CHALMERS
UNIVERSITY OF TECHNOLOGY

Interventional-twin audits for post-probe decision routing under hidden-factor shift

Downloaded from: https://research.chalmers.se, 2026-10-16 14:17 UTC

Citation for the original published paper (version of record):
Chen, Z., Yin, Z., Liu, S. et al (2026). Interventional-twin audits for post-probe decision routing under hidden-factor shift. Machine Learning with Applications, 25.
http://dx.doi.org/10.1016/j.mlwa.2026.100998

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Machine Learning with Applications 25 (2026) 100998

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Interventional-twin audits for post-probe decision routing under hidden-factor shift

Zhuo Chen [a,*], Zhenyun Yin [b], Shuhao Liu [c], Zhi Ling [d], Yu Yan [d]

[a] Chalmers University of Technology, Gothenburg, Sweden
[b] Eötvös Loránd University, Budapest, Hungary
[c] Harbin University of Science and Technology, Harbin, China
[d] Agnes AI, Kiwair Co., Ltd., Singapore

ARTICLE INFO

Dataset link: https://github.com/KleinChen42/MLWA_project

Keywords:
Interventional evaluation
Post-probe decision routing
Shortcut auditing
Controlled hidden-factor shift
Model selection

ABSTRACT

Interactive systems can encounter states that look equivalent before a low-energy probe even though a hidden mass produces different motion and therefore calls for a different commit route. Aggregate success cannot reveal whether a router used this observed post-probe effect or a correlated pre-probe shortcut. We introduce an *interventional-twin audit* for this question. The audit predeclares observable channels, constructs pairs that match before probing but diverge in observed effects, freezes a non-oracle input schema, and evaluates cue reversal, directional effect swaps, leakage, route failure, and selective re-probing. Across 180 family–seed units in six fixed synthetic parameterizations, direct post-probe routing attains 0.871 reversed-cue success versus 0.500 without the effect channel. When the factual effect is replaced by its matched counterpart while visible history and probe action are fixed, the learned router attains 0.759 bilateral directional consistency versus 0.000 without probe-effect access. A CausalWorld Pushing construction provides public evidence: matched state and frozen-RGB cases establish post-probe input value, and a separately frozen 20-run full-contract audit passes directional-permutation and grouped leakage gates. A 20-run matched-pair reliability diagnostic preserves pair identifiers while effect noise and missingness increase; the overall paired gain remains positive under severe corruption, while the severe worst-mass interval crosses zero. An architecture–capacity–budget stress shows positive post-probe gains but cell-dependent architecture ordering. A calibration-frozen risk–coverage sweep finds no safe positive-coverage point under three shortcut stresses. The result is a controlled, reproducible workflow for testing effect-responsive routing and reporting where evidence becomes inconclusive.

1. Introduction

Interactive machine-learning systems sometimes must choose a consequential route before an action-relevant property is directly visible. Consider two otherwise matched simulated blocks with the same declared pose and goal but different hidden masses. A shared low-energy push can produce different displacement, so a low- versus high-gain commit has a different declared target. A policy that follows a correlated pre-probe cue may appear reliable when the cue agrees with the target and fail when the cue is reversed. This is a partially observed decision pattern: information-gathering actions can alter what a system knows about the state as well as the state itself (Bajcsy, 1988; Kaelbling et al., 1998). History-conditioned policies are a practical response (Hausknecht & Stone, 2015), but high final success alone does not identify which part of the history controls the final route.

We address this evidential gap with an *interventional-twin audit*. Each accepted pair matches on a prespecified set of pre-probe observable channels, receives the same probe, and has a measurable post-probe effect gap together with distinct declared route targets. The protocol freezes a non-oracle input schema and combines cue reversal, pre-probe proxy tests, forbidden-key scans, directional effect swaps, and risk–coverage accounting. It asks a narrower question than whether a policy is generally capable: under a declared access contract, does its route respond to newly observed post-probe evidence rather than a pre-probe shortcut?

The route-target provenance is evidence-block specific. The locked six-fixed-parameterization block is a controlled synthetic routing diagnostic with configured targets; it does not record executed physical candidate commits. Public CausalWorld blocks instead construct targets

* Corresponding author.
E-mail addresses: zhuoc@chalmers.se (Z. Chen), s5owd3@inf.elte.hu (Z. Yin), 2420610137@stu.hrbust.edu.cn (S. Liu), zhi.ling@kiwair.com (Z. Ling), yu.yan@kiwair.com (Y. Yan).

https://doi.org/10.1016/j.mlwa.2026.100998
Received 19 July 2026; Received in revised form 24 August 2026; Accepted 31 August 2026
Available online 5 September 2026
2666-8270/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Table 1
Evidence map: contract, threat addressed, statistical unit, and bounded claim role. Results are not pooled across the controlled synthetic and public CausalWorld contracts.

Block	Contract and primary threat addressed	Primary unit	Bounded claim role
Locked synthetic frontier	Visible/action history and observed effect; configured route target. Tests cue reversal and directional swaps against no-probe and shuffled-target controls.	180 family–seed units; fixed-family seed interval	Controlled input-use and directional-swap audit
Public matched input-value check	CausalWorld state or frozen-RGB features, proprioception, observed effect; candidate-commit target. Tests matching and pre-probe access.	10 runs per representation; matched twins	Public-backend input-value check; learned rows saturate
Public matched-twin directional audit	State, proprioception, observed effect; candidate-commit target. Tests within-stratum effect permutation and a five-classifier all-fields leakage gate.	20 independent runs; five mass-pair strata and matched test twins	Bounded public contract portability
Public reliability frontier	State, proprioception, observed effect and mask; candidate-commit target. Tests effect-quality boundary under noise and missingness.	20 independent training runs; held-out masses nested	Non-saturated information-quality frontier, not pairwise twins
Candidate-selection boundary	Locked candidate records and configured target. Separates exact-matched nulls from declared shortcut-conflict stress.	Family–seed units within each frozen candidate pool	Conditional selection diagnostic; no generic advantage
Calibration-frozen risk–coverage	Twelve candidates; disjoint calibration, selection, and locked-test seeds. Tests when uncertainty should trigger commit versus abstention.	Family–condition–held-out seed	Safe commit only under the strict matched scenario in this archive

from executed candidate commits under their stated simulator contract. This separation prevents a superficially uniform label field from being mistaken for a common physical outcome measure. In both settings, hidden factors, oracle routes, and rule outputs are excluded from non-oracle inference.

The paper makes four contributions. First, it specifies an operational interventional-twin audit with accepted-pair, target-provenance, and input-access contracts. Second, it reports a locked six-fixed-parameterization comparison that isolates the value of declared post-probe effects against no-probe and shuffled controls. Third, it introduces directional within-twin effect swaps, proxy checks, and selective risk–coverage reporting as evidence-path diagnostics. Fourth, it reports three complementary public CausalWorld evidence blocks: saturated matched input-value checks, a frozen 20-run matched-twin directional audit, and a same-pair effect-reliability frontier that exposes a severe worst-mass failure boundary. A confirmatory architecture–capacity–budget stress resolves cell-specific paired contrasts, while a calibration-frozen unknown-shortcut frontier identifies when the audit should defer rather than commit. The result is a controlled evaluation methodology for compact routers under declared access contracts. Deployment and broader learned-objective comparisons remain natural next evaluations under richer task-specific designs. Table 1 is the paper's evidence map. It keeps the controlled synthetic and public CausalWorld blocks separate because their route targets, statistical units, and threats differ. Reading the table before the results makes the intended logic explicit: a matched-twin block tests the declared effect-to-route path, a leakage block tests the declared pre-probe access boundary, and a reliability block shows where the same signal becomes less useful.

2. Related work

Partially observed control, active sensing, and identification. POMDPs formalize sequential decisions from action–observation histories rather than fully observed states (Kaelbling et al., 1998), and history-conditioned policies are one practical response to partial observability (Hausknecht & Stone, 2015). Active perception and motor sensing use actions to acquire task-relevant information (Bajcsy, 1988; Margolis et al., 2023). Dual control makes the coupled control-and-information role of an action explicit (Feldbaum, 1960); recent work develops active acquisition, exploration, and system-identification objectives for POMDPs and manipulation (Li & Oliva, 2025; Memmel et al., 2024; Xie et al., 2024). Our contribution is different: it neither proposes an exploration objective nor claims a new identification algorithm. It audits whether a subsequent commit changes with the observed consequence of a fixed probe under a declared access contract.

Relation-based testing and local decision boundaries. Behavioral testing asks whether a model exhibits prespecified behaviors beyond aggregate held-out accuracy (Ribeiro et al., 2020). Contrast sets inspect local decision boundaries through small meaningful perturbations whose labels change (Gardner et al., 2020), while metamorphic testing checks declared input–output relations when a conventional test oracle is incomplete (Xie et al., 2011). Interventional twins synthesize these ideas for a sequential decision: matched instances share a probe, their observed effects must diverge, the route-target provenance is stated, and a within-pair effect swap tests whether the router responds in the declared direction. This is an evaluation-and-evidence contract, not a claim that probing itself is novel.

Evaluation under distribution shift and model scale. WILDS demonstrates why in-distribution accuracy is insufficient under natural distribution

shifts, while DomainBed shows that selection rules and comparable protocols materially affect domain-generalization conclusions (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021). Testable learning formalizes a related separation between a condition check and the guarantee that follows only when that check accepts (Klivans et al., 2024). Decision-focused shift evaluation further shows that the worst shift for average prediction need not be the worst shift for the downstream decision (Ren et al., 2024). At foundation/generalist-policy scale, HELM and RoboArena likewise emphasize explicit scenarios, multiple outcomes, and reproducible comparison protocols instead of a single aggregate score (Atreya et al., 2025; Liang et al., 2023). Our audit applies this transparency principle to a much narrower local question: whether a frozen route changes with matched post-probe evidence. It is neither a foundation-model benchmark nor a substitute for broad task evaluation.

Causal interventions, counterfactual evaluation, and controlled simulation. CausalWorld exposes controlled interventions over manipulation factors such as mass and appearance (Ahmed et al., 2021). Causal representation and causal-RL research motivate interventions as a way to stress assumptions about transfer and decision mechanisms (Deng et al., 2023; Schoelkopf et al., 2021). Recent interventional causal-representation work sharpens the distinct identifiability questions posed by finite samples and general environment families (Acarturk et al., 2024; Ng et al., 2025). Counterfactual RL methods instead focus on credit assignment or off-policy evaluation under an explicit causal model (Mesnard et al., 2021; Oberst & Sontag, 2019). Causal-confusion work uses targeted environment or expert interventions to identify variables that support an imitation policy (de Haan et al., 2019). Our aim is complementary: we freeze a hidden-factor decision contract and measure the response of a declared router to observed-effect interventions, shortcut reversal, and input-schema audits. We do not recover a causal graph or infer a causally sufficient internal representation.

Shortcut risk, selection, and privileged supervision. Shortcut learning and underspecification show that conventional validation can leave predictors with similar apparent performance but different behavior under shift (D'Amour et al., 2022; Geirhos et al., 2020). Domain- generalization studies further emphasize that model selection is part of an evaluation protocol rather than an implementation detail (Gulrajani & Lopez-Paz, 2021; Ye et al., 2021). Generalized distillation allows train-time information that is unavailable at inference (Lopez-Paz et al., 2016). Recent self-distillation for time-series classification transfers class-conditional information bidirectionally across network levels to improve supervised representations (Xiao et al., 2026). That setting is adjacent rather than equivalent: our focus is whether a frozen sequential router changes its commit with an observed post-probe effect under matched interventions. Decision-focused learning further distinguishes predictive accuracy from downstream decision quality (Wilder et al., 2019). The present audit therefore records aligned and reversed outcomes, declared access, failure, and targeted swaps; it does not claim that a distilled objective universally dominates alternative learned routers.

3. Interventional-twin audit

3.1. Running example and audit question

Consider two simulated pushing episodes that begin with the same declared cube pose, target, camera-derived features, and proposed low-energy probe. One cube has a low mass and the other a high mass, but mass is hidden from the router. Before probing, the allowed input records match. The shared push then produces different displacements, and executed candidate commits favor different follow-up routes. The audit asks one observable question: when the visible and action histories are held fixed and the measured displacement is exchanged between the matched episodes, does the router change toward the counterpart route? Table 3 carries this pair through every protocol step. Thus, an *accepted twin* is a matched pair that passes the declared construction checks; an *effect swap* replaces only its measured post-probe evidence; and a *route target* is a training/evaluation label whose source is recorded separately from inference inputs.

3.2. Decision setting

Let o_0 denote configured pre-probe observable channels, a_p a probe action, e its observed effect features, and a_c a commit route. A hidden factor z changes the transition induced by a_p and can change the appropriate commit route, but z is excluded from non-oracle inference. An accepted twin pair (i, j) satisfies

$$V(o_0^i, o_0^j) = 1 \tag{1}$$

under declared tolerances, while the same probe produces an effect gap

$$D(e^i, e^j) \geq \delta_e. \tag{2}$$

The pair is retained only when evaluation confirms that the appropriate commit routes differ under the stated task contract. This retention rule defines the population on which the audit asks whether a router uses the observed effect; it is not itself evidence that a learned router succeeds, and the experiments do not estimate the prevalence of accepted pairs in an unconstrained task population.

3.3. Inference and audit boundary

A non-oracle router is

$$\hat{a}_c = f_\theta(h_{\text{vis}}, h_{\text{act}}, e), \tag{3}$$

where the visible and action histories, together with the observed probe effect, are the permitted inputs. Hidden factors and oracle labels are evaluation-only; route targets are train-only. Their provenance is block-specific: configured synthetic targets in the locked six-fixed-parameterization diagnostic and candidate-commit outcome targets in public CausalWorld blocks. Transparent-reference predictions remain separate evaluation rows. The audit serializes declared non-oracle input keys, scans them for enumerated forbidden keys, and tests whether configured pre-probe inputs decode the hidden factor above a declared chance ceiling.

The inference record for each non-oracle prediction contains the declared input keys and a forbidden-key count. Route-target provenance is declared separately for each evidence block. The locked six-fixed-parameterization diagnostic uses a configured synthetic route contract, whereas the public CausalWorld bridge derives a route target from executed candidate commits under its stated simulator contract. Thus, the hidden factor may index data construction and evaluation groups, but neither z nor an oracle route is passed to f_θ at inference.

3.4. Functional identification scope

The audit makes a functional statement about a declared input–response path, not a full causal-representation claim. Let $h = (h_{\text{vis}}, h_{\text{act}})$ denote the permitted pre-effect history. The following conditions are checked for each accepted pair: (i) the declared matching validator accepts the pre-probe fields; (ii) the probe action is shared; (iii) the observed effects meet the declared gap; (iv) the route targets differ under the stated target provenance rule; and (v) an effect swap changes no permitted history field other than the supplied effect record.

Proposition 1 (*Contract-level functional identification*)**.** *For an accepted-pair distribution satisfying conditions (i)–(v), directional agreement after an observed-effect swap identifies a dependence of the evaluated router's declared input–output function on the supplied effect record within that contract. A router that ignores the effect record is invariant to the swap. The proposition does not identify a causal graph, recover a latent representation, or establish the same dependence outside the validated pair distribution.*

Table 2
Mapping the contract-level proposition to standard causal language. "Enforced" means enforced for the serialized evaluation record, not established for an unrestricted task population. "Diagnosed" denotes a falsification or measurement check rather than proof of the assumption.

Contract condition	Standard causal analogue	Audit status	Scope of the resulting statement
(i) Declared pre-probe match	Conditional exchangeability over declared observed fields	Enforced and diagnosed	Matching and leakage tests reduce measured imbalance; they do not rule out an unrecorded pre-probe correlate.
(ii) Shared probe action	Treatment-version consistency and pair-local no interference	Enforced in the serialized pair	Both twins receive the same declared probe. Cross-episode interference is absent in the simulator execution, but broader SUTVA is not inferred.
(iii) Measurable effect gap	Intervention relevance and effect-measure validity	Threshold-enforced; sensitivity diagnosed	The supplied effect records are distinguishable under D; this does not make the chosen extractor sufficient for every latent physical change.
(iv) Distinct route targets	Outcome/label validity under declared provenance	Enforced within each block	Synthetic targets follow the configured rule; public targets follow executed candidate commits. The two provenance paths are not pooled.
(v) Effect-only record swap	Consistency of the evaluated input intervention	Enforced by record comparison	The functional contrast changes only the effect argument of the frozen router input. It identifies input dependence, not a mediated physical causal effect.
Accepted-pair support	Positivity/overlap within the audit population	Enforced by construction ; population overlap unestablished	Both matched effect records occur in the accepted-pair support. No prevalence or positivity claim is made outside that support.

Reason. Conditions (i)–(v) hold the declared history fixed while replacing the effect argument to f_θ. The swap comparison can therefore distinguish the evaluated functional response from one that ignores that argument. It cannot rule out every unmeasured correlate or establish a mechanism inside the learned representation.

Table 2 connects these conditions to standard causal terminology. The strongest warranted conclusion is deliberately functional: within accepted-pair support, a frozen router changes its output when its declared effect input changes. Matching, leakage tests, and positive controls make violations observable in several measured channels, but do not turn the comparison into an estimate of a population causal effect.

3.5. Route-only illustrative learner

The route-only student receives the declared route target y_c and minimizes

$$\mathcal{L}_{\text{route}}(\theta) = \text{CE}\big(f_\theta(h_{\text{vis}}, h_{\text{act}}, e), y_c\big)\,. \tag{4}$$

It is evaluated against matched learned alternatives under the declared route-target source and input contract. The synthetic and public-CausalWorld label paths are reported separately rather than pooled.

3.6. Operational protocol

Algorithm 1: Interventional-twin audit.

1. **Predeclare the contract.** Specify observable pre-probe fields, the hidden-factor intervention, probe action, effect extractor, route-target provenance, equivalence/divergence thresholds, failure event, permitted inference keys, and leakage ceiling.
2. **Construct and validate twins.** Generate paired instances, retain only pairs satisfying $V = 1$ and $D \geq \delta_e$, and verify that the declared route contract differs across the twins. A configured synthetic block records the analytical target rule; a public physical simulator can instead execute candidate commits and record their outcomes.
3. **Freeze data and access.** Materialize aligned, cue-reversed, and held-out groups; freeze train/validation/test partitions and record the non-oracle input schema before model comparison.
4. **Fit matched comparators.** Train no-probe, post-probe, negative-control, transparent-reference, and oracle rows under their declared input, split, and budget conditions; architecture differences are reported explicitly rather than inferred from row names.
5. **Evaluate outcomes.** Report aligned and cue-reversed success, worst-hidden-group success, declared route failure for configured synthetic targets (or executed terminal failure for public commits), paired differences, and the hierarchical statistical unit.
6. **Audit the evidence path.** Test pre-probe decodability, scan serialized input schemas for forbidden keys, and apply effect masking, swapping, shuffling, or corruption.
7. **Audit decisions under uncertainty.** When a router may defer or re-probe, report commit coverage, conditional risk among covered commits, overall failure, and probe cost separately.

3.7. Outcome and reliability metrics

For an evaluation set D with split indicator s_n, hidden group z_n, correctness c_n, and irreversible-failure indicator u_n, cue- reversed success and worst-group success are

$$S_{\text{rev}} = \frac{\sum_n \mathbf{1}[s_n = \text{rev}]c_n}{\sum_n \mathbf{1}[s_n = \text{rev}]}, \qquad S_{\text{wg}} = \min_z \frac{\sum_n \mathbf{1}[z_n = z]c_n}{\sum_n \mathbf{1}[z_n = z]}. \tag{5}$$

For the configured synthetic block, the same quantity is called declared route failure; for public executed commits it is terminal failure. In either case,

$$F_{\text{all}} = |D|^{-1} \sum_n u_n. \tag{6}$$

Table 3
One matched low-mass/high-mass pushing pair carried through the seven audit steps. The example translates each protocol term into an observable operation; it is not an additional empirical result.

Step	Protocol term	Pushing-pair operation	Frozen evidence
1	Declare contract	Match pose, goal, and allowed camera/proprioceptive fields; vary mass; use one low-energy push.	Input schema, tolerances, effect distance, candidate commits, failure event.
2	Accept twins	Keep the pair only if pre-probe fields match, the shared push yields displacement gap $D \geq \delta_e$, and executed commits favor different routes.	Pair identifier, validator outcomes, effect gap, target provenance.
3	Freeze access	Put the pair in one split and expose only visible/action history plus observed displacement to non-oracle routers.	Split manifest and forbidden-key scan.
4	Fit comparators	Train matched no-probe and post-probe routers; keep shuffled-target, hand-rule, and oracle rows explicitly separate.	Configuration, seed, budget, and training log.
5	Score outcomes	Test the original pair, cue-reversed cases, and the weakest mass stratum.	Success, terminal failure, paired differences, independent run identifier.
6	Intervene on evidence	Hold permitted history fixed and exchange the two observed displacement records.	Route-change and bilateral-consistency records; pre-probe decoder audit.
7	Handle uncertainty	When the effect is weak or corrupted, compare commit with re-probe/defer.	Coverage, conditional commit risk, overall failure, and probe cost.

For a selective router with commit indicator q_n, coverage and conditional risk use different denominators:

$$C = |D|^{-1} \sum_n q_n, \qquad R_{\text{commit}} = \frac{\sum_n q_n u_n}{\max(1, \sum_n q_n)}. \tag{7}$$

We therefore never substitute R_{commit} for F_{all}. For paired effect records (e_n, e_n^{swap}), targeted decision sensitivity is the route-change rate

$$I_{\text{effect}} = |D|^{-1} \sum_n \mathbf{1}\left[f_\theta(h_n, e_n) \neq f_\theta(h_n, e_n^{\text{swap}})\right]. \tag{8}$$

Finally, the conditional model-selection analysis reports

$$\Delta_{\text{sel}} = S_{\text{locked}}(m_{\text{audit}}) - S_{\text{locked}}(m_{\text{aligned}}), \tag{9}$$

on a test set untouched by either selection rule.

Two-twin denominator example. Suppose a cue-reversed set contains two accepted twins and the router routes one correctly. Then $S_{\text{rev}} = 1/2$. If it commits on only one of the two cases and that commit fails, coverage is $C = 1/2$ and conditional commit risk is $R_{\text{commit}} = 1$, whereas overall irreversible failure is $F_{\text{all}} = 1/2$. Reporting all three values prevents a selective router from appearing safe merely because it abstains more often. The complete numerator, denominator, and statistical-unit guide is retained in the Supplement so that these definitions remain auditable without interrupting the main methodological argument.

3.8. Audit outcomes

Primary outcomes are cue-reversed and worst-hidden-group success together with declared route failure in the configured synthetic block. Public simulator blocks separately report executed terminal failure when applicable. Counterfactual effect swaps and corruption measure targeted input sensitivity. A selection audit compares conventional aligned-success selection with audit-aware selection on a locked interventional test set. These are controlled evaluation outcomes, not proof of causal sufficiency of an internal representation (see Fig. 1).

4. Experimental protocol

4.1. Locked synthetic frontier

The principal controlled comparison fixes six fixed synthetic coefficient parameterizations, the cue-suppressed input contract, partitions, optimizer budget, and hidden width. It contains 180 family–seed units (six fixed parameterizations and 30 paired seeds per parameterization). The primary non-oracle contrast is no-probe MLP versus direct-effect MLP. We also report a paired-contrast MLP (ITCL-Distill), a residualized-effect MLP, and a shuffled-target training-label negative control. A hand effect rule and a hidden-state oracle are diagnostic references, not learned comparators. The equal-input and route-only historical rows are separately seeded fits with the same input and base learner as the direct-effect row; they are documented in the supplement rather than counted as independent architecture or component-removal baselines. The statistical unit is the family–seed unit. The primary fixed-family interval resamples paired seeds inside each declared parameterization; it is not a superpopulation interval over arbitrary task families.

The architecture boundary extension addresses the fixed-width, single-budget, easy-regime limitation without using confirmatory outcomes to design the test. A discovery stage on seeds 0–11 identifies six cells spanning small, medium, and large capacity; half, standard, and double budget; and easy, intermediate, and hard regimes. MLP, GRU, and Transformer routers are then evaluated once on disjoint seeds 12–41. Each architecture–cell row contains 180 paired family–seed units. Primary contrasts are direct-effect minus no-effect for cue-reversed success, worst-hidden-group success, failure, and bilateral directional consistency. A locked paired reanalysis compares architecture gains on identical family–seed units, with fixed-family bootstrap intervals, sign-flip tests, and Holm adjustment across the 18 cue-reversed architecture contrasts. Architecture rankings are not prespecified claims.

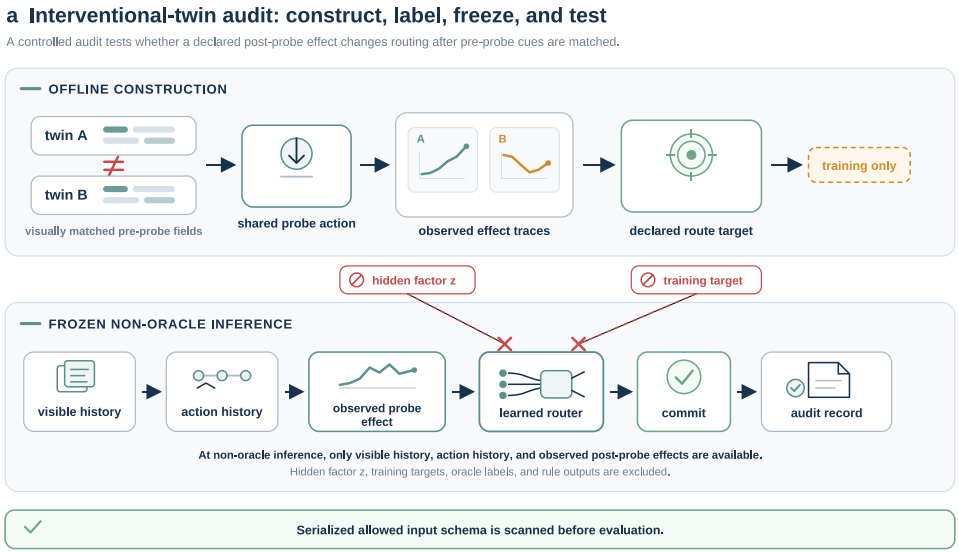

Fig. 1. Interventional-twin audit contract. Offline construction locks matching, effect-gap, and route-target provenance. The locked six-fixed-parameterization target is configured synthetically; public CausalWorld targets derive from executed candidate commits. At frozen non-oracle inference, the router receives only declared visible/action history and the observed probe effect; hidden factors and training targets are excluded. The audit records schema scans, cue reversal, declared route failure for configured targets (or terminal failure for public commits), and directional effect swaps.

4.2. Public CausalWorld pushing cases

The matched input-value cases fix initial pose, goal, and declared pre-probe observations while intervening on hidden mass. A low-energy probe produces displacement features, and low- and high-gain commits are prevalidated for different masses. The state-plus-proprioception and frozen-RGB-plus- proprioception cases each use ten independent runs and 300 requested matched configurations per run. For frozen RGB, a fixed ImageNet-pretrained ResNet-18 transforms declared camera frames; the audit records pixel and feature matching before probing, nonlinear pre-probe mass decoding, and enumerated forbidden input-schema keys. Learned post-probe rows saturate in these matched cases, so these rows test the value of the declared post-probe input rather than rank learned objectives.

A later, separately frozen full-contract matched-twin audit makes the directional test executable on the public backend. It is not an expansion of a successful canary: after the two-run moderate-corruption canary missed its directional gate, the follow-up changed the primary condition to light corruption, increased the construction budget, equalized the route-only width, and added effect-permutation and grouped leakage controls. Each of 20 independent runs requests 60 configurations in each of five hidden-mass-pair strata, then splits accepted pairs 60/20/20 after construction. It uses the same-width no-probe, direct-effect, ITCL-Distill, residualized-effect, and shuffled-target rows; a hand-effect reference and a commit-outcome oracle are marked separately. The primary condition uses effect noise $\sigma = 0.003$ and missingness 0.05. At test time, the direct-effect router is re-evaluated after a within-run, within-stratum nonidentity and non-twin permutation of observed effects. The primary statistical unit is the independent run. A convergence-audited, condition-grouped five-classifier leakage suite (regularized logistic, same-width MLP, random forest, histogram gradient boosting, and kNN) over all 56 declared pre-probe dimensions, construction retention, post-probe positive controls, and serialized forbidden-key scans are prespecified gates. Provenance metadata is tested only as an excluded schema canary. This block tests public matched-twin contract portability, rather than broad simulator generalization. The complete field-level protocol diff and its predeclaration implications are reported in the protocol-difference table in Supplement Section S0.8. The auxiliary non-saturated public frontier uses disjoint interleaved training and held-out mass grids. Unlike the matched input-value cases, it evaluates individually sampled held-out-mass records rather than reconstructing pairwise shared-probe twins. Its role is to test whether the post-probe input-value direction persists in a non-saturated public-backend routing problem; it complements rather than substitutes for the matched-twin checks above. Twenty confirmatory runs follow a five-run gate without changing the masses, router budgets, or corruption schedule. Same-width no-probe, direct-effect, residualized-effect, and shuffled-label routers are compared with a transparent effect reference and a separately marked commit-outcome oracle. Four predeclared regimes use Gaussian effect noise and missingness $(0, 0)$, $(0.003, 0.05)$, $(0.006, 0.15)$, and $(0.012, 0.35)$. Each router is trained and evaluated under the same regime, so this is an information-quality frontier rather than a clean-trained sensor-shift test. The primary public contrast is one direct-effect minus no-probe delta per independent training run ($n = 20$); held-out mass configurations are nested within the run. The historical nested run/configuration bootstrap is retained only as a secondary sensitivity analysis. Hidden mass and commit outcomes are excluded from non-oracle features, and the reported scan concerns enumerated forbidden keys in serialized non-oracle input schemas. These are public simulated manipulation cases, not physical-deployment evaluations.

The R2 same-pair reliability extension is separate from that auxiliary held-out-mass frontier. It reconstructs one physical matched-pair archive, preserves exact pair IDs across clean, light, moderate, and severe corruption, and fits each router independently in 20 runs. The primary unit remains the run. Direct-effect versus equal-width no-probe differences are computed within run on the same pairs for overall success, worst-mass success, terminal failure, and bilateral directional consistency. This design supports a paired degradation boundary; it is not a clean-trained deployment shift.

4.3. Selection and selective re-probing analyses

We report a stable cue-suppressed selection analysis, a separately prospectively specified two-candidate shortcut-conflict stress, and an expanded validation-frozen candidate-pool check. The two-candidate stress compares a cue-only no-probe MLP with an effect-route MLP. The expanded check evaluates 13 MLP/GRU candidates spanning three frozen width bands, with and without the post-probe effect input

Table 4
Locked synthetic route-target frontier. Values are equal-weight means [95% fixed-family, paired-seed intervals] over six fixed parameterizations and 30 paired seeds per parameterization. The primary learned comparison is direct post-probe effect versus no-probe. ITCL-Distill adds the paired contrast term at the same width. The shuffled-target row shuffles route targets during training while retaining the effect inputs; it is a training-label negative control, not an effect-permutation test. Hand and oracle rows are diagnostic references.

Method	Reversed-cue success	Worst-group success	Declared route failure
No-probe MLP	0.500 [0.500, 0.500]	0.448 [0.442, 0.453]	0.500 [0.500, 0.500]
Direct-effect MLP	0.871 [0.867, 0.874]	0.853 [0.849, 0.856]	0.129 [0.126, 0.132]
ITCL-Distill MLP	0.871 [0.867, 0.874]	0.852 [0.849, 0.856]	0.129 [0.126, 0.132]
Residualized-effect MLP	0.865 [0.861, 0.869]	0.817 [0.810, 0.824]	0.136 [0.133, 0.139]
Shuffled-target control	0.499 [0.479, 0.520]	0.464 [0.444, 0.484]	0.499 [0.479, 0.519]
Hand effect reference	0.664 [0.659, 0.668]	0.630 [0.625, 0.635]	0.338 [0.334, 0.342]
Oracle hidden reference	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]

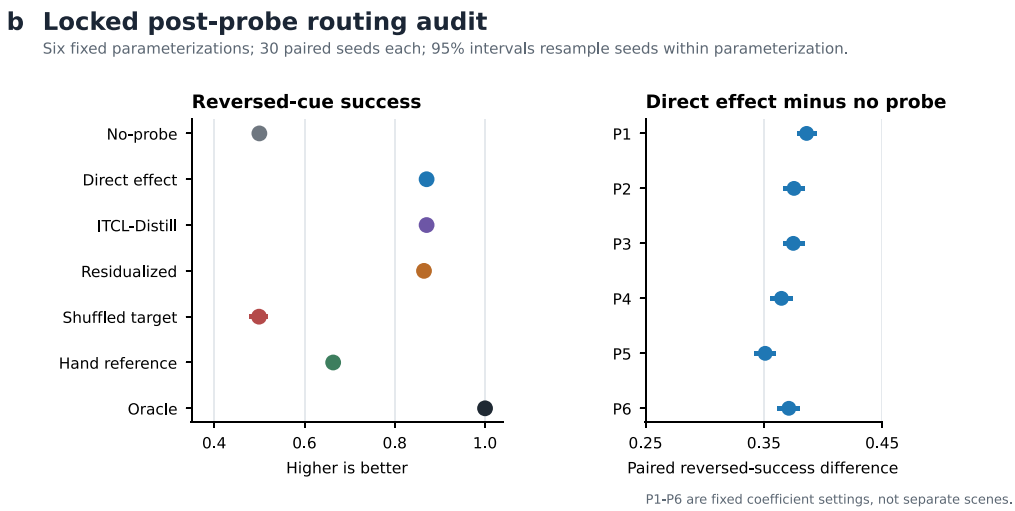

Fig. 2. Locked post-probe routing audit. Left: reversed-cue success for the primary learned contrast, the paired-contrast objective variant, adverse training-label control, and diagnostic references. Right: within-parameterization paired direct-effect minus no-probe differences across the six fixed coefficient settings (P1–P6). All aggregate intervals use an equal-weight mean over fixed parameterizations and bootstrap paired seeds within each setting.

plus a shuffled-target negative control. It contains three strict exact-matched effect-gap conditions and nine explicitly relaxed shortcut/imbalance conditions; the two scopes are reported separately and are never pooled. A separate declared effect-corruption analysis perturbs archived public-case effects, then compares forced commits with a validation-calibrated selective re-probe rule. Selection and sensor-reliability protocols are reported separately rather than pooled.

The unknown-shortcut selection extension evaluates 12 candidates under four separately reported scenarios. Calibration seeds 0–4 set robust-score, margin, disagreement, and selective-commit thresholds; seeds 5–9 choose the candidate; locked seeds 10–19 are accessed only after selection. The robust-with-reprobe rule commits only when all frozen thresholds pass and otherwise re-probes or abstains. Its primary outcomes are locked success, selection regret, catastrophic-selection frequency, and selective coverage. The calibration-frozen risk–coverage reanalysis derives four additional threshold policies exclusively from calibration records, preserves the same selection and locked-test splits, and reports coverage together with conditional catastrophic risk. A policy is considered a safe partial-coverage point only if coverage is positive and the family-bootstrap upper 95% bound on catastrophic selection is at most 0.10; locked outcomes never tune a threshold.

4.4. Unified multiplicity and scaling diagnostics

A single primary-hypothesis ledger covers 42 prespecified rows in six families: the original architecture replication, pre-probe leakage probes, public directional audit, R2 architecture boundary, R2 paired public reliability, and strict selection rows. Holm correction is applied within each family; Benjamini–Hochberg across all 42 rows and max-$|t|$ effect-dimension, and vectorized-compute diagnostics are reported separately as bounded planning measurements rather than universal operating constants.

5. Results

5.1. Auditing post-probe input value under a locked frontier

In the locked six-fixed-parameterization synthetic target comparison, the no-probe MLP reached mean reversed-cue success of 0.500 and worst-hidden-group success of 0.448. The same-width direct-effect MLP reached reversed success of 0.871 and worst-group success of 0.853; the paired-contrast ITCL-Distill MLP reached 0.871 and 0.852, respectively. The residualized-effect MLP reached 0.865 reversed success, whereas the shuffled-target control returned to approximately 0.5 on both reversed success (0.499) and declared route failure (0.499). Thus, the locked frontier supports the value of the declared post-probe effect channel against no-probe and shuffled controls. The direct-effect and paired-contrast learners are effectively tied under this contract, making the input-value conclusion stronger than any ranking between the learned routing objectives. Fig. 2 and Table 4 summarize these results.

5.2. A non-saturated architecture boundary separates input value from ranking

We next froze six architecture–capacity–budget–difficulty cells after a 12-seed discovery stage and evaluated them only on disjoint confirmatory seeds 12–41. Each cell contains 180 family–seed units.

Table 5
Confirmatory architecture boundary with absolute cue-reversed success and paired input-value gain. Each architecture cell is reported as direct-effect success/(direct minus matched no-effect gain) over the same 180 family–seed units. The final column summarizes pairwise gain contrasts supported after Holm correction within the primary metric; ≈ denotes a null-compatible pair. Ordering changes across cells, so the table supports architecture-dependent boundary behavior rather than a universal winner.

Capacity/budget/difficulty	MLP	GRU	Transformer	Paired gain ordering
Medium/standard/easy	0.869/(+0.369)	0.869/(+0.369)	0.851/(+0.351)	GRU ≈ MLP > Tr.
Large/double/intermediate	0.756/(+0.256)	0.775/(+0.275)	0.729/(+0.229)	GRU > MLP > Tr.
Small/half/intermediate	0.627/(+0.127)	0.661/(+0.161)	0.732/(+0.232)	Tr. > GRU > MLP
Large/double/hard	0.583/(+0.083)	0.604/(+0.104)	0.569/(+0.069)	GRU > MLP > Tr.
Medium/standard/hard	0.618/(+0.118)	0.621/(+0.121)	0.610/(+0.110)	GRU > MLP > Tr.
Small/half/hard	0.544/(+0.044)	0.551/(+0.051)	0.579/(+0.079)	Tr. > GRU ≈ MLP

Table 6
Directional within-twin effect swap under the locked synthetic route-target contract. The swap changes only the observed post-probe effect; visible history and probe action remain fixed. This is a controlled input-response diagnostic, not a physical candidate-commit result or an internal causal-representation test.

Variant	Route change	Counterpart route	Bilateral correct
No-probe MLP	0.000 [0.000, 0.000]	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]
Direct-effect MLP	0.775 [0.770, 0.780]	0.871 [0.868, 0.874]	0.759 [0.754, 0.763]
ITCL-Distill MLP	0.775 [0.770, 0.780]	0.871 [0.868, 0.874]	0.759 [0.754, 0.763]
Residualized-effect MLP	0.758 [0.752, 0.763]	0.864 [0.861, 0.867]	0.743 [0.738, 0.749]
Shuffled-target control	0.366 [0.346, 0.386]	0.501 [0.481, 0.521]	0.184 [0.163, 0.206]

Across MLP, GRU, and Transformer routers, the paired direct-effect minus no-effect difference in cue-reversed success is positive in every cell (Table 5). The gains range from 0.044 [0.038, 0.051] for the small, half-budget, hard MLP to 0.369 [0.365, 0.373] for the medium, standard-budget, easy GRU. The relative ordering changes across cells: for example, the Transformer has the largest gain in both small half-budget cells, whereas the GRU has the largest gain in the large double-budget cells. The confirmatory test was not consulted when choosing cells.

A paired architecture-to-architecture reanalysis uses the same family–seed units for every contrast and jointly reports absolute direct-effect success and direct-minus-no-effect gain. Sixteen of 18 cue-reversed gain contrasts remain supported after Holm correction, but the ordering reverses across boundary cells: GRU leads in the large double-budget cells, whereas Transformer leads in the small half-budget cells. The two null-compatible pairs are MLP versus GRU in the medium/standard/easy and small/half/hard cells. This establishes a non-saturated, architecture-dependent input-value surface rather than a universal architecture winner. Exact paired intervals, standardized effects, and adjusted tests are reported in the full paired table in the Supplement.

5.3. Directional swaps test the declared effect-to-route path

Table 6 replaces each factual post-probe effect by its accepted counterpart while retaining visible history and probe action. The direct-effect and paired-contrast rows each change route on about 0.775 of these interventions and attain bilateral directional consistency of 0.759. By contrast, no-probe routing is invariant (bilateral consistency 0.000), and the shuffled-target control attains 0.184. This is a controlled input-response diagnostic for the configured synthetic route target: it shows that permitted post-probe effects can alter routing in the expected paired direction. Because the target is configured, this block is an input-response diagnostic rather than a physical candidate-commit or internal-representation analysis.

5.4. Selection and risk coverage depend on declared shortcut stress

On the synthetic frontier after the shortcut cue is suppressed, audit-aware and conventional selection produce nearly identical reversed-test outcomes (audit minus conventional success 0.0015, 95% interval $[-0.0083, 0.0125]$). Under the separately prospective two-candidate shortcut-conflict stress, audit-aware selection improves locked success by 0.8118 (95% interval from a hierarchical bootstrap, $[0.7890, 0.8250]$) and changes the selected candidate in 98.9% of family–seed units.

In a validation-frozen 13-candidate pool, all three strict exact-matched audit-minus-conventional intervals cross zero; positive differences occur only in separately declared relaxed shortcut/imbalance stresses. The selection result is therefore conditional rather than generic; complete pool-specific statistics remain in the Supplement.

To make the unknown-shortcut decision operational, a further 12-candidate study used disjoint calibration seeds 0–4, selector-fitting seeds 5–9, and locked test seeds 10–19. In the strict matched scenario, the robust selector reached 0.759 locked success [0.744, 0.774], with regret 0.011 and catastrophic-selection rate 0.050.

We then swept five threshold policies derived from calibration seeds 0–4 only, selected candidates on seeds 5–9, and evaluated once on locked seeds 10–19. In the strict matched scenario, all policies retain full coverage with catastrophic-selection rate 0.050 (upper 95% bounds 0.089–0.094). In the three shortcut-stressed scenarios, permissive thresholds retain some or all coverage but have catastrophic rates from 0.610 to 0.900; stricter thresholds reduce coverage, sometimes to zero, without yielding a prespecified positive-coverage operating point whose catastrophic upper bound is at most 0.10. Fig. 3 thus converts abstention from an apparent missing result into a locked risk–coverage boundary: commit is supported only under the strict matched contract in this archive.

5.5. Public simulation separates input value from an externality claim

Table 9 and Fig. 4 first report two matched CausalWorld Pushing input-value checks. Fig. 4 uses worst-mass success, whereas Table 9 reports overall success. In the state-plus-proprioception case, no-probe routing attains 0.458 worst-mass success, whereas the evaluated post-probe rows attain 1.000. In the frozen-RGB-plus-proprioception extension, the corresponding no-probe worst-mass success is 0.000 and the post-probe value is 1.000. Maximum pre-probe pixel and fixed ResNet-18 feature differences are zero, the maximum nonlinear pre-probe mass-decoder accuracy is 0.500, and serialized non-oracle input scans contain zero enumerated forbidden keys. These matched cases demonstrate post-probe input value but cannot rank learned post-probe objectives.

An earlier frozen two-run matched-twin directional canary did not meet its predeclared bilateral-consistency gate and was not promoted. We report that result rather than treating it as a public replication.

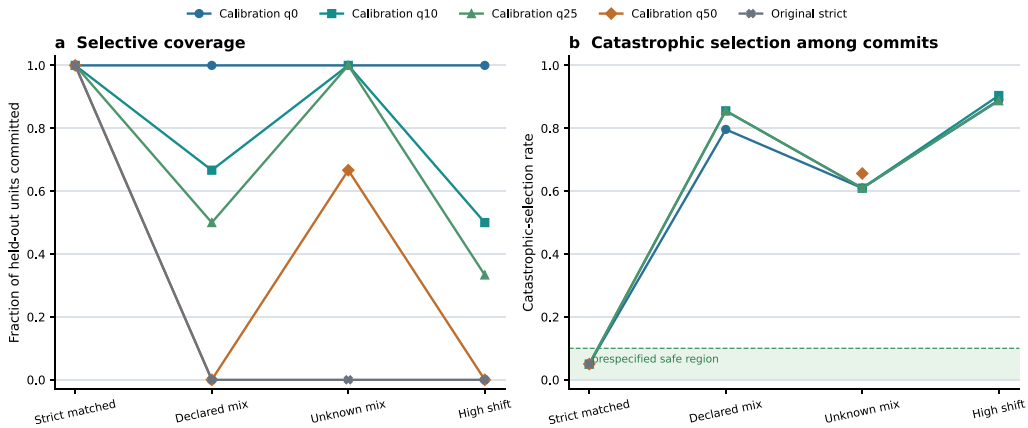

Fig. 3. Calibration-frozen risk–coverage boundary on locked held-out seeds. Left: selective coverage for five policies whose thresholds were derived without locked-test outcomes. Right: catastrophic-selection rate among committed units; the shaded band is the prespecified ≤ 0.10 safe region. Missing points denote zero committed units, not missing data. No tested policy attains positive safe coverage in the three shortcut-stressed scenarios.

Table 7
Frozen public CausalWorld matched-twin directional audit (20 independent training runs; primary light-corruption contract). Entries are means [95% percentile-bootstrap intervals over independent training runs]. Direct effect is additionally evaluated after a within-run, within-stratum nonidentity/non-twin effect permutation; the final row reports its paired bilateral-consistency difference. The shuffled-target row permutes route targets during training while retaining factual effect inputs; it is distinct from the test-time effect-permutation control. Terminal commit failure refers to the executed public candidate-commit contract.

Method	Success	Worst-mass success	Terminal failure	Bilateral consistency
No-probe	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]	0.000 [0.000, 0.000]
Direct effect	0.948 [0.940, 0.955]	0.896 [0.879, 0.912]	0.052 [0.045, 0.060]	0.896 [0.880, 0.911]
ITCL-Distill	0.947 [0.939, 0.955]	0.890 [0.873, 0.906]	0.053 [0.046, 0.061]	0.894 [0.879, 0.909]
Residualized effect	0.946 [0.938, 0.954]	0.892 [0.875, 0.908]	0.054 [0.045, 0.062]	0.893 [0.877, 0.909]
Shuffled-target control	0.438 [0.372, 0.509]	0.354 [0.279, 0.429]	0.562 [0.490, 0.627]	0.087 [0.026, 0.173]
Hand-effect reference	0.868 [0.852, 0.885]	0.781 [0.760, 0.802]	0.132 [0.115, 0.149]	–
Commit-outcome oracle	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]	–
Direct effect minus within-stratum effect permutation (bilateral consistency): 0.662 [0.637, 0.685]				

A later, separately frozen full-contract audit used a changed primary corruption condition, larger construction budget, equal-width route-only row, five held-out mass strata, within-stratum nonidentity effect permutations, and 20 independent training runs. The direct-effect row achieved bilateral directional consistency of 0.896 (lower 95% bound 0.880), while the paired direct-minus-permutation lower bound was 0.637. Every requested stratum was retained and the forbidden non-oracle key count was zero. A convergence-audited all-fields leakage suite over 20 independent runs and 56 declared pre-probe dimensions gave balanced accuracy of 0.500 for regularized logistic, same-width MLP, random-forest, histogram-gradient-boosting, and kNN probes (all primary Holm-adjusted $p = 1.000$). In contrast, the post-probe positive controls ranged from 0.829 to 0.869, and an excluded provenance-metadata canary decoded mass perfectly. This separates an auditable declared-input boundary from an implausible "no possible leakage" claim. Table 7 reports the full directional comparison. This audit supports a bounded public matched-twin contract-portability result; objective ranking and broader simulator generalization require separate non-saturated matched designs. The protocol-difference table in Supplement Section S0.8 enumerates every retained and changed protocol field; the two stages are not pooled.

The supplementary public-statistics table reports the independent-run paired effects, intervals, standardized effects, and Holm-adjusted sign-flip tests calculated from the reconciled raw archive. The corresponding nested twin-count and run-count stability summaries describe the archived evidence budget rather than extrapolating beyond it.

5.6. Same-pair effect reliability exposes a severe failure boundary

The confirmatory public reliability study preserves identical pair identifiers across all four corruption regimes and uses the independent training run as the statistical unit. Direct-effect overall success decreases from 1.000 under clean effects to 0.952, 0.785, and 0.621 under light, moderate, and severe corruption. Relative to the equal-width no-probe router, the paired overall gain remains 0.121 [0.099, 0.143] under severe corruption, and terminal failure is reduced by the same amount. The worst-mass gain, however, narrows from 0.500 under clean effects to 0.029 [−0.002, 0.058] under severe corruption; that interval is null-compatible. Bilateral directional gain also falls from 1.000 to 0.363 [0.329, 0.397]. Fig. 5 and Table 8 therefore show both the retained average input value and the subgroup boundary where the audit becomes inconclusive. The learned direct, equal-input route-only, and residualized routers remain close and are not ranked by this block.

Unified multiplicity accounting. All prespecified primary tests are recorded in one 42-row ledger. Holm adjustment is applied within each of six declared hypothesis families, and a Benjamini–Hochberg adjustment across all 42 rows is reported as a sensitivity analysis. Thirty-four directional alternatives remain supported after both adjustments; eight rows are null-compatible or otherwise non-positive, including the intended pre-probe leakage nulls and strict-selection nulls. The full ledger and max-$|t|$ simultaneous intervals, where common units are available, are reported in the Supplement rather than converting null evidence into positive claims.

Table 8
Public matched-twin reliability with identical pair IDs preserved across corruption regimes. Entries are run-level paired direct-effect minus equal-width no-probe differences with 95% intervals; the failure column is sign-reversed so positive values denote failure reduction. Each row uses 20 independent CausalWorld training runs.

Regime	Δ overall	Δ worst mass	Failure reduction	Δ bilateral
Clean	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]	1.000 [1.000, 1.000]
Light	0.452 [0.442, 0.460]	0.404 [0.390, 0.419]	0.452 [0.443, 0.460]	0.906 [0.888, 0.923]
Moderate	0.285 [0.270, 0.301]	0.177 [0.144, 0.204]	0.285 [0.269, 0.301]	0.619 [0.594, 0.643]
Severe	0.121 [0.099, 0.143]	0.029 [−0.002, 0.058]	0.121 [0.099, 0.143]	0.363 [0.329, 0.397]

Table 9
Public CausalWorld matched input-value checks. These deliberately saturated cases test whether the declared post-probe channel can carry useful information; they do not rank learned post-probe objectives. Worst-mass values appear in Fig. 4. The separate matched-pair reliability frontier is reported in Fig. 5 and Table 8.

Input condition	Method	Overall success	Failure
State + proprioception	No-probe MLP	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]
	Direct-effect MLP	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]
Frozen RGB + proprioception	No-probe MLP	0.500 [0.500, 0.500]	0.500 [0.500, 0.500]
	Direct-effect MLP	1.000 [1.000, 1.000]	0.000 [0.000, 0.000]

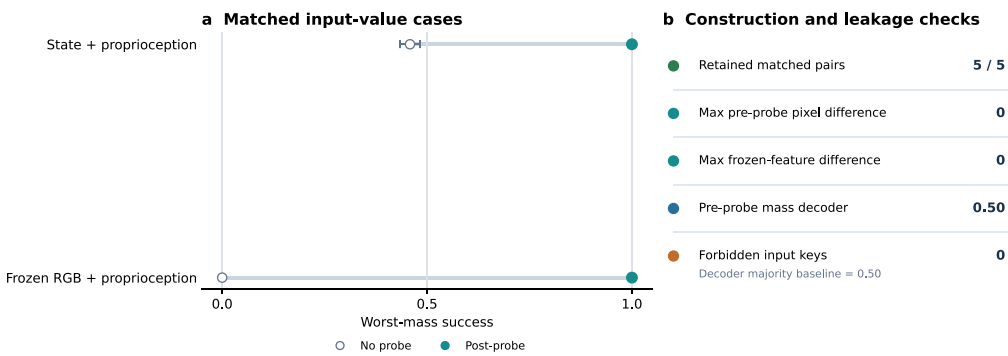

Fig. 4. Public CausalWorld Pushing construct-validity check. Five hidden-mass strata pass physical matched-twin preflight: declared pre-probe fields match, a shared probe yields a measurable effect gap, and candidate commits are distinct. The lower-left panel reports *worst-mass success* (the minimum success across mass strata), rather than the overall-success values in Table 9; the lower-right panel reports construction and leakage checks. Frozen RGB uses a fixed ResNet-18 feature extractor. Learned post-probe rows saturate in this check, so it does not rank learned objectives.

6. Discussion

6.1. What the audit establishes

The central contribution is an evaluation method: an interventional-twin contract makes a router's *declared input access* and its response to post-probe effect interventions inspectable. Under the locked controlled synthetic contract, observed effects improve routing relative to the matched no-probe input, while directional swaps distinguish effect-responsive routers from no-probe and shuffled-target controls. The separately frozen 20-run public CausalWorld matched-twin audit supplies the same type of evidence under its declared candidate-commit protocol: construction, grouped pre-probe leakage, and within-stratum effect-permutation gates all pass. Together, these results support controlled effect-to-route auditing under the evaluated contracts.

6.2. When the audit fails or becomes inconclusive

The audit is designed to expose limits as well as positive input value. The original public matched-input cases saturate among learned post-probe rows, so they establish input value but cannot rank learned objectives. An earlier separately frozen two-run directional canary did not meet its bilateral-consistency gate and was not promoted as a replication. Likewise, the direct-effect and paired-contrast MLPs are effectively tied in the locked frontier, and the direct-routing advantage narrows as effect noise and missingness increase. These outcomes are interpretable audit results: a pass requires a declared directional response against the appropriate controls, not only high aggregate success.

The architecture and selective-routing closures sharpen two further limits. Paired architecture contrasts are often statistically resolved, but their ordering changes with capacity, budget, and difficulty; this is evidence for a boundary surface, not an architecture league table. Likewise, calibration-only threshold sweeps do not recover a safe positive-coverage operating point under the declared, unknown-mixed, or high-shift shortcut stresses. In those regimes, abstention is the supported output of the audit rather than a deferred success claim.

6.3. Practical construction guidance and cost

The audit can be applied as a four-stage decision procedure. First, before fitting routers, declare the pre-probe matching fields, their tolerances, the effect representation and distance, the route-target provenance, the failure event, and the independent statistical unit. Calibrate the effect-gap threshold from measurement repeatability or a construction-only sample; do not lower it after seeing router outcomes. Second, retain and report the complete pair funnel. Proceed only when a shared probe yields measurable effects, candidate routes differ under the declared provenance rule, pre-probe decoders remain at their stated null, and a post-probe positive control is informative. Third, freeze calibration, selection, and test data. Select across aligned, reversed, and proxy-suppressed validation conditions; if no candidate

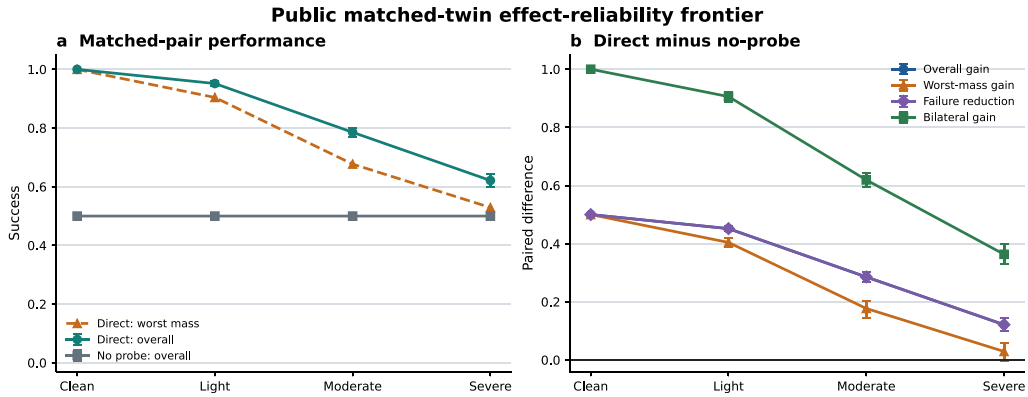

Fig. 5. Public matched-twin effect-reliability frontier over 20 independent runs. The same pair identifiers are preserved across clean, light, moderate, and severe effect-corruption regimes. Left: absolute direct-effect and equal-width no-probe performance. Right: run-level paired direct-effect minus no-probe differences; positive failure reduction means fewer terminal failures. Under severe corruption, the worst-mass interval crosses zero, exposing a prespecified failure boundary rather than a universal robustness claim.

satisfies the declared risk bound under an unknown or mixed shortcut, defer or re-probe rather than silently selecting the aligned winner. Fourth, expand independent runs or twins only until the prespecified directional interval and leave-one-stratum checks stabilize, then report coverage, conditional commit risk, overall failure, and probe cost on the untouched test set.

The supplementary scaling analysis reports the tested threshold, twin-count, effect-dimension, and runtime ranges that instantiate these stages. Those values are planning evidence for the present contracts, not universal constants. In a new application, failure to obtain overlap, a measurable positive control, or stable independent-unit intervals is itself an inconclusive audit outcome and should prevent a positive effect-use claim.

Concretely, the archived 20-run public pairs retain all 1200 pairs at $\delta_e = 0.005$, 1042 at 0.010, 124 at 0.015, and none at 0.020. Increasing retained twins per mass stratum from 2 to 12 narrows the paired success interval width from 0.030 to 0.016 in this archive. In a separate Monte Carlo calibration, a max-standardized-coordinate effect detector holds the null false-positive rate at 0.05, while detection power decreases from 0.979 at one dimension to 0.754 at 64 dimensions. A vectorized 10^5-pair, 64-dimensional distance pass takes 8.964 ms and 49.6 MiB peak traced memory on the reported CPU microbenchmark. At $d = 64$, measured time increases from 0.060 ms for 10^3 pairs to 0.643 ms for 10^4 and 8.964 ms for 10^5; at $N = 10^5$, the peak traced memory increases from 1.53 MiB at $d = 1$ to 49.6 MiB at $d = 64$.

These measurements give concrete, bounded planning rules. In this archive, $\delta_e = 0.010$ is the largest tested threshold that retains at least 80% of pairs while preserving the observed success difference; 0.015 is too selective for the same evidence budget. Two twins per stratum are the minimum tested count with a positive paired interval, but eight are required to reach the prespecified width target of 0.020 and twelve reach 0.016. For vector effects, standardize each coordinate, calibrate the maximum-coordinate threshold on null/repeated measurements, and increase the evidence budget when the calibrated power falls below the application's target. These rules must be recomputed for a new sensor and task; they are not hardware- or task-independent guarantees.

6.4. Scope and limitations

The evaluated visual representation is a fixed RGB feature extractor rather than an end-to-end learned perception system. The current evidence therefore concerns controlled simulation, fixed family parameterizations, declared input schemas, and the bounded public Causal-World construction. It does not establish physical deployment, raw-perception robustness, a complete causal graph, or universal dominance of one learned routing objective. A natural next use of the method is a richer public task whose matched-twin construction remains non-saturated, evaluated with the same frozen acceptance, provenance, and intervention checks.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During preparation and revision of this work, the authors used generative-AI tools solely for grammar checking and language editing. The authors reviewed and approved all revisions and take full responsibility for the article's content. Generative-AI tools were not used to generate scientific claims, select or fabricate data, run experiments, compute statistics, or determine the conclusions.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2026.100998.

Data availability

Code, manuscript source, generated figures and tables, configuration snapshots, and source data supporting the reported results are available at https://github.com/KleinChen42/MLWA_project, release `v1.1.0`. The release includes the raw-to-aggregate archive for the 20-run public audit, SHA-256 source-data manifests, environment requirements, and step-by-step reproduction commands in `REPRODUCTION.md`. The code is released under the MIT License.

References

Acarturk, E., Varici, B., Shanmugam, K., & Tajer, A. (2024). Sample complexity of interventional causal representation learning. In *Advances in neural information processing systems*: *Vol. 37*.

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Wüthrich, M., Bauer, S., Schölkopf, B., & Bengio, Y. (2021). CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International conference on learning representations*.

Atreya, P., Pertsch, K., Lee, T., Kim, M. J., Jain, A., et al. (2025). RoboArena: Distributed real-world evaluation of generalist robot policies. In *Proceedings of machine learning research*: *Vol. 305*, *Proceedings of the 9th conference on robot learning* (pp. 336–364).

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*(8), 966–1005.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, *23*(226), 1–61.

de Haan, P., Jayaraman, D., & Levine, S. (2019). Causal confusion in imitation learning. In *Advances in neural information processing systems*.

Deng, Z., Jiang, J., Long, G., & Zhang, C. (2023). Causal reinforcement learning: A survey. arXiv preprint arXiv:2307.01452.

Feldbaum, A. A. (1960). Dual control theory. I. *Avtomatika I Telemekhanika*, *21*(9), 1240–1249.

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., Dasigi, P., Dua, D., Elazar, Y., Gottumukkala, A., Gupta, N., Hajishirzi, H., Ilharco, G., Khashabi, D., Lin, K., Liu, J., Liu, N. F., Mulcaire, P., Ning, Q., ... Zhou, B. (2020). Evaluating models' local decision boundaries via contrast sets. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 1307–1323).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. arXiv preprint arXiv:2007.01434.

Hausknecht, M., & Stone, P. (2015). Deep recurrent Q-learning for partially observable MDPs. arXiv preprint arXiv:1507.06527.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134.

Klivans, A., Stavropoulos, K., & Vasilyan, A. (2024). Testable learning with distribution shift. In *Proceedings of machine learning research*: *Vol. 247*, *Proceedings of the 37th conference on learning theory* (pp. 2887–2943).

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., et al. (2021). WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of machine learning research*: *Vol. 139*, *Proceedings of the 38th international conference on machine learning* (pp. 5637–5664).

Li, Y., & Oliva, J. (2025). Towards cost sensitive decision making. In *Proceedings of machine learning research*: *Vol. 258*, *Proceedings of the 28th international conference on artificial intelligence and statistics* (pp. 3601–3609).

Liang, P., Bommasani, R., Lee, T., et al. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Lopez-Paz, D., Bottou, L., Schölkopf, B., & Vapnik, V. (2016). Unifying distillation and privileged information. arXiv preprint arXiv:1511.03643.

Margolis, G. B., Fu, X., Ji, Y., & Agrawal, P. (2023). Learning to see physical properties with active sensing motor policies. In *Proceedings of machine learning research*: *Vol. 229*, *Proceedings of the 7th conference on robot learning* (pp. 2537–2548).

Memmel, M., Wagenmaker, A., Zhu, C., Fox, D., & Gupta, A. (2024). ASID: Active exploration for system identification in robotic manipulation. In *The twelfth international conference on learning representations*.

Mesnard, T., Weber, T., Viola, F., Thakoor, S., Saade, A., Harutyunyan, A., Dabney, W., Stepleton, T. S., Heess, N., Guez, A., Moulines, E., Hutter, M., Buesing, L., & Munos, R. (2021). Counterfactual credit assignment in model-free reinforcement learning. In *Proceedings of machine learning research*: *Vol. 139*, *Proceedings of the 38th international conference on machine learning* (pp. 7654–7664).

Ng, I., Xie, S., Dong, X., Spirtes, P., & Zhang, K. (2025). Causal representation learning from general environments under nonparametric mixing. In *Proceedings of machine learning research*: *Vol. 258*, *Proceedings of the 28th international conference on artificial intelligence and statistics* (pp. 3700–3708).

Oberst, M., & Sontag, D. (2019). Counterfactual off-policy evaluation with gumbel-max structural causal models. In *Proceedings of machine learning research*: *Vol. 97*, *Proceedings of the 36th international conference on machine learning* (pp. 4881–4890).

Ren, K., Byun, Y., & Wilder, B. (2024). Decision-focused evaluation of worst-case distribution shift. In *Proceedings of machine learning research*: *Vol. 244*, *Proceedings of the 40th conference on uncertainty in artificial intelligence* (pp. 3076–3093).

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4902–4912).

Schoelkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalantidis, Y., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634.

Wilder, B., Dilkina, B., & Tambe, M. (2019). Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the thirty-third AAAI conference on artificial intelligence*.

Xiao, Z., Xing, H., Qu, R., Li, H., Feng, L., Zhao, B., & Wan, Q. (2026). Ensemble transitive bidirectional decoupled self-distillation for time-series classification. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *56*(3), 1587–1601.

Xie, A., Bhamidipaty, L. M., Liu, E. Z., Hong, J., Levine, S., & Finn, C. (2024). Learning to explore in POMDPs with informational rewards. In *Proceedings of machine learning research*: *Vol. 235*, *Proceedings of the 41st international conference on machine learning* (pp. 54414–54429).

Xie, X., Ho, J. W. K., Murphy, C., Kaiser, G. E., Xu, B., & Chen, T. Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, *84*(4), 544–558.

Ye, H., Xie, C., Cai, T., Li, R., Li, Z., & Wang, L. (2021). Towards a theoretical framework of out-of-distribution generalization. *Vol. 34*, In *Advances in neural information processing systems*.