WHAT MUST BE LEARNED TO ADAPT? DECISION-DIRECTED IDENTIFICATION UNDER ENVIRONMENT SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretraining is useful only when the knowledge it supplies rules out alternatives that would change the downstream decision. Yet adaptation is commonly described by overall source–target distance or by how much uncertainty pretraining removes, quantities that can be dominated by variation irrelevant to acting well. We formulate post-shift adaptation as *decision-directed identification*: after observing a source artifact, an agent should distinguish only target hypotheses that require incompatible decisions. This formulation yields a residual information radius and an instance-wise lower bound on target interaction. We operationalize the principle with Decision-Directed Adaptation (DDA), which probes decision-conflicting alternatives and stops once all plausible targets share an ϵ-good decision. Controlled experiments show that DDA can require substantially fewer probes than random exploration, full system identification, and decision-agnostic information gain. They also show that environmental distance can reverse the true ordering of adaptation difficulty and that retaining abundant nuisance information may provide little decision value. Robot experiments across seven Panda-to-XArm PickCube shifts and more than 200,000 held-out episodes expose the same practical structure: zero-shot transfer ranges from sufficient to ineffective, scratch learning and fine-tuning exchange advantage across shifts, and action history resolves an important latency-induced information deficiency. The resulting perspective turns transfer from a vague similarity judgment into a concrete question: which decision conflicts remain, and what evidence is needed to eliminate them?

1 INTRODUCTION

A pretrained agent deployed in a new environment may encounter different dynamics, observations, embodiment, or control latency. The usual question is whether the source and target are "similar." Yet similarity is not the same as usefulness. A large change in a nuisance variable may leave the optimal policy unchanged, whereas even a small change near a decision boundary can change the optimal policy. Likewise, a source artifact may accurately reveal many properties of a target while failing to reveal the one property that changes the required action.

A robot example makes the distinction concrete. A source policy may tolerate a noticeable rescaling of end-effector commands because the same closed-loop strategy remains successful, so adaptation is unnecessary at a declared success threshold. A seemingly modest action delay is different: the current observation no longer identifies which earlier command is being executed, and the policy lacks information needed to choose its next action. The first shift is large at the interface yet decision benign; the second creates an information-state deficiency. Treating both only as "domain distance" hides the mechanism that determines the required target interaction.

This paper asks a sharper question: *what must still be learned before the agent can make a good decision?* Our answer separates environment uncertainty from decision uncertainty. Two target models need not be distinguished merely because their parameters differ. They need to be distinguished when they admit no common ϵ-good decision. Source knowledge is useful to the extent that it removes such conflicts or makes the surviving conflicts easy to probe.

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

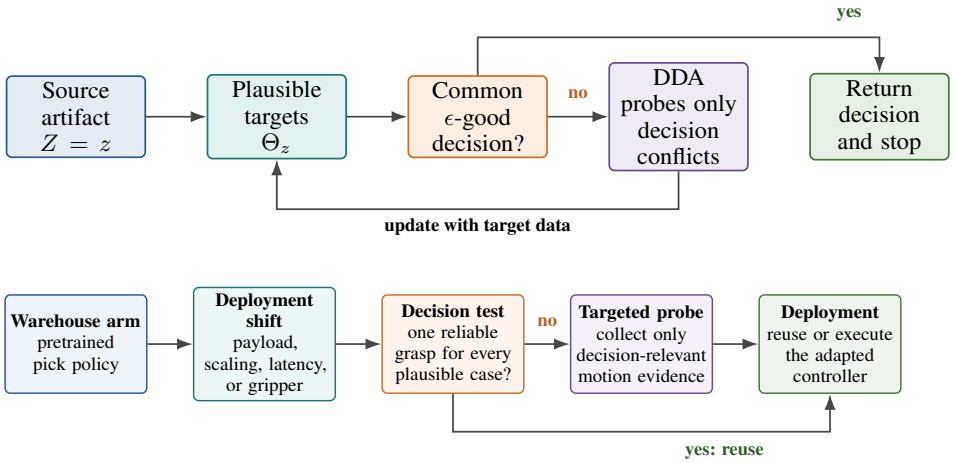

Figure 1: Decision-directed adaptation. The source first restricts which targets remain plausible. The learner ignores residual differences that share a good decision, probes only decision-conflicting alternatives, and stops before full environment identification. The lower boxed workflow instantiates the same logic for warehouse manipulation.

Figure 1 summarizes the decision-directed loop and its warehouse-manipulation interpretation. This question is increasingly important because modern pretrained systems are expected to serve many downstream tasks and robot embodiments. Recent theory proves substantial transfer gains in single-index models and relates success to feature-space overlap in high-dimensional regression (Jones-Mccormick et al., 2025; Tahir et al., 2025). Recent robot systems transfer across viewpoints, platforms, and even manipulation, navigation, locomotion, and aviation embodiments (Chen et al., 2025; Doshi et al., 2025). These advances show that reuse is possible, but they do not by themselves reveal whether a particular source has removed the uncertainty that changes the downstream decision, or how many targeted observations remain necessary. A framework for that missing quantity is valuable both for theory and for deciding when a deployed agent should act immediately, adapt, or collect a diagnostic probe.

This view connects sequential experimental design (Chernoff, 1959; Garivier & Kaufmann, 2016), transfer learning (Taylor & Stone, 2009; Rakelly et al., 2019), and decision-aware modeling (Ferns et al., 2004; Grimm et al., 2020). It is also related to the Decision-Estimation Coefficient (DEC), which characterizes general interactive decision problems (Foster et al., 2021). Our focus is deliberately narrower and conditional: a fixed downstream decision problem, an explicit source artifact, and the target experiments remaining after an environment shift. The information-radius formula itself is classical in spirit; our contribution is to make *post-source decision conflict* the organizing object for transfer and to test its consequences from exact finite problems to robot control. This positioning matters for falsifiability: the framework predicts zero interaction when all surviving targets share a good decision, and it predicts difficulty when only hard-to-probe conflicts remain, regardless of conventional distance.

We make four contributions: First, we formulate decision-directed identification under shift and define a residual information radius conditioned on source knowledge. Second, we give an instance-wise change-of-measure lower bound, a zero-interaction condition, and a monotonicity result. Third, we give DDA, a likelihood-based method that ignores model distinctions irrelevant to the final decision and uses a common-good-decision stopping rule. Fourth, we evaluate the framework in controlled finite and LQR problems and conduct a seed-aware ManiSkill robot study.

2 Decision-directed identification

Let Θ denote a family of possible target environments and let $\theta \in \Theta$ be the unknown environment actually encountered after the shift. Before collecting target data, the learner observes a source artifact $Z = z$. Here, Z is the random information produced by pretraining and z is its observed value; it may contain pretrained weights, source trajectories, a simulator, or a learned representation. In the

hard-information case, z leaves the compatible set $\Theta_z = \Gamma(z) \subseteq \Theta$, where Γ maps the artifact to all targets that remain possible. Soft source information plays the same role through a prior or initial likelihood over Θ.

Target interaction proceeds sequentially. At round t, the learner selects an experiment a_t from the experiment set $\mathcal{A}$ and observes an outcome Y_t drawn from $P_\theta^{a_t}$, the observation distribution of experiment a_t under target θ. An experiment can be a diagnostic query, an exploratory control input, or a short robot rollout. The learner stops after the random number τ of experiments and returns a decision $\hat{d}$ from the decision set $\mathcal{D}$. For any candidate decision d, $L_\theta(d)$ is its loss in target θ, and

$$L_\theta^\star = \min_{d' \in \mathcal{D}} L_\theta(d'), \qquad \mathcal{D}_\epsilon(\theta) = \{d \in \mathcal{D} : L_\theta(d) \le L_\theta^\star + \epsilon\} \tag{1}$$

define the optimal loss $L_\theta^\star$ and the set $\mathcal{D}_\epsilon(\theta)$ of ϵ-good decisions. The nonnegative tolerance ϵ states how much suboptimality is acceptable; when $\epsilon = 0$, only optimal decisions qualify. The goal is not necessarily to recover θ, but to return some $\hat{d} \in \mathcal{D}_\epsilon(\theta)$ with probability at least $1 - \delta$, where $\delta \in (0, 1/2)$ is the allowed failure probability.

Decision conflict: We write $\theta \perp_\epsilon \theta'$ when $\mathcal{D}_\epsilon(\theta) \cap \mathcal{D}_\epsilon(\theta') = \varnothing$. Statistical differences that preserve a common good decision create no identification obligation: the learner can act well without deciding which model is true. Conversely, $\theta \perp_\epsilon \theta'$ means that no decision is acceptable in both targets, so a correct learner must eventually distinguish them. The relevant alternatives after observing z are

$$\mathrm{Alt}_\epsilon(\theta; z) = \{\theta' \in \Theta_z : \theta \perp_\epsilon \theta'\}. \tag{2}$$

Thus $\mathrm{Alt}_\epsilon(\theta; z)$ contains only candidates θ' that survive the source artifact and require decisions incompatible with those good for θ.

Residual information radius: For a distribution w over experiments, the worst separation rate from a decision-conflicting alternative is

$$\mathcal{I}_\epsilon(\theta \mid z) = \sup_{w \in \Delta(\mathcal{A})} \inf_{\theta' \in \mathrm{Alt}_\epsilon(\theta; z)} \sum_{a \in \mathcal{A}} w(a) D_{\mathrm{KL}}(P_\theta^a \| P_{\theta'}^a). \tag{3}$$

Here, $\Delta(\mathcal{A})$ is the probability simplex over experiments and $w(a)$ is the fraction allocated to experiment a. The divergence $D_{\mathrm{KL}}(P\|Q)$ is the Kullback–Leibler (KL) divergence, which quantifies how well one observation from P separates it from Q. The weighted sum is an average information rate. The infimum selects the hardest surviving conflict, while the supremum selects the allocation that makes this hardest conflict as distinguishable as possible. We call $\mathfrak{R}_\epsilon(\theta \mid z) = 1/\mathcal{I}_\epsilon(\theta \mid z)$ the *residual decision complexity*, with $1/\infty = 0$. A large information radius means rapid discrimination and therefore small residual complexity. Unlike a geometric source–target metric, this quantity depends jointly on source knowledge, decision tolerance, available experiments, and downstream loss.

Proposition 1 (Decision-conditioned lower bound)**:** *Assume a finite target family and mutually absolutely continuous observation laws, meaning that an event cannot be impossible under one candidate but possible under another. Any algorithm satisfying* $\Pr_\vartheta[\hat{d} \in \mathcal{D}_\epsilon(\vartheta)] \ge 1 - \delta$ *for every* $\vartheta \in \Theta_z$*, with* $\delta < 1/2$*, obeys*

$$\mathbb{E}_\theta[\tau] \ge \frac{\mathrm{kl}(1-\delta, \delta)}{\mathcal{I}_\epsilon(\theta \mid z)} \qquad \textit{for every } \theta \in \Theta_z. \tag{4}$$

Here, $\Pr_\vartheta$ and $\mathbb{E}_\vartheta$ denote probability and expectation when ϑ is the true target, and $\mathrm{kl}(p, q) = p \log(p/q) + (1-p)\log((1-p)/(1-q))$ is binary relative entropy. The proof (Appendix A) applies the standard adaptive change-of-measure argument to each conflicting alternative. The numerator is the evidence required to make the error probability at most δ; the denominator is the greatest evidence obtainable per experiment against the hardest relevant alternative. Target data are therefore necessary only for alternatives that both survive the source artifact and demand incompatible decisions.

Corollary 1 (Decision sufficiency and monotonicity)**:** *If* $\bigcap_{\theta \in \Theta_z} \mathcal{D}_\epsilon(\theta) \neq \varnothing$*, zero target observations suffice. Moreover, if* $\Theta_{z_1} \subseteq \Theta_{z_2}$ *then* $\mathfrak{R}_\epsilon(\theta \mid z_1) \le \mathfrak{R}_\epsilon(\theta \mid z_2)$ *for every* $\theta \in \Theta_{z_1}$.

The first statement is operationally important: a member of the intersection is already acceptable for every compatible target. The second statement compares artifacts z_1 and z_2: if z_1 rules out at least as many targets as z_2, it cannot make the remaining decision problem harder. Continuing to fine-tune a decision-sufficient source policy may still alter performance, but that measures optimization stability rather than information required for transfer.

3 DECISION-DIRECTED ADAPTATION

DDA instantiates the formulation with likelihood confidence sets. Let $p_\theta(y \mid a)$ be the likelihood of outcome y after experiment a in target θ, and let $\ell_0(\theta; z)$ encode the source artifact. After t target observations, the cumulative log likelihood is $\ell_t(\theta) = \ell_0(\theta; z) + \sum_{s=1}^{t} \log p_\theta(Y_s \mid a_s)$. We retain

$$\mathcal{C}_t = \left\{ \theta : \ell_t(\theta) \geq \max_{\theta'} \ell_t(\theta') - \beta_t \right\}, \quad \beta_t = \log \frac{2|\Theta|(t+1)^2}{\delta}. \tag{5}$$

where $\mathcal{C}_t$ is the set of targets still statistically plausible, $|\Theta|$ is the number of candidate targets, and β_t is a confidence threshold calibrated using δ. DDA stops when $\bigcap_{\theta \in \mathcal{C}_t} \mathcal{D}_\epsilon(\theta)$ is nonempty. If the true target $\theta^\star$ remains in $\mathcal{C}_t$, every decision in this intersection is ϵ-good for the truth even if the environment itself is not fully identified. When no common decision exists, DDA selects the experiment that maximizes the weakest symmetric KL separation among conflicting plausible pairs:

$$a_t \in \arg\max_a \min_{\substack{\theta, \theta' \in \mathcal{C}_t \\ \theta \perp_\epsilon \theta'}} \frac{D_{\mathrm{KL}}(P_\theta^a \| P_{\theta'}^a) + D_{\mathrm{KL}}(P_{\theta'}^a \| P_\theta^a)}{2}. \tag{6}$$

The operator $\arg\max$ returns an experiment attaining the largest score; the inner minimum protects against the hardest current conflict. Averaging the two KL directions makes the pairwise score symmetric. Equation (6) is a tractable one-step surrogate for the globally optimized allocation in Equation (3); we do not claim universal instance optimality. Algorithm 1 gives the pseudocode in a formal algorithm environment. The comparison methods are random probing, round robin, full system identification that separates every plausible pair, and a posterior-weighted total-information rule. All methods share the likelihood update, confidence set, and stopping rule, so the comparison isolates experiment selection.

4 RELATION TO PRIOR WORK

Active testing and pure exploration: Sequential design already allocates measurements to separate alternatives (Chernoff, 1959), and fixed-confidence best-arm identification derives instance-dependent KL lower bounds and matching strategies (Garivier & Kaufmann, 2016; Kaufmann et al., 2016). Information-directed sampling targets decision regret rather than raw entropy (Russo & Van Roy, 2014). These lines of work establish the importance of choosing measurements according to what remains statistically ambiguous. Our formulation uses the same change-of-measure machinery but changes the alternative set twice: first by conditioning on the source artifact and then by removing targets that share an ϵ-good decision. Consequently, two statistically different models may remain unresolved by design, rather than because the learner failed to identify them.

Interactive complexity and decision-aware models: DEC provides a broad minimax characterization of interactive decision making (Foster et al., 2021; 2023). Value equivalence asks when learned models are interchangeable for planning (Grimm et al., 2020); bisimulation metrics compare MDP states through reward and transition behavior (Ferns et al., 2004). Most recently, generalized bisimulation extends this idea across MDPs and derives policy-transfer and metric-estimation guarantees (Tao et al., 2025). Such metrics quantify structural similarity between environments; they do not condition the required target-data budget on a source artifact, an acceptable-decision tolerance, and the experiments available to the learner. We instead ask how many *new target observations* are needed after an explicit piece of side information. Value equivalence describes when models support the same planning computations; our radius quantifies how difficult it is to establish enough such equivalence from target data. DEC addresses much broader classes of interactive problems and minimax regret. Equation (3) is therefore best viewed as a conditional, fixed-confidence specialization that makes transfer and remaining data acquisition explicit, not as a replacement for DEC, bisimulation, or value equivalence.

Transfer and adaptation: Transfer RL studies when prior tasks improve a target task (Taylor & Stone, 2009; Zhu et al., 2020); meta-RL methods infer latent task context from experience (Finn et al., 2017; Rakelly et al., 2019); domain randomization seeks robust policies by varying training conditions (Tobin et al., 2017). Recent statistical theory proves polynomial or exponential benefits of pretraining in single-index models (Jones-Mccormick et al., 2025) and characterizes transfer

through feature overlap in deep linear regression (Tahir et al., 2025). These results explain when particular representations or training dynamics help, but they concern static prediction models rather than adaptive experiment selection for a downstream decision. In robotics, RoVi-Aug synthesizes robot and viewpoint variation, and CrossFormer trains one policy across highly heterogeneous embodiments (Chen et al., 2025; Doshi et al., 2025). They establish impressive empirical transfer, but do not define the decision-relevant uncertainty left for a new target or the interaction cost of resolving it. Earlier and recent approaches therefore typically specify how to construct a reusable learner, policy, metric, or latent state. Decision-directed identification supplies an orthogonal diagnostic: given any such artifact, it asks which target distinctions can still change the acceptable decision and which interventions reveal them. This perspective also separates two effects often conflated in empirical transfer studies. A source policy may provide a favorable optimization initialization without eliminating statistical ambiguity, or it may already be decision sufficient and require zero target interaction even if subsequent fine-tuning is unstable.

5 CONTROLLED EXPERIMENTS

We first isolate the statistical mechanism in models where the hypothesis, experiment, and decision sets are known exactly. All methods use the same likelihood update, confidence rule, and stopping rule; only the probe selection changes. Results aggregate 20 random seeds, 50 repetitions per target, and all target hypotheses. Table 1 compares the no-artifact case. Bold entries mark the smallest mean interaction count in each problem; the advantage is not created by an easier stopping test because that test is identical across methods.

Finite nuisance problem: There are 16 hypotheses, two decisions, and three Bernoulli probes: one reveals the nuisance, one the decision label, and one mixes both. With no artifact, DDA requires 26.62 probes on average, compared with 45.17 (random), 45.31 (round robin), 79.42 (system ID), and 240.29 (total information). Thus a probe can be highly informative about the complete parameter θ yet poor for deciding: total information repeatedly queries the easy eight-valued nuisance, while DDA concentrates on the binary distinction that changes the decision. Relative to random probing, DDA uses about 41% fewer interactions; relative to full system identification it uses about 66% fewer. Exact decision information yields immediate stopping, whereas exact nuisance information reduces DDA only modestly, from 26.62 to 24.11 probes. The contrast shows that the amount of uncertainty removed by a source and the usefulness of that removal are different quantities.

Linear control: We cross three scalar dynamics values that determine the optimal LQR feedback gain with four nuisance values and offer control-relevant, nuisance-only, and mixed Gaussian excitations. At $\epsilon = 0.01$, DDA requires 8.20 probes, versus 9.36 (total information), 20.93 (round robin), 22.33 (random), and 40.66 (system ID). Total information is closer here because the mixed probe happens to carry useful control information, but DDA still gives the smallest mean. Learning the nuisance exactly does not change DDA's 8.20-probe mean; learning the control-relevant coordinate makes target probing unnecessary. The experiment connects the finite construction to a recognizable control objective: a learner need not reconstruct every dynamics parameter when several models induce the same acceptable feedback controller.

Distance inversion and representation failure: Table 2 isolates two failures of common transfer heuristics. The parameter-near target straddles a policy boundary and requires 152.86 DDA probes, whereas the parameter-far target changes only a nuisance coordinate and requires none. In the representation construction, retaining the decision feature yields 1.23 probes and 99.98% correctness, whereas retaining only the nuisance reaches the 100-probe budget and remains at chance.

These are controlled constructions, not universal claims about every learned representation. Their logical role is sharper: they disprove the sufficiency of parameter distance or retained information volume as general descriptions of adaptation difficulty. A transfer theory must also ask whether the remaining uncertainty can change an acceptable decision.

Table 1: Mean target probes without a source artifact. Lower is better; the best mean in each row is bold.

Problem	DDA	Random	Round robin	System ID	Total info
Finite nuisance	**26.62**	45.17	45.31	79.42	240.29
Linear control	**8.20**	22.33	20.93	40.66	9.36

Table 2: Structural counterexamples. Bold values expose the reversal: greater distance can be free, while retaining nuisance information can be useless.

Construction	Condition	Target probes	Correctness
Distance	close, distance 0.05	152.86	99.98%
Distance	far, distance 10.05	**0.00**	**100.00%**
Representation	decision feature retained	**1.23**	**99.98%**
Representation	nuisance feature only	100.00 (budget)	50.00%

6 Robot adaptation under environment shifts

Protocol: We use state-based PickCube-v1 in ManiSkill3 (Tao et al., 2024) and transfer two Panda source policies to XArm6-Robotiq. A canonical policy interface exposes six Cartesian end-effector deltas and one gripper command. The XArm gripper is mapped affinely from the shared $[-1, 1]$ convention to its native $[0, 0.81]$ range; interface, reset-isolation, action-history, and checkpoint-resume gates are verified before training. Policies are trained with PPO (Schulman et al., 2017) for at most 9,994,240 target interactions. We study seven shifts: action scaling, observation noise, observation dropout, latency of two or four steps, a combined hard shift, and a randomized interface.

For each feed-forward condition we train three scratch policies and six fine-tuned policies (two source seeds by three adaptation seeds). Latency conditions additionally use a four-action-history observation, with the source network expanded by zero-padding so its update-0 function is unchanged. Each selected best and final checkpoint is evaluated on 1,024 fresh episodes. In total, the robot experiments contain 204 evaluation jobs and 208,896 episodes. Following recommendations for few-run RL evaluation (Agarwal et al., 2021), the independently trained policy is the unit of analysis; held-out episodes measure a fixed policy and are not treated as additional training seeds. Table 3 reports held-out means in two panels: feed-forward policies for all shifts and History-4 policies for the two latency shifts where memory directly addresses delayed actions. Update 0 is zero-shot transfer from two source policies, feed-forward scratch uses three policies, and feed-forward fine-tuning uses six policies. Bold marks the largest observed mean for each shift across the methods evaluated, not a significance claim.

Table 3: Panda-to-XArm PickCube success. Entries are policy means on 1,024 held-out episodes. History-4 was evaluated only for latency and is therefore reported in a separate panel rather than as missing entries. The largest observed mean per shift is bold.

Panel A: Feed-forward policies (all shifts)

Shift	Update 0	Scratch	Fine-tune
Action scale 0.75	.939	**1.000**	.999
Observation noise	.870	**.980**	.960
Observation dropout	.631	**.962**	.910
Latency 2	.112	.665	.818
Combined hard	.125	.215	**.730**
Randomized interface	.135	.099	**.158**
Latency 4	.013	.016	.034

Panel B: Four-action-history policies (latency shifts)

Shift	Update 0	History-4 scratch	History-4 fine-tune
Latency 2	.114	.894	**.928**
Latency 4	.015	**.527**	.269

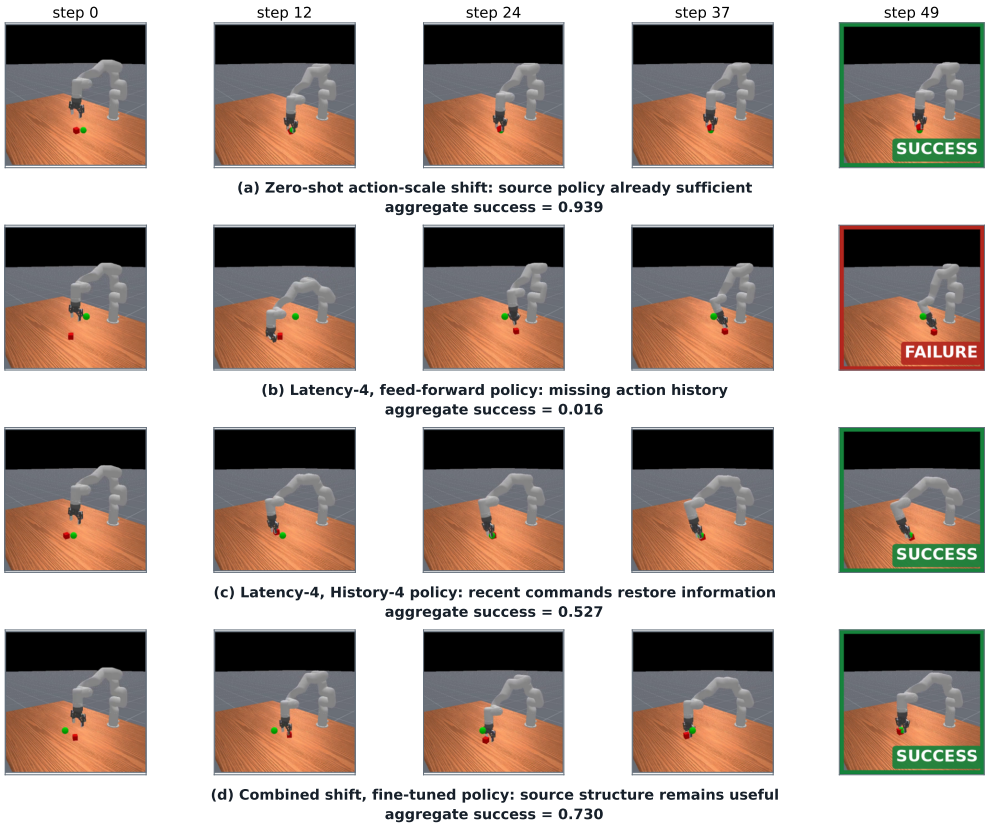

Figure 2: Representative deterministic held-out PickCube episodes: (a) a zero-shot action-scale shift, (b) latency 4 with a feed-forward policy, (c) latency 4 with History-4, and (d) fine-tuning under the combined shift. Each row uses the first evaluation seed in a prespecified range with the indicated success or failure outcome. The values beneath the strips are aggregate means from Table 3, not estimates from the displayed episode. The contrast between (b) and (c) illustrates how recent action history can restore information hidden from a feed-forward policy.

Shift structure dominates a single transfer verdict: Table 3 shows a span from essentially solved at update 0 (action-scale mean 0.939) to nearly impossible for the feed-forward policy (latency-4 mean 0.013). Scratch reaches 0.962 under observation dropout and 0.980 under observation noise; fine-tuning reaches 0.910 and 0.960, respectively. Thus pretraining is not uniformly beneficial even when the source policy is competent: for both observation corruptions, scratch training obtains the larger mean. On the combined shift, fine-tuning has a much larger mean than scratch (0.730 versus 0.215), suggesting that source structure can matter when several perturbations co-occur. Yet policy-level intervals are wide ($[0.428, 1.000]$ for fine-tuning and $[0, 1]$ for scratch) because only three matched adaptation seeds are available. We therefore treat this as evidence of heterogeneous transfer behavior, not definitive method superiority. The randomized interface remains difficult for both (0.158 fine-tuning and 0.099 scratch), showing that neither initialization alone nor additional PPO updates necessarily resolves a continually changing control interface.

The operational answer also depends on the required success threshold. The two action-scale update-0 policies score 0.966 and 0.913: both are sufficient at a 0.90 requirement, but only one at 0.95. A single source–target label therefore cannot determine an interaction count without a declared decision tolerance. This is the empirical counterpart of ϵ in Equation (1): changing what counts as "good enough" changes whether the source is already sufficient and, consequently, whether target adaptation is needed at all.

Latency is partly an information-state problem: With feed-forward observations, latency 2 reaches 0.665 from scratch and 0.818 after fine-tuning, while latency 4 remains near zero. Four-action history

raises latency-2 means to 0.894 and 0.928, and latency-4 means to 0.527 and 0.269 (Table 3 and Figure 2). History converts latency 4 from nearly universal failure into occasional high success, although the policy-seed intervals remain wide: $[0, 1]$ for scratch and $[0, 0.539]$ for fine-tuning. The effect is nevertheless mechanistically informative. A feed-forward observation cannot reveal which command is currently being executed after a delay, whereas recent action history supplies part of the missing information state. Adaptation failure can therefore reflect insufficient observability, not simply an optimizer that needs more samples. The learning-rate ablation in Appendix D reinforces this point: smaller updates are not a universal substitute for supplying the information needed to make the decision process effectively Markovian.

7 Discussion and limitations

The controlled results support the paper's central mechanism: full environment identification can spend samples on nuisance variables, geometric distance can misorder adaptation difficulty, and decision-relevant source information can eliminate target sampling. This matters operationally because a transfer benchmark should distinguish three events that final return alone conflates: the source is already acceptable, target interaction supplies missing decision-relevant information, or optimization merely changes a competent policy. Reporting update-0 performance, a declared success tolerance, and the interactions needed to cross that tolerance makes these cases visible. The robot results further show that the required information can depend on the shift: under latency, action history is more consequential than simply choosing scratch training or fine-tuning.

The paper has three limitations: Our lower bound assumes a finite, well-specified target family, and the greedy DDA rule is not claimed to attain it. The robot evidence is restricted to state-based PickCube, Panda-to-XArm transfer, two source policies, and three adaptation seeds, so wide policy-level intervals preclude broad method rankings. Most importantly, we neither estimate $\mathcal{I}_\epsilon$ from ManiSkill rollouts nor compare DDA with probing baselines there. The robot study therefore supports the qualitative distinctions among decision sufficiency, observability, and optimization, but not the stronger claim that an empirical residual radius predicts robotic adaptation cost.

8 Conclusion

This paper replaces the broad question "Are source and target similar?" with the operational question "What must still be learned to act well?" Our formulation answers by conditioning the target alternatives on source knowledge and retaining only pairs that require incompatible ϵ-good decisions. The resulting residual information radius connects those surviving conflicts to the target interactions that any successful learner must collect, while DDA turns the same principle into a probe-selection and stopping procedure.

The experiments show why this change of viewpoint matters. In controlled problems, decision-directed probing avoids spending samples on nuisance identification, environmental distance can rank adaptation difficulty in the wrong order, and a decision-sufficient source can eliminate target interaction altogether. The robot experiments reveal complementary practical consequences: the value of pretraining depends strongly on the shift, scratch learning and fine-tuning are not uniformly ordered, and delayed control can require a richer information state rather than merely more optimization.

The central message is therefore simple: successful adaptation does not require learning everything that changed. It requires learning exactly enough to rule out the plausible alternatives that would change the decision. Making the *decision tolerance*, *surviving conflicts*, and *informative experiments* explicit offers a sharper foundation for measuring transfer and designing agents that know when to act, when to probe, and when to stop adapting.

AI use statement

Generative AI tools were used to assist with writing this paper to improve the readability and with generating portions of the research code. The authors reviewed and verified all AI-assisted text and code and take responsibility for the final content.

Ethics statement

The empirical study uses simulated robots and contains no human subjects, personal data, or sensitive datasets. Decision-directed adaptation could improve the efficiency of deployed autonomous systems; as with other robot learning methods, unsafe deployment without task-specific validation is a potential misuse. Our experiments are limited to simulation and do not claim real-world safety.

Reproducibility statement

The experiment archive records the implementation revision, resolved configurations, random seeds, package versions, source checksums, run manifests, validation curves, best/final checkpoint hashes, and the 204-job held-out evaluation inventory. Appendices C–E specify the synthetic models, robot protocol, aggregation unit, uncertainty analysis, and verification checks. Every result table in the submission package is traceable to its source CSV and validated aggregation output.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. RoVi-Aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 209–233, 2025.

Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3): 755–770, 1959. doi: 10.1214/aoms/1177706205.

Ria Doshi, Homer Rich Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 496–512, 2025.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169, 2004.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1126–1135, 2017.

Dylan J. Foster, Sham M. Kakade, Jian Qian, and Alexander Rakhlin. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Dylan J. Foster, Noah Golowich, Jian Qian, Alexander Rakhlin, and Ayush Sekhari. Model-free reinforcement learning with the decision-estimation coefficient. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Taj Jones-Mccormick, Aukosh Jagannath, and Subhabrata Sen. Provable benefits of unsupervised pre-training and transfer learning via single-index models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28350–28376, 2025.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340, 2019.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

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

Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: A theory of transfer learning in high-dimensional regression. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58142–58168, 2025.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: GPU parallelized robotics simulation and rendering for generalizable embodied AI. *arXiv preprint arXiv:2410.00425*, 2024.

Zhenyu Tao, Wei Xu, and Xiaohu You. A generalized bisimulation metric of state similarity between markov decision processes: From theoretical propositions to applications. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Matthew E. Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685, 2009.

Joshua Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Zhuangdi Zhu, Kaixiang Lin, Anil K. Jain, and Jiayu Zhou. Transfer learning in deep reinforcement learning: A survey. *arXiv preprint arXiv:2009.07888*, 2020.

A Proofs

A.1 Proof of Proposition 1

Fix $\theta \in \Theta_z$ and any $\theta' \in \mathrm{Alt}_\epsilon(\theta; z)$. Because the two ϵ-good decision sets are disjoint, the event $E = \{\hat{d} \in \mathcal{D}_\epsilon(\theta)\}$ has $\Pr_\theta(E) \geq 1 - \delta$ and $\Pr_{\theta'}(E) \leq \delta$. The adaptive transportation lemma for sequential experiments gives

$$\sum_{a \in \mathcal{A}} \mathbb{E}_\theta[N_a(\tau)] D_{\mathrm{KL}}(P^a_\theta \| P^a_{\theta'}) \geq \mathrm{kl}(1-\delta, \delta), \tag{7}$$

where $N_a(\tau)$ is the number of times experiment a is selected. If $\mathbb{E}_\theta[\tau] = 0$, no conflicting alternative can exist and the result is immediate. Otherwise set $w(a) = \mathbb{E}_\theta[N_a(\tau)]/\mathbb{E}_\theta[\tau]$. Taking the infimum over conflicting θ' and using the definition of $\mathcal{I}_\epsilon$ yields

$$\mathrm{kl}(1-\delta, \delta) \leq \mathbb{E}_\theta[\tau] \inf_{\theta' \in \mathrm{Alt}_\epsilon(\theta; z)} \sum_a w(a) D_{\mathrm{KL}}(P^a_\theta \| P^a_{\theta'}) \tag{8}$$

$$\leq \mathbb{E}_\theta[\tau] \mathcal{I}_\epsilon(\theta \mid z), \tag{9}$$

which proves Equation (4).

A.2 Proof of Corollary 1

If the intersection of good-decision sets is nonempty, any element of it is ϵ-good for every compatible target, so no observation is required. For monotonicity, $\Theta_{z_1} \subseteq \Theta_{z_2}$ implies $\mathrm{Alt}_\epsilon(\theta; z_1) \subseteq \mathrm{Alt}_\epsilon(\theta; z_2)$. For each allocation w, the infimum over the smaller set is no smaller. Taking the supremum preserves the inequality, so $\mathcal{I}_\epsilon(\theta \mid z_1) \geq \mathcal{I}_\epsilon(\theta \mid z_2)$ and the reciprocal complexities have the claimed order.

A.3 Validity of the stopping rule

Suppose the confidence sequence satisfies $\Pr_\theta(\theta \in \mathcal{C}_t \ \forall t) \geq 1 - \delta$. At stopping, DDA returns $d \in \cap_{\vartheta \in \mathcal{C}_\tau} \mathcal{D}_\epsilon(\vartheta)$. On the confidence event, $\theta \in \mathcal{C}_\tau$, hence $d \in \mathcal{D}_\epsilon(\theta)$. Therefore the stopping rule is (ϵ, δ)-correct. This statement separates correctness of stopping from efficiency of the greedy selection rule.

B Algorithmic details

Algorithm 1 Decision-Directed Adaptation (DDA)

Input: target family Θ; experiment set $\mathcal{A}$; observation models $\{P_\theta^a\}$; loss functions $\{L_\theta\}$; source artifact z; tolerances ϵ and δ.
Initialize: set $t \leftarrow 0$ and initialize $\ell_0(\theta; z)$ for every $\theta \in \Theta$ from the source artifact.

1. Construct the plausible-target set $\mathcal{C}_t$ using Equation (5).
2. Form the common-good-decision set $G_t = \bigcap_{\theta \in \mathcal{C}_t} \mathcal{D}_\epsilon(\theta)$.
3. If $G_t \neq \varnothing$, return a decision $\hat{d} \in G_t$ and stopping time $\tau = t$.
4. Otherwise select probe a_t using the conflict-restricted criterion in Equation (6).
5. Observe Y_t, update every candidate by $\ell_{t+1}(\theta) = \ell_t(\theta) + \log p_\theta(Y_t \mid a_t)$, set $t \leftarrow t + 1$, and repeat.

Output: an ϵ-good decision on the confidence event and the number τ of target experiments used.

The system-ID baseline replaces the conflict-restricted pair set with all plausible pairs. Total information averages symmetric KL over all plausible pairs using products of posterior weights. Random samples uniformly, and round robin selects the least-used experiment. All methods share the same confidence and stopping rules, isolating the value of experiment selection.

C Controlled experiment specification

Finite model: Each hypothesis is a pair (b, n). The binary variable $b \in \{0, 1\}$ is the decision-relevant label, while $n \in \{0, \ldots, 7\}$ is an eight-valued nuisance that does not change the optimal decision. Choosing decision b gives zero loss and choosing $1 - b$ gives unit loss. Each probe returns a Bernoulli random variable, so its displayed mean is the probability of observing one. The nuisance probe has mean $0.05 + 0.90n/7$; the decision probe has mean $0.25 + 0.50b + 0.04(n/7 - 0.5)$; and the mixed probe has mean $0.15 + 0.45b + 0.30n/7$, clipped to $[0.01, 0.99]$. We use $\epsilon = 0$ (the returned decision must be optimal), $\delta = 0.05$ (at most five-percent error), and a 250-probe budget.

Linear control: Each target is (A, n), where the scalar dynamics coefficient $A \in \{0.65, 0.90, 1.10\}$ changes the optimal controller and the nuisance $n \in \{-0.90, -0.30, 0.30, 0.90\}$ does not. The system evolves as $x_{t+1} = Ax_t + u_t + \xi_t$: x_t is the scalar state, u_t is the control input, and ξ_t is process noise. Decisions are infinite-horizon scalar linear quadratic regulator (LQR) gains with state-cost weight $Q = 1$ and action-cost weight $R = 0.1$; loss is normalized excess LQR cost above the target-optimal gain. Gaussian probes observe noisy versions of A, n, or the mixture $0.65A + 0.35n$, with variances 0.015, 0.003, and 0.025, respectively. We use $\epsilon = 0.01$, $\delta = 0.05$, and a 250-probe budget.

Table 4 reports every source-artifact condition for both controlled problems.

Table 4: Mean target probes in the controlled experiments. "Dec." and "nuis." denote source information about the decision-relevant and nuisance coordinates.

Problem	Artifact	DDA	Total info	System ID	Random	Round robin
Finite	none	26.62	240.29	79.42	45.17	45.31
Finite	nuisance known	24.11	24.11	24.11	42.12	42.74
Finite	noisy nuisance	24.11	181.05	29.03	42.15	42.74
Finite	decision known	0	0	0	0	0
Linear	none	8.20	9.36	40.66	22.33	20.93
Linear	nuisance known	8.20	8.20	8.20	22.30	20.93
Linear	noisy nuisance	8.20	9.34	16.79	22.30	20.93
Linear	decision known	0	0	0	0	0

Representation construction: Four hypotheses cross a binary decision with a binary nuisance. Under the full representation, one Gaussian feature reveals the decision and one the nuisance. Under the collapsed representation, the decision feature is replaced by a constant while the nuisance feature is preserved. Variance is fixed so that the full-representation DDA complexity is 0.1. The collapsed problem is unidentifiable for the decision: all methods hit the 100-probe budget and attain chance correctness.

D Robot implementation and additional results

D.1 Interfaces and scientific gates

The canonical state vector includes task geometry, end-effector state, and robot-state features padded to a fixed width. Feed-forward observations have dimension 74. History-4 observations append three additional seven-dimensional actions plus validity indicators for a total dimension of 98. Migrating a source checkpoint copies the first 74 input columns and initializes the 24 new columns to zero; an equivalence test on 257 samples requires actor and value errors below 10^{-7} and observed zero error.

The native XArm Robotiq controller accepts gripper coordinates in $[0, 0.81]$. For canonical command $g \in [-1, 1]$, the native command is $0.405(1 - g)$. Verification checks confirm matching Cartesian signs, shared external action bounds, and temporal-action bookkeeping. Partial-reset tests verify that the latency queue is reset only for terminated environments. Resume tests verify that an update-0 best checkpoint is preserved. All gates passed before the reported training and held-out evaluation stages.

D.2 Training and selection

PPO uses 2,048 vectorized environments, 20 rollout steps, and therefore 40,960 target interactions per update. Training runs 244 updates. Standard learning rate is 3×10^{-4}; the low-rate ablation uses 3×10^{-5}. Validation occurs at update 0 and every five updates. A checkpoint is selected only by validation success (return as the tie-breaker) and is then evaluated on 1,024 fresh episodes. Update-0 source policies are evaluated separately; if they already meet a declared success threshold, target sample complexity is zero regardless of later fine-tuning behavior.

D.3 Learning-rate ablation

We repeat fine-tuning for observation dropout and the randomized-interface shift at one tenth of the standard learning rate, keeping all six source/adaptation seed pairs matched. This isolates optimization step size from source-policy and environment-shift effects. Table 5 shows that the smaller rate does not improve either tested shift; bold marks the higher mean.

Table 5: Fine-tuning learning-rate ablation on held-out success. Differences are low rate minus standard rate; p values are exact paired sign-flip tests and are uncorrected for multiple comparisons.

Shift	Standard LR	Low LR	Difference	p value
Observation dropout	**.910**	.848	$-.061$	.0312
Randomized interface	**.158**	.136	$-.022$	.8438

On observation dropout, the mean decreases from 0.910 to 0.848; on the randomized interface it changes from 0.158 to 0.136. Because multiple comparisons were performed and only six paired policies contribute to each test, we treat the results as exploratory. Their defensible conclusion is negative but useful: reducing the step size is not a general explanation for, or remedy to, the shift-dependent transfer failures in the main study.

E AGGREGATION AND UNCERTAINTY

The held-out evaluation manifest contains 204/204 completed jobs and no failures: 126 feed-forward jobs, 24 low-learning-rate jobs, 36 history-4 jobs, 14 feed-forward update-0 source evaluations, and four history-4 update-0 source evaluations. Each job has 1,024 episodes, for 208,896 episodes total. The aggregation records contain 49 policy-summary rows and 190 unique checkpoint hashes.

Policy seeds are the inferential unit. Reported policy-level intervals use Student-t intervals when at least three policies are present. Matched comparisons use exact sign-flip tests; no multiplicity-corrected significance claim is made. Episode-level Wilson intervals are retained in the evidence archive only to describe evaluation noise for a fixed policy.