AUDITING ORACLE HEADROOM: DECISION-LEVEL CROSS-REALIZATION VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Instance-wise oracle upper bounds are read as headroom for adaptive acquisition, but when unacquired outcomes are estimated by a learned evaluator, the same realization both chooses the oracle's per-case plan and scores it. Gap stability cannot detect this: the gap can be identical under two realizations even when every per-case decision reverses. We therefore validate the decision rather than the gap. Cross-realization transfer holds the case and the source-selected plan fixed, changes the evaluator realization, and scores that plan against the target realization's own fixed baseline. Under an idealized zero-mean independent-error model, the gap is optimistic and transfer is conservative; in a controlled generator with realization noise alone, the gap grows without bound while transfer stays at zero. Across four grounded world-model contracts, transfer depends on the registered perturbation: positive transfer is supported in the three development regimes under same-data retraining, but only in DINO-WM/PushT under post-hoc pipeline re-instantiation. An oracle-headroom claim is only as strong as the evaluator change it has survived.

1 INTRODUCTION

Evaluation is often treated as a fixed measurement procedure. But when the informative diagnostics are themselves costly, evaluation becomes a decision problem: which evidence to acquire for a given instance, how to spend a limited budget, and whether later tests should depend on what earlier ones returned. Adaptive evaluation is the natural response. When is such adaptivity actually warranted?

Instance-wise oracle upper bounds are used in adaptive acquisition to quantify potential headroom (Guney et al., 2025). A clairvoyant picks a different plan for each instance, knowing the outcome being measured, and its margin over a single population-optimal plan is read as room an adaptive evaluator could pursue. Prior work asks whether such oracles read information a deployable policy could not (Valancius et al., 2024). We hold the oracle construction fixed and ask a different question. The oracle selects its plan and scores that same plan under one realization of the evaluator — the random draw that produced the loss surface being minimized. Plan rankings belonging to that realization are therefore credited as adaptive opportunity. Would the same plan still be useful under another realization?

We validate the decision rather than the gap. A *case* is one evaluation instance, a *plan* is an affordable set of diagnostics to run on it, and an *evaluator realization* is one draw of the learned evaluator that scores what was not run. The *fixed baseline* is the single best plan for the whole population under that realization, selected without using the evaluation groups on which it is scored. Write O for the same-realization oracle gap, where the plan is chosen under realization a and scored under a, and $T_{a\to b}$ for the *cross-realization transfer* obtained by re-scoring that same case-specific plan under a different realization b, against b's own fixed baseline. What T returns is relative to which evaluator change is made, so it measures the portability of an oracle opportunity rather than a property the model has. We register two changes: retraining the estimator on the same data with a new seed, and re-instantiating the training pipeline so the training-group realization and the optimization randomness both move.

The transfer stage is neither of the two checks it is most easily confused with. Nested cross-validation changes the case being evaluated and tests generalization to new samples; this audit keeps the case fixed and changes the evaluator — the case axis $x \to x'$ versus the evaluator axis $a \to b$.

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.

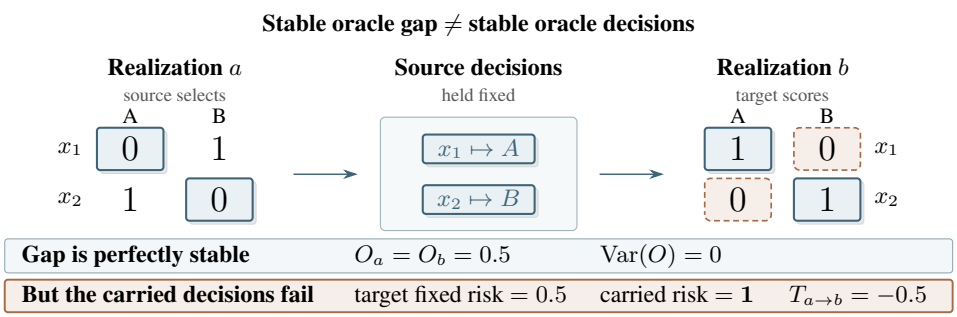


Figure 1: **Why oracle-gap variance is insufficient.** Both evaluator realizations have the same fixed-plan risk and the same oracle gap, $O_a = O_b = 0.5$, so the scalar gap has zero cross-realization variance. Yet the case-specific plans the oracle chooses under a are the wrong plans under b, giving $T_{a\to b} = -0.5$. Cross-realization transfer validates the decisions that create the gap rather than the gap magnitude.

Scalar stability of O, even where it is available, cannot establish decision stability: the aggregate keeps the magnitude and discards which plan was chosen for which case (Section 3.3). And where work on predictive multiplicity and underspecification establishes that valid realizations can disagree, we take that as given and ask the downstream question of whether such disagreement costs the oracle-attributed decision its value (Section 2).

We use world-model diagnostics as a grounded test bed because unacquired outcomes require learned evaluation, which creates a meaningful evaluator-realization axis, while the diagnostics carry real acquisition costs and a grounded target still exists to score against. Four evaluation contracts instantiate the audit. Under retraining, transferred decision value is positive in the three development regimes, most clearly in DINO-WM/PushT; under pipeline re-instantiation it is positive in PushT and the remaining intervals include zero. These four contracts are grounded stress tests of the audit rather than a prevalence estimate.

We make two methodological contributions. **(i) Decision-level cross-realization validation.** We introduce an estimand defined by four choices: the same case, the source-selected plan identity preserved, the evaluator realization changed, and scoring against the target realization's own fixed baseline. Figure 1 shows why scalar gap stability cannot substitute — O can be identical across realizations while every per-case decision reverses and transferred value is negative. **(ii) Controlled mechanism and grounded audit.** A controlled additive-noise construction illustrates the failure mode: realization-specific ranking noise alone creates large same-realization oracle gaps while transferred decision value stays near zero. Four grounded learned-evaluation contracts then instantiate the audit, transfer is examined under several registered evaluator changes, and source ensembling provides an audit-side variance-reduction diagnostic.

2 Related Work

Adaptive acquisition, experimental design, and adaptive evaluation. Learning what to measure per instance is mature, spanning active feature acquisition (Li & Oliva, 2021), amortized and semi-amortized sequential Bayesian experimental design (Foster et al., 2021; Hedman et al., 2025), and in-context measurement selection (Kobayashi et al., 2026); evaluation itself has been made adaptive through dynamic multi-round evaluation and adaptive benchmarking against an evolving ability profile (Zhang et al., 2025; Tian et al., 2026; Ding et al., 2026). These optimize or instantiate an acquisition policy; we study a precondition for interpreting the headroom such a policy is motivated by.

Selection bias, oracle gaps, and validation. Instance-wise oracle upper bounds also appear in adaptive feature acquisition, where an empirical oracle is evaluated as an idealized upper bound on acquisition performance (Guney et al., 2025), and where a separate line asks whether such oracle supervision is informationally legal for a deployed policy (Valancius et al., 2024). This audit builds

on classical selection/evaluation separation. Taking an argmin over noisy estimates and reading the winner's own estimate is the optimizer's curse (Smith & Winkler, 2006), the maximization bias that motivates double estimators in value learning (van Hasselt, 2010); selection overfits itself unless selection and assessment are separated (Cawley & Talbot, 2010); and inference after a data-driven selection needs its own treatment (Berk et al., 2013). The contribution is the decision-level estimand for instance-wise oracle-headroom claims: the same case, the selected plan identity preserved, the evaluator realization changed, and scoring against the target realization's own fixed baseline. Nested cross-validation changes the case and asks whether a selection procedure generalizes; this audit keeps the case fixed, changes the evaluator realization, and carries a per-instance decision across rather than a scalar estimate. A recent fixed-pool diagnostic for test-time collaboration reaches a related refusal to read an oracle gap as attainable gain (Hu, 2026); there the question is whether a verification signal identifies the good candidate in a fixed pool, here whether a clairvoyant's per-instance plan retains value on the same case under another registered realization.

Multiplicity and underspecification. That equally valid models can disagree is established, and this audit builds on it: the Rashomon effect names the coexistence of many models that fit about equally well (Breiman, 2001); predictive multiplicity measures conflicting individual predictions among competing models of comparable accuracy (Marx et al., 2020); model multiplicity examines the discretion this leaves in model selection (Black et al., 2022); underspecification shows that one pipeline can return predictors with equivalent held-out performance that behave very differently in deployment (D'Amour et al., 2022), structurally the closest of these to our nuisance construction; and allocation multiplicity asks whether such model sets deliver correspondingly diverse downstream allocations (Jain et al., 2025). These literatures motivate why valid evaluator realizations may disagree, and our narrow nuisance is closest to that setting; the broad stress test asks a stronger pipeline-robustness question and does not require the retrained evaluators to be equally accurate. Either way our audit is downstream of the phenomenon, and asks not whether valid realizations disagree but whether that disagreement costs the oracle-attributed decision its target-side value (Section 3.3).

World-model diagnostics. World-model evaluation increasingly uses capability-specific diagnostic suites — observable simulator contracts, intervention-fidelity audits, and decision-metric alignment of latent planning scores (Co et al., 2026; Vakalis, 2026; Wang et al., 2026) — rather than aggregate task metrics. Those ask which capabilities to measure; ours is the acquisition-level question of whether apparent instance-specific headroom survives a registered evaluator perturbation (Section 3.3).

3 METHOD

3.1 EVALUATION SETUP

An evaluation case x is a decision instance of the system under test, carrying a hidden diagnostic profile Y_x recording how it responds to each probe in a registered library $\mathcal{P}$. Each probe carries a deterministic, dependency-aware cost, so a *plan* s — an ordered, affordable subset of probes — costs $c(s)$, and only plans with $c(s) \leq B$ run at budget B over a grid $\mathcal{B}$ fixed in advance. Acquired entries of Y_x are observed exactly, unacquired entries come from a registered evaluator reading only the legal history, and every policy including the clairvoyant is scored with that same predictor — the clairvoyant may inspect the hidden profile to choose *what to acquire*, never to predict. Writing $L_x^{(a)}(s, B)$ for the resulting profile error (lower is better) when plan s is scored on case x at budget B under realization a, the risk of a rule π is the normalized trapezoidal area over the grid, $\mathcal{R}_a(\pi) = \mathbb{E}_x[\mathrm{AUC}_{B\in\mathcal{B}}\, L_x^{(a)}(\pi(x, B), B)]$.

Registered nuisances. A nuisance realization changes a registered source of randomness in how the evaluation loss is produced, while preserving the case, probe library, cost contract, target profile, budget grid and evaluation question. We register two, and they ask different questions. The *narrow* nuisance varies the training seed of one frozen profile-estimator family with the training population, architecture, folds and targets all held fixed, so it is a within-family stochastic-retraining perturbation. The *broad* nuisance additionally varies the training-group realization, so it is a **training-pipeline perturbation stress test**: it re-instantiates the registered evaluator-training pipeline and

asks whether oracle-attributed case-specific decisions survive that. It is not a Rashomon set and is not restricted to evaluators of equal predictive quality — mechanical validity and predictive-quality equivalence are different conditions, and the broad stress test enforces the former. Cases, folds, targets, estimator architecture, probe contract and budgets are fixed under both (Supplement §C), and every reliability statement below names the nuisance it was computed under.

Choosing and fixing the nuisance. Register the evaluator variation at the invariance level the claim asserts: a claim about retraining stability registers the optimization seed, one meant to survive independent training-data realizations also varies the data. The estimator family is part of the same contract — condition on it when one family is deployed, freeze the selection procedure when that choice is part of replication, report stratified T when several remain admissible. **The primary contract, family included, is fixed before transfer outcomes are inspected**; a nuisance chosen afterwards is a sensitivity analysis.

3.2 Oracle opportunity

Fix a realization a. The *realization-specific fixed plan* π_{fix} minimizes expected loss at each budget under a, selected on that realization's training groups and never on the evaluation groups where it is scored. The per-instance clairvoyant π_{oracle} instead picks, for every case and budget, the affordable plan that turns out best on that case, $s_x^{*,a}(B) = \arg\min_{s:\, c(s)\le B} L_x^{(a)}(s, B)$. The gap between them,

$$O_a \;=\; \mathcal{R}_a(\pi_{\mathrm{fix}}) \;-\; \mathcal{R}_a(\pi_{\mathrm{oracle}}),$$

is the same-realization *oracle gap*: a hindsight upper bound rather than anything a legal evaluator could obtain. The validity problem is that one realization plays both roles: a selects the per-instance plan *and* scores it, so any variation in plan ranking particular to a is credited to adaptivity.

3.3 Cross-realization reliability

We break the coupling by separating selection from scoring: take the clairvoyant plan chosen under a source realization a and evaluate it under a distinct target realization $b \neq a$,

$$T_{a\to b} \;=\; \mathcal{R}_b(\pi_{\mathrm{fix}}^{(b)}) \;-\; \mathbb{E}_x\Big[\,\mathrm{AUC}_B\; L_x^{(b)}(s_x^{*,a}(B), B)\Big].$$

where $\pi_{\mathrm{fix}}^{(b)}$ denotes the fixed plan selected for target realization b. A positive $T_{a\to b}$ means a plan chosen with hindsight under a still beats the strongest fixed plan registered for b, on b's own loss. The plan is transferred but the yardstick is not: the comparison is against b's fixed plan, never a's, since the source's baseline would let movement of the best fixed plan inflate or depress transfer. We write $T = \mathbb{E}_{a\neq b}[T_{a\to b}]$ for the average over the $K(K-1)$ ordered source–target pairs, with $K = 3$ under the narrow nuisance and $K = 24$ under the broader one; those pairs are not independent replicates, so uncertainty is group-aware throughout (Section 3.4). The remainder between the target's own oracle gap and the transferred value is non-negative target-side selection regret; its exact loss form is given in Supplement §D.

What the two nuisances estimate. The narrow realization set is a fixed historical finite set $\mathcal{A} = \{a_1, \ldots, a_K\}$ with $K = 3$, and the estimand is the average over its ordered pairs, $T_{\mathcal{A}} = \frac{1}{K(K-1)} \sum_{a\in\mathcal{A}} \sum_{\substack{b\in\mathcal{A}\\ b\neq a}} T_{a\to b}$. Inference is conditional on that set: the group bootstrap represents uncertainty over evaluation groups, not over a population of seeds. The broad nuisance instead has a registered randomization procedure Q — a training-group resampling draw with an optimization-randomness draw, evaluation groups fixed — with the pipeline-level estimand $T_{\mathcal{Q}} = \mathbb{E}_{A,B\overset{\mathrm{iid}}{\sim}\mathcal{Q}}\left[T_{A\to B} \mid A \neq B\right]$., of which the 24 realizations are a finite sample. $T_{\mathcal{A}}$ asks whether decisions survive one fixed seed set, T_Q whether they survive re-instantiating the pipeline.

Figure 1 shows why this is not the same as asking how much O itself varies. Variance of O measures the stability of the gap's magnitude; T measures the stability of the decisions that produce it, because it keeps which plan was selected for which case. In this paper *reliability* means that retained target-side decision value under a registered evaluator perturbation, conditional on the evaluator contract, including the estimator family when the contract fixes it. Exact plan agreement is

a different quantity: it measures identity stability rather than value stability, and the two can move apart.

Case-correspondence controls, which answer a different question from $T > 0$, are reported in Supplement §D.

Which way each quantity is biased. Suppressing the budget argument for notation, write the realized loss as a stable surface plus realization-specific evaluator error, $L_x^{(a)}(s) = \mu_x(s) + \varepsilon_x^{(a)}(s)$, and let H be the oracle decision opportunity on the stable μ-surface. Assume the realization error has zero mean given the case and plan, that source and target errors are independent, and that the evaluation reference is selected independently of the error it is scored on. Then

$$\mathbb{E}[O] \;\geq\; H \;\geq\; \mathbb{E}[T].$$

O is optimistic because one noise realization both selects and scores the winner. T is conservative because a plan selected from a noisy source realization need not be the stable optimum, and it is then scored independently. Two readings follow within this model. A positive T is conservative evidence of stable decision value, and a non-positive T is not evidence that H itself is non-positive. Our target-specific fixed baseline is selected on training groups and scored on disjoint evaluation groups; Supplement §H gives the proof and the corresponding expectation-level treatment of the reference term. The sandwich is a mechanism result under the stated independence conditions. Distinct stochastic retrainings do not by themselves establish source–target error independence, so the audits below report contract-conditional transfer under their registered evaluator perturbations. The conservative lower-bound reading is reserved for the independent-error model. Outside the model the residual is $-\mathbb{E}_x[\varepsilon_x^{(b)}(s_x^{*,a})]$, which moves $\mathbb{E}[T]$ upward toward $\mathbb{E}[O]$ as source and target errors share structure (Supplement §H), so neither reading is protected and transfer is reported as a contract-conditional quantity.

When does the audit have bite? It is informative exactly when the loss surface the oracle minimizes depends on an evaluator realization — learned imputers and profile predictors, stochastic proxy scorers, retrained or Monte-Carlo evaluators, learned surrogate metrics. If the oracle instead minimizes a fully observed deterministic ground truth, so $L_x^{(a)} = L_x$ for every a, there is no realization axis and the audit returns $T = O$ by construction. Budgeted world-model diagnostics are a useful test bed for the opposite reason: unacquired outcomes are estimated by a learned evaluator, which creates a real realization axis, while diagnostics differ in cost and a grounded target still exists to score against.

Legal gain. Realized gain of the tested legal policies is defined and reported separately in Supplement §§C and E; it is not used to establish oracle-headroom validity.

3.4 Protocol and inference

Under the narrow nuisance we use $K = 3$ estimator-seed realizations of one frozen estimator family, fitted with grouped cross-fitting; groups are the source trajectory or native episode. Target-specific fixed plans are selected on training groups within each fold, and every quantity is computed on the same cases, folds, plan space, grid and loss. Uncertainty is a paired group-aware percentile bootstrap over evaluation groups, conditional on the frozen realizations, and does not represent uncertainty over future seeds. Under the broad nuisance, where the $K = 24$ realizations are themselves drawn and give $24 \times 23 = 552$ ordered pairs, we report a *two-way* interval that resamples evaluation groups together with source and target realization identifiers; it is primary wherever a broad-nuisance number appears, with the group-only interval beside it in Supplement §G. The two target different quantities: the group-only interval conditions on the realized set and captures evaluation-group sampling alone, an interval for $T_{\mathcal{A}_{24}}$, while the two-way interval approximates the uncertainty in estimating T_Q from finitely many groups and finitely many pipeline realizations, with no asymptotic claim and no suggestion that 24 draws exhaust Q. Narrow intervals are group-conditional by construction, so interval widths are not compared across nuisances: point estimates compare the definitions, and support is read under each estimand's own contract. Intervals are paired group-aware percentile-bootstrap summaries. On the frozen additive-noise null, the registered percentile group bootstrap gives empirical coverage 0.952, 0.948 and 0.956 at the 21-, 24- and 32-group layouts used by this

study (Supplement §G). The extension from 12 to 24 realizations followed the registered seed sequence and was completed before any aggregate beyond $K = 12$ was inspected. The dense grids of per-pair, per-family and per-realization values in the supplement are descriptive sensitivity analyses.

Scope and reading conventions. These hold throughout and are not repeated with each result. Intervals are pointwise paired group-aware 95% bootstrap intervals, unadjusted for multiplicity; an effect is *supported* when its interval lies strictly on one side of zero. T is conditional on the registered nuisance and the evaluator contract. Narrow results are conditional on the three historical estimator seeds and are the primary audit; the broad pipeline stress test, the factorization and the family and policy sensitivities are post-hoc, and the broad nuisance is a pipeline perturbation rather than an equal-quality model set. Only the DreamerV3 narrow-nuisance validation was prospectively specified.

We evaluate four regimes spanning three architecture families: DINO-WM (Zhou et al., 2025) on PushT (latent visual), TD-MPC2 (Hansen et al., 2024) on walker-walk and cheetah-run (latent-state decision-centric), and DreamerV3 (Hafner et al., 2025) on walker-walk (RSSM latent dynamics). Walker appears under two families, giving partial task control, but not a controlled architecture isolation: their probe libraries, cost axes and estimator protocols differ. Each regime registers four diagnostic probes with dependency-aware costs and a frozen reachable budget grid, yielding 129, 252, 259, and 72 feasible plans for PushT, TD-MPC2 walker, TD-MPC2 cheetah, and DreamerV3 respectively (Supplement §§B and C). The DINO-WM and TD-MPC2 analyses are grouped development analyses; DreamerV3 uses a different protocol, stated wherever its numbers appear: episodes are split group-wise before any probe response exists, everything is fitted on development groups only, and confirmatory aggregates follow an irreversible unlock (Supplement §F); there we audit a registered finite-horizon evaluation construct on the frozen world model rather than its deployed actor. Its estimator family matches the others while its fitting topology and evidence population do not, so rows are comparable in status but not in magnitude. Every robustness analysis — broad nuisance, factorization, estimator family, synthetic control, neural policies — is post-hoc; only the DreamerV3 validation is prospective, and its confirmatory population has been re-used by them.

4 Empirical Results

4.1 A realization-noise mechanism can inflate O without transferable value

Figure 1 is a construction; a controlled generator shows the same divergence arising from noise alone. Each case carries one stable preferred plan among eight candidates, with advantage δ, and every plan loss is perturbed by independent realization noise of scale σ, so the signal-to-noise ratio is δ/σ (Supplement §G); $\delta = 0$ removes all stable case-specific structure. The gap nevertheless grows without bound as the noise scale grows — from 0.14 at $\sigma = 0.1$ to 2.28 at $\sigma = 1.6$ — while transfer stays at zero and the oracle's plan choices agree only at chance (Figure 2a). Across all thirty frozen configurations the gap explains almost none of the transferred value; adding the signal-to-noise term accounts for most of it, and the retained fraction T/O is ordered by the signal-to-noise ratio alone (Figure 2b,c).

4.2 Cross-realization reliability in grounded case studies

Each of the four contracts fixes its own probe library, cost axis, budget grid and profile-error loss in advance. Profile-loss units are regime-specific, so comparisons are within regime.

Under the fixed historical within-family retraining contract (Table 1), transfer is supported in the three development regimes — most clearly in PushT, with walker's lower endpoint close to zero — while DreamerV3's interval includes zero. The question these answer is whether the oracle-selected decision retains target-side value when the same evaluator design is retrained. A non-positive single-source transfer means the source-selected decision does not beat the target's fixed baseline under that audit. In the independent-error model that is not evidence that the stable oracle opportunity itself is non-positive. For the registered contracts, source–target error independence is not established: source and target share the evaluator design, and under the narrow nuisance the training data as well. If their errors share structure, Supplement §H gives the direction — the residual can move

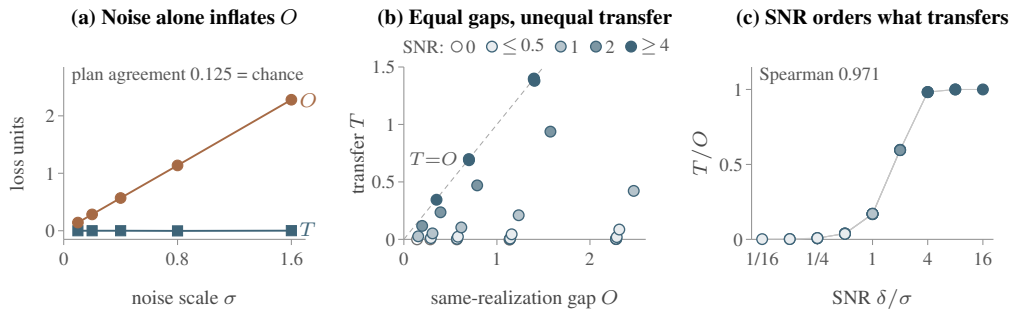


Figure 2: **Realization noise inflates the oracle gap without creating transferable value.** (a) With no stable case-specific signal ($\delta = 0$), the gap O grows with the noise scale σ while transfer T stays at zero and oracle-plan agreement sits at chance. (b) Across the thirty configurations, similar gaps carry very different transfer; O alone explains $R^2 = 0.018$ of T. (c) For $\delta > 0$, the retained fraction T/O is ordered by $\mathrm{SNR} = \delta/\sigma$ alone (shading as in (b)). A controlled post-hoc mechanism demonstration: O and T *can* diverge under realization noise; it does not model the four audited regimes.

Table 1: **Transferred decision value under two registered evaluator changes.** O_{seed} is the same-realization oracle gap. T_{seed} is the transferred decision value under the **primary** audit, retraining one frozen estimator family on the same data across $K{=}3$ historical seeds, with its paired group-aware interval. T_{broad} is the post-hoc *stress test* that re-instantiates the training pipeline ($K{=}24$), with the two-way interval that is primary for that estimand. Loss units are regime-specific and are not compared across rows. †prospectively pre-specified under retraining only. All intervals are pointwise 95%.

		primary: retraining, $K{=}3$	stress test: re-instantiation, $K{=}24$
Regime	O_{seed}	T_{seed}	T_{broad}
DINO-WM / PushT	0.0188	$+0.0145\,[+0.0114, +0.0175]$	$+0.0085\,[+0.0056, +0.0111]$
TD-MPC2 / walker	0.0069	$+0.0046\,[+0.0004, +0.0117]$	$+0.0019\,[-0.0013, +0.0097]$
TD-MPC2 / cheetah	0.0073	$+0.0030\,[+0.0015, +0.0046]$	$-0.0028\,[-0.0116, +0.0005]$
DreamerV3 / walker†	0.00138	$+0.00029\,[-0.00005, +0.00069]$	$-0.00016\,[-0.00064, +0.00023]$

$\mathbb{E}[T]$ toward $\mathbb{E}[O]$. We therefore read the estimate as portability under the registered change rather than invoking the bracketing guarantee.

Selecting from an ensemble of source realizations. Instead of choosing the per-case oracle plan from one source realization, consensus-source selection chooses it from the average of all non-target source realizations,

$$s_x^{*,-b}(B) = \arg\min_{s:\, c(s)\le B} \frac{1}{K-1} \sum_{\substack{a\in\mathcal{A}\\ a\neq b}} L_x^{(a)}(s, B).$$

The resulting decision is evaluated on target b. Consensus-source point-estimate transfer is higher in every regime: $+0.0155\,[+0.0124, +0.0185]$ in PushT, $+0.0050\,[+0.0007, +0.0121]$ in walker-walk, $+0.0035\,[+0.0020, +0.0052]$ in cheetah-run and $+0.00044\,[+0.00008, +0.00085]$ in DreamerV3, whose single-source interval includes zero while its consensus interval does not. Source averaging yields higher point-estimate transfer in all four regimes. The pattern is consistent with part of single-source pessimism coming from realization-specific ranking variation, while a shared case-specific preference survives across fits. Consensus-source selection requires several oracle loss realizations, so it is an audit-side construction rather than a legal evaluation-time policy, and it enters no gain claim.

Under the post-hoc pipeline re-instantiation stress test. A new training-data realization with fresh optimization randomness moves the estimates: PushT stays supported and the other three

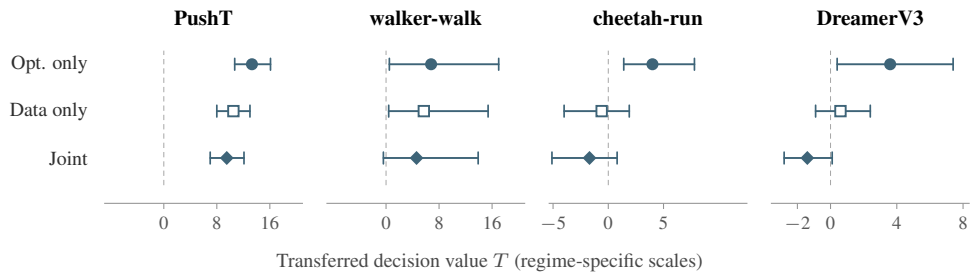


Figure 3: **Transfer across the factorized broad-design arms.** The three categorical positions are the registered optimization-only, data-only and joint perturbations; point estimates place both single-factor arms above the joint arm in all four regimes. The joint arm changes both registered components, while no ordering is assumed between the two single-factor arms. Intervals are paired group-conditional 95% intervals, and all factorized analyses are post-hoc. Axis values are in units of 10^{-3} (10^{-4} for DreamerV3). Full contrasts and construction details are in Supplement §G.

two-way intervals include zero. Per-realization results are in Supplement §G. Case-correspondence controls are reported in Supplement §D.

Transfer across the factorized broad-design arms. The two contrasts against the joint arm have a nested-design interpretation: the joint arm changes both registered components, whereas each single-factor arm changes only one. There is no analogous a priori ordering between the optimization-only and data-only arms. Empirically, the full point-estimate ordering $T_{\text{opt}} > T_{\text{data}} > T_{\text{joint}}$ holds in all four regimes (Figure 3); the joint–opt and joint–data contrasts exclude zero in all four, while data–opt excludes zero in three of four. Interval support is retained across all three arms in PushT, across optimization-only and data-only in walker-walk, and under optimization-only alone in cheetah-run and DreamerV3. We read this as alignment with the registered perturbation structure, not as a monotonic consequence of the theory, because changing the training-data realization can also change the fitted loss geometry. The pattern is qualitatively compatible with the shared-error direction in Supplement §H, but the factorized arms do not identify shared error. These post-hoc results show that the audit has empirical dynamic range across the registered perturbations.

Cross-family sensitivity. Changing estimator family changes the evaluator loss surface itself rather than drawing another realization within one fixed-family contract. Cross-family transfer is directional. In PushT it shows a target-evaluator pattern: both transfers into the neural evaluator are supported positive, $+0.0067\,[+0.0027, +0.0108]$ from masked ridge and $+0.0075\,[+0.0043, +0.0108]$ from bagged trees; both transfers into masked ridge are negative in point estimate, $-0.0027\,[-0.0061, +0.0005]$ from bagged trees and $-0.0057\,[-0.0087, -0.0029]$ from the neural family, the latter supported negative; and both transfers into the bagged-tree evaluator are positive in point estimate but unsupported, $+0.0016\,[-0.0034, +0.0067]$ and $+0.0033\,[-0.0001, +0.0067]$. This pattern is consistent with cross-family T being target-evaluator-relative. The full directional 3×3 matrix is in Supplement §I.

5 Discussion

What T measures, and which way it errs. T validates the value of a source-selected oracle decision under another evaluator realization. In the independent-error model the two quantities bracket the stable opportunity from opposite sides: the same-realization gap is optimistic because one noise draw both selects and scores the winner, and transfer is conservative because the source-selected plan need not be the stable optimum. Within that model a positive transfer is conservative evidence of stable decision value. The registered audits do not establish source–target error independence: source and target share the evaluator design, and under the narrow nuisance the training data as well. If source and target errors share structure, the residual in Supplement §H can move $\mathbb{E}[T]$ toward $\mathbb{E}[O]$. The reported transfer is therefore not guaranteed to be a conservative bound on H. Table 1 is therefore read as portability of the oracle-selected decision under the registered evalua-

tor change, conditional on the contract. Figure 3 shows how transfer changes across the registered factorized broad-design arms.

Single-source transfer can be pessimistic. The consensus-source result gives the empirical counterpart. Forming the oracle decision from the average of the non-target realizations yields higher point-estimate transfer in all four contracts; in DreamerV3, the single-source interval includes zero whereas the consensus-source interval does not. The pattern is consistent with realization-specific ranking variation perturbing per-case plan selection, while a shared case-specific preference survives across fits.

For audit practice. When several evaluator fits are available, report single-source transfer and optionally form the oracle decision from an ensemble of source realizations, to see whether a shared case-specific preference survives realization noise. This ensemble oracle is an audit construction, not a legal policy, and enters no gain claim. For an evaluator that is trained once and frozen permanently, the same-realization oracle gap remains the realization-specific quantity of interest; for headroom claims intended to survive retraining, independent replication or evaluator replacement, T under the matching registered change is what measures that portability.

Limitations and next steps. The conservative bracketing result is conditional on the independent-error and reference assumptions; the empirical audits instead measure transfer under registered evaluator perturbations. The broad two-way interval is a post-hoc finite-resampling summary and lies outside the null-coverage calibration, which covers the narrow group bootstrap. The evidence is conditional on four evaluation contracts and fixed evaluator families, and T establishes decision portability rather than a threshold of practical importance or a recoverable policy gain. Extending the audit across evaluator quality, model families, and external oracle-headroom settings is a natural next step.

6 Conclusion

Same-realization oracle gaps can reward realization-specific plan rankings, so gap magnitude alone does not validate the decisions that produce it. Cross-realization transfer tests whether those decisions retain value under another evaluator realization. For headroom claims intended to survive evaluator retraining or replacement, this decision-level validation complements the same-realization oracle gap.

AI use statement

AI-assisted tools were used for limited editorial purposes, including language polishing, readability improvement, and writing refinement. No AI-assisted tool was used to generate synthetic data, formulate or prove mathematical claims, design the methods or experiments, implement the analysis code, or determine the interpretation of the reported results.

All scientific content, experimental decisions, derivations, implementations, numerical results, and conclusions were independently verified and approved by the authors. The authors take full responsibility for the content of this paper and its supplement.

Reproducibility statement

Every paper-facing quantity is produced by frozen scripts from frozen inputs. We will release the code and environment specification, the group splits and evaluation-case manifests, the registered training-data realization manifests and optimization seeds, the estimator and policy configurations, the bootstrap seeds and method-freeze manifests, and the scripts that regenerate every table and figure, together with the stored loss tensors wherever redistribution is permitted. Section 3 states the evaluation contract, the registered nuisances and the inference procedure in full, and the appendices give the per-pair, per-realization and per-family values behind every aggregate.

REFERENCES

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. doi: 10.1214/12-AOS1077.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 850–863, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3531146.3533149.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001. doi: 10.1214/ss/1009213726.

Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

Peterson Co, Sicheng Hu, Chunxuan Jiao, Hongyang Cheng, Yulin Luo, Yijie Xu, Sixiang Chen, Zhongxia Zhao, Zihao Wang, DaFeng Chi, Peidong Liu, YuTong Chen, Henghua Liu, Zhihao Yuan, Huizhu Jia, Yuzheng Zhuang, Tianle Zhang, Liang Lin, Huajie Tan, and Shanghang Zhang. WorldSimProbe: Diagnosing simulator faithfulness in action-conditioned world models for embodied manipulation, 2026. URL https://arxiv.org/abs/2608.09298.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL http://jmlr.org/papers/v23/20-1335.html.

Xuanwen Ding, Chengjun Pan, Zejun Li, Jiwen Zhang, Siyuan Wang, and Zhongyu Wei. AutoJudger: An agent-driven framework for efficient benchmarking of MLLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15009–15034, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.685. URL https://aclanthology.org/2026.acl-long.685/.

Adam Foster, Desi R Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3384–3395. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/foster21a.html.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B Kolachalama. Active feature acquisition via explainability-driven ranking. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20748–20765. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/guney25a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/cf73d57b6dcda32b293df7c2d5341f49-Paper-Conference.pdf.

Marcel Hedman, Desi R. Ivanova, Cong Guan, and Tom Rainforth. Step-DAD: Semi-amortized policy-based Bayesian experimental design. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22904–22923. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/hedman25a.html.

Jie Hu. Oracle gap and signal fidelity: A fixed-pool diagnostic for test-time collaboration, 2026. URL https://arxiv.org/abs/2607.17531.

Shomik Jain, Margaret Wang, Kathleen Creel, and Ashia Wilson. Allocation multiplicity: Evaluating the promises of the Rashomon set. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 2040–2055, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3715275.3732138.

Yuta Kobayashi, Zilin Jing, Jiayu Yao, Hongseok Namkoong, and Shalmali Joshi. Learning-to-measure: In-context active feature acquisition. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZumBSRsRbE.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6765–6774. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006. doi: 10.1287/mnsc.1050.0451.

Shulin Tian, Ziqi Huang, Fan Zhang, Hongyuan Zhu, Yu Qiao, and Ziwei Liu. Open evaluation agent: Efficient and promptable evaluation of visual generative models, 2026. URL https://arxiv.org/abs/2608.09666.

Donna Vakalis. The intervention gap in latent world models, 2026. URL https://arxiv.org/abs/2608.29998.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for non-greedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024. URL https://proceedings.mlr.press/v235/valancius24a.html.

Hado van Hasselt. Double Q-learning. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems*, volume 23, pp. 2613–2621. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/091d584fced301b442654dd8c23b3fc9-Paper.pdf.

Jiawei Wang, Ke Rui, Yushen Zuo, Yichun Feng, and Minglei Li. Decision-metric alignment in latent world models: Diagnostics and action-conditioned objectives for MPC planning, 2026. URL https://arxiv.org/abs/2608.18746.

Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7561–7582, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.374. URL https://aclanthology.org/2025.acl-long.374/.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.