive. Original 1M Concat seeds 1 and 2 were not planned. Policies trained on newly collected data are reported separately.

Bank	Method	BC seed	Original set	Interior set
Original 1M	FiLM	42	$190.0\ [106.5, 273.5]$	$15.3\ [-53.6, 84.3]$
Original 1M	FiLM	1	$255.9\ [189.3, 322.5]$	$33.9\ [-38.1, 105.9]$
Original 1M	FiLM	2	$73.0\ [-9.3, 155.3]$	$2.1\ [-78.7, 82.9]$
Original 1M	Concat	42	$244.9\ [171.8, 318.0]$	$-36.0\ [-102.1, 30.2]$
3M	FiLM	42	$-114.0\ [-191.0, -37.1]$	$-293.9\ [-368.5, -219.2]$
3M	FiLM	1	$-30.7\ [-121.8, 60.4]$	$97.6\ [-4.5, 199.6]$
3M	FiLM	2	$-96.0\ [-178.0, -14.1]$	$-274.1\ [-385.6, -162.6]$
3M	Concat	42	$-27.3\ [-89.8, 35.1]$	$-71.1\ [-155.6, 13.4]$
3M	Concat	1	$-329.4\ [-430.9, -227.9]$	$157.3\ [37.7, 276.9]$
3M	Concat	2	$-156.2\ [-249.3, -63.1]$	$-127.8\ [-255.7, 0.04]$

F CENSUS UNITS AND UNANCHORED COMPARISONS

A complete follow-up group has two methods, one bank, one BC seed, both assignments, and both test sets. Assignment comparisons fix the test set. Test-set comparisons fix the assignment. The 47 complete groups give 94 comparisons for each type of change: 36 each from A and 3M, 16 from the original 1M bank, and six from B. B has no planned FiLM or Physics-blind controls. Exact ties are not reversals. The comparisons share policies and are not independent trials.

Table 9: HalfCheetah follow-up: reversals / eligible comparisons. CI counts require opposite signs of both conditional 95% intervals. Panel changes alter physics and reset seeds.

	Assignment		Evaluation panel	
Pair	Mean	CI	Mean	CI
Concat / Gated-LoRA	7/20	5/20	7/20	4/20
Concat / FiLM	5/14	4/14	5/14	2/14
Concat / Physics-blind	3/14	1/14	11/14	9/14
Gated-LoRA / FiLM	8/14	4/14	6/14	4/14
Gated-LoRA / Physics-blind	3/14	1/14	3/14	2/14
FiLM / Physics-blind	5/18	3/18	13/18	8/18
Total	31/94	18/94	45/94	29/94

Conditional intervals. For paired episode returns, define

$$Z_e = \frac{1}{13} \sum_{k=1}^{13} (X_{m,k,e} - X_{n,k,e}).$$

The interval uses the mean and sample standard error of the twenty Z_e with a t_{19} multiplier. This keeps the dependence from shared resets. A supported reversal needs opposite signs with both endpoint intervals excluding zero. These pointwise intervals treat policies and test settings as fixed. They do not include variation across banks, training runs, or sampled test settings. They are not corrected for multiple comparisons. Unresolved cases stay in the denominator.

Archived four-environment results. The original test sets combine endpoint values across physical parameters. Re-evaluating frozen checkpoints for 41,600 episodes reproduces all 2,080 saved configuration means and standard deviations within 10^{-5}.

The balanced three-seed subset has 36/120 supported reversals. Extra HalfCheetah and Ant seeds add 13/40. Nine of ten method pairs have at least one supported reversal.

All unanchored seed comparisons. The table reports Concat minus Gated-LoRA under each assignment.

Environment	Seed	Own gap [95% interval]	Best gap [95% interval]	Reversal to
HalfCheetah	1	-380.3 [$-424.5, -336.0$]	$+264.1$ [$+215.3, +313.0$]	Concat
HalfCheetah	2	-333.2 [$-376.3, -290.0$]	$+123.1$ [$+81.8, +164.3$]	Concat
HalfCheetah	42	-581.0 [$-614.8, -547.1$]	$+214.2$ [$+163.7, +264.6$]	Concat
HalfCheetah	789	-389.8 [$-459.2, -320.4$]	$+287.0$ [$+222.8, +351.2$]	Concat
HalfCheetah	2024	-585.7 [$-628.3, -543.0$]	$+283.5$ [$+233.9, +333.0$]	Concat
Ant	1	-337.3 [$-393.5, -281.2$]	-400.0 [$-481.6, -318.4$]	None
Ant	2	-391.5 [$-437.6, -345.4$]	-200.7 [$-270.6, -130.7$]	None
Ant	42	-216.8 [$-257.5, -176.1$]	-72.2 [$-144.6, +0.3$]	None
Ant	789	-272.8 [$-312.1, -233.5$]	-443.8 [$-532.0, -355.5$]	None
Ant	2024	-294.4 [$-337.5, -251.4$]	-318.2 [$-379.8, -256.5$]	None
Walker2d	1	$+174.4$ [$+111.7, +237.2$]	-77.1 [$-133.4, -20.9$]	Gated-LoRA
Walker2d	2	$+189.3$ [$+101.6, +277.0$]	-80.7 [$-173.8, +12.4$]	None
Walker2d	42	$+142.8$ [$+94.5, +191.2$]	$+150.7$ [$+74.6, +226.8$]	None
Hopper	1	$+516.2$ [$+491.9, +540.5$]	$+82.9$ [$+60.3, +105.4$]	None
Hopper	2	$+405.6$ [$+369.3, +442.0$]	$+197.2$ [$+154.9, +239.6$]	None
Hopper	42	$+370.0$ [$+325.5, +414.6$]	$+438.5$ [$+398.1, +478.9$]	None

Matched FiLM comparisons. The original-bank extension requires retraining to reproduce Concat weights exactly before adding FiLM. Only Hopper passes. Its six new unanchored assignment comparisons contain four supported reversals. Re-evaluating archived Ant checkpoints reproduces all 650 saved means and standard deviations, but retraining Concat on newly collected data does not reproduce the original weights. On these new data, the old and new training functions do match. A separate study therefore trains both Concat and FiLM on the same new data in the other three tasks, regardless of the earlier replay results. Its nine comparisons stay separate from the archived census. The three paired fits use BC seeds 42, 1, and 2, in that order. Exact means and intervals are retained with the figure data in the supplement.

Evaluation runtime. All census re-evaluations use four CPU threads. An initial two-thread run changed eight HalfCheetah Gated-LoRA Own configuration means by an average absolute difference of 59.6. We stopped that run and excluded it from the census. The four-thread replay gives the agreement reported above.

G Separating physics changes from reset changes

We retain the fitted policies and cross two physics sets with reset blocks 2000–2019 and 12000–12019. Each cell uses thirteen configurations and twenty episodes per configuration. Original targets combine training-axis endpoint values. Interior targets combine values inside those ranges. All thirteen original and eight interior targets lie outside the convex hull of the eleven training configurations.

Policy ordering. For Concat/Gated-LoRA under Own and Best, the crossed evaluation includes bank A at all three BC seeds and bank B at seeds 42 and 1. Changing physics at fixed resets gives five mean reversals in twenty comparisons, two supported by intervals. Changing resets at fixed physics gives one mean reversal and none supported. That gap changes from -5.3 to $+11.1$, with both intervals including zero. Matched resets use paired blocks. Separate reset blocks use Welch intervals for independent blocks.

Physics-blind control. The bank-A Best check compares Concat and FiLM with Physics-blind across all three training seeds. Some results that changed physics and resets together were known before we fixed this extension. Each gap is tested under all four physics/reset combinations with a pointwise t_{19} interval. This compares complete policies, rather than isolating a single physical coordinate. Evaluations use four CPU threads. Full gap values remain in the supplementary records.

H Scope and exploratory checks

The original targets were used during development. Exposure records track when later results became available. Exploratory checks of training stages, updates, action variation, and projection did not

establish a common mechanism. The projection check was affected by thread count. A fixed physics-based predictor passed only 1/26 external comparisons, in a partly observed case. Distance checks found no shared threshold or monotone rule. These checks provide no validated prediction for new banks.

I Independent test sets and evaluation uncertainty

I.1 Sampling and interval construction

Design. The registered study evaluates 48 fixed Concat/FiLM policies (four tasks, three training seeds, two assignments). Each task has eight independent test sets of 32 configurations and a separate 256-configuration reference, all drawn from one distribution. Each configuration picks two or three physics axes uniformly without replacement, picks a lower or upper endpoint for each, and moves from nominal toward it by a fraction drawn uniformly from 0.25 to 1. Four resets per configuration give 98,304 episodes, and each reset block is shared across compared policies, seeds, and assignments.

Two intervals. Let Y_{kr} be the paired return difference at configuration k and reset r, with C configurations and R resets, and let $\bar{Y}_k$ and s_k^2 be its within-configuration mean and variance. The fixed-physics interval uses standard error $\sqrt{\sum_k s_k^2/R}/C$ with Welch degrees of freedom $(\sum_k v_k)^2/\big(\sum_k v_k^2/(R-1)\big)$, where $v_k = s_k^2/R$. The configuration-aware interval uses $\mathrm{SD}(\bar{Y}_k)/\sqrt{C}$ with $C-1$ degrees of freedom, which already includes reset noise. The reference interval uses the configuration-aware rule with 256 configurations.

Registered primary result. A ranking conflicts with the reference when both intervals exclude zero with opposite signs. The conflict rate of configuration-aware intervals is lower than that of fixed-physics intervals by 0.52 percentage points, with 95% bootstrap interval $[0.00, 5.73]$ over whole test sets and reference configurations. Fixed-physics intervals give one conflicting ranking among 192 comparisons, and configuration-aware intervals give none. Because the interval includes zero, we do not conclude that conflicts are rarer. The reference itself supports a ranking in only 7 of 24 comparisons and is unresolved for 67 of the 100 fixed-physics claims and four of the nine configuration-aware claims, so the nine claims do not show better policy selection. The share of comparisons with a supported ranking differs by 47.40 percentage points $[41.67, 53.14]$.

Checks added after the study. Table 10 counts comparisons in which different test sets support opposite rankings. Bonferroni checks keep the original means, standard errors, and Welch degrees of freedom and change only the critical value to $t_{1-0.05/(2m)}$ for family size m. The deletion check first reproduces all 192 original results from raw returns, then removes each of the 32 configurations in turn. Of the 100 original claims, 92 keep their mean sign under every deletion and 50 keep interval support. By mean signs alone, 11 of 24 reversals survive every deletion. Configuration-aware intervals give no supported reversal. The checks overlap, so we do not combine them, and all estimates treat the expert bank and trained policies as fixed.

Table 10: Comparisons in which independent test sets support opposite rankings, out of six fixed Concat/FiLM comparisons per task. The last row counts supported orderings among 192 intervals (for deletion, the 100 original claims that keep support). Columns after the first are separate checks added after the study.

Task	Pointwise	Bonferroni within comparison	Bonferroni across 192	Survives every deletion
HalfCheetah	3/6	2/6	1/6	1/6
Ant	2/6	2/6	1/6	1/6
Walker2d	1/6	1/6	1/6	1/6
Hopper	6/6	5/6	4/6	2/6
Total	12/24	10/24	7/24	5/24
Supported orderings	100/192	68/192	45/192	50/100

I.2 REVERSAL-RISK ANALYSIS

Table 11 compares the fixed-set and distribution-level targets on the same 192 comparisons.

Table 11: Two evaluation targets for the same 192 comparisons. Both use pointwise 95% intervals with expert banks and fitted policies fixed. This study is separate from the 49/160 census of Section 3.2.

Target	Variation included	Supported	Unresolved
Fixed test set	Reset episodes	100/192	92/192
Physics distribution	Configurations and resets	9/192	183/192

This analysis was added after the study. For each of the 24 comparisons, we fit the decomposition in Equation 5 to the paired Concat-minus-FiLM gaps on the separate 256-configuration reference. The reference shares no configurations or resets with the eight test sets. The reset variance is the mean within-configuration variance of the gaps. The physics variance is the variance of configuration means minus its reset share. With mean gap μ, $s_T^2 = \sigma^2_{\text{phys}}/C + \sigma^2_{\text{reset}}/(CR)$, and fixed-set standard error $s_F = \sigma_{\text{reset}}/\sqrt{CR}$, a test-set mean is supported positive or negative with probabilities

$$p_+ = 1 - \Phi\left(\frac{c\, s_F - \mu}{s_T}\right), \qquad p_- = \Phi\left(\frac{-c\, s_F - \mu}{s_T}\right).$$

The risk score is the probability that eight independent sets contain at least one of each, $1 - (1 - p_+)^8 - (1 - p_-)^8 + (1 - p_+ - p_-)^8$, with $C = 32$ and $R = 4$. The critical value c uses the same significance levels as the pointwise and Bonferroni analyses but approximates the Welch degrees of freedom by 96. The observed Welch intervals have between 4.1 and 56.9 degrees of freedom (median 20.1). The separate resampling check below applies the exact Welch interval rules.

Table 12: Observed and predicted reversals by task. Predictions sum the risk scores of the six comparisons in each task.

	Pointwise		Bonferroni	
Task	Observed	Predicted	Observed	Predicted
HalfCheetah	3	3.3	1	1.2
Ant	2	2.7	1	0.5
Walker2d	1	3.6	1	1.7
Hopper	6	5.4	4	4.7
Total	12	15.0	7	8.0

Discrimination and calibration. The risk score ranks comparisons with observed reversals above those without them, with an AUC of 0.90 for pointwise reversals and 0.81 after Bonferroni correction. Without Hopper, the values are 0.85 and 0.78. Within tasks, the pointwise AUC is 0.91 over 22 pairs. The Brier score is 0.169, compared with 0.250 for a constant rate. Comparisons with scores below 0.5 had no observed reversal (mean score 0.28, six comparisons). Scores from 0.5 to 0.8 averaged 0.62 against an observed rate of 0.55, and scores above 0.8 averaged 0.93 against 0.86. The 24 comparisons share tasks and policies, so these values are descriptive.

Uncertainty of the reference estimates. Resampling reference configurations within each task gives a median standard-error inflation of 3.4 with a 95% bootstrap interval of $[3.1, 3.7]$. The predicted numbers of reversals range over $[12.7, 15.3]$ for pointwise intervals and $[6.5, 8.3]$ after correction.

Where the model is less accurate. The model predicts more supported claims than observed: 118 versus 100 for fixed-set intervals and 17.7 versus 9 for configuration-aware intervals. Pseudo test sets resampled from the reference, analyzed with the exact interval rules, give the same pattern (116 and 19 claims), while their reversal counts match the observations (15 and 6, with 95% ranges $[11, 19]$ and $[3, 9]$). Walker2d reversals are over-predicted. For individual comparisons, the predicted agreement between a subsample of k reference configurations and the full reference sign has mean absolute errors of 0.06 to 0.13 for k from 8 to 128. We therefore use the model to rank reversal risk, not to prescribe a number of configurations.

I.3 Expert assignment and test physics

This analysis was added after the study. For each task and BC seed, the Own and Best Concat/FiLM gaps share configurations and reset seeds. We form $\Delta_{c,r} = \left(R_{\mathrm{Concat,Best}} - R_{\mathrm{FiLM,Best}}\right) - \left(R_{\mathrm{Concat,Own}} - R_{\mathrm{FiLM,Own}}\right)$ for configuration c and reset r, which keeps the covariance between the four policies. With $\bar{\Delta}_c$ the reset mean and s_c^2 its within-configuration variance,

$$\mathrm{Var}_c\left(\bar{\Delta}_c\right) = \sigma^2_{\mathrm{between}} + \frac{1}{R}\,\mathbb{E}_c\left[s_c^2\right],$$

and we report $\sigma^2_{\mathrm{between}}$ and its share of $\mathrm{Var}_c(\bar{\Delta}_c)$. This share describes the variance of the estimated per-configuration changes. It is not a share of total return or of the full assignment effect. Intervals are pointwise bootstrap intervals over configurations, with expert banks and fitted policies fixed.

Table 13: Assignment-induced change in the Concat/FiLM gap across physical configurations. Test sets: the 256 configurations of the eight test sets. Reference: the separate 256-configuration reference.

		Test sets				Reference
Task	Seed	Mean Δ	SD between	SD reset	Share	Share
HalfCheetah	42	45.1	395.6	183.4	0.82	0.86
HalfCheetah	1	56.7	423.0	170.8	0.86	0.84
HalfCheetah	2	68.6	390.9	196.5	0.80	0.86
Ant	42	116.8	695.3	301.1	0.84	0.82
Ant	1	−143.4	514.7	297.8	0.75	0.72
Ant	2	28.2	643.2	295.7	0.83	0.84
Walker2d	42	233.8	681.9	246.1	0.88	0.91
Walker2d	1	−34.7	700.9	270.3	0.87	0.93
Walker2d	2	179.7	661.2	277.5	0.85	0.92
Hopper	42	55.3	988.8	172.7	0.97	0.98
Hopper	1	−122.1	1183.0	202.5	0.97	0.98
Hopper	2	66.7	1133.8	197.1	0.97	0.98

In every grid, the between-configuration standard deviation exceeds both the absolute mean change and the reset-noise standard deviation of a configuration mean, and the bootstrap interval for $\sigma^2_{\mathrm{between}}$ excludes zero. The same holds on the reference in all 12 grids. The reference confirms that between-configuration variation remains, not that the mean change has the same sign: for HalfCheetah seed 1, the mean is $+56.7$ on the test sets and -30.8 on the reference. Under a normal model for the per-configuration effects, the estimated share of configurations whose change has the opposite sign to the mean ranges from 0.37 to 0.48. This auxiliary estimate depends on the normal assumption and does not recover the direction of any single configuration.

I.4 Equal-budget selection check

A preceding registered study compares 8 configurations with 16 episodes each against 32 configurations with 4 episodes each, using the same 48 policies and sampling distribution. Both allocations cost 128 episodes per policy. Actual runtime may differ. Selection is scored on an independent 64-configuration reference with eight episodes each. The 512 paired selection subsets reuse one 32-by-16 pool, so they are not independent. The normalized broad-minus-narrow selection gain is 0.00337, with nested bootstrap interval $[-0.00368, 0.00680]$. It does not meet the registered criterion. This study therefore does not show better policy selection or identify the best way to divide the evaluation budget.

I.5 Choosing between fixed policies from physics

This exploratory analysis was added after the study. For each of the 24 fixed Concat/FiLM comparisons, a selector predicts the configuration-mean Concat-minus-FiLM gap from the five physical parameters and runs the policy with the higher prediction. It is fitted on the 256 test-set configurations and evaluated on the disjoint 256 reference configurations, and then in the reverse direction. We fixed three rules in advance: Gaussian kernel ridge (bandwidth 1, penalty 1, standardized inputs),

25-nearest-neighbour averaging, and linear ridge. The baseline is the better of the two fixed policies on the evaluation configurations, chosen with hindsight.

Table 14: Selection by physics versus the better fixed policy chosen with hindsight. Counts are out of 24 comparisons. Intervals are pointwise 95% or Bonferroni-corrected over 24 comparisons, with configurations resampled and policies fixed.

Training data	Rule	Gain > 0	95% CI > 0	Bonferroni > 0	Median gain
Same policies	Kernel ridge	24	21	16	64.1
Same policies	25-NN	23	14	7	43.9
Same policies	Linear ridge	20	6	4	20.3
Other two BC seeds	Kernel ridge	17	8	3	21.8
Other two BC seeds	25-NN	15	6	2	10.6
Other two BC seeds	Linear ridge	15	8	3	19.9

With the same policies, the reverse direction gives 24, 22, and 22 positive gains for the three rules. No interval for the same-policy selectors lies below zero. A selector fitted on the other two BC seeds transfers only weakly, so the rule must be fitted for the policies being deployed. The selector needs evaluations of both policies on development configurations, and deployment keeps both policies. It is not a confirmatory result. Confirmation requires freezing a rule and evaluating it on newly sampled configurations.

J PHYSICS INPUTS AND EXPERT LABELS

These analyses were added after the main studies. Unless stated otherwise, they use the 48 fixed Concat and FiLM policies of the independent test-set study (four tasks, three BC seeds, two assignments). The simulator always runs the true physics. Only the physics vector given to the policy changes. Reset seeds are disjoint from all earlier evaluations.

J.1 WRONG AND NOMINAL PHYSICS INPUTS

On the first 64 reference configurations, each policy receives the true physics, the true physics plus Gaussian noise with 0.5 or 1.0 times the per-axis standard deviation of the reference configurations, or the nominal setting. Noise of 0.5 standard deviations lowers the median return by 8.8% (Concat) and 7.0% (FiLM). Noise of 1.0 standard deviation lowers it by 20.3% and 26.6%. The nominal input is better than the true input for all 12 HalfCheetah and all 12 Walker2d policies, by a median of 28% and 34 to 45%. It is worse for Ant (about 6%) and Hopper (13 to 17%). The physics selector of Appendix I.5, applied to the noisy input with 0.5 standard deviations, still gains over the better fixed policy in 21 of 24 comparisons. With the nominal input it always picks the same policy and gains nothing.

J.2 WHERE THE TRUE INPUT HELPS

Each policy is evaluated with the true and with the nominal input at the ten non-nominal training settings, at ten single-axis settings halfway between the nominal setting and each training endpoint, and on the 64 multi-axis configurations above. Table 15 reports the median benefit of the true input. Intervals are 95% t intervals over configurations within each policy. At the nominal setting both inputs are identical and give identical returns, which checks the implementation.

The benefit is largest at the training settings in every task. In HalfCheetah and Walker2d it turns negative for all 12 policies of each task already at the single-axis halfway settings, so the losses do not require combined parameters.

J.3 ACTION GAP BETWEEN LABEL EXPERTS

For each task, assignment, and axis direction, let e_0 label the nominal setting and e_k label the training endpoint k. We roll out both assignments' nominal label experts for two episodes at the halfway physics and pool the visited states. On these common states the action gap is $J =$

Table 15: Benefit of the true over the nominal physics input, in percent of the true-input return. Median over six policies per row, with the number of policies whose interval lies below zero (Neg.) or above zero (Pos.).

		Training settings		Single-axis halfway		Multi-axis	
Task	Student	Median	Neg./Pos.	Median	Neg./Pos.	Median	Neg./Pos.
HalfCheetah	Concat	$+22.7$	0/0	-22.6	0/0	-28.2	6/0
HalfCheetah	FiLM	$+21.8$	0/0	-32.6	2/0	-28.3	6/0
Walker2d	Concat	$+0.3$	0/1	-26.6	3/0	-44.6	4/0
Walker2d	FiLM	-4.1	0/0	-26.9	3/0	-34.2	5/0
Ant	Concat	$+24.3$	0/1	$+3.1$	0/0	$+6.2$	0/1
Ant	FiLM	$+26.1$	0/2	$+13.3$	0/0	$+5.6$	0/0
Hopper	Concat	$+25.8$	0/3	$+17.1$	0/1	$+16.9$	0/3
Hopper	FiLM	$+24.0$	0/3	$+10.0$	0/1	$+12.7$	0/1

$\left(\mathbb{E}_s\|\pi_{e_k}(s) - \pi_{e_0}(s)\|^2\right)^{1/2}$, with clipped actions. A direction counts as a loss when the median benefit of the true input at its halfway setting, over the six policies of that task and assignment, is below -10%. This descriptive threshold is not a test.

Losses occur in 23 of the 64 directions whose label expert changes and in none of the 16 directions without a change (one-sided Fisher exact test, $p = 0.002$). The gap J separates losses from other directions better than the change of expert identity (AUC 0.89 against 0.64), also within the directions with a change (0.84). Losses occur in 11 of 12 directions with $J > 1.5$ and in none of 16 with $J \leq 0.5$. Within tasks, the AUC is 0.75 (HalfCheetah), 0.68 (Walker2d), and 0.57 (Hopper). Ant has one loss. The gap therefore explains the location of losses well in HalfCheetah and Walker2d but not in Hopper.

Equations 7 and 8 concern action errors. At a halfway setting where both neighbors use the same expert and the nominal-input action is close to that expert's action, $v \approx 0$, so any change blended in from a differently labeled endpoint increases the squared action error. The label interventions below test the effect on return. In HalfCheetah under Best, the students change their actions by about half of the endpoint change at these settings. This view is only a guide, because a blended action can fail in closed loop even when it is close to the label.

J.4 Label interventions

We retrain HalfCheetah Concat students under Best with the original states, physics inputs, BC seeds, epochs, batch size, and thread settings. Only the action labels of some training settings change. *Remove* relabels the four gravity and damping endpoints with the nominal label expert, so the label expert no longer changes along these directions. *Create* relabels the mass-increase endpoint with a different expert whose return there is 97.8% of the original label expert's return and whose action gap is $J = 1.59$. Predictions were written in the analysis code before training. Evaluation reuses the configurations and reset seeds of the original students, so differences are paired by episode. Table 16 summarizes the changes in true-input return.

Remove largely restores the targeted halfway settings and trades away endpoint return. After Remove, the true input still trails the nominal input at the gravity-decrease and damping-decrease halfway settings for some seeds, and returns at the 0.75 position stay low there, so the losses are reduced rather than eliminated. Create lowers the return at the predicted halfway setting in every seed, while the mean change over all other single-axis settings stays small. The intervention changes the labels, not only the action gap, so it shows that the label source can cause these failures. It does not show that the gap is their only cause. The study covers one task and one architecture.

J.5 Denser training values

We also trained HalfCheetah and Walker2d Concat and FiLM students under Best on five values per axis instead of three, adding the ten halfway settings. Each halfway setting is labeled by the bank expert with the highest return there over ten episodes. Demonstrations per setting fall from 50 to 26 to keep the number of episodes fixed. The number of states is matched in HalfCheetah but is 16% larger

Table 16: Change in true-input return after the label intervention, relative to the original students. Targeted rows give the range over three BC seeds (and four directions for Remove), with paired 95% t intervals over episodes. Rows marked (mean) give the range over seeds of the average change across all settings in the row. For multi-axis configurations, intervals use the 64 configuration-mean differences.

Intervention	Settings	Change	Intervals
Remove	Targeted halfway settings	$+680$ to $+1{,}220$	12/12 above zero
Remove	Targeted endpoints	$-1{,}323$ to -344	10/12 below zero
Remove	Other single-axis settings (mean)	-63 to $+31$	
Remove	Multi-axis configurations (mean)	$+193$ to $+335$	3/3 above zero
Create	Targeted halfway setting	-695 to -413	3/3 below zero
Create	Targeted endpoint	-277 to -80	2/3 below zero
Create	Other single-axis settings (mean)	-30 to $+33$	
Create	Multi-axis configurations (mean)	$+54$ to $+71$	3/3 include zero

in Walker2d, because the chosen experts survive longer. Denser training removes the multi-axis losses in HalfCheetah. Across the six policies, the benefit of the true input on multi-axis configurations moves from -20 to -31% to $+3$ to $+8\%$, and the true-input return rises by about 30%. Walker2d FiLM improves as well, and Walker2d Concat is mixed. New losses appear next to the new label changes, for example at the 0.75 position of the HalfCheetah gravity-decrease direction and at the Walker2d gravity-increase endpoint, where the new halfway label introduced a second expert. Under the noisy inputs of Appendix J.1, the denser students are less sensitive than the original students in HalfCheetah and more sensitive in Walker2d, so denser coverage does not change sensitivity in a consistent direction.

K A REGISTERED REACHING-TASK EXTENSION

Reacher-v4 tests a two-joint arm reaching a random goal. We fixed the design after the locomotion studies but before selecting experts or training students. We report all 18 students and 36,864 episodes separately from the locomotion census.

Experts, data, and fitting. The archived bank has seven experts: nominal and the two endpoints of mass $(0.75, 1.3)$, damping $(0.5, 2)$, and armature $(0.5, 2)$. Each records 507,904 PPO steps. Expert seeds were not recorded. Open-loop checks confirm that these axes affect dynamics. The tested gravity/friction changes have no effect. We exclude the other 17 archived experts, including target-trained policies. Training cross-evaluation uses $7 \times 7 \times 20 = 980$ episodes. Best selects the largest column mean, breaking ties by index. Fifty 50-step episodes per Own expert give 17,500 common states, relabeled under both assignments. Physics normalization uses the seven training configurations.

Concat, FiLM, and Physics-blind use the same hidden blocks and BC optimizer settings as in the locomotion studies. The conditioned models receive three physics inputs. Both assignments use the same initialization and minibatch order for each method and BC seed (42, 1, 2). Fits use 200 epochs and the same AdamW optimizer, batch size 512, cosine schedule, gradient clipping, and model selection by training MSE. Actions are clipped to the environment bounds.

Evaluation and primary result. Eight independent sets of 32 configurations and an independent 256-configuration reference use the same sampling rule as the locomotion study, restricted to the three active axes. All policies share configurations and four resets per configuration. Reset blocks differ across configurations and sets. Resets change the goal and initial state. We keep raw negative returns.

Let $D_a = C_a - F_a$, where C and F denote Concat and FiLM under assignment a. The primary measure is $I = D_{\text{BEST}} - D_{\text{OWN}}$. Paired episode differences are averaged over three fixed seeds and four resets within each configuration. The two-sided reference interval uses 256 configuration blocks and t_{255}, with no favored direction. It includes zero, so the primary criterion is not met. This does not show that the methods are equivalent.

Table 17: Reacher reference comparisons, averaged over three fixed BC seeds. Intervals account for configuration and paired-reset sampling, conditional on the archived expert bank and fitted students.

Contrast	Mean raw-return gap	95% interval
D_{OWN}	-0.315	$[-0.419, -0.212]$
D_{BEST}	-0.241	$[-0.279, -0.203]$
I (primary)	$+0.074$	$[-0.039, +0.187]$

Table 18: Reacher improves under Best while retaining its ordering. Mean raw return on the independent reference, averaged over three fits. Larger values are better.

Policy	OWN	BEST	Change
Concat	-8.664	-7.127	$+1.537$
FiLM	-8.349	-6.886	$+1.463$
Physics-blind	-7.619	-6.711	$+0.908$

Stable ordering and a physics-blind advantage. Physics-blind leads both conditioned students at all three seeds and both assignments. All twelve gaps keep their signs after the planned Bonferroni correction across 18 method-pair/seed/assignment reference gaps. Both conditioned policies improve more than Physics-blind at every seed. The six secondary interactions stay positive after correction across nine pair/seed interactions. There are no assignment reversals in nine reference comparisons by means, pointwise intervals, or adjusted intervals. None of eighteen fixed pair/seed/assignment comparisons has opposite supported signs across the eight test sets. Among 144 pointwise intervals, fixed-physics intervals support 124 negative gaps and leave 20 unresolved. Configuration-aware intervals give 115 and 29. Concat/FiLM Own seeds 1 and 2 have mixed mean signs, but no interval supports the positive direction.

The plan, recorded results, and analysis code are supplied with the paper.