Consequence Audits for Regression Diagnosis in Learned Dynamics

Anonymous authors
Paper under double-blind review

Abstract

A learned dynamics update can reduce prediction error while introducing errors that have larger effects on a downstream controller. We evaluate these effects by inserting a model prediction into a reference simulator and comparing paired finite-horizon continuations under a fixed controller. Each violation yields a reproducible consequence witness; comparing labels before and after an update identifies regressions. We also study how to obtain these witnesses from a finite model bank under a simulation budget, using the signed responses of queried predictions to prioritize subsequent queries. In a post-hoc analysis of a six-task repair study, 3,931 of 4,221 regressions under uniform repair have lower one-step mean squared error (MSE). A separate fresh-source experiment finds 7,286 failure sources versus 6,569 for MSE round-robin at the same simulation budget. This discovery advantage is not consistent across three new tasks, and witness weighting does not establish an incremental repair benefit. In a repeated-use test, consequence-based selection yields lower return than cost-matched selection rules. The audit thus exposes model-update regressions that prediction error alone can miss, while its selection performance depends on how the model is used.

1 Introduction

Learned dynamics are commonly compared through prediction errors on held-out transitions. A model update, however, changes both the size and the direction of its errors: a smaller residual can have a larger effect on a controller. We ask which predictions become harmful after an update and how their consequences can be reproduced.

Task-dependent model quality is central to objective-mismatch analyses, value-aware learning and value equivalence (Lambert et al., 2020; Farahmand et al., 2017; Grimm et al., 2020). We study this discrepancy through paired evaluations of native model predictions, the outputs of the learned models under each task's stated interface. At a fixed source state and action, we obtain the reference successor and a learned successor, then continue both in the reference environment under the same controller. The prediction and paired continuations form a consequence witness when a predefined tolerance or survival condition is violated. Comparing these labels before and after an update identifies regressions and relates them to changes in prediction error.

Obtaining each witness requires simulation, so evaluating every source–model pair can be costly. We therefore study distinct-source discovery within a finite model bank, a fixed collection of trained predictors. Once a violation has been found at a source, additional failures there do not increase this objective. For unresolved sources, queried continuations supply signed responses—reference return minus candidate return—that can help rank the remaining predictions. Our acquisition rule first probes residual directions and then fits a local response model to allocate further queries.

We contribute a paired audit that identifies and reproduces model-update regressions, and a budgeted search that reuses queried responses to find witnesses across a finite model bank. A local Taylor bound relates response approximation to probe geometry and curvature. We evaluate regression diagnosis retrospectively and test discovery, repair weighting and repeated-use selection prospectively.

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.

2 Consequence audits of native predictions

State and intervention. A source s specifies a physical state x_s, a fixed policy π_s, and action a_s. Let f denote the ground-truth (GT) simulator transition with its auxiliary source context suppressed, and $\hat{f}_m$ the mth predictor. Here a model prediction is its output under the stated task interface. After the reference transition, η_s records the integration snapshot, policy clock and retained context common to both branches. The model branch replaces only the physical successor coordinates. Both branches then follow reference dynamics under π_s. For a valid initial state, $J_H^\pi(z;\eta) = \sum_{h=0}^{H-1} r_h$ is the undiscounted H-step continuation return, with zero rewards after absorption. We define

$$z_s = f(x_s, a_s), \quad \hat{z}_{sm} = \hat{f}_m(x_s, a_s), \quad L_{sm} = J_H^{\pi_s}(z_s;\eta_s) - J_H^{\pi_s}(\hat{z}_{sm};\eta_s). \tag{1}$$

The loss is defined for valid predicted successors. Its return begins after the replaced transition, excluding that transition's immediate reward. The interface uses deterministic policies and saved integration snapshots; fixed-shape policy inference, zero-residual checks and official-environment replay verify continuation consistency.

Eligibility and violation. The eligible source set $\mathcal{S}_E$ contains sources reached before termination whose reference continuation is valid and survives. We report both generated and eligible source counts. Let T_H^π indicate absence of unhealthy termination, $\tau_s > 0$ be a fixed tolerance, and $\mathcal{V}$ the analytic valid/healthy state set. With **1** denoting the indicator function, the violation label is

$$Y_{sm} = \begin{cases} 1, & \hat{z}_{sm} \notin \mathcal{V}, \\ \mathbf{1}\{T_H^{\pi_s}(\hat{z}_{sm};\eta_s) = 0 \ \vee \ L_{sm} > \tau_s\}, & \hat{z}_{sm} \in \mathcal{V}. \end{cases} \tag{2}$$

Invalid predictions are assigned a violation without evaluating an undefined return. For valid predictions, the label tests unhealthy termination or a loss above tolerance. Successful goal completion is absorbing success, and environment time limits are treated separately from physical failure. A negative L_{sm} denotes a higher return than the reference branch and can pass this one-sided test; the target therefore measures conditional deterioration, not absolute return agreement.

Magnitude alone does not determine the label. For valid, surviving continuations, consider affine loss $L(v) = v_1 - v_2$ in residual coordinates. Correcting $(1, 1)$ to $(1, 0)$ reduces squared error from 2 to 1 but increases loss from 0 to 1, creating a violation when $0 < \tau < 1$. Removing one error component removes cancellation in this example, even under an absolute-loss target.

Witnesses and update review. A consequence witness records the source, action, model identity, prediction and simulator snapshot. A simulated violation additionally records the paired continuations; an analytically invalid prediction records the failed validity condition without a candidate rollout. On a common eligible source set, an update causes a regression when $Y_{sm}^{\text{before}} = 0$ and $Y_{sm}^{\text{after}} = 1$; a fix is the reverse. Comparing paired labels and MSE changes identifies predictions that improve in fit while deteriorating under the audit. The same comparison within model–policy contexts measures changes in mean MSE and violation rate.

Discovery utility. We count eligible sources at which at least one violation has been found, including analytically invalid predictions. The first witness at a source increases this count by one; additional witnesses there add nothing, so the source is retired from search. Distinct sources can share an underlying defect. Exhaustive evaluation of the bank identifies all sources with a violation (oracle failure sources); discovery recall is the fraction of these sources found within the budget.

3 Finding witnesses under a simulation budget

Two-direction response acquisition. A campaign shares one query budget across 128 sources. We first credit and retire sources with analytically invalid predictions, without a candidate rollout. For the rest, let $u_{sm} = D^{-1}(\hat{z}_{sm} - z_s)$, with wrapped angular differences for Pendulum. The diagonal D contains training-state standard deviations, floored at 10^{-3}; physical-state MSE is $\|u_{sm}\|_2^2$ divided by the number of coordinates. We first query the largest residual norm at each unresolved source. Responses with $Y_{sm} = 0$ are acceptable and initialize a local ridge fit; O_s indexes these queried

predictions. With $\epsilon = 10^{-6}$ and $\lambda = 10^{-4}$, define

$$r_s = \max\{\max_m \|u_{sm}\|_2, \epsilon\}, \quad \phi_{sm} = u_{sm}/r_s, \quad A_s = \lambda I + \sum_{i \in O_s} \phi_{si}\phi_{si}^\top, \quad b_s = \sum_{i \in O_s} \phi_{si} L_{si}/\tau_s. \tag{3}$$

Dividing loss by τ_s places responses on the scale of the violation threshold. The second probe maximizes the unexplained-direction score h_{sm} among unqueried candidates. After these two passes, subsequent queries maximize the fitted response q_{sm} across the campaign:

$$h_{sm} = \|\lambda A_s^{-1}\phi_{sm}\|_2, \qquad q_{sm} = \phi_{sm}^\top A_s^{-1} b_s. \tag{4}$$

Each acceptable response updates the fit at its source; a violation retires that source immediately. Both passes and subsequent queries use the same budget. The passes share a source order; geometry-score ties favor larger residual norms, then a fixed random permutation. Only queried responses enter the fit. Twelve prespecified live-simulator campaigns reproduce the saved-oracle query prefixes and labels (Appendix I.7).

The response model has no intercept because zero residual gives zero loss. The second probe maximizes the norm of the ridge residual component, combining direction and magnitude; it can select a collinear candidate. Ridge regularization keeps the inverse well-defined. The following bound describes its role in response approximation.

Response approximation. At one source write $g(v) = J_H^\pi(z;\eta) - J_H^\pi(z + Dv;\eta)$, so $g(0) = 0$. Suppose g is twice differentiable on valid-return segments from the origin to the probes and candidate, $\|\nabla g(0)\| \le G$, and the Hessian operator norm is at most M. Let X_s stack the observed $\phi_{si}^\top$ rows. The coefficients $\alpha = (X_s X_s^\top + \lambda I)^{-1} X_s \phi_{sm}$ express the prediction as a weighted sum of observed responses. The normalized ridge prediction satisfies

$$\left| \frac{L_{sm}}{\tau_s} - q_{sm} \right| \le \frac{G r_s}{\tau_s} h_{sm} + \frac{M r_s^2}{2\tau_s} \left(\|\phi_{sm}\|_2^2 + \sum_i |\alpha_i| \|\phi_{si}\|_2^2 \right) =: \varepsilon_{sm}. \tag{5}$$

The first term links the second-probe score to unmeasured response directions; the curvature term also depends on the coefficient magnitudes. For example, when $g(v) = v_2$, a probe at $(1, 0)$ returns zero although $(0, 1)$ has loss one. Querying the latter measures the missing component. Appendix B derives the bound from Taylor expansion and the ridge residual identity. It applies on the stated smooth segments, excluding termination boundaries and the survival indicator. We use scores for ranking and verify violations by simulation; G and M are not estimated.

Simulation cost. The total campaign query cap is b times its number of eligible sources, limited by the number of eligible valid pairs. Queries are pooled, so an individual source can receive more than b. We charge source-generation transitions, $H + 1$ transitions per generated source for its successor and reference continuation, and H per queried candidate. Absorbed rollouts retain their allocated charge; actual nonabsorbed transitions and wall-clock time are reported separately. Training and exhaustive oracle construction are research setup costs.

Search controls. MSE round-robin cycles unresolved sources, querying each source's remaining predictions in decreasing MSE order. Diversity gives equal weight to MSE percentile and distance from previously queried source contexts. Direction-only round-robin continues the geometry-based probe rule without response prioritization. One-probe response omits the second geometry pass and fits the same signed ridge model; magnitude-response instead scales the absolute first response by the ratio of candidate residual norm to the floored probe norm. For new-task discovery, a Gaussian process (GP) uses the same two probes and ranks candidates by posterior probability of exceeding tolerance (Rasmussen & Williams, 2006). Its kernel is chosen on historical initial-bank data. All controls share eligibility, retirement and simulation accounting.

4 Experimental design

Study 1 tests discovery in the initial bank; Study 2 evaluates supervised repair and provides paired updates for regression diagnosis; Study 3 tests discovery on new tasks; Study 4 evaluates model selection for repeated use.

Models, controllers and sources. The initial suite uses Gymnasium (Towers et al., 2024): Pendulum, MountainCarContinuous, and four MuJoCo tasks—Swimmer, Hopper, Walker2d and HalfCheetah (Todorov et al., 2012). Per task, we fit MLP, residual MLP and three-member probabilistic-ensemble dynamics (Chua et al., 2018) at three training seeds, saving checkpoints after 256, 1,024 and 4,096 optimizer updates. This yields 27 predictors; each ensemble contributes its mean. We call final checkpoints mature, giving nine mature models per task. Public PPO (Schulman et al., 2017), SAC (Haarnoja et al., 2018) and TD3 (Fujimoto et al., 2018) controllers from Stable-Baselines3 and RL Baselines3 Zoo (Raffin et al., 2021; Raffin, 2020) define three fixed task–controller cells per task; legacy locomotion actors are adapted to the current simulator interfaces.

Study 1 evaluates the fixed acquisition on the development study's fitted bank using 36,864 fresh sources. Each cell has 2,048 independent trajectories: half contribute reset states and half contribute one intermediate state at steps 32–192. Study 2 applies four supervised repair strategies to the 54 mature models and tests them on another 36,864 sources. All 216 repairs share initial weights within each model, target pools and 1,024 optimizer updates. Half of each minibatch replays training data; half samples eligible Study 1 sources. Uniform sampling assigns equal weights. MSE-priority gives tenfold weight to each model's highest-error 10% of sources; geometry-witness and consequence-witness give tenfold weight to sources found by direction-only and response search, respectively, at $b = 3$. Witness weights are shared across models and incur extra acquisition cost. The regression analysis is post-hoc.

Study 3 applies the unchanged rule to Reacher, InvertedPendulum and InvertedDoublePendulum, with three new PPO seeds per task, new dynamics banks and 18,432 sources. Reacher predictions preserve static goal context. Physical-state audits use $H = 80$. Tolerance is 5% of the larger of absolute reference return and Hr_{scale}, where r_{scale} is the median absolute training reward, floored at 0.1. Appendix A gives the complete training and sampling settings.

Repeated-use selection. Study 4 compensates a one-step full-state sensor delay: a fixed policy acts on the model's predicted current state while the reference plant evolves normally. Each campaign audits all nine mature models at 32 validation sources, selects the lowest violation rate and breaks ties by MSE then model index. Full-budget MSE selects the lowest mean prediction error using additional reference transitions; direct-return selection chooses the highest validation deployment return. Both receive comparable simulation budgets.

Across nine tasks and 27 fixed policy contexts, each of six campaigns uses 16 fresh paired test resets: 28,512 rollouts on 2,592 reset IDs. Deployment lasts 200 steps, except Reacher (50) and MountainCarContinuous (1,000). For each reset, we divide the consequence-minus-comparator return difference by the larger of the absolute fresh-observation policy return and deployment horizon times r_{scale}. We average these normalized differences equally over resets, campaigns and contexts, then multiply by 100 to report percentage points (pp). The primary requires at least 1 pp gain and a positive 99.99% paired-campaign bootstrap lower endpoint against both comparators.

A separate experiment evaluates three released TD-MPC2 checkpoints (Hansen et al., 2024) on DM Control cartpole-swingup (Tunyasuvunakool et al., 2020) under a reduced planning budget. At the first planning call, the intervention substitutes a one-step latent dynamics prediction for the encoded next observation; physical state is preserved and later calls use observed feedback. Selection compares audit violations, latent MSE and validation return. Each selected checkpoint includes the encoder, dynamics, reward, value and policy, so this comparison concerns complete planner bundles.

Criteria and inference. The primary discovery budget is $b = 3$. Success requires at least 5 pp equal-cell mean (macro) recall gain and a positive lower confidence endpoint for gain per generated source against both MSE round-robin and diversity. Zero-oracle cells contribute zero contrast to the mean and have undefined recall. Repair risk is the eligible-pair violation rate over all nine mature models, averaged equally across task–controller cells. Consequence-witness repair must reduce this risk by at least 0.25 pp, with a positive lower interval endpoint, against both uniform and MSE-priority repair.

Paired campaigns are resampled within fixed cells, retaining paired models together; intervals condition on the tested banks and controllers. The prespecified interval levels for Studies 1–3 are 99.9609375%, 99.990234375% and 99.9951171875%, respectively, reflecting sequential allocation across confirmation rounds and comparisons. Appendix A gives the resampling settings.

5 Results

5.1 Model updates reveal regressions hidden by MSE

Uniform supervised repair reduces aggregate violations while introducing new regressions. Table 1 shows that the latter include lower-MSE predictions under all four repair strategies. Many of these predictions were also within the absolute consequence tolerance before repair. The discrepancy therefore persists for predictions initially close to the reference return, beyond those accepted through the one-sided allowance for improvements.

Table 1: Post-hoc regressions across repair strategies. Lower MSE compares the same source–model pair before and after repair. Abs.-tolerable is the subset of lower-MSE regressions whose original absolute loss is within tolerance. Sources gives the number of distinct sources with a lower-MSE regression.

Strategy	Regressions	Lower MSE	Abs.-tolerable	Sources
Uniform	4221	3931	3630	2414
MSE priority	4306	3850	3535	2391
Geometry witness	4058	3754	3418	2303
Consequence witness	4115	3789	3466	2346

The same discrepancy appears after aggregation by model and controller. Under uniform repair, mean MSE improves in 156 of 162 contexts, but 18 of these have a higher violation rate. Figure 1 shows the paired changes for all four strategies. This post-hoc analysis describes the evaluated updates. It identifies counterexamples to MSE improvement as sufficient evidence of consequence improvement. Across fixed task–controller cells, a separate descriptive analysis finds MSE–risk ranking reversals in 16/18 full-bank and 11/18 mature-bank cells.

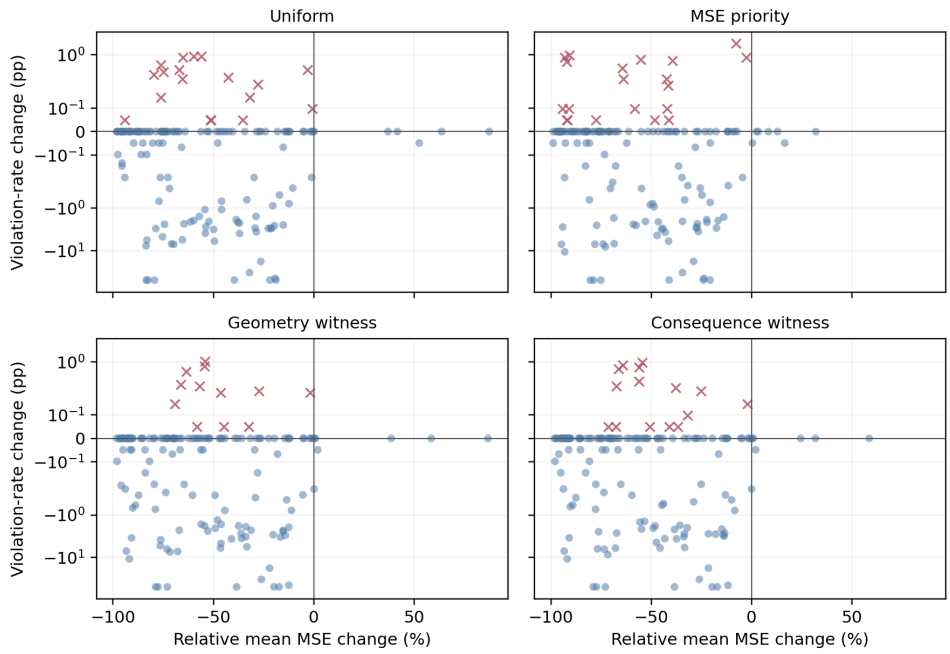


Figure 1: Paired changes in mean MSE and violation rate for 162 model–policy contexts per repair strategy. Red crosses indicate lower MSE with higher violation rate. The vertical scale is linear within 0.2 pp and logarithmic outside. Markers describe observed changes without a significance test.

Inspecting an individual regression. Figure 2 shows a HalfCheetah update that reduces MSE while changing a within-tolerance continuation into a violation. The case was selected post-hoc by source/model ID from qualifying regressions.

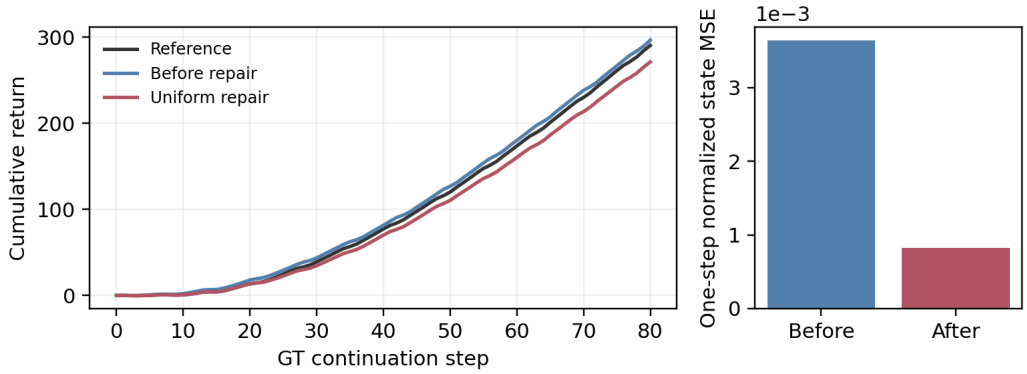


Figure 2: HalfCheetah regression example. Left: cumulative reference-continuation returns. Right: standardized one-step MSE, which falls from 3.64×10^{-3} to 8.26×10^{-4} after uniform repair. Signed loss rises from -6.201 to 19.119, crossing tolerance 14.528; both continuations survive.

To examine which residual components affect this witness, we replace one coordinate of the repaired successor with its reference value while holding the others fixed. All 18 corrections weakly reduce MSE, yet five increase consequence loss and eleven resolve the violation (Figure 3). These synthetic interventions show how reducing different error components can have opposite consequences under the same controller. Appendix I gives the selection procedure, replay controls and target/tolerance sensitivity.

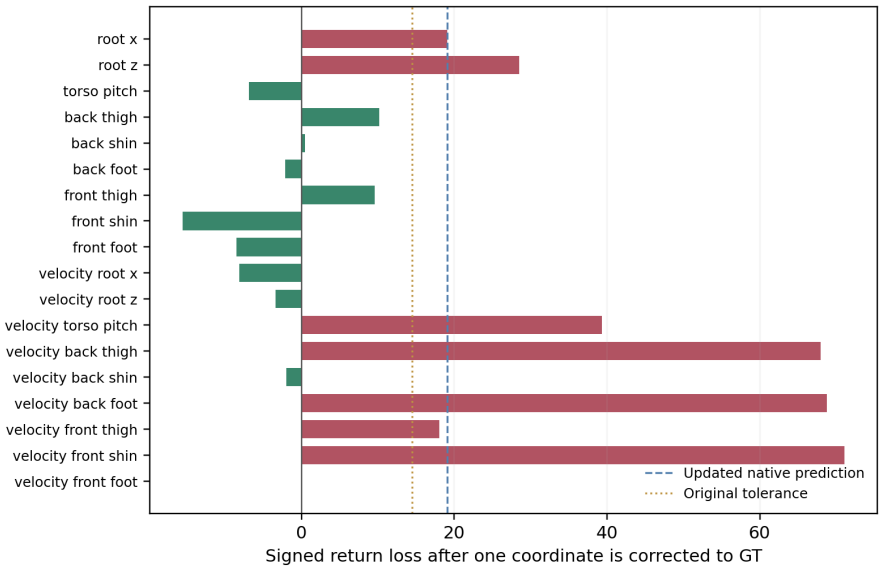


Figure 3: Single-coordinate reference corrections of the repaired successor in Figure 2, with all remaining coordinates fixed. Bars show signed loss; green/red indicate satisfaction/violation of the one-sided target. The dashed blue line is the native repaired prediction and the dotted orange line is its tolerance. These interventions use reference coordinates rather than a learned repair.

5.2 Response reuse increases discovery in the initial suite

Study 1 finds 7,286 of 9,210 oracle failure sources with two-direction response, compared with 6,569 for MSE round-robin and 6,104 for diversity (Table 2). All three use the same charge of 13,487,092 reference transitions, including source generation and reference continuations; two-direction response makes 106,605 candidate queries.

Table 2: Failure sources discovered at budget $b = 3$, from 6,144 generated sources per task. Elig.: eligible sources; Oracle: sources with at least one violating prediction; Dir. RR: direction-only round-robin.

Task	Elig.	Oracle	Two-dir.	MSE RR	Diverse	Dir. RR
Pendulum	6144	13	13	10	10	10
MountainCar	4816	29	26	26	27	25
Swimmer	6144	1815	1795	1564	1632	1749
Hopper	6144	6	2	1	0	0
Walker2d	6143	2582	2231	1882	2072	2161
HalfCheetah	6144	4765	3219	3086	2363	3141

Against MSE round-robin, macro recall gain is 12.85 pp and discovery gain per generated source is 1.945 pp (CI [1.562, 2.330]). Against diversity, the corresponding gains are 13.50 pp and 3.206 pp (CI [2.683, 3.762]). These satisfy the prespecified joint discovery criterion. The macro-recall interval against round-robin spans zero, $[-0.868, 23.117]$ pp; the positive interval concerns per-source discovery.

The second probe pass and subsequent adaptive queries together account for 4,122 discoveries, including 769 at sources whose first response showed a return improvement. One-probe response finds 6,848 sources, direction-only round-robin 7,086, and magnitude-response acquisition 5,514. These descriptive controls suggest that signed response reuse contributes beyond the two geometry probes. Figure 4 shows how recall changes with budget: the response variants share the first probe, then differ in how they allocate further queries. Full-budget recall reaches one by exhaustive coverage. Appendix G reports all methods, subsets and budgets.

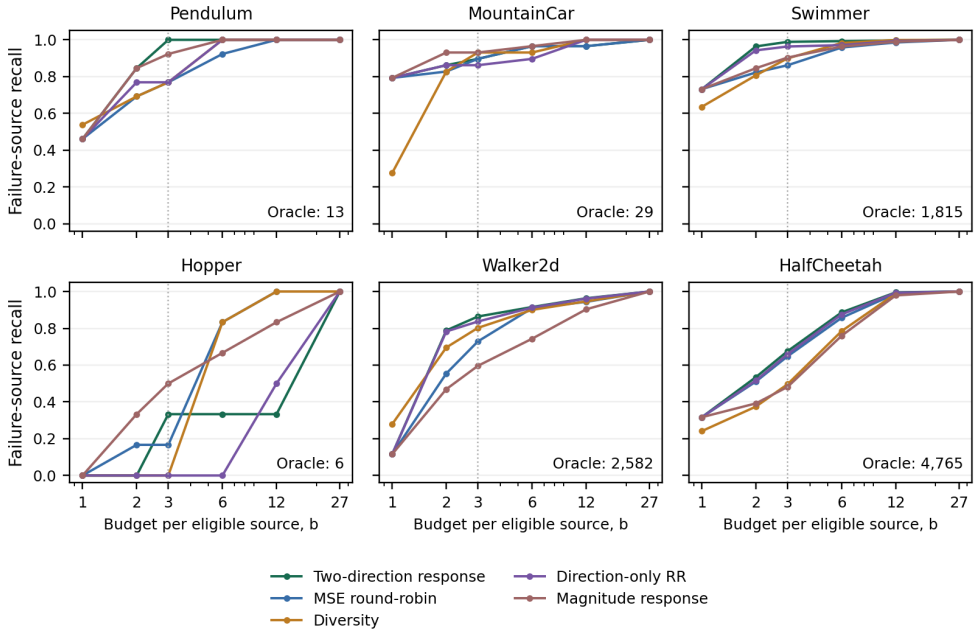


Figure 4: Failure-source recall versus campaign budget per eligible source in the initial bank. The dotted line marks $b = 3$; each panel reports its oracle failure-source count.

Discovery on new tasks. In Study 3, the unchanged two-direction response rule finds 338/1,078 sources, compared with 311 for MSE round-robin, 203 for diversity, 374 for response-GP and 382 for one-probe response, at 6,699,833 charged transitions. Against round-robin, macro gain is 4.481 pp and per-source gain is 0.146 pp (CI $[-0.179, 0.461]$), below the joint transfer criterion. The GP contrast is -0.195 pp per source (CI $[-0.467, 0.054]$). Thus the new bank provides no consistent advantage for the second probe or the ridge response model. Of the oracle failures, 1,042 are in InvertedDoublePendulum and 36 in Reacher; all three InvertedPendulum cells have none. Appendix I.6 gives the task breakdown and ablations.

5.3 Repair weighting and repeated-use selection

Witness weighting for repair. Consequence-weighted repair leaves 6,955 eligible-pair violations, compared with 7,257 for uniform and 7,194 for MSE-priority repair. Equal-cell risk reductions are 0.091 pp (CI $[-0.020, 0.199]$) and 0.072 pp (CI $[-0.051, 0.189]$), respectively. Neither establishes the prespecified 0.25 pp incremental benefit. The large improvement over unchanged models includes the shared benefit of additional data and optimization; witness weighting also incurs additional simulation cost.

Selecting models for repeated use. In the delay-compensation experiment, consequence selection yields lower normalized return than full-budget MSE by 1.111 pp (99.99% CI $[-2.354, -0.035]$) and than direct-return selection by 2.298 pp (CI $[-3.548, -1.209]$). Both intervals favor the baselines, so the consequence-based rule underperformed in this repeated-use setting.

In the separate TD-MPC2 experiment, all three selectors choose the same checkpoint in all six groups, yielding no incremental selection benefit. Appendix J reports checkpoint returns, per-task delay-compensation outcomes, costs and timing. The earlier selection experiment with audit risk as its endpoint is reported separately in Appendix H.

6 Related work

Value-aware model learning and value equivalence formalize task-dependent model quality, while objective-mismatch analyses examine the gap between prediction objectives and control performance (Farahmand et al., 2017; Grimm et al., 2020; Lambert et al., 2020). With fixed auxiliary context, our signed response is the difference in a single finite-horizon continuation value evaluated at the reference and predicted successors. Value equivalence instead requires Bellman-update agreement over specified policy and function classes. We use a one-sided tolerance and survival condition to identify individual violations, then compare paired labels across updates and search for witnesses under a simulation budget.

Simulation-based falsification searches for counterexamples using quantitative simulation feedback (Corso et al., 2021; Dreossi et al., 2019). We apply this approach to a finite bank of native predictions, with a distinct-source objective and explicit continuation costs. Active testing studies label acquisition for model evaluation and the treatment of selective observations (Kossen et al., 2021). Our adaptive search targets discovery, whereas model selection uses balanced observations of all candidates to estimate audit risk.

7 Discussion

A lower mean MSE can coexist with new source-level violations and higher violation rates in particular controller contexts. The audit supplies the source, prediction and continuation needed to reproduce each discrepancy. Coordinate interventions illustrate how these records support investigation of an update's effects on subsequent control.

Single insertion measures the effect of one successor under a fixed controller and reference continuation. The delay-compensation experiment repeatedly changes the policy input while preserving the plant state. In that setting, the consequence-based selection rule yielded lower held-out return than the cost-matched baselines. This comparison evaluates the complete selection rule, including finite validation data and its MSE tie-break. It establishes the rule's performance in the tested setting

without isolating target mismatch as the cause. Consequence labels remain useful for locating and reproducing update discrepancies.

8 Conclusion

Paired consequence audits reveal learned-dynamics updates that reduce MSE yet introduce controller-conditional regressions. Saved predictions and continuations make these discrepancies reproducible. Response reuse improves distinct-source discovery in the initial bank, with no consistent gain on the new tasks. The repair and selection studies do not establish corresponding downstream benefits.

Reproducibility statement

The supplementary material specifies source generation, model and controller revisions, intervention interfaces, acquisition rules and statistical criteria. Recorded predictions, paired returns and query traces support reconstruction of the reported labels and comparisons. A companion CPU replay package covers all four studies; simulator replay checks, execution costs and artifact provenance are documented separately.

AI use statement

We used generative AI tools to assist with manuscript editing and translation, identifying relevant literature, implementing and debugging code for the experimental setup, checking the validity of equations and mathematical derivations, and generating figures. These tools also assisted with research design, experiment execution, analysis, and mathematical exposition. The authors reviewed the resulting text, checked references against the original publications, and tested the AI-assisted code. The authors take responsibility for the final manuscript, implementation, and reported results.

References

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

Anthony Corso, Robert J. Moss, Mark Koren, Ritchie Lee, and Mykel J. Kochenderfer. A survey of algorithms for black-box safety validation of cyber-physical systems. *Journal of Artificial Intelligence Research*, 72:377–428, 2021. doi: 10.1613/jair.1.12716. URL https://arxiv.org/abs/2005.02979v3.

Tommaso Dreossi, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Hadi Ravanbakhsh, Marcell Vazquez-Chanlatte, and Sanjit A. Seshia. VerifAI: A toolkit for the formal design and analysis of artificial intelligence-based systems. In *Computer Aided Verification (CAV)*, volume 11561 of *Lecture Notes in Computer Science*, pp. 432–442, 2019. doi: 10.1007/978-3-030-25540-4_25. URL https://doi.org/10.1007/978-3-030-25540-4_25.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-Aware Loss Function for Model-based Reinforcement Learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 1486–1494, 2017. URL https://proceedings.mlr.press/v54/farahmand17a.html.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1587–1596, 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The Value Equivalence Principle for Model-Based Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1861–1870, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, Robust World Models for Continuous Control. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cf73d57b6dcda32b293df7c2d5341f49-Abstract-Conference.html.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5753–5763. PMLR, 2021. URL https://proceedings.mlr.press/v139/kossen21a.html.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective Mismatch in Model-based Reinforcement Learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120, pp. 761–770, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

Antonin Raffin. Rl baselines3 zoo. GitHub software repository, 2020. URL https://github.com/DLR-RM/rl-baselines3-zoo.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL https://jmlr.org/papers/v22/20-1364.html.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 026218253X. URL https://gaussianprocess.org/gpml/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109. URL https://doi.org/10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024. URL https://arxiv.org/abs/2407.17032.

Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 6:100022, 2020. doi: 10.1016/j.simpa.2020.100022. URL https://doi.org/10.1016/j.simpa.2020.100022.

A Detailed experimental protocols

Initial six-task bank. The initial suite is Pendulum-v1, MountainCarContinuous-v0, Swimmer-v5, Hopper-v5, Walker2d-v5 and HalfCheetah-v5, under Gymnasium 1.2.0/MuJoCo 3.3.3. Each task has plain MLP, residual MLP and three-member probabilistic-ensemble dynamics, three training

seeds and checkpoints at 256, 1,024, 4,096 updates: 54 fits and 162 predictor means. These are baseline dynamics fits using established probabilistic modeling principles (Chua et al., 2018), not PETS planning reproduction. Public SB3 PPO/SAC/TD3 controllers are fixed contexts (Raffin et al., 2021; Raffin, 2020); legacy V3 actors are adapted to V5, not original benchmark reproductions. Each task uses 65,536 PPO/SAC training transitions and 8,192 separately seeded validation transitions; TD3 is withheld from this original collection.

Development and source confirmation. The development study prospectively tested the earlier magnitude rule and failed. Using those outcomes, we specified the two-direction rule and ran Study 1 on 36,864 fresh sources with the same bank. Studies 1–3 use 2,048 sources per task/controller cell, one per independent trajectory: half resets and half anchors at steps 32–192. A source is eligible only if its reference continuation is valid and survives. We fix $H = 80$ and training-state scales with a 10^{-3} floor, and use

$$\tau_s = 0.05 \max\{|J_H^{\pi_s}(z_s; \eta_s)|, H r_{\text{scale}}\}, \qquad r_{\text{scale}} = \max\{0.1, \text{median}_{\text{train}}\, |r|\}. \tag{6}$$

The tolerance is an operational convention. All source generation and reference/probe continuations are charged. Canonical 128-row actor arithmetic, official-implementation checks, saved integration state and zero-residual pilots guard against numerical artifacts (Appendix D).

Controls and fixed discovery decision. Controls include MSE round-robin, state diversity, uniform order, disagreement, magnitude and other signed-response variants, and direction-only round-robin with identical first two probes. All share source retirement, analytic invalid-state credit and caps $b \in \{1, 2, 3, 6, 12, 27\}$. The two primary baselines are MSE round-robin and diversity at $b = 3$. Let U_c count discovered sources, F_c oracle failure sources and C task–controller cells. The equal-cell recall contrast is

$$\Delta_{\text{macro}} = \frac{1}{C} \sum_{c=1}^{C} \begin{cases} (U_c^{\text{ours}} - U_c^{\text{base}})/F_c, & F_c > 0, \\ 0, & F_c = 0. \end{cases} \tag{7}$$

Here $C = 18$ for Study 1 and $C = 9$ for Study 3. Zero-oracle cells contribute zero contrast and have undefined recall. Success requires macro recall gain at least 5 pp and a positive lower interval endpoint for discovery gain per generated source against both baselines. Paired 128-source campaigns are resampled within each fixed cell; intervals condition on the fitted banks/controllers. Study 1 uses 200,000 draws and 99.9609375% intervals. Other budgets and subsets are descriptive.

Study 2: witnesses as repair data. Study 2 applies four prespecified repair strategies to all 54 mature models. Each is copied into unchanged, uniform replay, MSE-priority, geometry-witness and consequence-witness conditions. Repairs share initial weights, normalization, 1,024 optimizer updates, learning rate 10^{-4} and batch 512. Half each minibatch replays original training data; half samples eligible Study 1 sources. Priority sources receive weight 10 instead of 1. MSE uses each model's top 10%; witness conditions use sources actually discovered at cap 3 in the mature nine-model bank. Zero witnesses implies uniform sampling. All four repairs retain the original supervised dynamics objective; there are 216 repairs, not 216 independently initialized models.

Fresh Study 2 sources are never used for fitting. We test 36,864 sources against 45 predictors/task (1,658,880 pairs), preserving fixes and regressions. Primary consequence-witness repair must reduce equal-cell, nine-model-average eligible-pair risk by at least 0.25 pp versus both uniform and MSE-priority, with positive paired-campaign interval lower endpoints. Resampling retains all paired models together (400,000 draws, 99.990234375% intervals). This matches optimizer updates and target pools, *not GT information cost*: witness methods pay extra continuations. TD3 is included in the repair pool and is no longer a held-out repair controller.

Study 3: new tasks and a response-GP control. The unchanged ridge rule is then evaluated on Reacher-v5, InvertedPendulum-v5 and InvertedDoublePendulum-v5. For each, we train PPO from scratch at three fixed seeds for 1,048,576 transitions each, without performance-based selection. A new 27-predictor bank/task uses the same architecture/seed/checkpoint design. Dynamics data use policy seeds 0/1; seed 2 is held out from collection. Test sources total 18,432 with 497,664 native pairs. Reacher goal coordinates and velocities are preserved as known static context; its H80 audit continues the physical dynamics beyond the training TimeLimit 50. This explicitly wrapped successor, rather than an unconstrained prediction of context, is the evaluated model.

We add a finite-bank response-GP control using the same two paid geometry probes and Gaussian posterior conditioning (Rasmussen & Williams, 2006). Among lengthscales 0.25,0.5,1, historical Study 1 selects 0.5; its 7,306 development discoveries exceed ridge's 7,286. That comparison is development only. Study 3 fixes kernel, noise, rule, models and criteria before fitting/test outcomes. GP subsequently ranks unqueried predictions by posterior $P(\tilde{g} > 1)$ for normalized signed response $\tilde{g} = L/\tau$; this is an acquisition heuristic, not calibrated failure probability. Study 3 retains the discovery primary (800,000 draws, 99.9951171875% intervals); the GP contrast is prespecified secondary.

Historical model choice. A separate Study 1 test froze MSE-based and consequence-based model choices from the exhaustive development study before viewing Study 1 outcomes. Its mature-bank primary required 0.5 pp equal-cell risk reduction and a positive interval lower endpoint. The plan was frozen after raw generation began but before outcome inspection. We report its outcome and information cost (Appendix H).

Study 4: repeated-use selection. The normalized-return primary is specified in Section 4. Model choices are fixed before held-out tests. We use 200,000 paired-campaign bootstrap draws, with quantiles 0.00005 and 0.99995. The released TD-MPC2 checkpoint comparison (Hansen et al., 2024) is a separate descriptive experiment with 95% intervals (Appendix J).

B Response approximation and query invariance

For arbitrary probe vectors u_i and coefficients α_i, the smoothness assumptions in Section 3 imply

$$\left| g(v) - \sum_i \alpha_i g(u_i) \right| \le G \left\| v - \sum_i \alpha_i u_i \right\|_2 + \frac{M}{2} \left(\|v\|_2^2 + \sum_i |\alpha_i| \|u_i\|_2^2 \right). \tag{8}$$

Taylor's theorem gives $g(w) = \nabla g(0)^\top w + R_w$ with $|R_w| \le (M/2)\|w\|^2$. Therefore

$$g(v) - \sum_i \alpha_i g(u_i) = \nabla g(0)^\top \left(v - \sum_i \alpha_i u_i \right) + R_v - \sum_i \alpha_i R_{u_i}. \tag{9}$$

Cauchy–Schwarz and the triangle inequality prove Equation 8. With X_s as in Section 3 and $y_s = (L_{si}/\tau_s)_{i \in O_s}$, the ridge coefficients satisfy

$$\alpha = (X_s X_s^\top + \lambda I)^{-1} X_s \phi_{sm}, \qquad \tau_s q_{sm} = \sum_i \alpha_i L_{si},$$

$$u_{sm} - \sum_i \alpha_i u_{si} = r_s \lambda A_s^{-1} \phi_{sm}. \tag{10}$$

The first identity follows from the primal–dual ridge identity; the second follows from $A_s = \lambda I + X_s^\top X_s$. Substitution into Equation 8, followed by division by τ_s, gives Equation 5. The direction term is proportional to the score used for the second probe, while the curvature term also depends on the coefficient magnitudes. The regularized residual operator is not generally an orthogonal projection. For a valid, surviving candidate, $q_{sm} - \varepsilon_{sm} > 1$ implies a loss violation and $q_{sm} + \varepsilon_{sm} \le 1$ implies loss within tolerance. Likewise, $q_{sm} - q_{sn} > \varepsilon_{sm} + \varepsilon_{sn}$ implies $L_{sm} > L_{sn}$. These conditional implications describe response accuracy under the stated smoothness bounds; they do not establish a discovery-rate guarantee.

The acquisition receives numeric responses and binary labels only through a query callback. Its fixed features include all residuals, but not their unqueried consequences. Induction proves identical query prefixes for oracles agreeing on observed responses. Every queried pair is unique; positives retire their source. A full budget must recover every eligible source with at least one violation. We verify these identities and independently reconstruct the selected ridge fit using the dual system.

C Exact controls

We compare the development controls below with two-direction response, its one-probe variant, exploration bonuses 0.5 and 1.0, and direction-only round-robin. The bonus is

$\beta \max(1, \max_{i \in O_s} |L_{si}/\tau_s|) h_{sm}$. Direction-only round-robin cycles unresolved sources with the same geometry probe rule, without numeric response prioritization.

For a bank of K candidates in d residual dimensions, let $r_{sm} \in \{0, \dots, K-1\}$ be descending scaled-MSE rank within source. Let $p(s)$ denote the first queried model at source s, and $\ell_s = L_{s,p(s)}$ its observed loss. All fixed priority ties use the same seeded random pair permutation. Source retirement is common.

- MSE ranks all remaining pairs by $\|u_{sm}\|_2^2/d$. Round-robin ranks by $-r_{sm}$.
- Uniform uses only the seeded random permutation.
- Diversity combines the empirical percentile of pair MSE with distance from previously audited source contexts, each with weight $1/2$. Context is the normalized GT successor/action representation. Euclidean distance is divided by the square root of its dimension and capped at 1; before any query the distance is 1. An already visited context therefore has distance 0.
- Absolute-response shares the first pass and ranks by $\log(1 + |\ell_s|/\ell_0) - \log(1 + r_{sm})$, with $\ell_0 = 0.05 H r_{\text{scale}}$.
- Signed-projection shares the first pass and ranks by $\ell_s u_{sm}^\top u_{s,p(s)} / \max(\|u_{s,p(s)}\|_2^2, 10^{-12})$.
- Ensemble disagreement is member-mean variance in standardized learned-delta coordinates, averaged over output dimensions. Deterministic predictors receive zero in the pooled bank; the ensemble-only subset is the comparable-interface analysis.
- Bank disagreement uses $\log(1 + 10^6 \operatorname{mean}_j \operatorname{Var}_m u_{smj}) - \log(1 + r_{sm})$. It is a fixed descriptive control, not a calibrated probability.

The alternatives, query budgets and primary baselines were specified before the confirmation outcomes.

D Data, dynamics fits and controller adaptation

For each of the initial six tasks, 64 parallel data trajectories alternate PPO and SAC controllers. Actions receive independent Gaussian noise with standard deviation 0.15 times the action bound; 15% of rows instead receive uniform random actions, followed by action clipping. Terminal trajectories and the configured collection time limits are reset with nonoverlapping deterministic seed sequences. Training and validation use separate seeds. Inputs and one-step physical deltas are standardized on training data only, using floors 10^{-3} and 10^{-4}, respectively. Pendulum input uses cosine, sine and velocity; its angular delta uses the wrapped convention. Global translation coordinates are set to zero in locomotion inputs while the predicted output remains the physical delta. The confirmation residual scale is training-state standard deviation, distinct from the learned-delta standardization. V20 uses the same collection design with native PPO seeds 0/1; its task-specific interface is in Appendix I.

The MLP has two hidden layers of width 128 and SiLU activations. The residual MLP has an input projection, two LayerNorm residual blocks with residual factor 0.5, a final LayerNorm and output projection. Each of the three ensemble members is a two-hidden-layer Gaussian predictor; the mean is the candidate, and log variance is clamped to $[-10, 2]$ during Gaussian negative-log-likelihood training. Members independently bootstrap the minibatch. All fits use AdamW with learning rate $5 \cdot 10^{-4}$, weight decay 10^{-5}, batch size 512 and gradient-norm clipping at 10. Checkpoints are taken after 256, 1,024 and 4,096 updates regardless of validation quality.

Table 3: Physical successor dimensions and controller contexts. Public controller checkpoints are fixed, not policy-training repetitions.

Task	State dim.	Action dim.	Fixed controllers
Pendulum	2	1	PPO, SAC, TD3
MountainCarContinuous	2	1	PPO, SAC, TD3
Swimmer	10	2	PPO, SAC, TD3
Hopper	12	3	PPO, SAC, TD3
Walker2d	18	6	PPO, SAC, TD3
HalfCheetah	18	6	PPO, SAC, TD3

Policy assets retain exact public repository revisions, original configuration, normalization statistics and weight hashes. Extracted actors are checked against SB3 deterministic outputs. PPO actions are clipped; SAC/TD3 apply their squashing and action scaling. The PPO Swimmer and Walker2d policies receive their remaining-time feature. Normalization is frozen. These checks establish implementation agreement for the adapted evaluation, not reproduction of legacy V3 benchmark returns.

The MuJoCo batch implementation saves full integration snapshots, restores auxiliary state and warm starts, and copies caller-owned snapshot buffers before stepping. Physical interventions replace position/velocity coordinates only. Independent official-environment replay checks states and rewards after intervention using the same recorded actions. Classic-control implementations match the official storage/observation precision. A short pilot checks repeated return agreement and zero-residual consistency before each full confirmation. Pre-freeze engineering failures and repairs are retained.

A V17 post-freeze pilot revealed a numerical confound: changing the actor inference batch shape changed float32 actions enough to alter an 80-step continuation even at zero residual. The largest observed discrepancy in the dedicated engineering check was 22.49 return units. Fixed 128-row padded actor inference reduced all nine checked controller/batch discrepancies to exactly zero, including the confirmation batch size. We preserved the original code/freeze, interrupted the two other confirmation runs, and reran every pilot/confirmation with the correction. Source IDs, thresholds, acquisition rules, budgets and primary criteria did not change, and no acquisition outcome was used to choose this correction. The first three already-collected datasets and fitted banks remain fixed with their original execution provenance; later data collection uses the corrected arithmetic. Interrupted data and their compute costs remain in the archive and are excluded from the confirmation analysis.

E Sampling, costs and prespecified sensitivity analyses

Each task/policy uses 2,048 unique source seeds. Even-indexed sources use reset state; odd-indexed sources use one independently assigned anchor in 32–192. This yields 1,024 sources per distribution/cell and no repeated anchors from the same trajectory. A reference that terminates unhealthily is excluded from conditional discovery, but its generation and reference acquisition remain charged. MountainCar goal completion is successful absorption. A state already at the goal is not stepped again in continuation.

Campaigns contain 128 sources of one task/policy, with a fixed half-reset/half-anchor mixture. There are 16 campaigns per cell. V18 has 288 campaigns and uses 200,000 bootstrap resamples within each cell. Its round allocation is 0.00078125, split across two primary comparisons, giving quantiles 0.0001953125 and 0.9998046875. V19 retains 288 campaigns with all repair conditions paired; V20 has 144 campaigns. Their interval levels are specified in Appendix A. The discovery criteria combine a macro point estimate and a per-source interval, with zero-oracle cells included. Intervals are conditional empirical bootstrap estimates.

The full 27-candidate bank is primary. Descriptive subsets use 9 candidates from one architecture or 9 final-checkpoint candidates across architectures, retaining the same query-per-eligible-source convention. Policy and reset/anchor profiles are subgroup counts from the original campaigns; they do not receive independently reallocated budgets. Dynamics MSE/violation tables describe fixed correlated checkpoints and do not treat them as independent statistical observations.

Predictions are generated by the saved learned networks without residual amplification. The V20 Reacher interface additionally preserves the stated static context coordinates. Reference rewards, candidate rewards, loss, termination, physical validity and source eligibility are saved separately. Features are saved before labels. Independent analysis verifies label construction, model forward passes, query traces, retirement, full-budget identity and cost arithmetic. Actual execution counts and conservative audit charges have different meanings and are reported separately.

F Development and confirmation

The artifact versions map to the studies as follows: development (V17), source confirmation (Study 1, V18), repair (Study 2, V19), new-task discovery (Study 3, V20), and repeated-use selection (Study 4, V21). Version identifiers below refer to these records.

The earlier Cartpole work developed the audit objective and candidate response rules. In a calibration-inclusive round, the selected evaluator found 215/273 failure sources while round-robin found 231. A later source-only evaluator found 270/309 against 261 for round-robin but failed its stronger frozen primary. The subsequent no-offline-calibration rule found 262/271 versus 229 for round-robin and 181 for diversity on 24,576 fresh sources, at the same 6,729,376 GT charge; its round-robin recall contrast was 12.18 percentage points with a 99.84375% bootstrap interval [5.11,19.85]. The first pass cost 21,181 queries and discovered 203 sources; the remaining stage discovered 59.

The development study extended the magnitude rule to six tasks and did not meet its primary criterion: 5,465/9,234 sources versus 6,575 for round-robin and 6,142 for diversity at the same charge. Its outcomes were then used to develop two-direction response, evaluated on fresh Study 1 sources with the same bank. The Cartpole and six-task studies have separate populations and analyses.

G Complete empirical tables

Table 4: All eighteen controller cells, including zero-oracle cells.

Task	Policy	Elig.	Oracle	Two-dir.	MSE RR	Div.	Dir. RR
Pendulum	PPO	2048	3	3	2	3	3
Pendulum	SAC	2048	6	6	5	4	4
Pendulum	TD3	2048	4	4	3	3	3
MountainCar	PPO	2048	8	6	6	7	5
MountainCar	SAC	1511	21	20	20	20	20
MountainCar	TD3	1257	0	0	0	0	0
Swimmer	PPO	2048	761	756	579	660	757
Swimmer	SAC	2048	310	301	263	262	292
Swimmer	TD3	2048	744	738	722	710	700
Hopper	PPO	2048	0	0	0	0	0
Hopper	SAC	2048	2	2	0	0	0
Hopper	TD3	2048	4	0	1	0	0
Walker2d	PPO	2048	1033	1027	675	902	1028
Walker2d	SAC	2047	361	106	105	107	76
Walker2d	TD3	2048	1188	1098	1102	1063	1057
HalfCheetah	PPO	2048	1532	932	923	537	918
HalfCheetah	SAC	2048	1387	798	671	646	731
HalfCheetah	TD3	2048	1846	1489	1492	1180	1492

Table 5: Prespecified nine-candidate architecture and mature subsets; descriptive.

Task	Subset	Oracle	Two-dir.	MSE RR	Div.	Dir. RR
Pendulum	mlp	7	7	7	7	7
Pendulum	residual	12	12	7	10	8
Pendulum	ensemble	5	5	4	4	5
Pendulum	mature	3	3	3	3	3
MountainCar	mlp	17	9	16	10	10
MountainCar	residual	21	20	20	20	21
MountainCar	ensemble	2	2	2	2	2
MountainCar	mature	13	9	13	13	8
Swimmer	mlp	229	221	168	182	198
Swimmer	residual	918	916	913	909	915
Swimmer	ensemble	1278	1275	1218	1258	1211
Swimmer	mature	397	392	375	369	386
Hopper	mlp	5	3	3	5	3
Hopper	residual	0	0	0	0	0
Hopper	ensemble	1	1	1	1	1
Hopper	mature	0	0	0	0	0
Walker2d	mlp	2326	2232	2027	2166	2197
Walker2d	residual	2261	2069	2099	2071	1607
Walker2d	ensemble	2191	2128	1922	2152	2094
Walker2d	mature	2222	2154	2092	2116	2115
HalfCheetah	mlp	3614	2995	2858	2616	2789
HalfCheetah	residual	3554	2973	2969	2611	2855
HalfCheetah	ensemble	3418	2856	2598	2215	2618
HalfCheetah	mature	3048	2402	2235	1860	2192

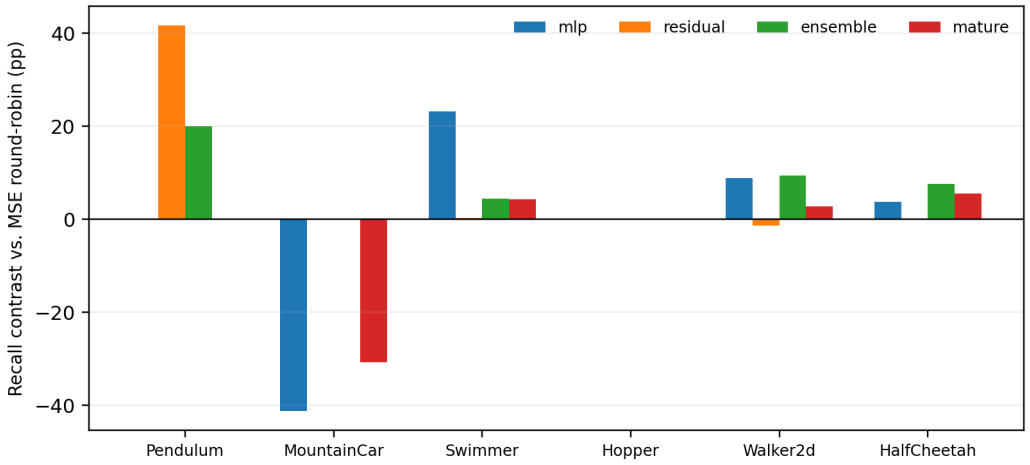


Figure 5: Recall difference between two-direction response and MSE round-robin within each nine-candidate subset at $b = 3$. Zero-oracle subsets contribute zero contrast.

Table 6: Development counts for one- and two-probe response rules and exploration bonuses. Two-direction response without a bonus was selected before Study 1.

Task	One	Two	Two +.5	Two +1	Dir. RR
Pendulum	19	19	19	19	14
MountainCar	24	24	24	24	24
Swimmer	1779	1830	1831	1831	1782
Hopper	1	1	1	1	0
Walker2d	1948	2249	2256	2241	2160
HalfCheetah	3155	3274	3282	3248	3164

Table 7: Pendulum: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	6	11	13	13	13	13
One-probe response	6	12	12	13	13	13
Two-direction +0.5	6	11	13	13	13	13
Two-direction +1.0	6	11	13	13	13	13
Direction-only RR	6	10	10	13	13	13
Magnitude response	6	11	12	13	13	13
MSE round-robin	6	9	10	12	13	13
MSE + diversity	7	9	10	13	13	13
MSE	7	9	10	13	13	13
Absolute response	6	10	11	12	13	13
Signed projection	6	10	11	12	12	13
Ensemble disagreement	4	4	4	5	6	13
Bank disagreement	7	7	9	12	13	13
Uniform	0	1	3	3	7	13

Table 8: MountainCarContinuous: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	23	25	26	28	28	29
One-probe response	23	24	26	26	26	29
Two-direction +0.5	23	25	26	28	28	29
Two-direction +1.0	23	25	26	28	28	29
Direction-only RR	23	25	25	26	29	29
Magnitude response	23	27	27	28	29	29
MSE round-robin	23	24	26	28	28	29
MSE + diversity	8	24	27	27	29	29
MSE	9	22	27	27	29	29
Absolute response	23	26	28	28	28	29
Signed projection	23	26	26	26	27	29
Ensemble disagreement	2	2	2	2	9	29
Bank disagreement	6	17	25	27	29	29
Uniform	1	2	3	8	14	29

Table 9: Swimmer: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	1328	1750	1795	1802	1809	1815
One-probe response	1328	1689	1737	1775	1806	1815
Two-direction +0.5	1328	1761	1799	1803	1811	1815
Two-direction +1.0	1328	1766	1797	1803	1811	1815
Direction-only RR	1328	1711	1749	1763	1803	1815
Magnitude response	1328	1535	1640	1749	1796	1815
MSE round-robin	1328	1493	1564	1739	1789	1815
MSE + diversity	1154	1465	1632	1784	1813	1815
MSE	1003	1347	1618	1782	1813	1815
Absolute response	1328	1509	1585	1748	1799	1815
Signed projection	1328	1490	1556	1664	1731	1815
Ensemble disagreement	958	1124	1239	1278	1482	1815
Bank disagreement	1306	1564	1668	1770	1813	1815
Uniform	266	457	621	1023	1487	1815

Table 10: Hopper: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	0	0	2	2	2	6
One-probe response	0	2	2	2	2	6
Two-direction +0.5	0	0	2	2	2	6
Two-direction +1.0	0	0	2	2	2	6
Direction-only RR	0	0	0	0	3	6
Magnitude response	0	2	3	4	5	6
MSE round-robin	0	1	1	5	6	6
MSE + diversity	0	0	0	5	6	6
MSE	0	0	1	5	6	6
Absolute response	0	1	1	5	5	6
Signed projection	0	2	2	3	5	6
Ensemble disagreement	1	1	1	1	2	6
Bank disagreement	0	1	3	5	6	6
Uniform	0	0	0	3	3	6

Table 11: Walker2d: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	298	2036	2231	2364	2487	2582
One-probe response	298	1699	1931	2201	2480	2582
Two-direction +0.5	298	2035	2229	2374	2492	2582
Two-direction +1.0	298	2037	2225	2386	2495	2582
Direction-only RR	298	2021	2161	2356	2482	2582
Magnitude response	298	1211	1539	1919	2332	2582
MSE round-robin	298	1431	1882	2348	2437	2582
MSE + diversity	720	1796	2072	2324	2444	2582
MSE	1252	1821	2100	2317	2444	2582
Absolute response	298	1567	1874	2342	2433	2582
Signed projection	298	1542	1812	2062	2461	2582
Ensemble disagreement	1363	2115	2136	2199	2403	2582
Bank disagreement	1219	1828	2226	2327	2475	2582
Uniform	1070	1729	2054	2289	2449	2582

Table 12: HalfCheetah: all method/budget counts.

Method	1	2	3	6	12	27
Two-direction response	1509	2547	3219	4225	4743	4765
One-probe response	1509	2519	3140	4145	4743	4765
Two-direction +0.5	1509	2524	3221	4232	4751	4765
Two-direction +1.0	1509	2488	3211	4263	4748	4765
Direction-only RR	1509	2460	3141	4172	4733	4765
Magnitude response	1509	1867	2293	3619	4664	4765
MSE round-robin	1509	2432	3086	4084	4710	4765
MSE + diversity	1149	1784	2363	3740	4712	4765
MSE	936	1566	2192	3710	4712	4765
Absolute response	1509	2223	2888	3970	4687	4765
Signed projection	1509	2403	2975	4006	4726	4765
Ensemble disagreement	1065	1523	2010	3592	4628	4765
Bank disagreement	1231	2133	2759	3905	4709	4765
Uniform	974	1831	2506	3704	4635	4765

The new full oracle used 79,993,334 actual nonabsorbed transitions and 84,584,932 conservative allocated GT opportunities. Operational query charges above are distinct from this research execution. The trained banks and their 442,368 data transitions are reused from V17.

H Supplemental model selection

The mature-bank comparison records 739 violations under consequence selection and 912 under MSE selection on 35,535 eligible sources. Equal-cell risk reduction is 0.469 pp (99.9609375% CI [0.165, 0.768]), below the prespecified 0.5 pp minimum. This experiment uses audit risk as its endpoint.

Table 13: Mature-checkpoint choices and fresh violations. M, R and E denote MLP, residual MLP and ensemble; digits denote training seeds. All checkpoints use 4,096 updates.

Task	Policy	MSE	Cons.	Eligible	MSE fail	Cons. fail
Pendulum	PPO	E2	E2	2048	0	0
Pendulum	SAC	E2	E2	2048	0	0
Pendulum	TD3	E2	E1	2048	1	1
MountainCar	PPO	E1	E1	2048	0	0
MountainCar	SAC	E1	E2	1511	0	0
MountainCar	TD3	E1	E1	1257	0	0
Swimmer	PPO	R0	M2	2048	21	2
Swimmer	SAC	R0	R0	2048	1	1
Swimmer	TD3	R2	R2	2048	0	0
Hopper	PPO	R0	R0	2048	0	0
Hopper	SAC	R0	R0	2048	0	0
Hopper	TD3	M2	M2	2048	0	0
Walker2d	PPO	R2	R2	2048	0	0
Walker2d	SAC	R0	M2	2047	25	12
Walker2d	TD3	R0	R2	2048	1	0
HalfCheetah	PPO	R0	R1	2048	175	123
HalfCheetah	SAC	R2	R1	2048	151	110
HalfCheetah	TD3	R0	R1	2048	537	490

The full-bank descriptive comparison has risk reduction +0.510 pp, the same-level interval [+0.206, +0.817] pp. Complete model identities and hashes, timing, per-campaign outcomes and conservative acquisition costs are retained in the supplemental protocol and results.

I Repair and new-task details

I.1 Discrete response-GP control

The GP is a standard Gaussian regression control on the same finite candidate bank. Its prior conditions a unit-amplitude squared-exponential kernel on the exact zero-origin response. With $\ell = 0.5$, noise variance $\sigma^2 = 10^{-4}$, and observed normalized responses y_O, define

$$k_0(v, w) = \exp\left(-\frac{\|v - w\|^2}{2\ell^2}\right) - \exp\left(-\frac{\|v\|^2 + \|w\|^2}{2\ell^2}\right), \tag{11}$$

$$\mu(v) = k_{vO}(K_{OO} + \sigma^2 I)^{-1} y_O, \tag{12}$$

$$v_*(v) = k_0(v, v) - k_{vO}(K_{OO} + \sigma^2 I)^{-1} k_{Ov},$$

$$q^{\mathrm{GP}}(v) = \Phi\left(\frac{\mu(v) - 1}{\sqrt{\max(v_*(v), 10^{-12})}}\right). \tag{13}$$

Here v, w are the same source-normalized residual features ϕ used by ridge, and Φ is the standard normal CDF. The nugget is a fixed modeling/regularization choice despite deterministic simulator responses. The posterior is used as an acquisition score without probability calibration, including near termination. Positive labels retire a source immediately; its response is unnecessary for subsequent discovery at that source. Acceptable responses update the GP by Gaussian conditioning. Both probe passes, order, geometry and tie permutation are shared with ridge. The remaining candidates are prioritized globally within the campaign.

The implementation uses sequential rank-one conditioning; a separate direct matrix-solve implementation reconstructs every new-task cap-3 query prefix. Synthetic checks also test full coverage,

retirement and invariance to unqueried responses. Historical lengthscale selection includes all six old tasks and three candidates; the chosen kernel is not revised on V20. Its confidence intervals are descriptive secondary contrasts; the stated sequential allocation is reserved for the two primary baselines.

I.2 Repair intervention and paired risk

For $M = 9$ mature initial models and strategy a in fixed cell c, define

$$\widehat{R}_c^{(a)} = \frac{\sum_{s \in \mathcal{S}_{E,c}} \sum_{m=1}^{M} Y_{sm}^{(a)}}{M|\mathcal{S}_{E,c}|}, \qquad \Delta_{\text{repair}}^{a,b} = \frac{1}{C} \sum_{c} (\widehat{R}_c^{(b)} - \widehat{R}_c^{(a)}). \tag{14}$$

An empty eligible cell contributes zero contrast and remains reported. Eligibility and reference continuations are common across strategies. A fix changes a positive unchanged-model label to zero; a regression changes zero to one. Therefore the unchanged-to-repaired failure-count reduction equals fixes minus regressions exactly, including analytic invalidity and unhealthy termination. The raw audit independently checks this identity for every campaign and strategy.

The source-level witness mask is shared across the nine models: it means at least one native mature prediction at that source was verified harmful under the stated budget. Each model receives the same source-level witness weights. Uniform weights are one; MSE-priority weights ten for the top 10% of eligible source MSE per model. The two witness strategies instead give weight ten to their actually discovered sources. Minibatch sampling uses the same random draws and initial seed across strategies, so identical masks yield identical repaired weights. We verify this property for zero-witness conditions. The original Gaussian objective and member bootstrap are retained for ensembles. No validation-based early stopping or repair selection is performed.

I.3 Newly trained policies and known context

Each new PPO policy uses two 64-unit Tanh hidden layers for actor and value networks, eight vectorized environments, rollout length 256, minibatch 256, ten epochs per update, learning rate $3 \cdot 10^{-4}$, discount 0.99, GAE 0.95, clip 0.2, value coefficient 0.5, zero entropy coefficient and gradient clipping 0.5. Observation/reward normalization is used during training; inference freezes observation statistics and clips normalized observations at 10. Policies are saved at 1,048,576 transitions at all three seeds. A separate fixed 32-reset check uses ordinary episode limits (Reacher 50, otherwise 1000), with no filtering or selection by return.

The physical state/action dimensions are Reacher 8/2, InvertedPendulum 4/1 and InvertedDoublePendulum 6/1. Reacher inputs use sine/cosine of arm angles, the remaining physical state and action. Double-pendulum inputs use sine/cosine of both angles, cart position, velocities and action. Inverted-pendulum inputs use physical state and action. Reacher goal positions and velocities are copied from the input as known invariant context. The corresponding validation-error and ensemble-disagreement contributions are zero under this same wrapper. Native successor residuals and all acquisitions use that evaluated output. Data collection resets Reacher at 50 steps; confirmation deliberately uses an 80-step physical continuation. No learned immediate-reward model is evaluated.

Pinned Gym source/XML hashes are checked against installed files. Native physics, rewards, observations and termination are compared with official environments; actor outputs and snapshot interventions are checked separately. Every test follows a repeat/cross-batch zero-residual pilot. The reproducibility records include protocol and dataset hashes, independent CPU prediction comparisons and engineering checks.

I.4 Repair outcomes

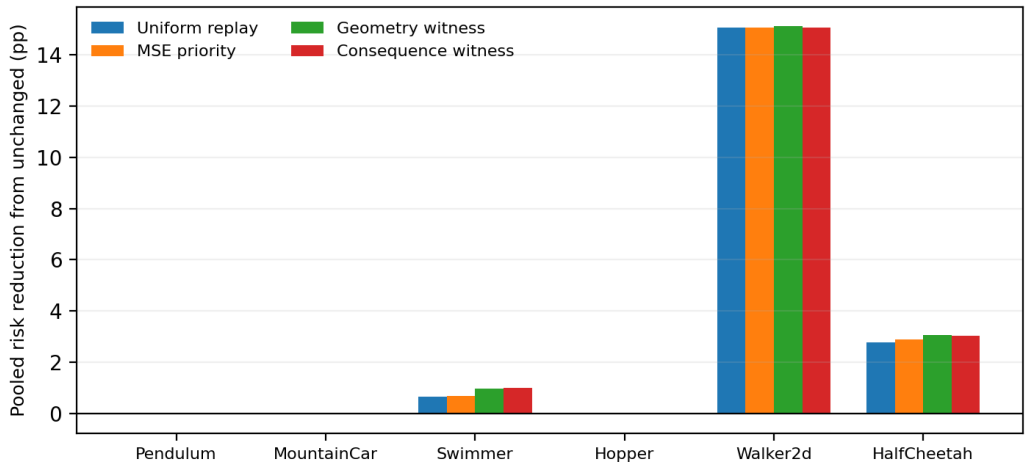


Figure 6: Task-level reduction in pooled violation risk relative to the unchanged model (pp). Positive values indicate fewer violations after repair.

Table 14: Repair totals across eligible pairs. Risk is the equal-cell mean (%).

Strategy	Failures	Fixes	Regressions	Risk
Unchanged	17507	0	0	5.2783
Uniform	7257	14471	4221	2.1887
MSE priority	7194	14619	4306	2.1701
Geometry witness	6912	14653	4058	2.0846
Consequence witness	6955	14667	4115	2.0979

Table 15: Eligible-pair failures after repair; identical denominator across strategies within each task.

Task	Pairs	Before	Uniform	MSE	Geom.	Conseq.
Pendulum	55296	6	3	3	1	1
MountainCar	43038	15	14	19	14	17
Swimmer	55296	622	256	251	80	76
Hopper	55296	0	0	1	0	0
Walker2d	55296	8584	245	243	225	254
HalfCheetah	55296	8280	6739	6677	6592	6607

Table 16: Consequence repair versus Uniform. Positive reduction favors consequence weighting.

Task	Policy	Pairs	Conseq.	Other	Red. pp
Pendulum	PPO	18432	1	1	+0.000
Pendulum	SAC	18432	0	0	+0.000
Pendulum	TD3	18432	0	2	+0.011
MountainCar	PPO	18432	3	2	-0.005
MountainCar	SAC	13464	14	12	-0.015
MountainCar	TD3	11142	0	0	+0.000
Swimmer	PPO	18432	32	43	+0.060
Swimmer	SAC	18432	8	69	+0.331
Swimmer	TD3	18432	36	144	+0.586
Hopper	PPO	18432	0	0	+0.000
Hopper	SAC	18432	0	0	+0.000
Hopper	TD3	18432	0	0	+0.000
Walker2d	PPO	18432	15	20	+0.027
Walker2d	SAC	18432	233	218	-0.081
Walker2d	TD3	18432	6	7	+0.005
HalfCheetah	PPO	18432	1322	1347	+0.136
HalfCheetah	SAC	18432	1233	1190	-0.233
HalfCheetah	TD3	18432	4052	4202	+0.814

Table 17: Consequence repair versus MSE priority. Positive reduction favors consequence weighting.

Task	Policy	Pairs	Conseq.	Other	Red. pp
Pendulum	PPO	18432	1	2	+0.005
Pendulum	SAC	18432	0	0	+0.000
Pendulum	TD3	18432	0	1	+0.005
MountainCar	PPO	18432	3	3	+0.000
MountainCar	SAC	13464	14	16	+0.015
MountainCar	TD3	11142	0	0	+0.000
Swimmer	PPO	18432	32	92	+0.326
Swimmer	SAC	18432	8	49	+0.222
Swimmer	TD3	18432	36	110	+0.401
Hopper	PPO	18432	0	0	+0.000
Hopper	SAC	18432	0	1	+0.005
Hopper	TD3	18432	0	0	+0.000
Walker2d	PPO	18432	15	29	+0.076
Walker2d	SAC	18432	233	212	-0.114
Walker2d	TD3	18432	6	2	-0.022
HalfCheetah	PPO	18432	1322	1277	-0.244
HalfCheetah	SAC	18432	1233	1243	+0.054
HalfCheetah	TD3	18432	4052	4157	+0.570

Table 18: Consequence repair versus Geometry witness. Positive reduction favors consequence weighting.

Task	Policy	Pairs	Conseq.	Other	Red. pp
Pendulum	PPO	18432	1	1	+0.000
Pendulum	SAC	18432	0	0	+0.000
Pendulum	TD3	18432	0	0	+0.000
MountainCar	PPO	18432	3	3	+0.000
MountainCar	SAC	13464	14	11	-0.022
MountainCar	TD3	11142	0	0	+0.000
Swimmer	PPO	18432	32	34	+0.011
Swimmer	SAC	18432	8	8	+0.000
Swimmer	TD3	18432	36	38	+0.011
Hopper	PPO	18432	0	0	+0.000
Hopper	SAC	18432	0	0	+0.000
Hopper	TD3	18432	0	0	+0.000
Walker2d	PPO	18432	15	16	+0.005
Walker2d	SAC	18432	233	201	-0.174
Walker2d	TD3	18432	6	8	+0.011
HalfCheetah	PPO	18432	1322	1327	+0.027
HalfCheetah	SAC	18432	1233	1189	-0.239
HalfCheetah	TD3	18432	4052	4076	+0.130

Witness acquisition costs 13,487,092 reference transitions, compared with 4,958,692 for common references alone. The matched repair updates therefore use different information budgets.

I.5 Post-hoc regression analysis

Table 1 reports pair-level counts for all strategies. The following tables give context-level changes and target/tolerance sensitivity; Figures 2 and 3 show the individual case.

Table 19: Post-hoc update review. The last column counts contexts with lower mean MSE and more violations; contexts are correlated.

Strategy	Contexts	MSE improved	Risk also worse
Uniform	162	156	18
MSE priority	162	155	20
Geometry witness	162	155	12
Consequence witness	162	156	15

Table 20: Post-hoc target/tolerance sensitivity: equal-cell violation risk (%). Multipliers rescale the original tolerance at fixed reference eligibility.

Target	Multiplier	Before	Uniform	MSE	Geom.	Conseq.
one-sided	0.5	10.002	5.520	5.557	5.151	5.157
one-sided	1.0	5.278	2.189	2.170	2.085	2.098
one-sided	2.0	2.532	0.521	0.509	0.494	0.474
two-sided	0.5	15.721	10.473	10.633	10.541	10.528
two-sided	1.0	7.973	4.335	4.353	4.292	4.350
two-sided	2.0	3.240	1.119	1.118	1.090	1.106

The sensitivity grid was specified after all V19 primary outcomes: one-sided versus absolute return difference, crossed with tolerance multipliers 0.5, 1 and 2, at fixed reference eligibility. Analytic invalidity and unhealthy termination remain violations. These are descriptive counts without new tests or tuning. The absolute-tolerance subset shows that the regression observation is not solely an artifact of accepting large fictitious improvements before repair.

The illustrative HalfCheetah case is source 1905000002, the first qualifying source/model in the ID-ordered list. It was selected after the repair analysis; the case and coordinate interventions were

fixed before their intervention outcomes. A coordinate correction replaces one component with its reference value while holding the other components at the repaired prediction. Loss increases use a threshold of 10^{-8}.

Native-before/native-after/full-GT controls agree with the saved audit; the full-GT control has zero loss within 10^{-7}. Additional allocated/actual GT: 1,680/1,680.

I.6 New-task discovery

Table 21: New-task failure-source counts at budget $b = 3$, aggregated over three PPO seeds per task.

Task	Elig.	Oracle	Two-dir.	MSE RR	Div.	GP
Reacher	6144	36	36	33	28	33
Inv. pendulum	6144	0	0	0	0	0
Double pendulum	5279	1042	302	278	175	341

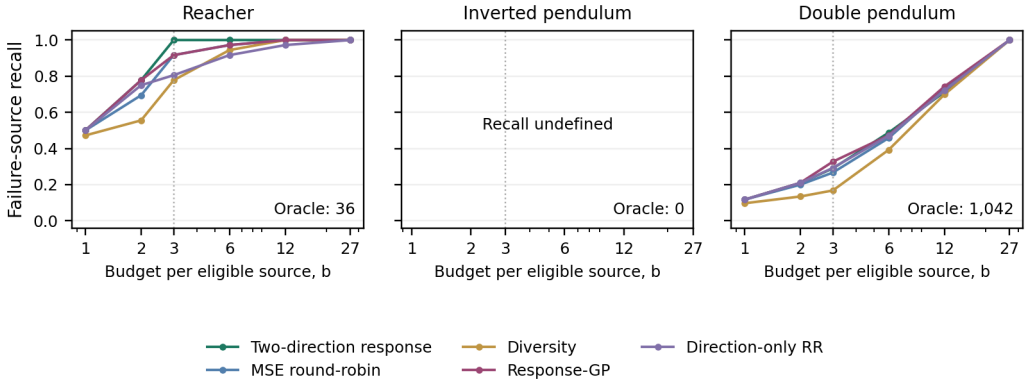


Figure 7: New-task failure-source recall versus campaign budget per eligible source. The dotted line marks $b = 3$. Oracle counts give the denominator; recall is undefined for InvertedPendulum.

Against diversity, the new-task macro recall gain is 14.692 pp and the per-source discovery gain is 0.732 pp (CI [0.383, 1.134]). The interval allocation is specified in Appendix A.

Table 22: New-task acquisitions at $b = 3$. All methods share 1,078 oracle failure sources, 52,701 queries and 6,699,833 charged reference transitions. Ablations are descriptive.

Method	Found
Magnitude response	343
Absolute response	317
Signed projection	371
MSE RR	311
MSE	178
Diversity	203
Ensemble disagreement	194
Bank disagreement	162
Uniform	264
One-probe response	382
Two-direction response	338
Two-direction +0.5	355
Two-direction +1.0	355
Direction-only RR	334
Response-GP	374

Table 23: New-task failure-source counts by PPO seed at $b = 3$.

Task	Seed	Elig.	Oracle	Two-dir.	MSE RR	Div.	GP
Reacher	0	2048	20	20	19	17	17
Reacher	1	2048	7	7	5	4	7
Reacher	2	2048	9	9	9	7	9
Inv. pendulum	0	2048	0	0	0	0	0
Inv. pendulum	1	2048	0	0	0	0	0
Inv. pendulum	2	2048	0	0	0	0	0
Double pendulum	0	1648	328	116	102	41	127
Double pendulum	1	1845	617	145	136	117	151
Double pendulum	2	1786	97	41	40	17	63

Table 24: Native-policy return and survival over 32 fixed resets per seed, using ordinary episode limits.

Task	Seed	Horizon	Mean return	SD	Survived
Reacher	0	50	-4.18	1.64	32
Reacher	1	50	-4.54	1.25	32
Reacher	2	50	-4.19	1.74	32
Inv. pendulum	0	1000	1000.00	0.00	32
Inv. pendulum	1	1000	1000.00	0.00	32
Inv. pendulum	2	1000	1000.00	0.00	32
Double pendulum	0	1000	7915.12	3347.55	27
Double pendulum	1	1000	7567.95	3388.92	25
Double pendulum	2	1000	7332.17	3829.06	25

Table 25: Discovery within nine-candidate architecture and mature-checkpoint subsets at $b = 3$.

Task	Subset	Oracle	Two-dir.	MSE RR	Div.	GP
Reacher	mlp	20	19	19	19	18
Reacher	residual	32	31	31	30	29
Reacher	ensemble	9	9	7	9	8
Reacher	mature	15	13	14	14	14
Inv. pendulum	mlp	0	0	0	0	0
Inv. pendulum	residual	0	0	0	0	0
Inv. pendulum	ensemble	0	0	0	0	0
Inv. pendulum	mature	0	0	0	0	0
Double pendulum	mlp	535	270	248	182	262
Double pendulum	residual	606	329	317	309	346
Double pendulum	ensemble	438	234	213	209	210
Double pendulum	mature	502	286	243	213	251

I.7 Execution costs and callback checks

V19 fresh test charges 137,676,606 GT, with 130,233,607 actual nonabsorbed transitions. V20 fresh test charges 42,296,873, with 40,694,823 actual transitions. New PPO training uses 9,437,184 transitions and dynamics collection 221,184. These are separate from operational discovery replay. Pilot, failed-attempt and engineering counts are preserved separately. A repair-study verification-path correction preceded all repaired checkpoints and fresh outcomes; the scientific protocol was unchanged. Its execution record is archived with the reproducibility materials.

Table 26: Live physics callback verification on the first 128 policy-seed-0 sources/task. Timing covers the continuation-query phase from cached, already acquired features; one engineering repetition, not an end-to-end speed benchmark.

Task	Method	Queries	Found	Seconds
Reacher	Two-direction	384	2	17.91
Reacher	MSE RR	384	2	16.65
Reacher	Diversity	384	2	16.62
Reacher	GP	384	2	16.81
Inv. pendulum	MSE RR	384	0	42.78
Inv. pendulum	Diversity	384	0	39.05
Inv. pendulum	GP	384	0	34.91
Inv. pendulum	Two-direction	384	0	35.30
Double pendulum	Diversity	324	3	14.31
Double pendulum	GP	324	7	13.20
Double pendulum	Two-direction	324	4	13.35
Double pendulum	MSE RR	324	4	14.00

The engineering campaign was fixed before V20 test outcomes. Every requested continuation was simulated when requested; saved candidate labels were loaded only after all four acquisitions finished. All 12 query prefixes and queried labels match the exhaustive saved-data oracle, with signed-loss differences below 10^{-7}. Additional allocated/actual query GT is 349,440/348,378. This verifies an executable oracle callback, not improved end-to-end throughput; source/model/reference acquisition and loading are excluded from timing. Hardware and other-user contention prevent interpreting cross-server differences as method effects.

J Repeated-use model selection and runtime measurements

Study 4 uses the previously specified models, policies and H80 audit target, with model choices fixed before held-out evaluation. The delay-compensation law is

$$\hat{x}_t = \hat{f}_m(x_{t-1}, a_{t-1}), \quad a_t = \pi(\mathcal{O}(\hat{x}_t; \eta_{t-1}), t), \quad x_{t+1} = f(x_t, a_t). \tag{15}$$

At $t = 0$ the policy receives a fresh observation; the equations apply for $t \geq 1$. The current true state is delivered only at the next decision. Unlike single insertion, only the policy input is estimated while the plant state is preserved; this tests decision transfer of the existing metric to a different use condition. Auxiliary integration state comes from the delayed source; native forward calculations reconstruct derived observations. Invalid numeric outputs trigger zero-action fallback. Validation conditions on reference survival, whereas testing retains all fresh resets and terminated episodes with absorbing zero rewards. The deployment horizon H_c^{dep} is 200, except Reacher 50 and MountainCar-Continuous 1000; the validation audit uses $H = 80$.

Risk selection uses a balanced complete audit of nine models at 32 sources, breaking ties by normalized MSE then model index. With no eligible sources it falls back to all-source MSE. Full-budget MSE redirects continuation cost to 801-step reference-controller trajectories after the same warmup. Direct-return selection gives each model floor(26,656/9) opportunities with paired resets and reports 7 unused opportunities. Absorbed opportunities are charged; actual GT counts are separate. Existing training is a reused setup cost.

Return contrast. Section 4 defines the episode-wise normalized return contrast used below. The exact bootstrap quantiles are in Appendix A.

Table 27: Task-wise return contrasts for consequence-based selection.

Task	MSE-budget gain (pp)	Direct-return gain (pp)	Different choices
Pendulum	-0.004	-0.012	13/18
MountainCarContinuous	-0.005	+0.005	12/18
Swimmer	-0.086	-0.562	15/18
Hopper	+1.791	-0.669	16/18
Walker2d	+0.791	-4.409	16/18
HalfCheetah	-9.712	-10.392	13/18
Reacher	+0.163	-0.009	18/18
InvertedPendulum	+0.000	+0.000	6/18
InvertedDoublePendulum	-2.936	-4.634	11/18

The last column counts different consequence/full-budget-MSE choices among 18 campaigns. Gains average the fixed controller contexts. The primary resamples six paired campaigns within each context for 200,000 draws; the extreme-tail endpoint has about ten draws of resolution.

Released latent model. We evaluate the three released TD-MPC2 checkpoints `cartpole-swingup-1.pt` through `cartpole-swingup-3.pt` on DM Control cartpole-swingup. They share an architecture and use 64 planning particles, 8 elites, 8 policy trajectories, 2 iterations and planning horizon 3. For checkpoint m, encoder e_m, latent dynamics d_m, observation o_t and executed action a_t, define

$$z^+_{\text{enc},m} = e_m(o_{t+1}), \qquad \hat{z}^+_m = d_m(e_m(o_t), a_t). \tag{16}$$

At an anchor, both branches restore the same physical state, planner state and random-number state. The reference branch initializes its first planning call with $z^+_{\text{enc},m}$; the intervention branch uses $\hat{z}^+_m$. Physical state is preserved, and subsequent planning calls in both branches use their current observations normally. We compare 30-step continuation returns with loss tolerance 3. Validation uses 6 groups of 4 nominal episodes per checkpoint, with anchors 25 and 100 in 120-step trajectories. Choices are fixed before 16 fresh test episodes per group and checkpoint (96 total), each of length 200. All lengths count control decisions, each with action repeat 2; rewards sum both environment steps. All selectors use the same validation data and are assigned its full simulation cost. Latent MSE depends on each checkpoint's coordinates.

Table 28: Held-out MPC return contrasts for TD-MPC2 checkpoint selection.

Comparator	MPC return gain	95% CI
latent_mse	+0.0000	[+0.0000, +0.0000]
validation_return	+0.0000	[+0.0000, +0.0000]

All three selection rules choose checkpoint 2 in all six groups. Test return means for checkpoints 1/2/3 are 128.868, 283.400, 282.330 (96 episodes each). There is 1 native validation violation in 144 pairs, confined to checkpoint 1. Identical choices give zero incremental selection benefit. Checkpoint selection chooses the whole planner bundle, including encoder, dynamics, reward, value and policy; each checkpoint generates its own nominal trajectories, so source distributions also differ. The comparison is therefore between complete planner bundles.

Full-pipeline timing. Four unchanged acquisitions run for three repetitions per task on the same H100 GPU 7, using policy 0, the mature nine-model bank, 32 sources and cap 3. Each method reloads models/policy and reacquires sources, predictions, H80 references and requested GT callbacks; order rotates. These timers exclude process launch, import/guard and local backup. Separate backup measurements cover NPZ publishing only, not total job turnaround. OS/runtime caches may be warm and GPU contention is not fully controlled. Retirement changes query counts, so this is not a compute-speed comparison at identical work counts. Below are median seconds; full ranges and query counts are released.

Table 29: Median acquisition-pipeline runtime in seconds.

Task	Two-dir	MSE RR	Diversity	GP
Pendulum	2.927	3.019	3.081	3.008
MountainCarContinuous	3.275	3.557	3.378	3.402
Swimmer	3.663	3.729	3.720	3.717
Hopper	4.275	4.181	4.193	4.168
Walker2d	4.606	4.676	4.653	4.599
HalfCheetah	3.642	3.876	3.727	3.647
Reacher	3.519	3.553	3.562	3.584
InvertedPendulum	3.261	3.234	3.197	3.233
InvertedDoublePendulum	3.745	3.198	3.341	3.293

Recorded costs and checks. Across 162 validation campaigns, operational charges and actual nonabsorbed transitions are shown below. Shared warmup makes unique executed validation cost 11,535,267 GT, distinct from summing selector costs. All-candidate testing uses 4,875,614 and timing 1,197,186 actual GT. Prior training is reused.

Table 30: Simulation costs of model-selection validation.

Selector	Allocated GT	Actual GT
Consequence	4,318,272	4,116,340
Full-budget MSE	4,318,272	3,747,901
Direct return	4,317,138	3,834,902
Common-source MSE	585,792	564,360

Post-hoc choice diagnostics find 91/162 campaigns with zero violations throughout the bank and 144/162 with a minimum-risk tie. Consequence and common-source MSE choose the same model in 129/162 campaigns.

Prespecified first-source/first-episode native replay passes on all nine tasks: paired-continuation returns and native Gym states match exactly; maximum reward discrepancy is 3.6×10^{-15} (5,816 additional GT). Eighteen TD-MPC2 saved-action replays have identical observations and at most 6.0×10^{-8} reward discrepancy from float32 storage (7,200 GT). These checks do not reproduce original training. The integrated CPU packet reproduces four studies from recorded outcomes; full raw/runtime archives remain separate.