ADAPTATION CHANGES MEASURED DATA COMPARISONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing robot demonstration data requires comparing the policies learned from them, often after adaptation to a target task. A success difference measured with one adapter does not reveal whether that comparison persists with another: the adapter may change both overall performance and the effect attributed to the source data. We address this ambiguity with a paired four-cell evaluation that crosses two source-data conditions with two learning operators, holds the remaining experimental choices fixed, and reports both absolute success and the change in the data contrast. In Meta-World, fixing Gaussian-mixture adaptation scales changes the perturbed-minus-clean effect by -5.99 percentage points (95% interval $[-8.94, -3.26]$), while improving clean-data success by 23.91 points. Coupled channel interventions distinguish reductions in the data contrast accompanied by degraded versus improved performance. A separate PyBullet direct-training comparison reduces the noisy-data penalty by 12.19 points ($[6.64, 18.59]$) under Huber rather than MSE loss. A deterministic-policy control using the same matched acquisition groups yields an interaction of -0.27 points ($[-0.96, 0.36]$). These results demonstrate adaptation-sensitive data comparisons and provide a protocol for checking a measured data advantage against the learning procedure intended to use it.

1 INTRODUCTION

Choosing which demonstrations to collect or reuse is a consequential decision in robot learning. Robot datasets now span diverse tasks and embodiments, and pretrained policies can serve as initializations for subsequent finetuning (Brohan et al., 2022; Vuong et al., 2023; Team et al., 2024). Studies of dataset composition, scaling and diversity accordingly ask which experience improves downstream manipulation (Saxena et al., 2025; Lin et al., 2025; Shi et al., 2026). In transfer settings, answering this question involves a source policy, a target-adaptation procedure and a rollout evaluation. We study a specific decision within this process: does a measured advantage of one source dataset over another persist when the target-adaptation rule changes?

Existing valuation and attribution frameworks already define data contributions relative to a learning task and a performance measure (Ghorbani & Zou, 2019; Ghorbani et al., 2020; Kwon & Zou, 2021). Temporal influence further tracks how training-data effects depend on the learning trajectory (Wang et al., 2025). These formulations motivate an empirical question about downstream reuse. Observing that one dataset produces higher final success under a given adapter does not distinguish a favorable source-data effect from an interaction with that adapter. To make that distinction, the same source-data comparison must be evaluated under a controlled change in subsequent learning.

A first challenge is to separate the data comparison from the quality of adaptation itself. Changing an adapter can alter the optimization of both source policies, even though their checkpoints and target demonstrations are unchanged. Learned-variance objectives illustrate the issue: their optimization can impair mean prediction (Seitzer et al., 2022; Stirn et al., 2023). If repairing that objective also removes a noisy-data advantage, the final success gap alone cannot explain whether the original data choice remains useful. This matters for reuse because a source dataset chosen with one adaptation rule may later be used with a different rule.

A second challenge is to interpret a change in the gap. Two datasets can become less distinguishable because an adapter improves the weaker branch, or because both policies become poor. Closed-

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.

loop success also measures more than action-prediction accuracy (Ross & Bagnell, 2010; Ross et al., 2011), and finite-run variability can obscure small contrasts (Henderson et al., 2018; Agarwal et al., 2021). Measuring the gap alone therefore misses the direction of the underlying performance changes.

A third challenge is to understand when an operator change should alter the comparison. A large improvement in adaptation need not produce a large interaction if it benefits both source policies equally. Conversely, source policies with similar initial action errors can respond differently to the same target updates. A useful diagnostic must therefore characterize their differing responses, rather than the adapter's average gain. We derive a local cancellation condition for a differentiable validation score and test whether its first-update alignment predicts the eventual success interaction. Selection remains a separate decision: detecting an interaction does not establish that a data selector improves the next policy.

We address these challenges with the paired four-cell evaluation in Figure 1. Two source-data conditions are each evaluated under two adapters. Within an experiment, source checkpoints, target demonstrations, update budgets, minibatch streams and evaluation scenes are paired across adaptation arms. We report all four absolute success rates, the source-data contrast within each adapter, and the change between those contrasts. Coupled interventions cross the responsibility and component-scale rules, testing whether a smaller contrast accompanies improved or degraded performance without assuming independent gradient channels. Parameter studies and paired simulator trajectories connect the aggregate comparison to individual outcomes.

The motivating Gaussian-mixture experiment gives a concrete result: fixing adaptation scales changes the perturbed-data advantage from $+5.33$ to -0.66 percentage points while raising clean-data success by 23.91 points. The fixed-scale contrast spans zero, so the finding is a change in the measured comparison rather than an established reversal of the population ranking. A separate PyBullet experiment finds that replacing MSE with Huber loss reduces the noisy-data penalty while both losses favor clean labels. This direct-training comparison tests loss choice; the deterministic-policy control uses the same matched acquisition groups and yields an interaction of -0.27 points ($[-0.96, 0.36]$).

Our contribution is an empirical demonstration that target adaptation can change a measured robot-data comparison, together with controls that make that change interpretable. The resulting protocol tests a data advantage under the learning procedure intended to use it; prospective selection is evaluated separately against its acquisition controls.

2 Related work

Valuation and selection. Shapley approximations, influence functions and group-effect analyses estimate contributions to a specified predictor (Jia et al., 2019; Koh & Liang, 2017; Koh et al., 2019). TracIn, datamodels and TRAK use checkpoints, subset models or scalable attribution (Pruthi et al., 2020; Ilyas et al., 2022; Park et al., 2023); memorization analyses connect examples to test performance (Feldman & Zhang, 2020). Forgetting, early-training scores and pruning study removable data (Toneva et al., 2018; Paul et al., 2021; Sorscher et al., 2022). Reweighting, adaptive valuation and curricula optimize example use (Ren et al., 2018; Shu et al., 2019; Yoon et al., 2020; **?**); group DRO and DoReMi optimize group performance or mixtures (Sagawa et al., 2019; Xie et al., 2023). We directly compare paired outcomes under two operators.

Robot data and learning operators. DataMIL, Re-Mix and mutual-information curation target robot-data selection (Dass et al., 2026; **?**; Hejna et al., 2025). Our Re-Mix-weighted acquisition probe and prospective tests assess selection separately from adapter sensitivity. Diffusion Policy and ACT use different action representations (Chi et al., 2025; Zhao et al., 2023), while MAML learns initializations for adaptation (Finn et al., 2017). Robust regression motivates alternatives to squared residuals (Barron, 2019); control benchmarks distinguish training performance from generalization (Packer et al., 2018). Our comparison fixes the data and evaluation conditions while intervening on a specified learning operator.

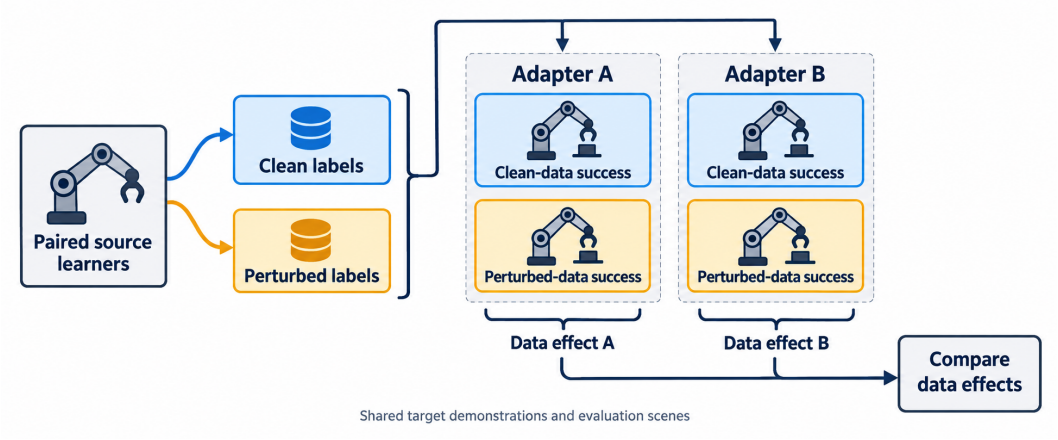

Figure 1: A data comparison is evaluated under two learning operators. Each column gives a within-operator data contrast D_a; their difference is the interaction I. Initial learners, update budgets and evaluation scenes are paired within the corresponding experiment. All four absolute success rates are needed to distinguish changing data preference from a floor or ceiling effect. This conceptual schematic contains no measurements.

3 Paired Data Comparisons

Let θ_G be the source policy after acquiring candidate data G for a fixed update budget. Within each learner, acquisition branches start from the same parent parameters and optimizer state; candidates are withheld from parent training and normalization. A target adapter F_a then uses demonstrations D_T. Write $S_a(G) = \mathbb{E}[C_T(F_a(\theta_G; D_T, \zeta))]$ for mean target success. Target data, update budgets, minibatch streams and evaluation scenes are paired. For two source-data conditions G_1, G_0 and two adapters a_1, a_0, report

$$D_a = S_a(G_1) - S_a(G_0), \qquad I = D_{a_1} - D_{a_0}. \tag{1}$$

The four absolute success rates show whether either adapter is competent; D_a gives the data preference within an adapter; I tests whether its magnitude changes. An interaction does not require a sign reversal. A sign reversal of point estimates does not establish opposite population rankings unless uncertainty supports both signs. For experiments that vary the training loss directly, that loss is the operator and no separate target adaptation is applied.

Reading the four outcomes. We test the measured preference under a specified change in adaptation, using pairing to hold the other experimental choices fixed. An intervention that repairs a poor adapter can remove an apparent data advantage, whereas an intervention that degrades both branches can compress their difference. Reporting I without the four success rates cannot distinguish these cases. The prospective selection experiment provides a separate check of whether choosing data improves the intended downstream comparison.

When should a comparison change? For a differentiable validation score q and a one-step adapter $F_a^\eta(\theta) = \theta + \eta v_a(\theta)$, define $h(\theta) = \nabla q(\theta)^\top [v_1(\theta) - v_0(\theta)]$. If ∇q is L-Lipschitz along the four update segments, the corresponding interaction satisfies

$$I_q = \eta[h(\theta_{G_1}) - h(\theta_{G_0})] + R, \qquad |R| \le \frac{L\eta^2}{2} \sum_{a,g} \|v_a(\theta_{G_g})\|_2^2. \tag{2}$$

Thus, the first-order interaction depends on how differently the operator change aligns with validation improvement at the two source checkpoints. Equal aligned changes cancel to first order, even when the adapter improves both policies. The four-term Taylor argument appears in Appendix O. It concerns a smooth score and one update; closed-loop success after multiple updates is evaluated separately.

4 EXPERIMENTAL SETUP

Tasks and policies. We use two disjoint, custom eight-task subsets of Meta-World (Yu et al., 2020), with 80 training and 20 validation scripted-expert trajectories per task, retaining demonstration failures. They are not official MT10 splits. The hard source suite comprises assembly, basketball, shelf-place, soccer, sweep-into, peg-insert-side, hammer, and handle-pull. The original future targets are reach, push, pick-place, and window-open. A separate extension adds door-open, drawer-open, drawer-close, and button-press; all eight targets are absent from the hard source suite. Policies use three 256-unit ReLU layers, 39 normalized state/goal coordinates, and eight task indicators. We evaluate both mean-squared behavior cloning and a five-component Gaussian-mixture likelihood head. Appendix A specifies the frozen arrays and corrected evaluation procedure.

Paired acquisition and adaptation. Each task's first 20 training trajectories form the base pool. Source branches start after 250 or 1,000 updates and receive 100 continuation updates, with 192 base and 64 candidate transitions per batch. Matched source groups balance task identity and trajectory-level expert success, then cross initial-state distance with clean or perturbed action labels. The original perturbation uses a fixed Gaussian noise bank with standard deviation 0.15 followed by action clipping; the replication adds two independently drawn banks. A base-only branch continues on the original data. Each source policy adapts independently using the first ten target demonstrations, 250 updates, fresh Adam at 3×10^{-4}, batch size 256, source normalization, and a shared zero target-task code. We pair minibatch sequences across source arms within each optimization repeat.

Controls and inference. The scale study reuses all ten GMM source learners and compares learned likelihood scales with scales fixed to 0.15 during target adaptation, preserving initial executed actions. Its primary is the fixed-minus-learned change in perturbed-minus-clean success. This follow-up was designed after the secondary label result and frozen before its new rollout outcomes. Separate learner cohorts test prospective selection and matched properties; follow-ups reuse their cohorts. Success means reaching the task criterion at least once within 200 steps, evaluated on 20 shared initializations per target. Crossed 95% intervals jointly resample learners and target-specific shared scenes, keeping each study's four targets fixed. Repeats, groups, and episodes are nested measurements. The original selector primary retains its prespecified learner interval. The appendices retain cohort identities, original primary endpoints, resource accounting and secondary contrasts for the reported experiments. Earlier replay-dose and temporal diagnostics remain archived in the research artifact.

Deterministic matched-group control. A new control uses the same 32 matched candidate groups, perturbations, source and target arrays as the GMM study, with the ten existing deterministic learners. It compares MSE with twice Huber loss (threshold 0.1) under the same 250-update target adaptation. A full replay exactly reproduces the reused MSE adapter parameters before paired evaluation. The earlier pooled-acquisition study and Re-Mix-weighted acquisition probes are reported in Appendix M; their acquisition construction is distinct from this matched-group control.

Second physics engine. A separate experiment uses the native Reach and Push tasks in panda-gym with PyBullet (Gallouédec et al., 2021). Ten newly initialized deterministic policies per task compare identical demonstration states with clean or corrupted action labels under MSE and Huber training. Here the operator changes training directly, rather than future-task adaptation. Each task supplies 400 scripted-controller trajectories, including failures. Independently per learner, each label row is selected with probability 0.3 to receive Gaussian noise of standard deviation 0.7 followed by clipping. This stronger contamination probes a different regime from the Meta-World source perturbation. All branches use a two-layer 128-unit MLP, 6,000 Adam updates, and 64 shared test scenes per task. Success is reached at any time within 50 simulator steps. Development uses disjoint scenes and fixes these settings before the clean/noisy test comparisons. Appendix N specifies features, controllers, pairing and development outcomes.

Table 1: Future-task success (%) on the paired GMM adaptation study. All six conditions use the same ten learners and fresh scene panel, averaged over two source states and two optimization repeats. Clean and perturbed rows average all 16 matched source groups; base-only continues training without candidate acquisition. Fixing the adaptation scale improves all three source-data conditions. Bold marks the larger macro mean within each source-data pair, not a separate significance test.

Source data	Adaptation scale	Reach	Push	Pick-place	Window	Macro
Base only	Learned	59.88	0.75	4.62	65.50	32.69
Base only	Fixed (0.15)	85.00	18.38	35.25	87.75	**56.59**
Clean acquisition	Learned	59.65	0.91	3.74	66.75	32.76
Clean acquisition	Fixed (0.15)	85.61	18.48	34.42	88.19	**56.67**
Perturbed acquisition	Learned	68.11	3.00	11.35	69.89	38.09
Perturbed acquisition	Fixed (0.15)	82.75	19.58	32.57	89.16	**56.01**

5 EXPERIMENTS

5.1 MAIN RESULTS

Table 1 reports clean, perturbed and base-only success under the paired learned/fixed-scale adaptation protocol.

Under learned scales, perturbed acquisition attains higher macro success than clean acquisition. Fixing the scale raises clean-label success substantially. The paired perturbation-effect interaction is -5.99 pp, with crossed 95% interval $[-8.94, -3.26]$ (Table 21). The fixed-scale perturbation effect is -0.66 pp with an interval spanning zero. The evidence therefore establishes a change in the comparison, rather than a reliable advantage for perturbed labels under both adaptation rules. The per-target columns expose substantial remaining differences in difficulty: push and pick-place remain weaker than reach and window-open.

5.2 ABLATION STUDIES

Adaptation and no-acquisition controls. The base-only rows test whether the scale intervention primarily remedies a weakness shared across source-data choices. Fixed scales improve base-only success by 23.91 pp, close to the 23.91 pp gain for clean acquisition. A preadaptation density diagnostic finds much larger changes in target-training likelihood than in executed-action error (Table 15). Learned-variance optimization is already known to impair mean prediction (Seitzer et al., 2022; Stirn et al., 2023); the paired intervention connects that dependence to the comparison of source experience. Fixing scales jointly changes gradient weighting, mixture responsibilities, and optimization trajectories.

Alternative learned-scale remedies. To test whether the result is confined to one scale replacement, we compare three predefined stabilizers: a likelihood scale floor of 0.05, a log-scale penalty toward 0.15, and a detached scale-head feature path. The last blocks scale-head gradients into the shared backbone while allowing the scale head to train; component assignments and features remain coupled. All settings retain the full network, matched source groups, ten learners, four targets and 250 updates. Source state 1,000 and repeat zero are fixed. All five operators, including reused learned/fixed controls, receive the same new scene panel. Table 2 reports every setting; Appendix P defines the losses and inference.

The floor and detached scale path raise observed clean success from 28.47% to 45.87% and 37.35%, respectively, while perturbed branches retain higher mean success. The log-scale penalty does not improve clean success. None of the three remedy interactions is resolved after multiplicity adjustment. Fixing scales yields the largest clean-data improvement and an interaction of -11.88 pp ($[-15.50, -8.34]$) on this panel. Thus, improved absolute adaptation performance and removal of the measured perturbed-data advantage are distinct outcomes.

Table 2: Stabilizers on identical GMM source branches and paired new scenes. Success is in percent; D is perturbed minus clean and I its change from learned scales (pp). Intervals on the three remedy interactions use Bonferroni-adjusted bootstrap quantiles; the fixed-scale interval is descriptive 95%.

Adapter	Clean	Perturbed	D	I	Interval
Learned	28.47	38.16	$+9.70$	—	—
Fixed 0.15	64.54	62.36	-2.18	-11.88	$[-15.50, -8.34]$
Floor 0.05	45.87	52.53	$+6.66$	-3.03	$[-8.50, +2.01]$
Log-scale penalty	27.12	38.48	$+11.36$	$+1.66$	$[-1.47, +5.01]$
Detached scale path	37.35	49.98	$+12.63$	$+2.94$	$[-1.20, +7.08]$

Table 3: Coupled responsibility/component-scale interventions at source state 1,000, repeat 0. Clean and perturbed columns are absolute macro success (%); D is perturbed minus clean (pp). Crossed 95% intervals use ten learners and shared scenes. All four original targets and every source group are included.

Operator	Clean	Perturbed	D	95% interval
LL	25.72	35.24	$+9.52$	$[+6.15, +12.74]$
FL	14.66	19.15	$+4.49$	$[+1.67, +7.37]$
LF	54.30	56.85	$+2.55$	$[-2.54, +7.38]$
FF	59.19	56.20	-2.98	$[-5.18, -0.94]$

5.3 Coupled responsibility and component-scale interventions

We cross two coupled training interventions: the responsibility rule assigning examples to mixture components and the component-scale rule weighting their gradients. L denotes learned scales and F a fixed scale of 0.15; the first letter selects responsibilities and the second component scales. The diagonal operators recover the original likelihood gradients. Both off-diagonal operators train the full network on the same source branches, target data and update budget. Appendix I gives their definitions and the preadaptation density diagnostic.

Changing the responsibility rule while retaining learned component scales gives the prespecified interaction of -5.03 pp ($[-8.84, -1.04]$). The complementary component-scale intervention gives -6.98 pp ($[-11.56, -2.30]$). Both absolute outcomes deteriorate under FL, whereas LF improves both; a smaller data contrast therefore does not itself indicate better adaptation. The nonadditivity contrast is -0.50 pp ($[-7.27, +5.96]$); this wide interval establishes neither additivity nor independent channels. Component scales change precision weighting and scale-head gradients, and predicted scales in LF can still change through shared features. These are coupled operator interventions, not an identified mediator. Secondary intervals are descriptive; Appendix I retains all conditional contrasts and base-only outcomes.

5.4 Matching the acquisition construction across policy studies

The deterministic control now uses exactly the same 32 matched groups and underlying arrays as the GMM study. It compares MSE and Huber adaptation on ten existing deterministic learners, preserving source branches, target demonstrations, update streams and the new evaluation scenes within each comparison. The group manifest hashes match across policy studies, and a full replay reproduces the reused MSE adapter parameters exactly. This removes pooled acquisition as an explanation for the within-deterministic comparison; the independently trained policy cohorts still differ.

The matched-group interaction is -0.27 pp ($[-0.96, +0.36]$). Appendix Q reports the distance strata and task-level contrasts. The earlier pooled control is retained separately in Appendix M.

5.5 A second engine with comparable clean-data performance

The PyBullet experiment tests direct behavior-cloning loss under stronger action-label contamination, using a separate robot and newly initialized learners. The two losses have similar clean-

Table 4: Deterministic matched-group control, source state 1,000 and repeat zero. Absolute macro success (%), perturbed-minus-clean D (pp), and crossed 95% intervals. The Huber-minus-MSE interaction is the primary contrast.

Adapter	Base	Clean	Perturbed	D	95% interval
MSE	73.12	72.96	73.07	$+0.11$	$[-0.25, +0.48]$
Huber	74.62	74.94	74.78	-0.16	$[-0.63, +0.28]$

Table 5: PyBullet results: clean / noisy absolute success (%) and the Huber-minus-MSE change in noisy-minus-clean success (pp). The macro interaction is primary; task-specific intervals are descriptive. Ten learner/corruption units and 64 shared scenes per task yield 5,120 simulator episodes.

Task	MSE	Huber	Interaction [95% interval]
Reach	100.00 / 89.06	100.00 / 97.97	$+8.91$ $[+3.91, +14.84]$
Push	61.56 / 36.56	60.47 / 50.94	$+15.47$ $[+4.38, +28.13]$
Macro	80.78 / 62.81	80.23 / 74.45	$+12.19$ $[+6.64, +18.59]$

data macro success, 80.78% for MSE and 80.23% for Huber, while noisy-data success rises from 62.81% to 74.45% (Table 5). The noisy-minus-clean effect is -17.97 pp ($[-22.27, -13.91]$) under MSE and -5.78 pp ($[-9.77, -1.48]$) under Huber. The prespecified interaction is $+12.19$ pp ($[+6.64, +18.59]$). A descriptive post-hoc comparison of clean-data success gives Huber minus MSE of -0.55 pp ($[-4.45, +3.13]$); it supplies no evidence of the large clean-data improvement seen in the GMM study.

The interactions are positive in both tasks. Push avoids the clean-data ceiling and gives $+15.47$ pp ($[+4.38, +28.13]$), while Reach has 100% clean-data success under both losses. Clean labels remain preferable under both operators: this is a change in effect magnitude, not a ranking reversal. The observed attenuation is consistent with the tolerance of Huber loss to large label errors in this direct-training comparison.

5.6 Replication and parameter sensitivity

The GMM comparison is repeated under two independently drawn label-noise banks, three fixed scales, and four additional future tasks. These studies retain the same ten source learners; they broaden interventions and tasks without increasing the independent learner count. The two new noise banks yield interactions of -15.53 and -14.76 pp, with crossed intervals excluding zero. Their fixed-scale noisy-minus-clean effects have intervals spanning zero (Appendix Table 19). The four additional targets yield a macro interaction of -15.76 pp ($[-25.92, -6.34]$); drawer-open remains variable and drawer-close is near ceiling (Appendix Table 20). These are within-Meta-World replications of the GMM finding.

Scale sensitivity. We vary fixed standard deviation over $\{0.05, 0.15, 0.30\}$ at source state 1,000/repeat zero, retaining all ten learners, groups, targets and 250 adaptation updates. Reused and new policies share a fresh scene panel. Figure 2 reports every predefined setting. The interactions range from -13.07 to -15.55 pp, with all descriptive crossed 95% intervals below zero. Clean-data success ranges from 41.20% to 67.13%, versus 27.20% under learned scales. Thus absolute adaptation performance is parameter-sensitive, while the signed data contrast decreases at every tested fixed scale.

5.7 Case studies and failure analysis

We categorize every clean-branch episode by paired success under learned and fixed-0.15 adaptation. Table 6 includes all outcomes, including episodes where the fixed-scale policy fails and the learned-scale policy succeeds. For each nonempty category, we replay the lexicographically first target, learner, source group, and episode. Each replay reproduces the original binary outcome. These post-result illustrations explain individual trajectories; category frequencies come from the complete evaluation.

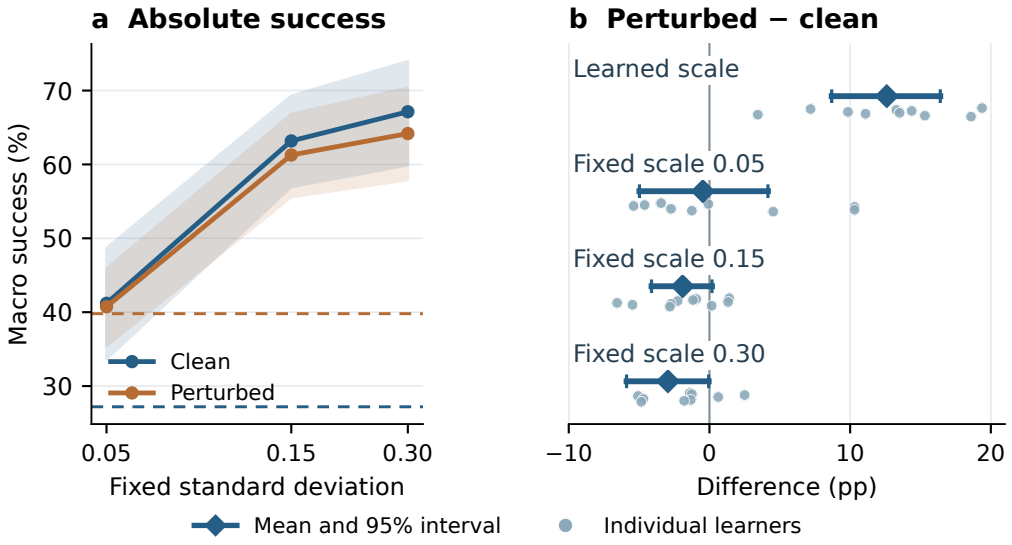

Figure 2: Adaptation-scale sensitivity at one fixed source state and repeat. Left: clean and perturbed absolute success, with crossed 95% bands; dashed lines are learned-scale means. Right: paired perturbation effects and crossed intervals; learner points are offset below the corresponding means for visibility. All scales are prespecified; intervals are descriptive.

Table 6: Paired clean-branch outcomes on the fresh scale-sensitivity panel. Counts cover all ten learners, 16 clean source branches, and 20 scenes per target at source state 1,000, repeat 0.

Target	Both succeed	Both fail	Fixed only	Learned only
Reach	1,078	478	1,458	186
Push	26	2,358	797	19
Pick-place	96	1,556	1,517	31
Window-open	2,020	60	1,095	25
Total	3,220	4,452	4,867	261

Across 12,800 paired episodes, fixed-only and learned-only outcomes account for 38.0% and 2.0%, respectively. These paired transitions show why an aggregate improvement does not imply an improvement on every trajectory.

6 Discussion

Data contrasts and adaptation quality. The original GMM comparison changes when the adapter's likelihood scales are fixed. Clean and base-only success rise by similar amounts, linking the original noisy-data advantage to how the source policy is subsequently adapted. The coupled channel interventions distinguish two ways of reducing the data contrast: switching the responsibility rule lowers both clean and perturbed success, whereas fixing component scales raises success. Reporting the four absolute outcomes therefore gives the interaction its operational meaning.

Interpreting the additional comparisons. The deterministic Meta-World control retains the GMM study's matched acquisition construction but gives an interaction of -0.27 pp ($[-0.96, 0.36]$). A large interaction is therefore not an automatic consequence of matching these data groups. The PyBullet direct-training comparison instead reduces the noisy-data penalty under Huber loss with similar observed clean-data performance. Together, these comparisons support measuring sensitivity for the specified operator and data construction. The local expansion explains first-order cancellation, while the retrospective gradient diagnostic does not establish prediction of long-horizon success (Appendix O).

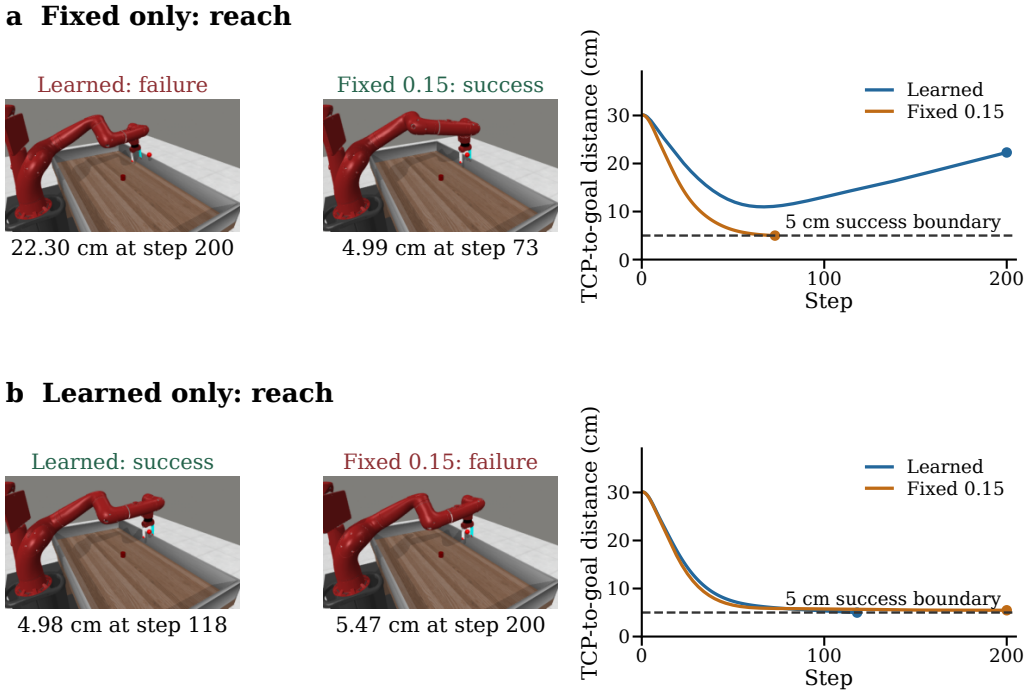

Figure 3: Contrasting paired trajectories from the prespecified representative-selection rule. Each case starts from the same scene under the two adaptation operators. Actual Meta-World/MuJoCo terminal frames are paired with the measured tool-center-point (TCP) distance to the goal. Reach success occurs when this distance is at most 5 cm; successful traces stop at the first crossing, while failed traces remain above the boundary for all 200 steps. The suspended goal sphere, rather than the tabletop block, defines the reach target. Operator colors identify curves; image labels report the outcome and terminal distance. Full distances, including near-threshold failures, appear in Appendix Table 14.

From a comparison to a selection decision. Prospective selection does not establish an incremental benefit over the group-identity control: both attain the same $+0.60$ pp mean gain, while the crossed interval spans zero. Two Re-Mix-weighted acquisition configurations also leave gains unresolved within their 100-update continuation window; these probes do not evaluate full mixture pretraining. Detecting an adapter interaction and improving a selection rule remain separate empirical questions.

7 Conclusion

Changing target adaptation can change the measured performance difference between source datasets. Paired GMM interventions demonstrate this effect and show why it must be read alongside absolute adaptation success. The four-cell protocol provides a direct way to test the stability of a data comparison under a specified operator change and to distinguish that question from whether data selection improves downstream learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, improving the clarity and organization of the manuscript. The authors take full responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of MuralAgent and its implementation throughout the main paper and appendix. The anonymized source code is also available at https://anonymous.4open.science/r/adaptation-E251.

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.

Jonathan T Barron. A general and adaptive robust loss function. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334. IEEE, 2019.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Shivin Dass, Alaa Khaddaj, Logan Engstrom, Aleksander Madry, Andrew Ilyas, and Roberto Martín-Martín. Datamil: Selecting data for robot imitation learning with datamodels. In *International Conference on Learning Representations*, volume 2026, pp. 1311–1336, 2026.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in neural information processing systems*, 33:2881–2891, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135, 2017.

Quentin Gallouédec, Nicolas Cazin, Emmanuel Dellandréa, and Liming Chen. panda-gym: Open-source goal-conditioned environments for robotic learning. *arXiv preprint arXiv:2106.13687*, 2021.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2242–2251, 2019.

Amirata Ghorbani, Michael Kim, and James Zou. A distributional framework for data valuation. In *International Conference on Machine Learning*, pp. 3535–3544. PMLR, 2020.

Joey Hejna, Suvir Mirchandani, Ashwin Balakrishna, Annie Xie, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, Dhruv Shah, Coline Devin, and Dorsa Sadigh. Robot data curation with mutual information estimators. *arXiv preprint arXiv:2502.08623*, 2025.

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.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 1167–1176, 2019.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1885–1894, 2017.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. *arXiv preprint arXiv:2110.14049*, 2021.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. In *International Conference on Learning Representations*, volume 2025, pp. 54877–54910, 2025.

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.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, pp. 27074–27113, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in neural information processing systems*, 33: 19920–19930, 2020.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International conference on machine learning*, pp. 4334–4343, 2018.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668, 2010.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, 2011.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Vaibhav Saxena, Matthew Bronars, Nadun Ranawaka Arachchige, Kuancheng Wang, Woo Chul Shin, Soroush Nasiriany, Ajay Mandlekar, and Danfei Xu. What matters in learning from large-scale datasets for robot manipulation. International Conference on Learning Representations, 2025.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *IEEE Transactions on Robotics*, 2026.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International conference on artificial intelligence and statistics*, pp. 5593–5613, 2023.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

Jiachen Tianhao Wang, Dawn Song, James Y Zou, Prateek Mittal, and Ruoxi Jia. Capturing the temporal dependence of training data influence. In *International conference on learning representations*, volume 2025, pp. 11236–11273, 2025.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100, 2020.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

A Data and evaluation provenance

Environment correction. The original collector passed a seed to `reset`. Meta-World 3.1.1 ignores that argument and, by default, samples its random vector through a different random-number path. Consequently, legacy per-episode numeric seed labels do not identify replayable initializations. The training and validation arrays remain separately stored, immutable datasets. An audit found no matching initial object/goal coordinates within 10^{-5} between their banks, for all sixteen tasks. This audit does not imply that the two distributions have disjoint support.

All early rollout comparisons affected by unpaired resets were withdrawn. Corrected evaluation explicitly calls the environment's seed method, enables its seeded random-vector path, disables frozen random vectors, and resets the NumPy seed before reset. A replay integrity test compares initial observations, simulator position/velocity states, and the next transition after an intervening episode for all sixteen tasks. All paired-arm output files are additionally checked for exact agreement of initial observations at the same evaluation seed. The legacy datasets are reproduced from archived arrays and their checksums, not from the invalid original seed labels.

Endpoint and stopping. Success means that the environment reports success at least once within the stated horizon. Early stopping after the first success preserves this endpoint. The associated cumulative reward is only a prefix return; no full-horizon-return comparison is made in the paper. The easy collection horizon was 200 steps and the hard collection horizon 500, with a short 15-step tail after first expert success. Failures are retained rather than silently replaced with successful episodes.

Development sequence. The initial easy and hard replay studies preceded the lower-dose diagnostic, the held-out acquisition extension, and the transported-gradient diagnostic. These follow-ups are explicitly exploratory and respond to pilot observations. The independent ten-seed selector comparison was specified after the pilots but before its training and test evaluation. Its code snapshot, protocol, checkpoint hashes, and all seed outcomes are retained. It confirms only the specified selector contrast; it does not retroactively convert exploratory findings into preregistered conclusions.

B Prospective selection protocol

B.1 Do experience descriptors predict future utility on new learners?

We fit three ridge predictors on the acquisition pilot and freeze them before training ten new parents (seeds 13–22). Static descriptors comprise expert-success fraction, trajectory length, base-relative centroid distance, and within-group state variance. A group-identity control uses sixteen indicators; a third model adds current candidate MSE, L1, and source-validation gradient alignment to that control. Each model includes an additive log-update-count term, standardizes features using pilot data only, and uses fixed ridge coefficient 1. Static and identity-only models have invariant group rankings; only learner features permit ranking adaptation. These descriptors do not causally isolate novelty, diversity, or learnability.

The test repeats the acquisition/adaptation protocol for all sixteen groups, two source states, two optimization repeats, and four fixed future targets: 680 source branches and 2,720 adapted policies. Predictions and checkpoint hashes are frozen before evaluation on a fresh 20-initialization panel per target. The primary endpoint compares static-descriptor selection with the expected outcome of uniformly choosing one candidate group, not the mixed transition-uniform acquisition control. Ties are averaged without consulting outcomes. Group identities, target tasks and data banks are shared with the pilot.

The primary effect is $+0.60$ pp (95% paired learner-seed t interval $[+0.04, +1.16]$; 8/10 positive seed differences), which has an interval above zero within this fixed candidate pool. The identity-only control has a secondary effect of $+0.60$ pp ($[-0.15, +1.34]$). Adding learner features changes its selected-arm outcome by -0.25 pp ($[-0.71, +0.21]$). The identity-only control achieves the same observed mean gain as static descriptors, so this result does not show that the descriptors outperform group identity; equal observed means do not establish equivalence. Secondary intervals are descriptive and not adjusted for multiple comparisons; they do not replace the frozen primary test. These

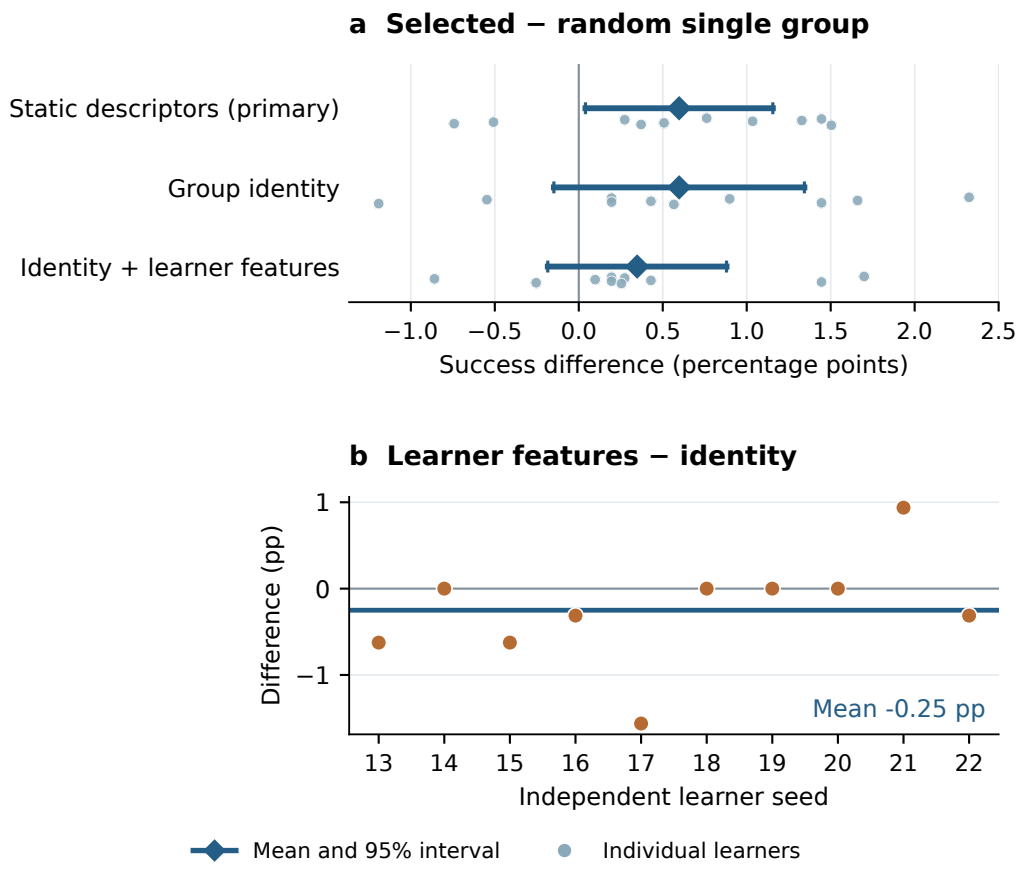

Figure 4: Prospective prediction on ten new acquisition learners. Top: seed effects and paired 95% t intervals; static descriptors are the frozen primary comparison. Bottom: the prespecified incremental contrast between learner-feature and group-identity selection. The candidate pool and target tasks are held fixed across development and confirmation.

results concern the specified predictors and cannot establish that no other experience properties are predictive. The observed 2.25 pp gain over mixed acquisition decomposes into 1.65 pp from choosing a uniformly random single group instead of mixed acquisition and 0.60 pp from which group is selected. This post-test accounting identity separates focused allocation from group choice; it does not identify a causal effect of a descriptor. The static predictor selects the long-trajectory hammer group (group 13) in every test context, so this experiment validates a frozen within-pool selection rather than transferable effects of individual descriptors. In a post-test sensitivity analysis, we jointly resample learner seeds and the common initialization panel within each fixed target (10,000 draws). The static-selector percentile interval is $[-0.17, +1.43]$ pp and includes zero. This descriptive sensitivity does not replace the prespecified fixed-panel test, but the small positive estimate is not robust to this additional source of uncertainty. A post-test base-only continuation control gives a static-selection gain of 0.75 pp, with a learner interval $[-0.20, 1.70]$ and crossed interval $[-0.28, 1.88]$. Thus, superiority to mixed acquisition does not establish a reliable benefit over continued training on the original data. An equal-group mixture follow-up separates group-weight changes from the fixed-group versus per-example mixture contrast; both components have crossed intervals including zero (Appendix C).

C PROSPECTIVE PREDICTION CONTRASTS

Table 7: All frozen future-utility contrasts. Differences are in percentage points; intervals are conditional on the fixed data and evaluation panels. Only the first contrast is primary.

Predictor	Comparator	Mean	95% interval
Static descriptors	Random single group	+0.60	$[+0.04, +1.16]$
Static descriptors	Mixed acquisition	+2.25	$[+1.02, +3.48]$
Group identity	Random single group	+0.60	$[-0.15, +1.34]$
Group identity	Mixed acquisition	+2.25	$[+1.38, +3.12]$
Identity + learner features	Random single group	+0.35	$[-0.18, +0.88]$
Identity + learner features	Mixed acquisition	+2.00	$[+1.28, +2.72]$
Identity + learner features	Identity selection	-0.25	$[-0.71, +0.21]$

Post-test no-acquisition comparator. After observing the frozen predictor results, we add 40 base-only source branches and 160 adapted policies for the same ten learners. The source update count, optimizer state, adaptation operator, and evaluation initializations match the acquisition arms. These comparisons were not prespecified primary tests. All five contrasts and both uncertainty summaries are retained; intervals are descriptive and unadjusted.

Table 8: Additional acquisition effects relative to continuing on the base data. All values are percentage points. Crossed intervals jointly resample learners and shared initialization panels within each fixed target.

Acquisition rule	Mean	Learner t interval	Crossed interval
Mixed acquisition	-1.50	$[-2.47, -0.53]$	$[-3.03, -0.06]$
Random single group	+0.15	$[-0.43, +0.74]$	$[-0.66, +1.00]$
Static descriptors	+0.75	$[-0.20, +1.70]$	$[-0.28, +1.88]$
Group identity	+0.75	$[-0.29, +1.79]$	$[-0.44, +1.94]$
Identity + learner features	+0.50	$[-0.35, +1.35]$	$[-0.59, +1.59]$

Static selection exceeds mixed acquisition by 2.25 pp, but exceeds base-only continuation by only 0.75 pp; both intervals for the latter include zero. The mixed arm is 1.50 pp below base-only continuation. Consequently, an advantage over mixed acquisition does not establish a reliable gain from acquiring new experience.

Post-test group-weight control. The earlier random-single comparison averages sixteen groups equally, whereas the mixed acquisition control weights groups by transition count. We therefore add an equal-group mixture: each candidate example first chooses a group uniformly, then a transition uniformly within that group. All 40 source branches and 160 adaptations retain the same parents, optimizer states, 192 base plus 64 candidate examples per update, 100 source updates, and target protocol. This follow-up was designed after inspecting the preceding outcomes; it does not replace the frozen primary endpoint. Let S, M_g, and M_t denote the uniformly random fixed-group, equal-group per-example mixture, and transition-uniform mixture policies. Their mean outcomes satisfy $C(S) - C(M_t) = [C(S) - C(M_g)] + [C(M_g) - C(M_t)]$. The first contrast changes how groups are mixed; the second changes group marginal weights. The first also changes coverage, repetition, gradient noise, and optimization trajectories, so it does not isolate a gradient-interference mechanism.

Table 9: Additional equal-group mixture controls. Percentage-point differences; all intervals are post-test, descriptive, and unadjusted.

Contrast	Mean	Learner t interval	Crossed interval
$M_g - M_t$	+0.69	$[+0.09, +1.28]$	$[-0.22, +1.59]$
$S - M_g$	+0.96	$[+0.14, +1.79]$	$[-0.07, +2.05]$
Static selector $-M_g$	+1.56	$[+0.38, +2.75]$	$[+0.16, +2.97]$
Identity selector $-M_g$	+1.56	$[+0.59, +2.54]$	$[+0.16, +2.97]$
Learner-feature selector $-M_g$	+1.31	$[+0.49, +2.13]$	$[+0.00, +2.56]$
M_g- base only	-0.81	$[-1.97, +0.35]$	$[-2.28, +0.62]$

C.1 Matched within-task properties and label fidelity

We next freeze a within-task comparison before training ten new learners (seeds 23–32). Candidate trajectories are paired by adjacent lengths within expert-success strata. The member farther from its nearest same-task base initial observation enters the higher-distance subset; the other enters the lower-distance subset. Distance is squared Euclidean distance averaged over 39 base-normalized state coordinates. Each arm contains 29 or 30 trajectories per task, with exactly matched trajectory-level success rates and closely matched mean lengths. This controls task identity but not every state attribute: distance is not semantic novelty, and transition-weighted success rates need not match exactly.

Each subset has clean and corrupted-label variants. Corruption adds one fixed Gaussian draw with standard deviation 0.15 to each action component and clips to $[-1, 1]$; clean and corrupted arms sample identical transitions. This changes labels while preserving state coverage. Clipping can change realized perturbation strength between groups. The 32 factorial cells and base-only control yield 1,320 source branches and 5,280 future-adapted policies. We retain the prior update/adaptation budgets, use a fresh 20-initialization panel per target, and freeze a primary high-minus-low distance contrast on clean labels before training. Its primary interval jointly resamples learners and shared target initialization panels; task identities, matched pools, and the noise realization remain fixed.

The primary future-success difference is $+0.25$ pp, with a prespecified crossed 95% interval $[-0.16, +0.69]$. This does not establish a signed advantage for either matched subset. The secondary clean-minus-corrupted contrast is -0.06 pp ($[-0.31, +0.19]$). All frozen contrasts and balance checks appear in Appendix D. Full cell, target, and state summaries are retained in the artifact.

D Matched-property experiment

The fidelity interaction is the higher-minus-lower distance effect with clean labels minus that effect with corrupted labels. All contrasts average the same fixed source tasks, target tasks, source states, and optimization repeats.

Table 10: All frozen factorial contrasts in percentage points. The first comparison is primary; the remainder are descriptive and not adjusted for multiple comparisons.

Contrast	Mean	Learner t interval	Crossed interval
Higher minus lower, clean	+0.25	$[-0.07, +0.57]$	$[-0.16, +0.69]$
Clean minus corrupted	-0.06	$[-0.27, +0.15]$	$[-0.31, +0.19]$
Distance by fidelity interaction	-0.03	$[-0.31, +0.24]$	$[-0.46, +0.41]$
Lower, clean minus base	-0.66	$[-1.30, -0.02]$	$[-1.58, +0.20]$
Lower, corrupted minus base	-0.61	$[-1.28, +0.05]$	$[-1.52, +0.26]$
Higher, clean minus base	-0.41	$[-0.98, +0.17]$	$[-1.32, +0.50]$
Higher, corrupted minus base	-0.33	$[-0.98, +0.31]$	$[-1.27, +0.61]$

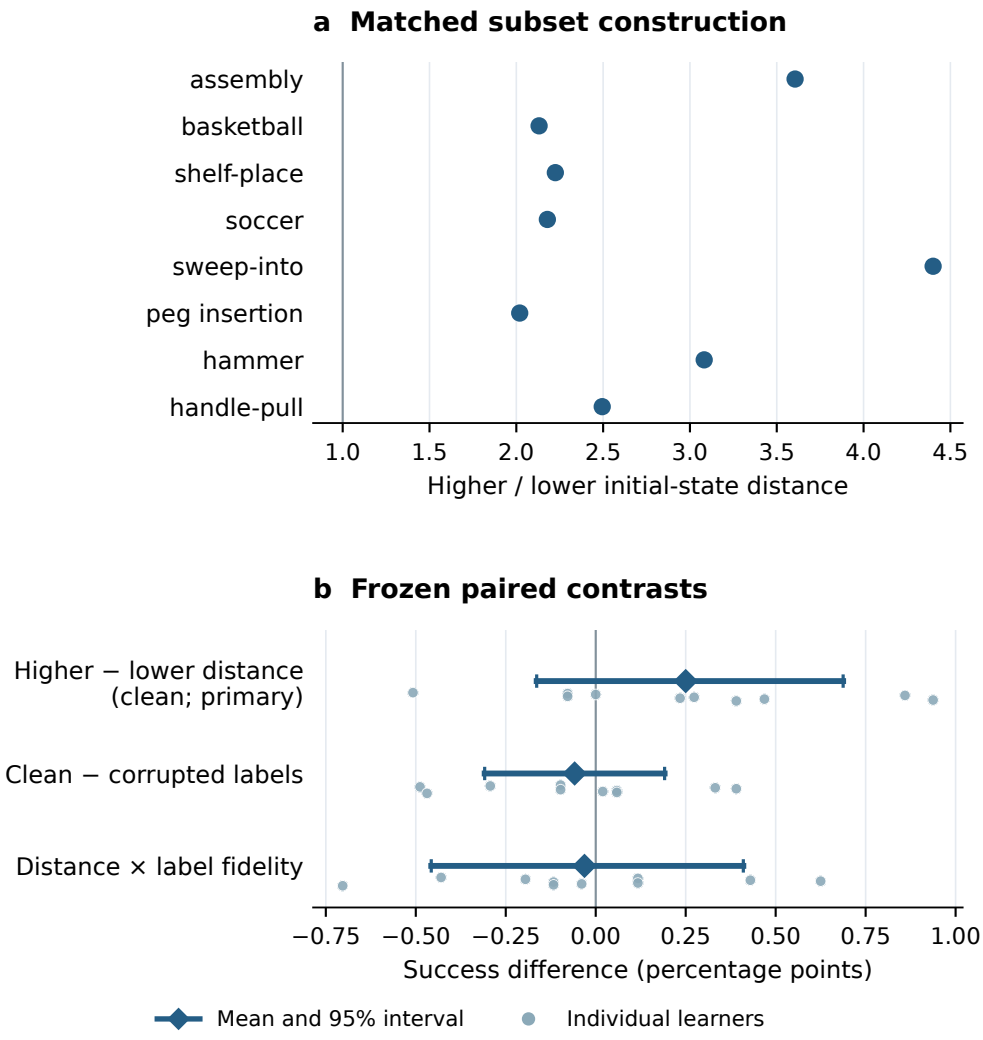

Figure 5: Matched-property study. Top: higher/lower initial-distance ratios, which differ by construction. Bottom: ten learner effects and crossed learner/initialization intervals for the primary distance comparison and two descriptive secondary contrasts. The same source tasks, future targets, and noise realization are held fixed.

Table 11: Matching and realized perturbation checks. L/H denotes lower/higher initial distance. Success is transition-weighted here; trajectory-level success fractions match exactly. RMSE is the realized action-label corruption, after clipping.

Source task	Traj./arm	Mean length L/H	Success L/H (%)	RMSE L/H
assembly	30	105.0/104.9	100.0/100.0	0.145/0.146
basketball	29	108.6/108.7	100.0/100.0	0.139/0.139
shelf-place	30	132.9/133.9	74.9/75.1	0.136/0.137
soccer	29	112.2/112.9	69.3/69.5	0.132/0.133
sweep-into	30	111.9/111.7	55.3/55.2	0.134/0.134
peg-insert-side	30	153.8/155.0	67.5/67.7	0.139/0.138
hammer	30	81.6/81.7	100.0/100.0	0.142/0.145
handle-pull	30	118.5/118.5	100.0/100.0	0.130/0.129

Action-component clipping fractions and every trajectory identifier are retained in the frozen artifact. Observational matching does not eliminate differences in object/goal configurations, subsequent state coverage, or action distributions. One fixed corruption realization is shared across all learners, so these intervals do not average over independently recollected data or noise banks.

E ARTIFACT ORGANIZATION

The artifact README links current entry points and the historical experiment map. Code, tests, immutable data manifests, frozen prediction protocols, per-seed summaries, and figure source tables accompany the manuscript. Full policy/optimizer checkpoints and per-episode records are retained with the experiment outputs. The included-path audit under `results/audit_final` identifies corrected evaluations and checks paired initializations; directory suffixes alone do not establish validity. Invalidated historical evaluations are excluded.

F GMM POLICY REPLICATION

After observing the unresolved deterministic-policy result, we freeze a new replication before its training and evaluation. We repeat the matched-property experiment with ten new learners (seeds 33–42), retaining all candidate subsets, source/target tasks, budgets, factorial arms, and contrast weights. The model retains the three-layer 256-wide state encoder but predicts a five-component diagonal Gaussian mixture. Component means use tanh; scales use softplus plus 0.01. Training minimizes exact mixture negative log likelihood. Execution deterministically takes the mean of the highest-weight component. This is an adapted GMM behavior-cloning probe: unlike the low-noise sampling option in robomimic (Mandlekar et al., 2021), it does not sample components at evaluation. It is not a reproduction of that system or an isolated causal test of architecture. Output parameterization, loss, and execution change jointly, and the new learner cohort is independent. Equal update counts do not match optimization difficulty or exact computation across objectives.

The frozen design uses 1,320 source branches, 5,280 adapted policies, and 20 fresh paired initializations per future task. The primary crossed bootstrap retains the original averaging weights and resamples learners and shared initializations (10,000 draws). Data pools and the corruption realization remain fixed. We do not pool cohorts or select tasks to strengthen a significance claim.

The GMM primary high-minus-low clean-distance difference is $+0.58$ pp, with crossed 95% interval $[-0.57, +1.86]$. The interval does not establish a signed distance advantage. The frozen secondary corrupted-minus-clean contrast is $+6.77$ pp ($[+4.59, +8.89]$); all ten learner means have this sign. It is larger at the later source checkpoint (11.07 pp versus 2.47 pp). These secondary intervals are unadjusted. Base-only GMM success is only 1.38% on push and 6.00% on pick-place, so this is not evidence from a uniformly strong alternative policy. The perturbation can alter density fitting and subsequent adaptation; regularization and optimization explanations remain hypotheses, not isolated mechanisms. This comparison uses a fixed noise bank. Figure 6 includes absolute success, so a small contrast is not interpreted without its performance context.

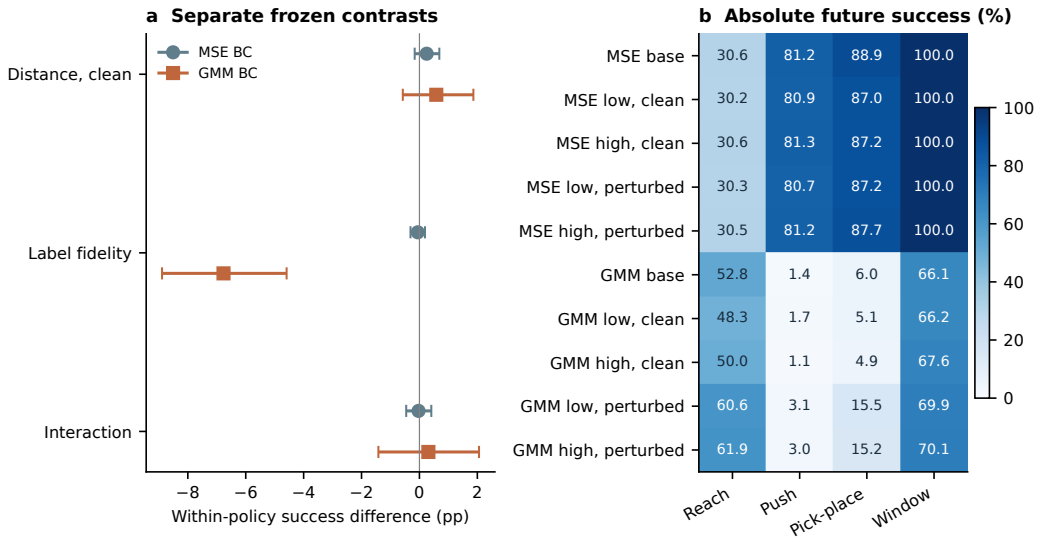

Figure 6: Cross-policy replication. Left: separately estimated within-policy effects and crossed 95% intervals; only clean distance is primary. Right: absolute success averaged over all learners, states and repeats; candidate rows also average all source tasks. The cohorts use distinct scene panels, so their apparent difference is not a paired architecture effect.

Table 12: All frozen GMM contrasts (pp). Secondary comparisons are descriptive and unadjusted.

Contrast	Mean	Learner t interval	Crossed interval
Higher minus lower, clean	+0.58	$[-0.76, +1.92]$	$[-0.57, +1.86]$
Clean minus corrupted	-6.77	$[-8.63, -4.91]$	$[-8.89, -4.59]$
Distance by fidelity interaction	+0.31	$[-1.55, +2.17]$	$[-1.42, +2.06]$
Lower, clean minus base	-1.22	$[-4.41, +1.97]$	$[-4.12, +1.61]$
Lower, corrupted minus base	+5.70	$[+2.68, +8.72]$	$[+2.81, +8.86]$
Higher, clean minus base	-0.64	$[-2.89, +1.62]$	$[-2.82, +1.48]$
Higher, corrupted minus base	+5.98	$[+2.52, +9.44]$	$[+2.73, +9.50]$

G Density diagnostic and adaptation-scale intervention

We next freeze a paired intervention on target adaptation, using the same ten source learners and every factorial branch. One arm reuses the learned-scale adapted policies. The other starts from the identical source checkpoint and replaces every Gaussian scale in the target-adaptation likelihood with 0.15. Mean outputs, mixture logits and the shared encoder remain trainable; the unused scale-output rows receive zero gradient with fresh Adam and no weight decay. Their unused predictions can still change through the encoder. The initial executed actions are identical. The fixed value matches the numerical label-noise amplitude by design, not a claim of optimality. Data, batch order, learning rate and 250-update budget match exactly. Both arms are re-evaluated on the same new 20-initialization panel per target; old-panel outcomes are not reused as controls.

The primary contrast is $(C_{\text{perturbed,fixed}} - C_{\text{clean,fixed}}) - (C_{\text{perturbed,learned}} - C_{\text{clean,learned}})$. Its interval jointly resamples learners and shared target initialization panels. This is a post-result-designed, evaluation-before-outcome-frozen intervention on the same learners and noise bank, not an independent learner or noise replication.

The primary interaction is -5.99 pp, with crossed 95% interval $[-8.94, -3.26]$. The interval supports a decrease in the signed perturbed-minus-clean contrast under fixed-scale adaptation. Figure 7 reports absolute performance; the table below gives every contrast. Changing the scale rule also

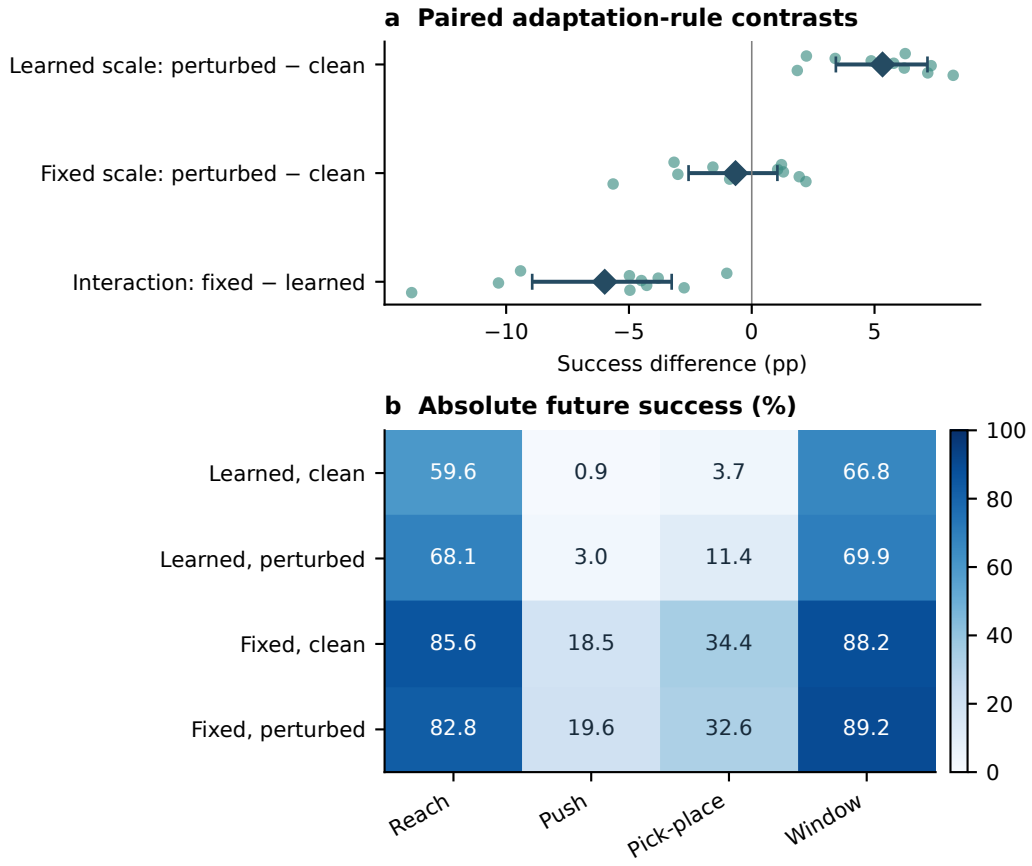

Figure 7: Original paired adaptation-scale intervention across both source states and repeats. Learner effects and crossed intervals are shown above absolute target success. Only the interaction is the frozen follow-up primary.

changes gradient weighting, mixture responsibilities and optimization trajectories; it does not by itself prove that source variance collapse mediates the effect.

Table 13: All scale-ablation contrasts (pp). Only the first is primary; the others are descriptive and unadjusted.

Contrast	Mean	Learner t interval	Crossed interval
Perturbation-benefit interaction	-5.99	$[-8.80, -3.17]$	$[-8.94, -3.26]$
Perturbed minus clean, learned	+5.33	$[+3.76, +6.89]$	$[+3.43, +7.16]$
Perturbed minus clean, fixed	-0.66	$[-2.55, +1.23]$	$[-2.57, +1.04]$
Fixed minus learned, clean	+23.91	$[+17.85, +29.98]$	$[+17.83, +30.24]$
Fixed minus learned, perturbed	+17.93	$[+14.21, +21.64]$	$[+13.68, +22.21]$
Fixed minus learned, base-only	+23.91	$[+15.94, +31.87]$	$[+16.56, +32.03]$

H Paired case-study trajectories

Representatives use the frozen lexicographic selection rule. The artifact stores every observation, action, reward, true simulator TCP center, goal, distance, success step, and requested frame at steps 0, 50, 100, and termination. Binary success, episode length, and initial state match the original

Table 14: All eight unchanged case replays. Distances are in cm. Minimum distance considers evaluated steps 1 through termination. A dash means no success within 200 steps. Four decimal places expose the threshold proximity of the fixed-scale failure in the both-fail case.

Category	Adaptation	First success	Minimum	Terminal
Both succeed	Learned	68	4.9754	4.9754
Both succeed	Fixed 0.15	45	4.9066	4.9066
Both fail	Learned	–	12.4490	25.4824
Both fail	Fixed 0.15	–	5.0093	5.1193
Fixed only	Learned	–	10.9751	22.2961
Fixed only	Fixed 0.15	73	4.9899	4.9899
Learned only	Learned	118	4.9812	4.9812
Learned only	Fixed 0.15	–	5.4709	5.4709

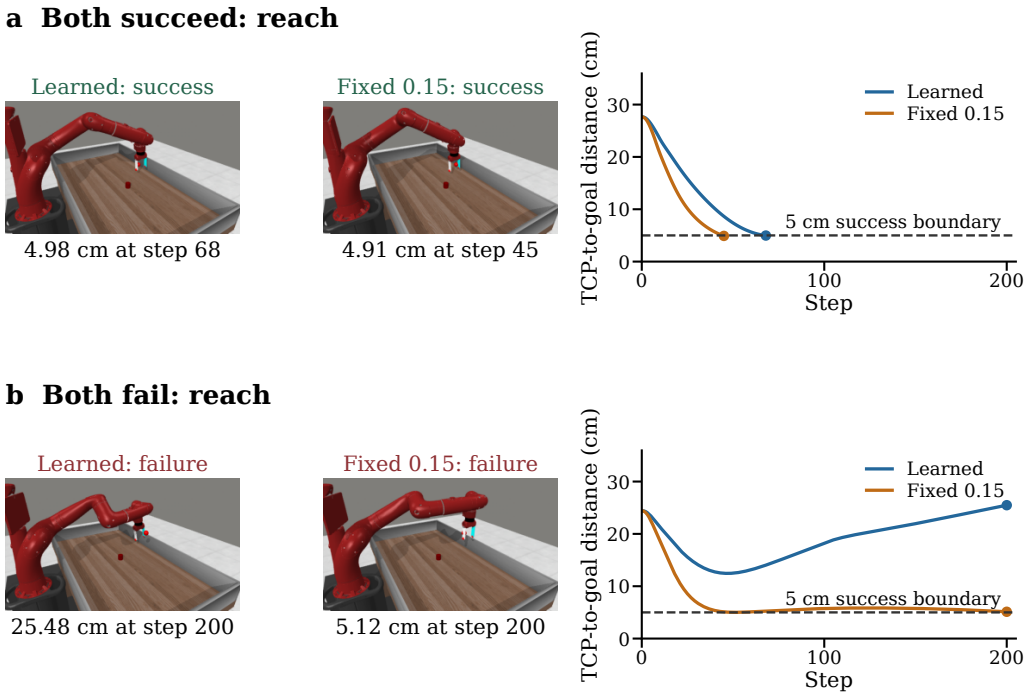

Figure 8: Agreement categories. Actual Meta-World/MuJoCo terminal frames accompany true TCP-to-goal distances. The dashed 5 cm boundary is the installed reach success criterion; successful traces end at their first crossing, and failures never cross within 200 steps.

evaluations. Reach success uses the installed Meta-World criterion $d \leq 5$ cm, checked after each action. The tabletop block is not the reach goal.

A success label can change for a small geometric difference near the threshold. Such cases reflect the simulator's binary criterion rather than qualitatively distinct motion. The closest fixed-scale approach in the both-fail case remains just outside the boundary, as the minimum-distance column shows. Success-once classification requires the full evaluated trace: a terminal distance alone would not establish failure.

I Responsibility/component-scale contrasts

After the GMM secondary result, we inspect the source checkpoints on identical batches of 256 transitions from each target's ten adaptation demonstrations, without using rollout outcomes in this diagnostic. The likelihood and executed-action errors are different quantities. Table 15 retains both,

Table 15: Preadaptation diagnostic, averaged over all learners, repeats, groups and target training batches. NLL uses the learned GMM density; action MSE uses deterministic highest-weight-component means.

Source state	Metric	Clean	Perturbed
250	Action MSE	0.346	0.339
250	Mean component std	0.222	0.232
250	Near-floor fraction	0.132	0.093
250	Target-training NLL	52.970	22.537
1000	Action MSE	0.384	0.371
1000	Mean component std	0.182	0.214
1000	Near-floor fraction	0.300	0.173
1000	Target-training NLL	242.924	53.122

together with the fraction of action-coordinate scales in the highest-weight component at or below 0.011 (the model floor is 0.01). These are post-test descriptive measurements on training inputs, not independent evidence of a causal mechanism. General optimization difficulties of learned-variance Gaussian regression are already established (Seitzer et al., 2022; Stirn et al., 2023).

The original scale intervention changes both posterior component assignments and the component loss. We cross their rules in a coupled channel intervention, without assuming that their training effects are additive. For predicted means μ_k, scales σ_k and mixture weights π_k, set $s_{L,k} = \sigma_k$ and $s_{F,k} = 0.15\mathbf{1}$, and define

$$\begin{aligned} r_k^a &= \mathrm{softmax}_k\big(\log \pi_k + \log \mathcal{N}(y; \mu_k, \mathrm{diag}(s_{a,k}^2))\big)\,, \\ \mathcal{L}_{ab} &= -\sum_k \mathrm{stopgrad}(r_k^a)\left[\log \pi_k + \log \mathcal{N}(y; \mu_k, \mathrm{diag}(s_{b,k}^2))\right]. \end{aligned} \tag{3}$$

The first letter selects the responsibility rule, and the second selects the component scale. Responsibilities are recomputed at each update. The diagonal gradients recover learned-scale (LL) and fixed-scale (FF) mixture likelihood training; the off-diagonal operators FL and LF bridge them. Their scalar surrogate values are not mixture NLLs. We retain the full model, source checkpoints, target data and adaptation budget. Both new operators train all source branches; all four operators are evaluated on a fresh common scene panel.

For $q_k(\theta) = \pi_k \mathcal{N}(y; \mu_k, \mathrm{diag}(s_k^2))$ and $r_k = q_k / \sum_j q_j$, differentiation gives

$$\nabla_\theta[-\log \sum_k q_k] = -\sum_k r_k \nabla_\theta \log q_k. \tag{4}$$

Using detached responsibilities on the right therefore reproduces the diagonal likelihood gradient at each update. This identity concerns gradients, not equality of scalar loss values. Off-diagonal rules recompute assignments from the current network and then apply a different component-scale rule; neither assignments nor the shared representation are held fixed across training. Full-network gradient and short optimizer-path checks cover both diagonal rules before the run.

The bridge study uses all 33 source branches (base-only and 32 factorial branches), ten source learners and four original future targets. It trains 2,640 new adapted models, reuses 2,640 diagonal controls and evaluates all 5,280 on 20 fresh scenes per target, for 105,600 nested episode records. The responsibility-rule primary is fixed before these outcomes; secondary contrasts use the same crossed bootstrap without multiplicity adjustment. All source, data, protocol and checkpoint hashes are retained. Original studies and this later-state follow-up have distinct endpoints and scene panels.

Table 16: Bridge contrasts (pp). The first row is primary. D_{ab} is the perturbed-minus-clean effect under operator ab.

Contrast	Mean	Crossed 95% interval
$D_{FL} - D_{LL}$	-5.03	$[-8.84, -1.04]$
$D_{LF} - D_{LL}$	-6.98	$[-11.56, -2.30]$
$D_{FF} - D_{LL}$	-12.51	$[-16.08, -8.84]$
$D_{FF} - D_{FL} - D_{LF} + D_{LL}$	-0.50	$[-7.27, +5.96]$
$D_{FF} - D_{LF}$	-5.53	$[-11.22, +0.39]$

Table 17: Base-only adaptation success (%) on the mechanism panel.

Operator	Mean	Crossed 95% interval
LL	$+26.75$	$[+20.50, +32.63]$
FL	$+13.88$	$[+8.75, +19.50]$
LF	$+54.00$	$[+47.00, +61.62]$
FF	$+60.25$	$[+53.50, +66.88]$

J ADDITIONAL-TARGET CONTRASTS

All 2,640 adapted policies are newly trained for these four additional targets. Each uses all 250 adaptation updates and is evaluated on 20 fresh paired scenes, giving 52,800 nested episode records. The data, source model, optimizer reset, minibatch ordering and target task code are paired between adaptation rules. The original source learners and one label-noise bank are reused. The table reports all macro contrasts; machine-readable outcomes include every learner, group and target.

Table 18: Additional-target macro contrasts (pp). Only the first row is primary.

Contrast	Mean	Crossed 95% interval
Perturbation-effect interaction	-15.76	$[-25.92, -6.34]$
Perturbed minus clean, learned	$+16.82$	$[+10.16, +23.77]$
Perturbed minus clean, fixed	$+1.06$	$[-4.09, +5.86]$
Fixed minus learned, clean	$+21.98$	$[+11.01, +33.36]$
Fixed minus learned, perturbed	$+6.22$	$[+1.62, +10.75]$
Fixed minus learned, base-only	$+18.88$	$[+7.75, +31.37]$

K GMM REPLICATION DETAILS

K.1 REPLICATION ACROSS LABEL PERTURBATIONS

To test dependence on the original noise draw, we generate two additional complete label-perturbation banks with the same Gaussian amplitude and clipping rule. We reuse the ten source parents at 1,000 updates and repeat zero, retrain every perturbed source branch for 100 updates, and apply both full 250-update adaptation procedures. Clean branches are shared controls. All three banks are compared on the same fresh scenes used by the parameter study. The two new banks require 320 source branches and 2,560 adapted policies.

Table 19: Replication across three fixed label-perturbation banks at source state 1,000 and repeat zero. Paired effects and crossed 95% intervals are in percentage points. Each bank uses the same ten learners and shared clean controls; banks are separate replications, not additional independent learners. Intervals are unadjusted across banks.

Noise bank	Perturbed − clean Learned scale	Perturbed − clean Fixed scale	Interaction
Original	$+12.60$ $[+8.69, +16.41]$	-1.91 $[-4.12, +0.18]$	-14.51 $[-18.42, -10.42]$
Independent draw 1	$+13.23$ $[+9.88, +16.78]$	-2.30 $[-4.73, +0.13]$	-15.53 $[-19.36, -11.88]$
Independent draw 2	$+12.74$ $[+8.84, +16.84]$	-2.02 $[-4.43, +0.28]$	-14.76 $[-18.99, -10.61]$

Both additional banks reproduce the negative adaptation interaction. In each bank, the perturbed-data advantage under learned scales becomes a smaller negative point estimate under fixed scales; all fixed-scale perturbation intervals include zero. Thus, replication supports a change in the comparison, without establishing that clean labels reliably dominate after fixing scales. The larger interactions than in Table 21 concern one later source state and repeat on a new scene panel; they do not replace the original full-matrix endpoint.

K.2 Four additional future tasks

We next test door-open, drawer-open, drawer-close and button-press, which were absent from the preceding GMM future-target evaluations. These tasks also belong to the project's earlier replay suite. Both learned and fixed-0.15 adaptation are trained afresh for every source branch, using the same ten source learners at state 1,000, repeat 0, the first ten target demonstrations and 250 updates. New scenes are paired across arms. The primary averages only these four additional tasks; it does not pool them with the original targets.

Table 20: Additional-task results. Absolute success is shown as clean / perturbed (%) for each adaptation rule. Interaction is the fixed-minus-learned change in perturbed-minus-clean success (pp). The new-task macro interaction is primary; per-task intervals are descriptive and unadjusted.

Target	Learned	Fixed 0.15	Interaction [95% interval]
Door open	50.06 / 73.41	97.88 / 96.22	-25.00 $[-37.03, -13.81]$
Drawer open	33.94 / 67.03	60.44 / 63.84	-29.69 $[-61.22, +2.69]$
Drawer close	97.97 / 99.69	97.78 / 99.28	-0.22 $[-4.91, +4.31]$
Button press	83.66 / 92.78	97.44 / 98.44	-8.12 $[-15.06, -1.94]$
Macro	66.41 / 83.23	88.38 / 89.45	-15.76 $[-25.92, -6.34]$

The additional-task primary interaction is -15.76 pp (95% interval $[-25.92, -6.34]$). Target-specific intervals are negative for door-open and button-press, while drawer-open has substantial learner variation and drawer-close is near ceiling under both rules (Table 20). Additional macro contrasts, including the base-only adaptation gain, appear in Appendix J; the machine-readable analysis retains all target-specific contrasts and absolute base-only outcomes.

Table 21: Control ablations separate data choice, allocation, and adaptation. Values are paired success differences (pp), with crossed 95% intervals resampling learners and shared scenes. Panels use different learner cohorts and cannot form a method leaderboard. *Frozen primary of the scale follow-up. †Original selector primary; its prespecified fixed-panel learner interval is reported in the text. Other rows and selector crossed intervals are descriptive and unadjusted.

Paired contrast	Role of control	Difference	95% interval
A. Adaptation control: ten GMM learners			
Perturbed − clean, learned scale	Data preference	$+5.33$	$[+3.43, +7.16]$
Perturbed − clean, fixed scale	Adaptation dependence	-0.66	$[-2.57, +1.04]$
Fixed − learned perturbation effect*	Paired interaction	-5.99	$[-8.94, -3.26]$
Fixed − learned, base only	Adaptation weakness	$+23.91$	$[+16.56, +32.03]$
B. Selection controls: ten independent acquisition learners			
Static − random single group†	Group choice	$+0.60$	$[-0.17, +1.43]$
Identity − random single group	Identity control	$+0.60$	$[-0.35, +1.51]$
Dynamic − identity	Learner features	-0.25	$[-0.97, +0.38]$
Static − mixed acquisition	Mixture comparator	$+2.25$	$[+0.66, +3.94]$
Random single group − mixed	Allocation control	$+1.65$	$[+0.46, +2.95]$
Static − base only	Acquisition benefit	$+0.75$	$[-0.28, +1.88]$

L SELECTION AND ALLOCATION CONTROLS

Selection, identity, and allocation controls. A separate prospective experiment freezes predictors before training ten new acquisition learners. Static descriptors, group identity, and identity plus learner features use the same candidate pool and target tasks. The original primary contrast, static selection versus a uniformly random single group, is $+0.60$ pp with prespecified learner interval $[+0.04, +1.16]$. Its crossed interval spans zero when shared-scene uncertainty is included. Identity selection reaches the same observed mean gain, and adding learner features changes the identity-selected outcome by -0.25 pp. Thus, these results do not establish an incremental benefit from the tested descriptors or dynamic features.

The comparator explains much of the apparent gain: static selection exceeds mixed acquisition by 2.25 pp, of which 1.65 pp already occurs for random single-group acquisition. Relative to base-only continuation, the static selector gains 0.75 pp with an interval spanning zero. Group choice, mixing, and acquiring new data are consequently separate experimental decisions. Full predictor definitions, the equal-group mixture control, and all contrasts appear in Appendix C. The original matched-distance primary comparisons also remain unresolved in both MSE and GMM cohorts (Appendices D and F); the label perturbation was a secondary comparison that motivated the later adaptation intervention.

M NON-GMM ACQUISITION AND MIXTURE PROTOCOL

The Meta-World extension reuses the ten MSE parent learners, seeds 23–32, at source update 1,000. New acquisition branches pool all 60 withheld trajectories per task. Task-uniform acquisition first selects one of eight source tasks uniformly, then samples uniformly over that task's candidate transitions. The Re-Mix-weighted acquisition probe changes candidate task probabilities; it does not rank transitions within a task. All source branches use 100 MSE updates of continued Adam, batch 256, with 192 base and 64 candidate transitions. Base-only uses 256 base transitions. Clean and noisy uniform branches share every sampled index. The latter adds a fixed Gaussian noise bank with standard deviation 0.15 to candidate labels and clips to $[-1, 1]$; base labels remain unchanged.

Each saved source branch adapts to each original future target using ten target trajectories, a zero eight-coordinate task code and source observation normalization. Both adapters use fresh Adam at 3×10^{-4}, batch 256 and 250 updates. Huber loss is scaled by two, with threshold 0.1, to match mean squared error within its quadratic region. All parameters are trained. The extension trains 40 source branches and 320 adapted policies and evaluates 6,400 episodes on a fresh shared panel. Source and target minibatch streams and the Meta-World corruption bank are fixed across learners.

The crossed interval therefore conditions on these streams and that bank while resampling learner and scene variation.

Re-Mix-weighted acquisition configuration. The reference and proxy networks each have three 512-unit layers with LayerNorm and dropout 0.4. Action coordinates are standardized separately per source task using candidate-training statistics and assigned to 48 Gaussian bins. Cross-entropy sums over action coordinates. Both models use AdamW at 3×10^{-4}, weight decay 10^{-4} and batch 256; domains are sampled uniformly during mixture fitting. The reference runs for at most 2,000 updates, with validation checks every 100. Training stops after any task fails to improve validation cross-entropy by 0.001 for three checks; the best macro-validation checkpoint is restored. The proxy runs for 1,000 updates.

The implementation follows the released Re-Mix update: positive per-example excess cross-entropy updates domain weights with exponentiated step size 0.2 and smoothing 0.05 toward uniform weights. Its domain reduction divides by the number of nonzero loss entries. Proxy gradients use unclipped excess loss; final sampling weights average the successive weights. Candidate and validation trajectories are disjoint, and target adaptation data or rollout outcomes are never used for mixture fitting. The artifact records every validation check, chosen checkpoint, weight trajectory and computation count.

The probe transfers domain weights to a 100-update acquisition window rather than using them throughout source pretraining. It uses state observations, task domains and uniform initial weights; the released dataset-domain implementation initializes weights in proportion to domain sizes. Reference training stops on a per-task patience criterion but restores the best macro-validation checkpoint. Thus the stopping trigger and checkpoint selection use different reductions: the selected checkpoint can follow deterioration on an individual task. The published reference-selection principle instead seeks a checkpoint before any domain overfits. Constant-rate training and the initial fixed proxy duration are further configuration differences. The additional reference/proxy training is charged as scoring work, while source continuation and target adaptation budgets remain equal across the evaluated data branches.

Table 22: All non-GMM Meta-World macro contrasts (pp), including unresolved results. The noise-effect interaction is primary; all other intervals are descriptive and unadjusted.

Contrast	Mean	Crossed 95% interval
noisy minus clean MSE	-0.88	$[-2.25, +0.50]$
Re-Mix minus uniform MSE	-0.25	$[-2.00, +1.63]$
clean minus base MSE	$+0.38$	$[-2.00, +2.75]$
noisy minus base MSE	-0.50	$[-3.00, +1.88]$
Re-Mix minus base MSE	$+0.13$	$[-2.25, +2.50]$
noisy minus clean huber	$+0.00$	$[-1.63, +1.63]$
Re-Mix minus uniform huber	$+0.25$	$[-1.88, +2.38]$
clean minus base huber	-0.13	$[-2.88, +2.62]$
noisy minus base huber	-0.13	$[-2.88, +2.75]$
Re-Mix minus base huber	$+0.13$	$[-1.88, +2.50]$
noise effect interaction	$+0.88$	$[-1.25, +3.00]$
Re-Mix effect interaction	$+0.50$	$[-1.88, +3.00]$
huber minus MSE base	$+4.88$	$[+1.37, +8.75]$
huber minus MSE clean	$+4.38$	$[+1.13, +7.88]$
huber minus MSE noisy	$+5.25$	$[+2.13, +8.75]$
huber minus MSE Re-Mix	$+4.88$	$[+1.63, +8.38]$

Table 23: Re-Mix task weights across ten source fits. Uniform task allocation is 0.125. Weights affect candidate sampling only; the base fraction remains 0.75.

Source task	Mean	Minimum	Maximum
Assembly	0.1139	0.1099	0.1192
Basketball	0.0998	0.0927	0.1051
Shelf place	0.0640	0.0598	0.0688
Soccer	0.1652	0.1610	0.1699
Sweep into	0.0740	0.0689	0.0771
Peg insert side	0.1411	0.1321	0.1463
Hammer	0.1295	0.1217	0.1353
Handle pull	0.2126	0.2024	0.2209

All ten short-budget reference fits reach and select the 2,000-update cap; none triggers the validation stopping rule. The ten mixture fits process 7,680,000 gradient examples and 17,533,600 additional reference/diagnostic forward examples. These scoring costs are additional to the equal downstream policy-update budgets. The reference losses are still improving at this cap, motivating the separately frozen training-budget follow-up below.

M.1 Acquisition-probe reference-budget and proxy-duration check

This follow-up was designed after inspecting the short-budget fit histories and downstream results. Its stopping and evaluation rules were frozen before its own test rollouts. It changes the reference cap to 10,000 updates, retains evaluation every 100 updates and the same any-domain patience of three checks, and trains the proxy for the number of updates in the selected reference checkpoint. Architecture, action binning, optimizer, domain-weight update, source continuation and target adaptation remain unchanged. The reference training and validation losses for the first 2,000 steps match the initial experiment exactly for every seed.

The ten references all trigger the validation stopping rule, after 2,900–4,500 updates. The selected checkpoint is the final checkpoint in every run: macro-validation selection does not enforce the published principle of selecting before any individual domain overfits. Ten new Re-Mix-weighted acquisition branches produce 80 new adapted policies. The 160 base and uniform-control policies are reused without retraining. All 240 policies are tested on 20 fresh shared scenes per target, with scene seed base 1,019,000, giving 4,800 episodes. The bootstrap resamples learners and independent scene panels per target for 10,000 draws (seed 810016). It conditions on the four task identities, demonstration pools and update streams. Results are not pooled with the earlier scene panel.

Table 24: All absolute results in the Re-Mix-weighted acquisition follow-up (success, %).

Adapter	Branch	Reach	Push	Pick-place	Window	Macro
MSE	Base only	12.50	76.00	76.50	100.00	66.25
MSE	Uniform clean	11.50	78.00	73.50	100.00	65.75
MSE	Re-Mix clean	12.00	76.00	74.00	100.00	65.50
Huber	Base only	20.00	81.50	79.50	100.00	70.25
Huber	Uniform clean	25.00	82.50	79.00	100.00	71.63
Huber	Re-Mix clean	21.00	82.50	80.50	100.00	71.00

Table 25: All macro contrasts in the Re-Mix-weighted acquisition follow-up (pp). The mixture-effect interaction is primary; other intervals are descriptive and unadjusted.

Contrast	Mean	Crossed 95% interval
Re-Mix minus uniform MSE	-0.25	$[-1.88, +1.13]$
clean minus base MSE	-0.50	$[-2.38, +1.37]$
Re-Mix minus base MSE	-0.75	$[-3.00, +1.00]$
Re-Mix minus uniform huber	-0.63	$[-2.50, +1.38]$
clean minus base huber	$+1.38$	$[-0.75, +3.75]$
Re-Mix minus base huber	$+0.75$	$[-1.38, +3.00]$
Re-Mix effect interaction	-0.38	$[-2.50, +1.88]$
huber minus MSE base	$+4.00$	$[+1.00, +7.63]$
huber minus MSE clean	$+5.88$	$[+2.38, +10.00]$
huber minus MSE Re-Mix	$+5.50$	$[+2.37, +9.25]$

Table 26: Actual mixture-fitting duration and the resulting average sampling weights. Reference, selected checkpoint and proxy duration coincide in every run. The fixed maximum is 10,000 updates; every run stops by validation patience.

Seed	Updates	Asm.	Bsk.	Shelf	Soc.	Sweep	Peg	Ham.	Handle
23	4500	0.122	0.124	0.072	0.166	0.088	0.151	0.125	0.152
24	4100	0.120	0.125	0.074	0.166	0.083	0.150	0.130	0.152
25	3300	0.117	0.125	0.070	0.163	0.090	0.141	0.139	0.154
26	2900	0.123	0.120	0.075	0.169	0.085	0.140	0.132	0.157
27	3000	0.119	0.126	0.077	0.163	0.086	0.142	0.132	0.156
28	3400	0.123	0.120	0.077	0.168	0.086	0.140	0.136	0.150
29	3300	0.121	0.120	0.075	0.160	0.083	0.148	0.135	0.159
30	3800	0.122	0.125	0.071	0.161	0.085	0.150	0.138	0.148
31	4300	0.122	0.119	0.069	0.167	0.086	0.151	0.133	0.152
32	3400	0.117	0.126	0.078	0.163	0.085	0.145	0.136	0.150

The ten longer fits process 18,432,000 gradient examples, 9,216,000 reference-forward examples during proxy training, and 26,952,480 diagnostic-forward examples. These costs are additional to the initial short-budget study and to downstream policy training. Neither the weighted-acquisition gain nor its operator interaction is resolved in this fresh-panel comparison of the 100-update continuation window.

N PyBullet experiment details

Official panda-gym 3.0.7 and PyBullet 3.2.7 provide the environment dynamics and success criteria. The custom reach controller moves proportionally toward the goal. The push controller approaches behind the object, descends and moves toward the goal. Training failures are retained. Input features concatenate native observations, desired goal and relative goal/object/end-effector positions, standardized using the common training observations. The policy has two 128-unit ReLU layers and three linear action outputs; actions are clipped at execution.

Each task uses 400 expert trajectories from scene seeds 10,000–10,399. The two data branches have identical states and counts. A learner-specific noise realization independently selects each action-label row with probability 0.3 and adds $\mathcal{N}(0, 0.7^2)$ noise and clips labels to $[-1, 1]$. Each pair of loss branches shares the exact corrupted labels. Learner initialization and minibatch order are paired across all four branches. MSE and twice Huber loss with threshold 0.1 each use 6,000 Adam updates at learning rate 10^{-3} and batch 256. There is no target fine-tuning stage.

Development and held-out evaluation. Before the scientific runs, 80 demonstrations and 2,500 updates gave development Reach success of 100% for both losses, but Push success of 31.25% and 9.375% for MSE and Huber. A second Push check with 400 demonstrations and 6,000 updates gave 59.375% and 53.125%. This clean-policy competence check used independent development

scenes and no corrupted-label comparison. The latter settings were then frozen for both tasks; both development rounds remain in the artifact.

Evaluation crosses ten initialized learners, two tasks, two label conditions and two losses with 64 shared test scenes (seeds 20,000–20,063), giving 5,120 simulator episodes. Native success reached at any step within the 50-step limit is primary; final success and distance are retained. The paired bootstrap resamples learner indices and shared scene indices, applying each draw to all branches and both fixed tasks. Thus uncertainty includes initialization/corruption and scene variation, conditional on the common collected demonstration pools.

O LOCAL INTERACTION AND PREADAPTATION DIAGNOSTIC

Let $\theta_g = \theta_{G_g}$ and $F_a^\eta(\theta_g) = \theta_g + \eta v_a(\theta_g)$, with paired update randomness. Suppose q is differentiable and its gradient is L-Lipschitz on each of these four segments. Taylor's first-order remainder gives

$$q(F_a^\eta(\theta_g)) = q(\theta_g) + \eta \nabla q(\theta_g)^\top v_a(\theta_g) + r_{ag}, \qquad |r_{ag}| \le \tfrac{1}{2} L \eta^2 \|v_a(\theta_g)\|_2^2.$$

Substituting into the four-cell contrast cancels both unadapted scores. The linear terms yield Equation 2, and the triangle inequality bounds $R = r_{11} - r_{10} - r_{01} + r_{00}$. If the linear term exceeds this bound in magnitude, its sign agrees with the one-step interaction. Equal aligned effects remove only the first-order term; they need not eliminate higher-order effects. Identical source checkpoints under paired updates give an exactly zero interaction. This is a local expansion, with no assumption that the channel interventions in Table 3 have additive effects.

For a preadaptation diagnostic, use the mixture-weighted mean $\bar{\mu}_\theta(x) = \sum_k \pi_{\theta,k}(x)\mu_{\theta,k}(x)$ and negative held-out action MSE as q. For fresh Adam without weight decay, the exact first-update direction is $v_a = -g_a/(|g_a| + \epsilon)$ coordinatewise, where g_a is the training-loss gradient on the shared first minibatch. Computing $\eta[h(\theta_1) - h(\theta_0)]$ requires validation gradients and the two loss gradients at the unmodified source checkpoints. It requires neither an adapted policy nor simulator rollouts, but does require the definitions of both operators. The mixture-mean proxy differs from executed modal-mean actions. ReLU boundaries and component switches preclude assuming global smoothness of the implemented policy, and the expansion supplies no bound for 250-step success.

Retrospective diagnostic evaluation. We score the ten existing GMM learners at source state 1,000 and repeat zero, before computing any adapted-model gradients. Adaptation uses target training trajectories 0–9; validation gradients use disjoint trajectories 10–19 from the same archived training pool. Each first minibatch contains 256 transitions from the original adaptation stream. All 330 source checkpoints remain unchanged. Scoring 1,320 source/target combinations produces 640 matched-pair scores with no optimizer step or new simulator rollout. The scores and source hashes are sealed before reading the archived learned/fixed-scale outcomes, but these learners had informed the study design; this is a retrospective check, not a prospective validation.

For each learner, average the 16 pair scores over the four fixed targets and compare with the corresponding final success interaction after 250 updates. Table 27 gives the prespecified Spearman correlation, 10,000 learner-bootstrap intervals and 100,000-permutation two-sided tests. The intervals condition on the reused scene panel. The first-update score does not establish a predictive association with the final success interaction. Its sign is not changed after observing the outcomes. Scoring takes 56.39 seconds with four CPU workers and processes 1,029,600 validation-gradient examples and 675,840 training-gradient examples. The artifact retains all pair-level scores and checks the 52,800 reused episode records against their shared initial scenes.

Table 27: Retrospective preadaptation diagnostic, ten learners. Correlations concern the 250-update success interaction, not the one-step smooth proxy in Equation 2.

Score	Spearman ρ	95% interval	Permutation p
First-Adam alignment	-0.479	$[-0.923, +0.280]$	0.167
Unadapted proxy gap	-0.321	$[-0.913, +0.491]$	0.368

P ALTERNATIVE GMM STABILIZER PROTOCOL

Let σ_k denote the predicted component standard deviation. The floor intervention evaluates the usual mixture likelihood with $\max(\sigma_k, 0.05)$ coordinatewise. The regularized intervention adds $0.1\ \mathrm{mean}_{x,k,d}(\log \sigma_{kd}(x) - \log 0.15)^2$ to the original negative mixture log-likelihood. The detached-path intervention computes scale logits from the same learned output-head weights applied to $\mathrm{stopgrad}(h_\theta(x))$, while means and mixture logits use $h_\theta(x)$. Its forward distribution initially matches the original; its scale-head parameters continue to train. This blocks one direct gradient path, without making heads independent or isolating a unique mediator. These are explicit ablations, not reproductions of named heteroscedastic-regression methods or tuned best-performing remedies.

All ten original GMM learners (seeds 33–42), 32 matched groups plus base, four targets, source state 1,000 and source repeat zero are included. The target data and 250-update fresh-Adam protocol match the original. All five operators are evaluated on 20 shared scenes per target, seed base 1,112,000, horizon 200. The new run trains 3,960 policies and evaluates 6,600, giving 132,000 episode records. Primary contrasts are the three remedy interactions relative to learned scales. Ten thousand crossed bootstrap draws resample ten learners and twenty scenes within each fixed target, seed 1,122,001. Their intervals use quantiles $0.05/6$ and $1 - 0.05/6$ for nominal 95% family coverage by Bonferroni; other 95% intervals are descriptive. No remedy setting is selected or retuned using these outcomes. The archive preserves source/input hashes, all outcomes and independent numerical checks.

Table 28: Additional GMM stabilizer outcomes: base-only absolute success and within-adapter perturbed-minus-clean effects, with descriptive crossed 95% intervals.

Adapter	Base (%)	D (pp)	Base interval	D interval
Learned	28.88	$+9.70$	$[+23.25, +34.13]$	$[+6.43, +13.09]$
Fixed 0.15	65.62	-2.18	$[+59.25, +72.12]$	$[-4.99, +0.31]$
Floor 0.05	45.38	$+6.66$	$[+39.12, +51.75]$	$[+3.24, +10.40]$
Log-scale penalty	29.88	$+11.36$	$[+24.12, +35.38]$	$[+7.74, +15.33]$
Detached scale path	39.88	$+12.63$	$[+34.38, +45.12]$	$[+9.10, +16.42]$

Q DETERMINISTIC MATCHED-GROUP PROTOCOL

The exact matched groups, source arrays and target arrays are shared with the GMM experiment. The deterministic source cohort comprises seeds 23–32, distinct from the GMM cohort. All 32 group branches and base-only continuation at state 1,000/repeat zero are retained. We reuse 1,320 verified MSE adapters and train 1,320 full Huber adapters; both receive fresh Adam at 3×10^{-4}, 250 updates, batch 256 and first-ten target demonstrations with the identical batch stream. Twice Huber with threshold 0.1 matches MSE's local curvature. The independent freeze verifies metadata and a full replay of an MSE adapter before reuse. Paired evaluation covers 2,640 policies and 52,800 episodes with 20 scenes per target (seed base 1,102,000), horizon 200. Crossed bootstrap uses 10,000 draws and seed 820,020, keeping target identities fixed. The primary is the change in perturbed-minus-clean success; all other contrasts are descriptive.

Table 29: Deterministic matched-group outcomes. Absolute success (%) is shown in clean/noisy/base order; the Huber-minus-MSE interaction is in pp. Macro is primary; task and distance intervals are descriptive. Window is at ceiling in every branch.

Task / stratum	MSE: C / N / B	Huber: C / N / B	Interaction [95% interval]
Reach	29.78 / 30.19 / 30.00	38.09 / 37.50 / 37.00	$-1.00\ [-2.53, +0.41]$
Push	90.12 / 90.00 / 89.50	92.28 / 92.25 / 92.00	$+0.09\ [-1.19, +1.00]$
Pick-place	71.94 / 72.09 / 73.00	69.38 / 69.38 / 69.50	$-0.16\ [-1.44, +1.19]$
Window	100.00 / 100.00 / 100.00	100.00 / 100.00 / 100.00	$+0.00\ [+0.00, +0.00]$
Macro	72.96 / 73.07 / 73.12	74.94 / 74.78 / 74.62	$-0.27\ [-0.96, +0.36]$
Distance 0	—	—	$-0.16\ [-0.97, +0.58]$
Distance 1	—	—	$-0.38\ [-1.34, +0.52]$