THE EXPERIMENT DEFINES WHAT IS MEASURABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Must causal representation learning reconstruct an entire latent world when the scientific task asks for one target-wise measurement? Two stochastic hard interventions on a known target define such a measurement before any representation model is chosen: their density ratio cancels every non-target mechanism, and the Bayes environment logit reads out the remaining scalar quantity. Observation determines what this latent estimand becomes. Under an arbitrary invariant observation channel, the observable ratio is a posterior average of the latent intervention ratio, yielding a coordinate, a quotient, or a posterior observable. Alignment across entities or acquisition conditions imposes a separate requirement. The intervention laws also control the semantics and conditioning of the measurement. Exponential tilts program a chosen target statistic; among monotone tilts, affine statistics uniquely optimize worst-case conditioning under a weak symmetric-KL budget. Controlled experiments recover the programmed statistic across linear, monotone, and non-injective choices and predict a renderer-induced quotient before training. Matched comparisons with ILCM, CITRIS, BISCUIT, iVAE, and GSCALE-I produce the corresponding crossed profile: target-wise classification favors the programmed quotient, whereas interfaces designed to recover state or intervention structure retain more signed-state and intervention information. Natural-image and chemical-sensor studies further separate within-context discrimination from alignment under unseen entities and acquisition drift. The first identification question is therefore which measurement the experiment induces, before attempting full latent reconstruction.

1 INTRODUCTION

Interventions are usually treated as supervision for recovering a latent causal model. Yet a controlled experiment often asks for a narrower object: one usable readout for a known target, and, when identification permits, one coordinate. Reconstructing every latent, an inverse renderer, and possibly the causal graph then exceeds the scientific question.

Consider two replacement distributions that randomize one target; the experiment records only high-dimensional observations. Between the resulting hard interventions, every non-target causal mechanism is identical. Their density ratio therefore reduces to the ratio of the two replacement laws. A Bayes environment classifier recovers that ratio:

$$\boxed{\text{paired intervention} \longrightarrow \text{target-wise density-ratio experiment} \longrightarrow \text{one scalar measurement.}}$$

The experiment defines the measurement; binary classification serves as its direct readout.

What survives observation is less immediate. A sensor or renderer may discard exactly the distinctions encoded by the intervention ratio. For an arbitrary invariant observation channel,

$$\frac{dP_i^{1,X}}{dP_i^{0,X}}(X) = \mathbb{E}_{P_i^0}\left[\frac{dq_i^1}{dq_i^0}(Z_i)\,\middle|\,X\right].$$

The intervention pair selects the latent estimand, whereas the observation structure determines whether its observable form is a coordinate, a quotient, or a posterior observable. Injectivity yields a causal coordinate; non-injectivity in the ratio or renderer induces a quotient; further information loss leaves a posterior observable. Even a valid local measurement may then fail to travel: alignment across entities, nuisance conditions, or acquisition drift is a separate property rather than a consequence of

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.

classification accuracy. Figure 1 summarizes these successive roles of intervention, observation, and context.

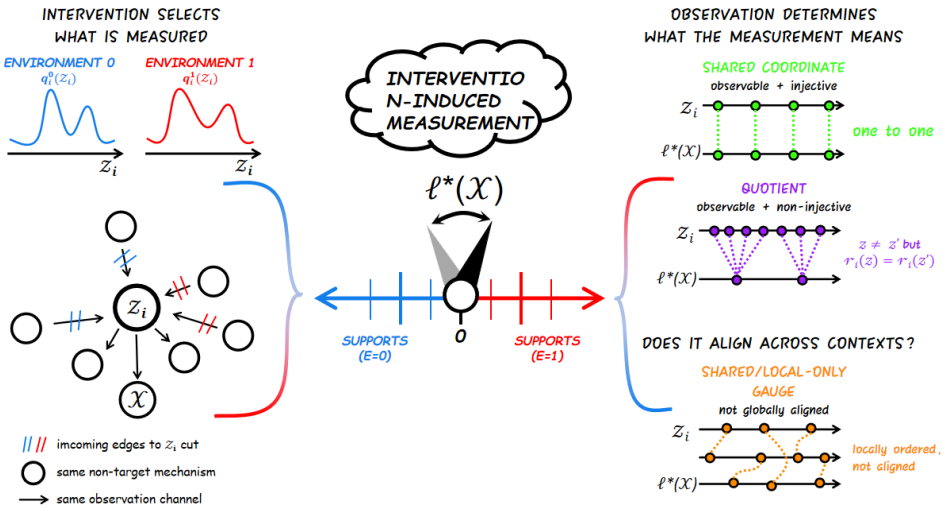


Figure 1: Paired interventions define a target-wise measurement before a representation model is chosen. The intervention laws select the latent density-ratio statistic, and an environment classifier reads out its observable form. Observation determines whether the visible form remains a coordinate, collapses to a quotient, or more generally becomes a posterior observable. Context separately determines whether the resulting gauge aligns across deployment conditions.

The intervention laws can also be chosen to program the measurement. If the second replacement law exponentially tilts the first by a statistic T_i, the Bayes logit measures a posterior log-moment of T_i and, when observable, exactly recovers T_i up to an affine transform. Hence intervention design is representation design. Among monotone exponential tilts, an affine tilt uniquely maximizes worst-case inverse conditioning per unit weak symmetric-KL budget. Identifiability alone is insufficient: an injective cubic coordinate is theoretically recoverable but locally ill-conditioned.

Relation to prior work. Most identifiability results fix an information interface and ask which latent representation it determines. Nonlinear ICA uses auxiliary variables or temporal structure to identify sources, while CRL uses interventions to recover causally related latents and, in some settings, their graph (Hyvärinen & Morioka, 2016; Khemakhem et al., 2020; Schölkopf et al., 2021; Ahuja et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023; Varici et al., 2024a; 2025). When single-node interventions, invertible observation, or complete observability are relaxed, the identifiable object can contract to a partial representation or causal abstraction (Bing et al., 2024; Varici et al., 2024b; Xu et al., 2024; Chen et al., 2024; Li et al., 2025). Grouped observations provide another interface, and cross-environment adaptation asks whether an identified representation remains reusable (Morioka & Hyvärinen, 2024; Talon et al., 2024). Across these results, the available contrast governs identification, observation can coarsen its object, and context can break reuse. What remains implicit is the design choice that precedes identification: which target-wise estimand an intervention contrast creates.

The present construction connects two strands usually studied apart. Binary classification recovers a density ratio (Gutmann & Hyvärinen, 2010; Menon & Ong, 2016), and environment labels can restore source identifiability in contrastive nonlinear ICA (Hyvärinen & Morioka, 2016). Target-matched hard interventions add the causal step: cancellation isolates the chosen target mechanism before any representation model is fitted. Observation then maps that latent ratio to a coordinate, quotient, or posterior observable, while context determines whether its gauge travels. The intervention pair therefore supplies more than supervision for a predetermined recovery objective: its replacement laws select the scalar estimand itself.

Interventional CRL clarifies what changes when the interface is designed for recovery rather than measurement. ILCM receives paired counterfactuals, CITRIS known-target temporal sequences,

and BISCUIT binary-interaction trajectories (Brehmer et al., 2022; Lippe et al., 2022; 2023). These interfaces are designed to retain signed state or intervention structure. Under matched images, environment-conditioned iVAE supplies the auxiliary-variable analogue, whereas GSCALE-I recovers a full representation and graph from score changes (Khemakhem et al., 2020; Varici et al., 2025). Evaluating their frozen representations against the programmed quotient, signed coordinate, and intervention target exposes the corresponding allocation of information: quotient-oriented contrasts reveal the programmed scalar, whereas recovery-oriented interfaces preserve signed state and intervention structure.

Contributions.

1. We derive the scalar measurement induced by paired hard interventions and characterize its exact form after arbitrary observation.
2. We distinguish observability from cross-context alignment and specify how a high-AUC coordinate can remain locally valid yet globally incompatible.
3. We establish that intervention choice programs representation semantics and prove affine optimality for weak-budget worst-case conditioning.
4. Controlled studies test programmed semantics, observational quotients, and contextual alignment, and separate target-wise measurement from full latent recovery under matched observations and interventions.

2 Paired interventions as causal measurements

2.1 Setup

Let $Z = (Z_1, \dots, Z_d)$ follow a Markovian structural causal model with density

$$p(z) = \prod_{k=1}^{d} p_k(z_k \mid z_{\mathrm{pa}(k)}).$$

For a known target i, environment $e \in \{0, 1\}$ is a stochastic hard intervention that replaces only the ith mechanism by a parent-free density q_i^e:

$$p_i^e(z) = q_i^e(z_i) \prod_{k \neq i} p_k(z_k \mid z_{\mathrm{pa}(k)}).$$

An arbitrary invariant Markov kernel $K(dx \mid z)$ generates the observations and admits stochastic renderers, observation noise, nuisance variables, and non-injective maps. We write $P_i^{e,X} = P_i^{e,Z} K$, assume common support, and define $r_i = dq_i^1 / dq_i^0$.

2.2 The measurement identity

Theorem 1 (Paired-intervention measurement)**.** *For the two hard interventions above,*

$$\frac{dP_i^{1,Z}}{dP_i^{0,Z}}(z) = r_i(z_i).$$

Moreover, under the reference environment P_i^0,

$$\boxed{\frac{dP_i^{1,X}}{dP_i^{0,X}}(X) = \mathbb{E}_{P_i^0}[r_i(Z_i) \mid X]} \quad P_i^{0,X}\text{-a.s.}$$

The first statement follows by direct cancellation of every non-target mechanism. For the second, for every bounded measurable f,

$$\mathbb{E}_{P_i^1}[f(X)] = \mathbb{E}_{P_i^0}[r_i(Z_i) f(X)] = \mathbb{E}_{P_i^0}[\mathbb{E}[r_i(Z_i) \mid X] f(X)],$$

and uniqueness of the Radon–Nikodym derivative completes the proof. No global invertibility, decoder, graph recovery, or recovery of Z_{-i} is required.

Let E be the binary environment label with prior π_e. The Bayes logit is

$$\ell_i^*(x) = \log \frac{dP_i^{1,X}}{dP_i^{0,X}}(x) + \log \frac{\pi_1}{\pi_0}.$$

We call $M_i(x) = \ell_i^*(x)$ the *intervention-induced causal measurement*.

Empirical readout. Given samples (X, E) from the two intervention environments, we fit a probabilistic classifier $\widehat{\eta}_i(x) \approx P(E = 1 \mid X = x)$ by binary cross-entropy and use the prior-corrected logit

$$\widehat{M}_i(x) = \log \frac{\widehat{\eta}_i(x)}{1 - \widehat{\eta}_i(x)} - \log \frac{\widehat{\pi}_1}{\widehat{\pi}_0}.$$

At the population optimum, this equals $\log dP_i^{1,X}/dP_i^{0,X}$. Thus minibatch balancing or unequal sample counts change only the known prior offset, not the measurement selected by the intervention laws. All implementations estimate this offset from the training split and freeze it thereafter.

Gauge and calibration. For an observable exponential tilt, $\widehat{M}_i$ targets $T_i(Z_i)$ up to the scale fixed by the tilt strength and an additive normalizing constant. Ordering claims therefore use the uncalibrated score, whose direction is fixed by the declared ordering of the two intervention laws. A numerical-coordinate claim additionally requires one affine calibration fitted on source validation data and frozen before transfer. If observation identifies only a quotient, no such calibration can restore distinctions within a fiber.

Boundary. Target-only cancellation is pair-specific: comparison with the observational regime or generic soft interventions generally retains parent dependence (Appendix D.11).

2.3 What is identified?

Corollary 2 (Observable-form trichotomy)**.** *Exactly three observable forms can arise:*

1. *If $r_i(Z_i)$ is $\sigma(X)$-measurable and r_i is injective, $M_i(X) = \log r_i(Z_i) + C$ is an invertible coordinate for Z_i.*
2. *If the observable ratio is non-injective in Z_i, M_i identifies only the level-set quotient $z \sim z' \Leftrightarrow r_i(z) = r_i(z')$ (or the corresponding quotient induced by observation aliasing).*
3. *If $r_i(Z_i)$ is not observable, M_i is the log of a posterior average, not a deterministic function of Z_i.*

Measurability is local to a context. Cross-context alignment imposes a second requirement: a quotient can be shared even when an injective coordinate remains context-local.

The exact equality between observed and latent ratios holds if and only if $r_i(Z_i)$ is $\sigma(X)$-measurable. Sigma-measurability is the minimal observation condition. Successful environment classification can coexist with poor coordinate recovery: AUC measures separability of $P_i^{0,X}$ and $P_i^{1,X}$, whereas coordinate recovery asks whether their likelihood ratio is a stable function of the target.

Many datasets additionally contain a context C, such as object identity, specimen, camera, or experimental batch. A classifier can induce a useful coordinate within every context yet assign a different numerical gauge to each.

Definition 3 (Shared numerical gauge)**.** *The paired measurement has a shared monotone gauge over contexts $c \in \mathcal{C}$ if there is one strictly monotone ϕ_i and context offsets b_c such that*

$$\ell_i^*(X, C = c) = \phi_i(Z_i) + b_c \quad \textit{a.s. for every } c \in \mathcal{C}.$$

Otherwise, context-wise discrimination does not define a transferable global coordinate.

Aligned ranks certify less than a shared numerical gauge. High context-wise Spearman establishes a common *orientation*, but remains compatible with context-specific monotone warps ϕ_c. Transferable ordering and numerical alignment must consequently be assessed separately.

3 DESIGNING THE RULER

Because the intervention pair defines the ruler, its laws are design variables. Choose a base replacement law q_i^0 and define

$$q_i^1(z_i) = \frac{\exp\{\lambda T_i(z_i)\} q_i^0(z_i)}{\mathbb{E}_{q_i^0} \exp\{\lambda T_i(Z_i)\}}.$$

Theorem 1 yields

$$\ell_i^*(X) = \log \mathbb{E}_{P_i^0}[\exp\{\lambda T_i(Z_i)\} \mid X] + C.$$

If $T_i(Z_i)$ is observable from X, then $\ell_i^*(X) = \lambda T_i(Z_i) + C$. Thus the intervention pair programs the meaning of the representation: a linear T programs an ordered coordinate, a symmetric T a quotient, and a monotone nonlinear T a warped coordinate.

Semantics is not enough; the programmed measurement must be invertible with finite data. Let P be a fixed non-degenerate base law on an interval and consider the symmetric pair

$$q_\lambda^\pm(dz) = \exp\{\pm\lambda T(z) - A(\pm\lambda)\} P(dz), \qquad A(s) = \log \mathbb{E}_P e^{sT(Z)}.$$

For increasing absolutely continuous T, define $a_T = \operatorname{ess\,inf} T'(z) > 0$. The paired log ratio has minimum slope $m_\lambda = 2\lambda a_T$. Let the intervention budget be the symmetric KL

$$J_\lambda(T) = D_{\mathrm{KL}}(q_\lambda^+ \| q_\lambda^-) + D_{\mathrm{KL}}(q_\lambda^- \| q_\lambda^+).$$

Theorem 4 (Local design optimality)**.** *If exponential moments exist near zero, then*

$$\lim_{\lambda \to 0} \frac{m_\lambda^2}{J_\lambda(T)} = \frac{a_T^2}{\mathrm{Var}_P[T(Z)]} \le \frac{1}{\mathrm{Var}_P(Z)}.$$

Equality holds if and only if $T(z) = az + b$ P-almost surely for some $a > 0$. Hence affine tilts uniquely maximize worst-case inverse conditioning per unit weak symmetric-KL budget.

The proof writes $T(z) = a_T z + h(z)$ with nondecreasing h and uses $\mathrm{Cov}(Z, h(Z)) \ge 0$; Appendix A gives details. The bound ranks $T(z) = z$ as optimal, a bi-Lipschitz warp as stable but less efficient, and $T(z) = z^3$ as worst-case inefficient despite its injectivity. A KL constraint without a fixed base law or coverage constraint remains ill-posed because it may concentrate nearly all mass on an arbitrarily small interval.

Learning the designed ruler. Under overlap, classifier regret controls measurement error through the conditioning induced by T. Appendix B states the exact inequalities and minimax equivalence and relates the target-wise guarantee to finite-sample interventional CRL (Acartürk et al., 2024).

4 EXPERIMENTS

4.1 MEASUREMENT CRITERIA

A scalar logit need not constitute a scientific measurement even when it classifies environments accurately. Its intervention law may encode a statistic alien to the scientific target, observation may preserve only a quotient, and deployment contexts may deny it a common gauge. Each study isolates one boundary and holds the other two fixed. Favorable identification alone does not guarantee easy empirical recovery, even in a simple physical system (Gamella et al., 2025).

Classifiers receive raw observations and binary environment labels; latent coordinates enter only the evaluation. Appendix C specifies architectures, sampling laws, preprocessing, and seed-level results. The complete ledger in Appendix D records robustness analyses, theorem checks, negative controls, downstream graph diagnostics, and discarded exploratory branches.

4.2 INTERVENTION SELECTS THE DIRECT READOUT

Programmed statistics under a fixed learner. Only T changes; the renderer, CNN, sample count, and paired Jeffreys budget (0.4) remain fixed. One binary learner recovers an ordered coordinate, a nonlinear monotone ruler, and a non-injective quotient, each with mean $|\rho(\ell, T)| > 0.97$ (Table 1). The quadratic design preserves quotient correlation at 0.972 ± 0.022 and reduces signed correlation to 0.070 ± 0.048. Under a matched budget, conditioning follows the affine $>$ warped $>$ cubic order established by Theorem 4; Appendix C reports the calibration and seed-level analyses.

Table 1: Shared-renderer estimand spectrum (five seeds). Only T changes; both environments contain 800 images and have matched Jeffreys budget.

$T(z)$	Identified object	AUC	$\lvert\rho(\ell, T)\rvert$	$R^2(\ell, T)$	$\lvert\rho(\ell, z)\rvert$
z	coordinate	$.646 \pm .023$	$.973 \pm .016$	$.944 \pm .027$	$.973 \pm .016$
$z + 0.5z^3$	monotone coordinate	$.643 \pm .022$	$.978 \pm .018$	$.916 \pm .045$	$.978 \pm .018$
z^2	sign quotient	$.656 \pm .017$	$.972 \pm .022$	$.933 \pm .020$	$.070 \pm .048$

CRL interfaces on a common SCM. A shared three-variable SCM instantiates four experimental interfaces: paired scale environments for the binary measurement, paired counterfactuals for ILCM, known intervention-target trajectories for CITRIS, and signed action trajectories for BISCUIT. Binary, CITRIS, and BISCUIT share a renderer; ILCM retains its native invertible observation adapter. A common evaluator probes the frozen representations on independent held-out data.

The interfaces preserve different invariants. Target-only cancellation makes the scale-pair score a direct readout of the programmed quotient. Paired and temporal supervision instead retains signed state and intervention identity. Quotient correlation, signed-coordinate correlation, and intervention-target accuracy resolve the distinction under one evaluator.

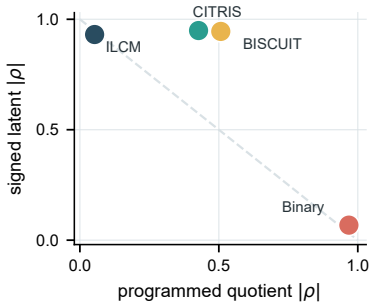


Figure 2: Estimand choice reverses the method ranking.

Table 2: Supervision-complete CRL bridge (mean $\pm$ s.d., three seeds) under one frozen-representation evaluator.

Method / experimental interface	Programmed z_0^2	Signed z_0	Intervention target
Binary / scale pair	$\mathbf{0.968 \pm 0.018}$	0.068 ± 0.045	0.396 ± 0.005
ILCM / paired counterfactual	0.053 ± 0.005	0.931 ± 0.019	0.532 ± 0.023
CITRIS / target trajectory	0.427 ± 0.129	$\mathbf{0.949 \pm 0.037}$	$\mathbf{0.598 \pm 0.013}$
BISCUIT / action trajectory	0.507 ± 0.040	0.945 ± 0.010	0.587 ± 0.017

Binary reaches 0.968 on the programmed quotient but only 0.068 on the signed coordinate. ILCM, CITRIS, and BISCUIT retain signed-coordinate scores above 0.93 and intervention-target accuracy above 0.53, while none matches the binary quotient score.

The reversal persists under strict pixel matching. Binary, an environment-conditioned iVAE (Khemakhem et al., 2020), and GSCALE-I use the same official images, 800 samples per hard-intervention environment, and five seeds. Their (programmed quotient, signed latent) scores are respectively $(0.746 \pm 0.081, 0.120 \pm 0.064)$, $(0.516 \pm 0.136, 0.336 \pm 0.108)$, and $(0.095 \pm 0.046, 0.800 \pm 0.148)$. The ranking again reverses with the estimand; full supervision, seed, and compute audits appear in Appendices D.9 and D.10.

4.3 Observation determines the identifiable object

The dSprites ellipse (Matthey et al., 2017) is invariant to a half-turn, so raw orientation is not observable. For the paired fibers and symmetric intervention laws, the observed log ratio reduces to the sum of the two aliased latent codes. Its five quotient levels are fixed before training. Across three seeds, the learned logit has absolute Spearman 0.124 ± 0.056 with the raw angle but 0.820 ± 0.019 with the predicted quotient (Figure 3). The model has not failed to recover angle; the experiment and renderer never identified angle in the first place.

4.4 Context determines alignment

smallNORB: aligned ordering across entities. Two randomized elevation environments are formed from raw smallNORB images (LeCun et al., 2004). Training and validation occupy

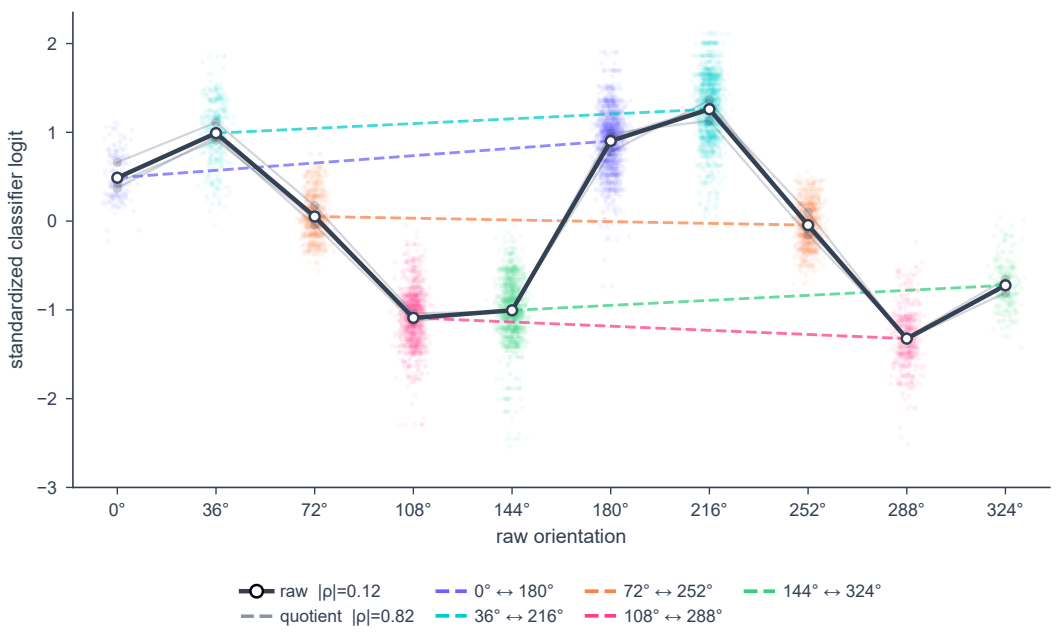


Figure 3: Observation determines the identifiable object. The dSprites ellipse aliases orientations separated by $180°$ in the evaluated setup. The classifier logit is weakly associated with the ten-level raw angle ($|\rho| = 0.12$) but strongly associated with the parameter-free five-level quotient predicted from the renderer ($|\rho| = 0.82$). Colors identify the five aliased angle pairs.

disjoint source objects; evaluation uses the official 25 unseen test instances. For each fixed object×azimuth×lighting context, we directly rank nine raw images by their logits without averaging over nuisances. Across five seeds and 13,500 strict contexts, mean Spearman is 0.9764 ± 0.0035, the seed-wise 5% quantile is 0.9200 ± 0.0139, 96.73% of contexts reach 0.9, and 99.985% retain the correct sign.

Targeted shortcut tests leave the ordering intact. A linear pixel probe reaches only 0.7687 mean strict Spearman and has a negative 5% tail. Foreground centering and scale normalization reduce the CNN score from 0.9711 to 0.9300 for seed 17 but leave it strongly positive. Low-level geometry contributes without accounting for the full coordinate. A lighting intervention evaluated against elevation yields -0.2628 mean Spearman; identical null environments yield -0.0249. Neither the evaluator nor the architecture manufactures monotonicity under these controls.

COIL-100: local discrimination without a shared gauge. Across entities, alignment requires more than repeatable local phase; the images must also admit a common visual origin. COIL-100 (Nene et al., 1996) contains motorized turntable images, but different objects do not share a visually aligned zero orientation. Complementary intervention pairs program $\cos\theta$ and $\sin\theta$. Across five seeds, the classifiers retain validation AUC 0.904 ± 0.006 and seen-object correlations $0.923/0.917$, yet correlations on held-out objects fall to $0.261/0.118$; $28.0\%/41.3\%$ even reverse sign. Global angle MAE is $80.52° \pm 2.66°$. An oracle object-specific affine gauge reduces it to $50.52° \pm 2.17°$ but cannot restore global alignment. Strong environment separability thus remains local evidence, not proof of a shared coordinate.

Chemical sensing: aligned ordering under temporal drift. Paired concentration policies are replayed on the UCI Gas Sensor Array Drift dataset (Fonollosa et al., 2015; Vergara, 2013). Six physical gases were dosed at controlled concentrations and measured by 16 chemical sensors over 36 months. For each gas, a binary classifier receives only the 128-dimensional sensor response and a batch-balanced randomized environment label; concentration implements the intervention policy but is hidden from the model. The policy $P(E = 1 \mid z, c) = \sigma(2\beta z + b_c)$ specifies one concentration slope and the known batch offset b_c. Batch balance fixes the offset and preserves the common slope.

Table 3: Cross-context alignment over five seeds. Environment AUC is measured on source validation data; coordinate statistics are computed within the stated held-out evaluation unit. The evaluation contains 13,500 smallNORB contexts, 70/30 COIL objects per phase, and 60 gas×month×seed contexts.

Setting	Evaluation unit	Env. AUC	Coordinate alignment / transfer
smallNORB	unseen object × azimuth × lighting	0.879 ± 0.003	mean 0.9764 ± 0.0035; 5% 0.9200 ± 0.0139; 99.985% positive
COIL-100 seen	seen object × phase	0.904 ± 0.006	cos/sin $0.923/0.917$
COIL-100 unseen	unseen object × phase	0.904 ± 0.006	cos/sin $0.261/0.118$; 28.0%/41.3% reverse
Gas policy	gas × held-out month	0.727 ± 0.044	mean/median $0.838/0.893$; 5% 0.642; 58/60 positive
Gas null	same gas × month units	0.529 ± 0.050	mean/median $-0.091/-0.183$; 25/60 positive

It is part of the estimand, not an incidental implementation choice. Batches 1–6 are source-disjoint training/validation data; batches 7 and 10, collected at months 21 and 36, are held out from model fitting, early stopping, and sign selection.

Across 6 gases $\times 2$ held-out batches $\times 5$ seeds, strict Spearman has mean 0.838, median 0.893, and 5% quantile 0.642; 58/60 contexts retain the correct direction. A concentration-independent null has AUC 0.529 ± 0.050 and only 25/60 positive contexts. A fixed linear classifier still reaches mean Spearman 0.701 and 59/60 positive contexts.

A stricter audit fits one affine logit-to-concentration map on source validation data, freezes it, and subtracts only the known policy offsets at test time. Held-out sample-level R^2 is positive in 50/60 contexts (median 0.472), and 54/60 slopes lie within a factor of two of the source slope. Two ammonia reversals drive the mean R^2 below zero. Ordering transfers broadly, but a shared numerical gauge transfers only partially. Because the study replays archived physical exposures, it does not constitute a prospective experiment.

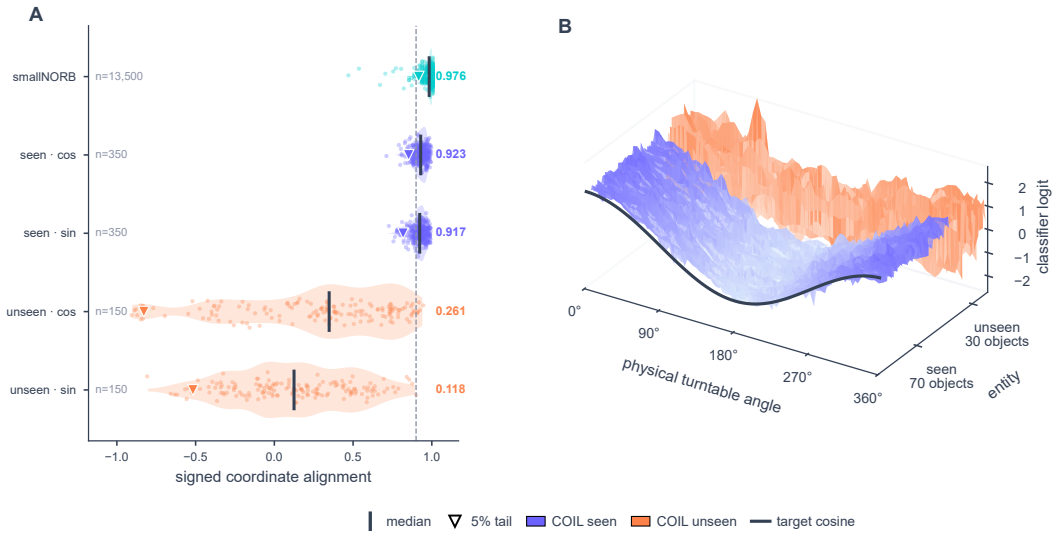


Figure 4: The cross-context alignment boundary over five seeds. **A:** correlations across all 13,500 strict smallNORB contexts concentrate near one, whereas COIL correlations collapse and often reverse on held-out objects despite strong seen-object values. **B:** saved COIL trajectories expose the context-dependent rotation field. Local loops persist, but their gauges do not align into one transferable coordinate.

Table 3 makes the contrast quantitative: smallNORB supports a shared order, COIL-100 only local gauges, and chemical sensing an order that largely survives drift. Appendix D.6 reports seed-level distributions, sample-efficiency, and controls.

5 DISCUSSION

The intervention pair chooses a latent statistic, the observation channel determines its observable form, and the classifier reads out the resulting measurement.

An injective observable ratio yields a coordinate; a non-injective ratio yields a quotient; further information loss leaves a posterior observable. Environment accuracy cannot distinguish these objects. COIL-100 makes the gap visible: AUC remains high while held-out objects often reverse coordinate direction. Evidence for a coordinate must therefore include context-wise orientation and tail behavior, not separability alone.

Identification within one context does not make a measurement portable. smallNORB admits a common order across unseen entities and nuisances, whereas COIL-100 admits only incompatible object-specific gauges. Chemical sensing occupies the middle ground: temporal drift largely preserves direction but weakens numerical calibration. Observable form and cross-context alignment are different properties because either may survive without the other.

The theory assumes a known target and paired hard interventions. It cannot recover information discarded by the sensor, nor can rank transfer certify a common numerical gauge. Paired and temporal CRL nevertheless enter through their native interfaces, while the iVAE/GSCALE-I audit holds the images fixed.

The principal empirical limitation is that the main real-data studies do not place the intervention laws inside a prospective data-collection loop. A definitive test requires a controllable physical, biological, or chemical system in which replacement policies are fixed before acquisition and source calibration remains frozen across later contexts. GPU telemetry provides an accessible physical testbed for an initial implementation of this protocol. Two randomized workload-replacement policies were fixed before collection and executed across 24 physical intervention blocks. All blocks completed without a safety stop. Commanded workload correlated with utilization ($\rho = 0.925$) and power ($\rho = 0.889$), but less strongly with temperature ($\rho = 0.370$), whose dynamics retain thermal memory. Session-internal cross-validation gave environment AUC 0.660, indicating prospective separability under fixed intervention laws. Because all blocks belong to one session, the pilot does not establish alignment across acquisition contexts. An independently collected session is the next test of alignment; the same design criterion applies to controlled chemical and biological experiments, where intervention laws must precede collection and evaluation must be reserved for unseen acquisition batches.

6 CONCLUSION

Paired hard interventions do more than supervise a representation learner: they define a causal measurement. Their likelihood ratio eliminates all non-target mechanisms. The observation channel fixes the observable form, and deployment context governs its alignment. Separating these roles turns intervention selection into representation design and preserves the distinction between observability and cross-context transfer. A binary classifier suffices because the experiment performs the causal reduction. Causal representation learning need not begin with reconstruction of the latent world. It can first determine what the experiment has made measurable.

AI USE STATEMENT

In this work, we used generative AI tools for AI-assisted writing and language polishing to improve readability, as well as for code review and correction. The authors reviewed and verified all AI-assisted text, claims, code, and corrections. The authors retain full responsibility for the final content of this work, including all material produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs supporting the theoretical results are provided in Appendices A and B. Experimental protocols, dataset construction and preprocessing, intervention laws, architectures, optimization settings, evaluation procedures, and audit criteria are documented in Appendices C and D. These appendices also report seed-level results and the controls needed to trace the aggregate claims in the main text.

REFERENCES

Emre Acartürk, Burak Varici, Karthikeyan Shanmugam, and Ali Tajer. Sample complexity of interventional causal representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39350–39385, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/458fa8ee331566383d8e74bdb647f829-Abstract-Conference.html.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 843–867, 2024. URL https://proceedings.mlr.press/v236/bing24a.html.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f8b7aa3a0d349d9562b424160ad18612-Abstract.html.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. CaRiNG: Learning temporal causal representation under non-invertible generation process. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7236–7259, 2024. URL https://proceedings.mlr.press/v235/chen24ai.html.

Jordi Fonollosa, Irene Rodríguez-Luján, and Ramón Huerta. Chemical gas sensor array dataset. *Data in Brief*, 3:85–89, 2015. doi: 10.1016/j.dib.2015.01.003.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18143–18169, 2025. URL https://proceedings.mlr.press/v267/gamella25a.html.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, 2010. URL https://proceedings.mlr.press/v9/gutmann10a.html.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 3765–3773, 2016.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo P. Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217. PMLR, 2020.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. II–104–II–111, 2004. doi: 10.1109/CVPR.2004.1315150.

Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the identifiability of causal abstractions. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3241–3249, 2025. URL https://proceedings.mlr.press/v258/li25g.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Stratis Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13557–13603. PMLR, 2022. URL https://proceedings.mlr.press/v162/lippe22a.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal representation learning from binary interactions. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 1263–1273, 2023. URL https://proceedings.mlr.press/v216/lippe23a.html.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. Online dataset, 2017. URL https://github.com/google-deepmind/dsprites-dataset.

Aditya Menon and Cheng Soon Ong. Linking losses for density ratio and class-probability estimation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 304–313, 2016. URL https://proceedings.mlr.press/v48/menon16.html.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36249–36293, 2024. URL https://proceedings.mlr.press/v235/morioka24a.html.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-100). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, February 1996. URL https://cave.cs.columbia.edu/Statics/publications/pdfs/Nene_TR96_2.pdf.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1039–1048. PMLR, 2020. URL https://proceedings.mlr.press/v124/squires20a.html.

Davide Talon, Phillip Lippe, Stuart James, Alessio Del Bue, and Sara Magliacane. Towards the reusability and compositionality of causal representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 296–324, 2024. URL https://proceedings.mlr.press/v236/talon24a.html.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2314–2322, 2024a. URL https://proceedings.mlr.press/v238/varici24a.html.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ca70528fb11dc8086c6a623da9f3fee6-Abstract-Conference.html.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025. URL https://www.jmlr.org/papers/v26/24-0194.html.

Alexander Vergara. Gas sensor array drift at different concentrations. UCI Machine Learning Repository, 2013.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/97fe251c25b6f99a2a23b330a75b11d4-Abstract-Conference.html.

Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55389–55433, 2024. URL https://proceedings.mlr.press/v235/xu24ac.html.

A Proofs

A.1 Exact observability and the observable-form trichotomy

Let $R(X) = dP_i^{1,X}/dP_i^{0,X}(X)$. Theorem 1 gives $R(X) = \mathbb{E}[r_i(Z_i) \mid X]$. If $r_i(Z_i)$ is $\sigma(X)$-measurable, the conditional expectation equals $r_i(Z_i)$. Conversely, if $R(X) = r_i(Z_i)$ almost surely, then the right-hand side is a measurable version of $r_i(Z_i)$, proving necessity. If r_i is injective, its measurable inverse recovers Z_i on its range. If not, any two values on one level set have identical Bayes logit, so a classifier restricted to the environment experiment cannot distinguish them through this measurement.

A.2 Proof of Theorem 4

The paired log ratio is

$$\phi_\lambda(z) = 2\lambda T(z) + A(-\lambda) - A(\lambda),$$

so $m_\lambda = 2\lambda a_T$. A direct exponential-family calculation gives

$$J_\lambda(T) = 2\lambda[A'(\lambda) - A'(-\lambda)] = 4\lambda^2 \operatorname{Var}_P[T(Z)] + O(\lambda^4).$$

Therefore

$$\lim_{\lambda \to 0} \frac{m_\lambda^2}{J_\lambda(T)} = \frac{a_T^2}{\operatorname{Var}_P[T(Z)]}.$$

Write $T(z) = a_T z + h(z)$, where h is nondecreasing. For an independent copy Z',

$$\operatorname{Cov}(Z, h(Z)) = \frac{1}{2}\mathbb{E}[(Z - Z')(h(Z) - h(Z'))] \geq 0.$$

Consequently,

$$\operatorname{Var}[T(Z)] = a_T^2 \operatorname{Var}(Z) + 2a_T \operatorname{Cov}(Z, h(Z)) + \operatorname{Var}(h(Z)) \geq a_T^2 \operatorname{Var}(Z).$$

Equality forces h to be constant on the support, hence T is affine. The same proof holds on ordered finite support with the minimum adjacent slope.

B Finite-classifier guarantees

Theorem 5 (Logistic regret to logit error)**.** *If both $\eta(X) = P(E = 1 \mid X)$ and $\widehat{\eta}(X) = \sigma(\widehat{\ell}(X))$ lie in $[\alpha, 1 - \alpha]$, then*

$$\frac{\alpha(1-\alpha)}{2}\mathbb{E}(\widehat{\ell} - \ell^*)^2 \leq \mathcal{L}(\widehat{\ell}) - \mathcal{L}(\ell^*) \leq \frac{1}{8}\mathbb{E}(\widehat{\ell} - \ell^*)^2.$$

Proof. Conditional logistic risk is the Bregman divergence of $A(\ell) = \log(1 + e^{\ell})$. Since $A''(\ell) = \sigma(\ell)(1 - \sigma(\ell))$ lies between $\alpha(1 - \alpha)$ and $1/4$ on the clipped range, Taylor's theorem followed by expectation gives the result. □

Theorem 6 (Minimax equivalence)**.** *Let $\mathcal{P}$ be a class of experiments satisfying $\ell^*_P(X) = \phi(Z)$ almost surely for a known increasing ϕ with $0 < m \le \phi'(z) \le M < \infty$. If $\mathfrak{R}_n^{\mathrm{logit}}$ and $\mathfrak{R}_n^{\mathrm{coord}}$ are the minimax squared prediction risks for the Bayes logit and coordinate, respectively, then*

$$\frac{\mathfrak{R}_n^{\mathrm{logit}}}{M^2} \le \mathfrak{R}_n^{\mathrm{coord}} \le \frac{\mathfrak{R}_n^{\mathrm{logit}}}{m^2}.$$

Proof. Any coordinate estimator $\widehat{z}$ induces $\widehat{\ell} = \phi(\widehat{z})$, whose squared error is at most M^2 times coordinate error. Conversely, project any logit estimator onto the range of ϕ and invert it; the resulting coordinate error is at most logit error divided by m^2. Taking the relevant infima and suprema proves both bounds. □

C Experimental details and audits

smallNORB. The official training objects are split into 20 optimization source objects and five source-disjoint validation objects. The official 25 test objects are never used for optimization or model selection. Each strict evaluation unit fixes object, azimuth, and lighting and ranks the nine elevation images. The main five-seed audit uses 8,000 optimization and 2,000 validation samples per environment. The sample-efficiency experiment instead treats $100, 500, 8000$ as total per-environment budgets and uses an 80/20 split. Its $n = 8000$ estimate is consequently not directly comparable with the main audit.

Table 4: Five-seed strict smallNORB audit.

Seed	Mean ρ	5% quantile	$\rho \ge 0.9$	Adjacent accuracy
17	0.9711	0.9000	95.26%	88.96%
23	0.9750	0.9167	96.22%	89.65%
47	0.9778	0.9167	96.59%	90.64%
71	0.9783	0.9333	97.96%	90.67%
101	0.9800	0.9333	97.59%	90.75%
Mean ± s.d.	0.9764 ± 0.0035	0.9200 ± 0.0139	$96.73\% \pm 1.08\%$	$90.13\% \pm 0.80\%$

COIL-100. Images are resized to 32×32. Seventy objects are used for training and 30 are held out. Each object contains 72 motor angles at 5° spacing. Two classifiers implement complementary phase tilts for cosine and sine. The five-seed pooled audit contains 700 seen and 300 held-out object-phase units and stores 72,000 trajectory values.

Chemical sensor array. The dataset contains 13,910 controlled physical exposures collected over 36 months (Fonollosa et al., 2015). Each observation is a 128-component summary of the steady-state and transient responses of 16 metal-oxide sensors. For each gas separately, we normalize log concentration with source statistics and assign each physical row once according to $P(E = 1 \mid z, \text{batch}) = \sigma(2\beta z + b_{\text{batch}})$ with $\beta = 1.5$. The batch-specific intercept balances environment frequency, which blocks acquisition batch as an environment shortcut and retains the shared concentration slope 2β. The designed Bayes logit therefore has the policy-level form $\phi(z) + b_c$. Optimization and validation rows are disjoint within every batch×concentration group. Batches 7 (month 21) and 10 (month 36) were predeclared as tests because both are later periods with at least five concentration levels for all six gases. Their concentrations instantiate the offline policy and evaluate its coordinate, but they are not used for model fitting, early stopping, calibration refitting, or sign selection.

Matched representation protocol. All methods use the same official GSCALE-I image pixels, 800 samples per hard environment, intervention targets, and seeds $\{11, 23, 47, 71, 101\}$. The binary method uses paired hard-intervention pixels and target identity. iVAE uses the observational and four

interventional pixel sets, with environment identity as its auxiliary variable. GSCALE-I uses the same pixels and additionally uses hard-versus-observational score ratios, as required by its implementation. iVAE has a convolutional encoder/decoder, a two-dimensional diagonal Gaussian posterior, and an environment-conditioned factorized Gaussian prior. We fixed $\beta = 10^{-4}$, 50 epochs, and patience 10 for all seeds; no latent test label selects a hyperparameter. Timing excludes shared data generation.

Table 5: Baseline coverage and protocol compatibility. Strict pixel matching holds observations fixed; the supervision-complete bridge supplies each method's native interface on a common SCM.

Method	Method-specific observation or supervision	Status
iVAE	Independent samples with an auxiliary environment variable	Strict same-pixel match
GSCALE-I	Observation and known-target hard-intervention environments	Strict same-pixel match
ILCM	Paired pre/post-intervention counterfactual observations	Native-supervision bridge
CITRIS	Temporal intervened sequences	Native-supervision bridge
BISCUIT	Binary-interaction trajectories	Native-supervision bridge
CRL from general environments	Many environments with general mechanism variation	Not run: only single-target environments
Reusable/compositional CRL	Source representations and a task-adaptation protocol	Not run: transfer diagnostic only

Reproducibility. All reported aggregate values are computed from saved seed-level JSON files. The figure source data, fixed context selections, leakage audit, and plotting scripts are included with the submission. No test coordinate label is used for training, early stopping, selecting a sign, or choosing displayed contexts.

D Complete experimental ledger

The ledger records every completed experiment relevant to the measurement framework. Confirmatory tests remain distinct from downstream and exploratory diagnostics. Unsuccessful branches constrain the scope of the evidence and do not enter the confirmatory claims.

D.1 Continuous and procedural-image pilots

Table 6: Initial tests of target-wise cancellation. Coordinate scores are mean rank correlations; graph results are downstream diagnostics.

Setting	Coordinate	Graph result	Interpretation
Continuous, 4 nodes, 3 DAGs $\times$ 3 seeds	≈ 0.97	8/9 exact	target cancellation
Continuous, 8-node mixed graph	0.937	AUC/AP/F1 0.994/0.978/0.909	scales beyond 4 nodes
Non-injective continuous channel	0.078	env. AUC 0.771	AUC is insufficient
Procedural 32$\times$32 images, 6 runs	0.977–0.983	6/6 exact	pixel-level readout

The non-injective continuous experiment uses paired Gaussian scale interventions. The learned logit has mean Pearson 0.938 (minimum 0.929) with the programmed squared factors, despite rank correlation 0.078 with the original signed factors. The apparent coordinate failure is an early quotient result: environment classification succeeds on the programmed squared factors.

D.2 SHAPES3D: RECOVERY, NUISANCE, AND SAMPLE SIZE

Shapes3D uses the official 480,000-image Cartesian product as the renderer. A discrete four-node SCM selects ordered factors, and opposite exponential tilts make the population log ratio affine in the target.

Table 7: Shapes3D confirmatory suite. The main and independent-pixel-nuisance rows each contain three DAGs and three seeds.

Condition	Samples/env.	Mean coordinate ρ	Downstream graph
Main hard interventions	1,800	0.979 ± 0.005	9/9 exact
Weak hard interventions	1,800	0.950	3/3 exact
Independent gain/offset/noise	1,800	0.974	9/9 exact
Low-sample hard interventions	200	0.904	5/9 exact
Low-sample hard interventions	400	0.957	4/9 exact

The non-monotone 5/9 versus 4/9 exact graph count at 200 versus 400 samples is variance in the second-stage graph score; we do not present it as a monotone graph sample-complexity curve. Coordinate recovery itself improves.

Native Shapes3D nuisance randomizes shape and scale at the latent level; it does not add post-render noise. Its hardest case is the collider:

Table 8: Native-nuisance collider sample curve (three-seed means).

Samples/intervention sign	Coordinate Spearman	Graph F1
100	0.797	0.444
200	0.884	0.556
400	0.937	0.778
600	0.953	1.000

At 600 samples, the three seed-wise coordinate scores are 0.967/0.946/0.947 and all graphs are exact. Exact graph recovery at this budget separates the population identity from finite visual sample complexity.

D.3 PROGRAMMED ESTIMANDS AND MATCHED INTERVENTION BUDGETS

The shared-renderer spectrum in Table 1 uses an 81-level target support on $[-2, 2]$, 3,000 fresh observational evaluation images, 20 training epochs, and seeds $7, 17, 23, 47, 101$. For every statistic, the two target replacement laws are $p_{\pm}(z) \propto \exp\{\pm\lambda T(z)\}$, with λ solved separately so that the paired Jeffreys divergence is exactly 0.4. Consequently the population log ratio is exactly $2\lambda T(z)$ (affine $R^2 = 1$) in all three conditions; only the programmed ruler changes. The finite-CNN seed audit is:

Table 9: Seed-level audit for the shared-renderer estimand spectrum. Entries separated by slashes follow seeds $7/17/23/47/101$.

Statistic	Programmed $\lvert\rho\rvert$ by seed	Signed-z $\lvert\rho\rvert$ by seed	Programmed affine R^2 (mean $\pm$ s.d.)
z	.950/.981/.993/.970/.974	.950/.981/.993/.970/.974	$.944 \pm .027$
$z + 0.5z^3$	.951/.988/.995/.987/.968	.951/.988/.995/.987/.968	$.916 \pm .045$
z^2	.977/.987/.986/.978/.934	.024/.089/.110/.013/.114	$.933 \pm .020$

Table 10: Programmed-coordinate tests before matching intervention strength.

Renderer/design	Intended estimand	Recovery	Consequence
Continuous scale pair	Z_i^2	Pearson 0.938	signed factor lost
Shapes3D quadratic	Z_i^2	Pearson 0.899	signed $\rho = 0.262$
Shapes3D cubic	Z_i^3	min. $\rho = 0.323$	ill-conditioned
Shapes3D $z + 0.5z^3$	warped coordinate	Pearson 0.921; min. $\rho = 0.890$	graph exact

The direct test of Theorem 4 assigns every statistic the same paired Jeffreys budget and 800 examples per intervention.

Table 11: Matched-budget intervention design (three seeds).

Budget	Statistic	Min. log-ratio slope	Programmed ρ	Inversion	Population efficiency
0.15	linear	0.6091	0.642 ± 0.211	0.188 ± 0.084	2.455
	warped	0.4468	0.656 ± 0.086	0.184 ± 0.033	1.324
	cubic	0.0095	0.270 ± 0.332	0.349 ± 0.140	0.000609
0.40	linear	1.0010	0.870 ± 0.031	0.084 ± 0.018	2.455
	warped	0.7328	0.800 ± 0.098	0.115 ± 0.045	1.324
	cubic	0.0156	0.356 ± 0.167	0.300 ± 0.060	0.000609

At the weaker budget, finite-CNN variation leaves linear and warped statistically tied. Population efficiency retains the exact ordering, whereas the finite-sample scores separate clearly at the moderate budget.

D.4 dSprites: observability, quotient prediction, and finite samples

Table 12: dSprites renderer boundary at 800 samples per intervention. The ellipse aliases orientations separated by π; the asymmetric heart removes that aliasing.

Renderer	DAG	Mean/min. coordinate ρ	Graph F1	Reduction F1
Ellipse	chain	0.772/0.183	–	–
	fork	0.732/0.020	–	–
	collider	0.763/0.147	–	–
Heart	chain	0.880/0.625	1.000	1.000
	fork	0.888/0.657	0.750	0.857
	collider	0.899/0.714	1.000	1.000

For the ellipse, the exact predicted five-level quotient is evaluated on 3,000 fresh observations per seed:

Table 13: Ellipse quotient audit (three seeds).

Seed	$\lvert\rho\rvert$ with raw angle	$\lvert\rho\rvert$ with quotient	Quotient Pearson
11	0.147	0.809	0.818
22	0.165	0.809	0.870
33	0.061	0.841	0.859
Mean $\pm$ s.d.	0.124 ± 0.056	0.820 ± 0.019	0.849 ± 0.027

The quotient affine-fit R^2 is 0.721 ± 0.046. Mean absolute pixel difference between π-paired images is 0.00260 for the ellipse versus 0.02606 for the heart. Rasterization leaves a weak raw-angle signal, but the quotient dominates.

Table 14: Asymmetric-heart chain sample curve (three seeds).

Samples/env.	Mean coordinate ρ	Weakest coordinate ρ	AUC
100	0.764 ± 0.029	0.325 ± 0.127	0.829 ± 0.022
200	0.773 ± 0.031	0.332 ± 0.196	0.830 ± 0.011
400	0.820 ± 0.030	0.410 ± 0.120	0.877 ± 0.005
800	0.906 ± 0.025	0.733 ± 0.096	0.891 ± 0.003

D.5 Direct regret-to-ranking theorem closure

The theorem-closure experiment fixes an independent Shapes3D evaluation mixture, clips probabilities at 0.05, and uses three seeds with a 3,000-sample-per-class reference classifier.

Table 15: Excess logistic risk predicts coordinate ordering error.

Nuisance	Samples	Estimated excess risk	Inversion	$\lvert\rho\rvert$
Fixed	100	0.0815	0.1310	0.832
	200	0.0119	0.0572	0.946
	400	0.0075	0.0377	0.960
	800	0.0066	0.0277	0.969
Native	100	0.0820	0.1352	0.824
	200	0.0145	0.0626	0.940
	400	0.0059	0.0404	0.959
	800	0.0024	0.0286	0.970

Across all 24 seed/sample/nuisance runs, excess risk and pairwise inversion have Pearson correlation 0.967 and Spearman correlation 0.881. Reference classifier inversion is 0.0072 (fixed) and 0.0116 (native). The comparison thereby targets the excess risk and ordering error in the finite-classifier theory, without using AUC as a surrogate for regret.

D.6 smallNORB sample efficiency and leakage controls

Table 16: Source-disjoint smallNORB sample efficiency (five seeds). Budgets include the 80/20 optimization/validation split.

Budget/env.	Strict mean ρ	5% quantile	Adjacent accuracy	AUC
100	0.820 ± 0.052	-0.183 ± 0.496	0.831 ± 0.015	0.691 ± 0.052
500	0.918 ± 0.032	0.617 ± 0.164	0.862 ± 0.033	0.809 ± 0.056
8,000	0.972 ± 0.003	0.900 ± 0.000	0.892 ± 0.009	0.878 ± 0.004

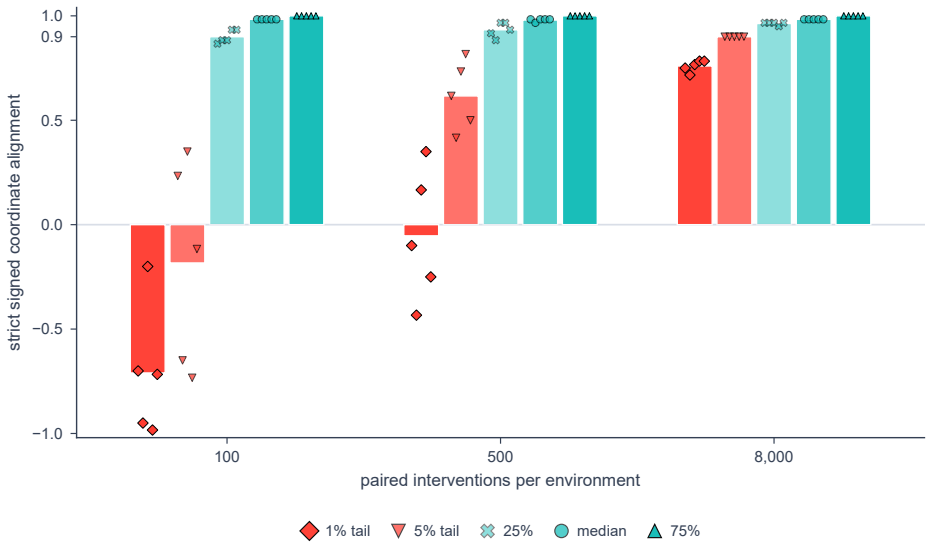


Figure 5: Strict-tail convergence on unseen smallNORB objects. Bars show context-level rank quantiles over five seeds; markers show individual seeds. Additional interventions primarily stabilize the difficult tail; the central monotone relation is already present.

Table 17: smallNORB leakage and target-specificity controls (seed 17).

Condition	AUC	Strict mean ρ	5% quantile	Adjacent accuracy
Raw CNN	0.884	0.971	0.900	0.890
Linear pixel probe	0.709	0.769	−0.150	–
Foreground normalized	0.857	0.930	0.750	–
Wrong target: lighting	0.901	−0.263	–	0.466
Identical null environments	0.532	−0.025	–	0.491

The previous 0.999 score averaged 108 nuisance images before ranking and is retained only as a conditional-mean pilot. The strict context-level audit serves as the primary estimate.

D.7 COIL-100 FIVE-SEED GAUGE AUDIT

Table 18: COIL-100 separates local discriminability from shared-gauge transfer.

Metric	Five-seed result
Environment AUC	0.904 ± 0.006
Seen-object cosine/sine correlation	0.923 / 0.917
Held-out cosine/sine correlation	0.261 / 0.118
Held-out negative fraction	28.0% / 41.3%
Global angle MAE	$80.52^\circ \pm 2.66^\circ$
Oracle object-gauge MAE	$50.52^\circ \pm 2.17^\circ$

Across the 15 dataset/phase/seed conditions in the gauge audit, validation AUC and held-out transfer are negatively associated (Pearson -0.878). The pooled association is descriptive, not causal. Its negative sign rules out classification strength alone as an explanation of transfer.

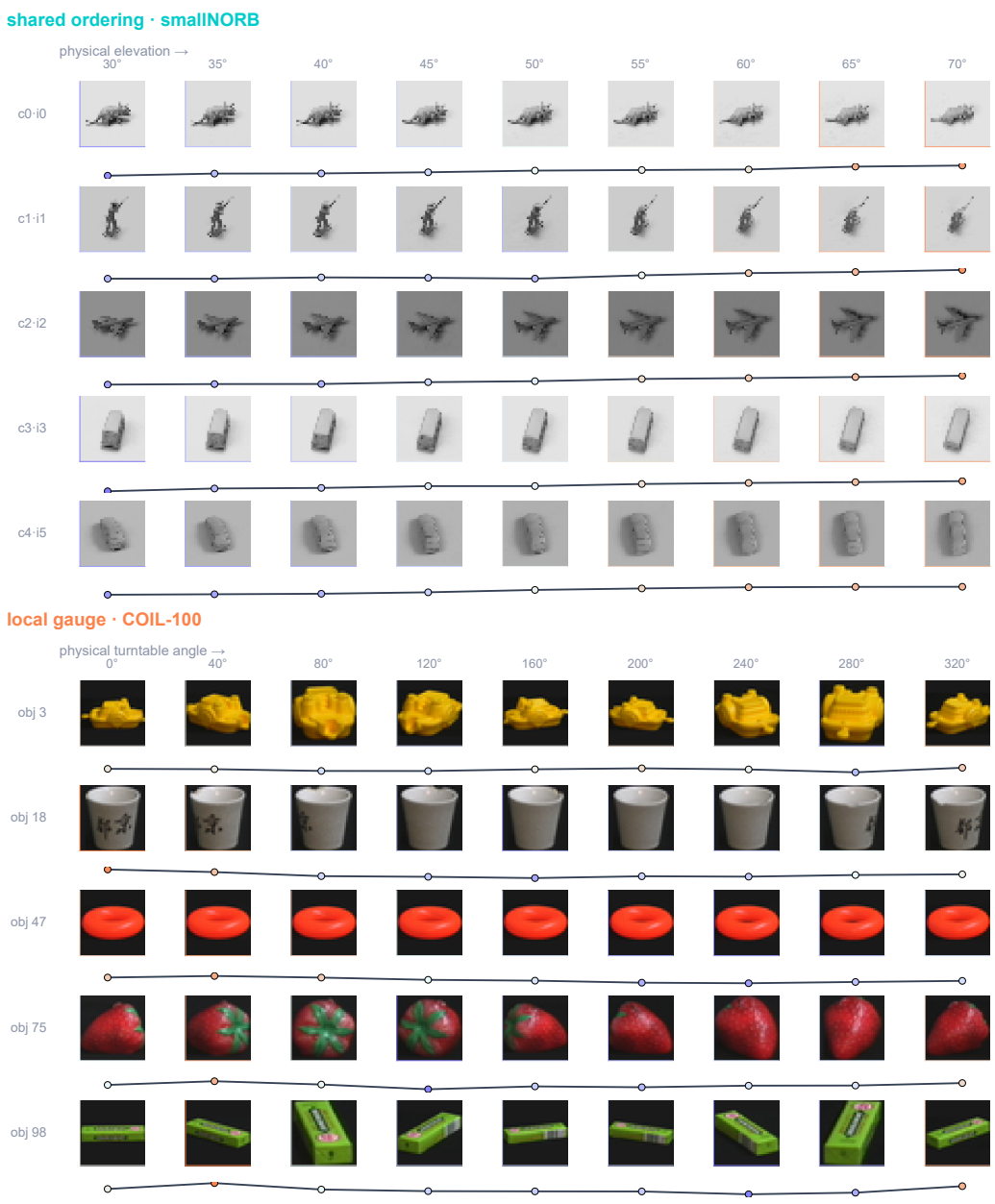


Figure 6: When a classifier logit becomes a physical coordinate, and when it does not. **Top:** on smallNORB, one elevation measurement transfers across held-out object instances and nuisance contexts. **Bottom:** on COIL-100, classifiers follow turntable phase for seen objects, but the visual gauge changes across unseen objects. Images are fixed audit contexts selected without reference to their outcomes.

D.8 CONTROLLED CHEMICAL EXPOSURES UNDER SENSOR DRIFT

Table 19: Concentration-coordinate transfer to months 21 and 36. Each row contains two held-out batches and five seeds. Spearman is computed directly over physical measurements without concentration-level averaging.

Gas	Mean ρ	s.d.	Median	Minimum	Positive
Ethanol	0.896	0.036	0.888	0.836	10/10
Ethylene	0.861	0.028	0.868	0.811	10/10
Ammonia	0.560	0.520	0.674	-0.580	8/10
Acetaldehyde	0.939	0.013	0.941	0.920	10/10
Acetone	0.914	0.061	0.942	0.829	10/10
Toluene	0.857	0.095	0.853	0.764	10/10
All strict contexts	0.838	0.245	0.893	-0.580	58/60

The held-out 5% quantile is 0.642 and 49/60 contexts exceed 0.8. The sole direction failure is ammonia at seed 101, where both held-out batches reverse despite positive source-batch coordinates; it is retained in every aggregate. A concentration-independent null produces 0.529 ± 0.050 validation AUC, mean/median held-out Spearman $-0.091/-0.183$, and 25/60 positive contexts. Random directions can have large absolute correlation because many physical sensor channels respond to concentration. Cross-context sign consistency supplies the decisive null comparison. A fixed $C = 1$ logistic-regression readout gives mean/median Spearman 0.701/0.766 and 59/60 positive contexts. The observed direction does not depend on nonlinear architecture.

Table 20: Source-only numerical-gauge audit. One affine calibration is fitted on source-validation rows; test batches permit only subtraction of the known policy offset, with no refitting or sign selection.

Metric	Held-out result
Normalized-concentration MAE, median	0.363
Sample-level R^2, median / positive	0.472 / 50 of 60
Level-mean R^2, median / positive	0.436 / 42 of 60
Held-out/source slope ratio, median	0.780
Slope sign retained / within $[0.5, 2]$	58 of 60 / 54 of 60

Mean sample-level R^2 is -0.127 because the two retained ammonia/seed-101 reversals have $R^2 = -20.09$ and -10.81. The positive median and negative mean distinguish broadly transferable ordering from only partial transfer of the numerical gauge.

D.9 SUPERVISION-COMPLETE CRL BRIDGE: SEED-LEVEL RESULTS

The bridge uses the graph $z_0 \to z_1 \to z_2$ and $z_0 \to z_2$, a shared 32-by-32 RGB renderer for Binary/CITRIS/BISCUIT, and the invertible 3D vector adapter required by the official ILCM scaling implementation. Training views are independently sampled from the same SCM; evaluation uses the same held-out latent trajectories, rendered through each method's stated observation adapter. The linear intervention probe is fitted on frozen train representations, selects its regularization on validation, and is evaluated once on test. CITRIS selected epochs 60/53/46 and BISCUIT 54/60/60 for seeds 7/17/23. ILCM's method-specific intervention head reaches 0.849 ± 0.008; Table 2 retains the common probe for every method.

Table 21: Seed-level common-metric audit for the supervision-complete bridge. Each slash-separated entry follows seeds 7/17/23.

Method	Programmed-z_0^2 $\lvert\rho\rvert$	Signed-z_0 $\lvert\rho\rvert$	Intervention balanced accuracy
Binary	.977/.947/.981	.025/.115/.064	.391/.396/.401
ILCM	.055/.056/.047	.917/.953/.923	.559/.515/.523
CITRIS	.382/.573/.326	.975/.966/.906	.592/.588/.613
BISCUIT	.516/.543/.463	.947/.934/.955	.581/.573/.606

D.10 Matched representation baselines: seed-level results and compute

Table 22: Matched seed-level estimand separation.

Seed	Binary signed	Binary quotient	iVAE signed	iVAE quotient	GSCALE signed	GSCALE quotient
11	0.175	0.679	0.502	0.262	0.785	0.128
23	0.066	0.776	0.309	0.610	0.624	0.131
47	0.104	0.827	0.331	0.530	0.693	0.121
71	0.057	0.807	0.167	0.524	0.951	0.071
101	0.200	0.643	0.373	0.654	0.947	0.025
Mean $\pm$ s.d.	0.120 ± 0.064	0.746 ± 0.081	0.336 ± 0.108	0.516 ± 0.136	0.800 ± 0.148	0.095 ± 0.046

Table 23: Matched-compute audit. Shared data generation (92.4 seconds on average) is excluded from all training times.

Method	Train + encode time	Stored model scalars	Relative size
Binary measurement	6.29 ± 1.02 s	628,130	$1.0\times$
iVAE	78.56 ± 16.25 s	300,443	$0.48\times$
GSCALE-I, four stages	119.79 ± 13.96 s	12,665,233	$20.2\times$

D.11 Assumption-breaking and downstream diagnostics

Why pairing matters. Comparison of one hard intervention with the observational environment gives

$$\frac{p_i^{\mathrm{hard}}(z)}{p^{\mathrm{obs}}(z)} = \frac{q_i(z_i)}{p_i(z_i \mid z_{\mathrm{pa}(i)})},$$

which generally depends on the parents of a non-root target. Likewise, a paired generic soft-intervention ratio retains parent dependence. Two hard replacement environments are sufficient for target-only cancellation because their common non-target factors cancel. The restriction belongs to the discriminative reduction; other CRL algorithms need not satisfy it.

Table 24: Hard-versus-soft intervention boundary on Shapes3D.

Intervention	Mean coordinate ρ	Fork F1	Collider F1
Paired hard	0.979	1.000	1.000
Paired soft	0.845	0.750	0.667

The chain remains exact under the soft intervention, which makes multiple graph motifs necessary for an informative assumption check.

Graph recovery lies downstream of the target-wise measurement. The primary response-signature learner is exact in all nine main Shapes3D runs, but UT-IGSP (Squires et al., 2020) is less reliable on learned logits:

Table 25: Downstream graph diagnostics on seed-11 Shapes3D.

Method/input	Chain F1	Fork F1	Collider F1
UT-IGSP, oracle factors	1.000	1.000	1.000
UT-IGSP, learned logits	0.667	0.889	0.500
UT-IGSP, rank-Gaussian logits	0.333	0.000	0.444
Training-free reachability, learned logits	6/9 exact over new seeds		

For DCDI-G (Brouillard et al., 2020) on the chain, oracle factors reach an exact exported DAG after 6,739 steps. Raw logits do not meet the acyclicity stopping rule within 12,000 steps and their last thresholded graph has F1 0.667. With official normalization, all true edges appear but one reverse edge remains (F1 0.857), so the graph is cyclic and correctly not exported. The outcome diagnoses optimization under a fixed budget and cannot rank representation quality.

D.12 Exploratory second-stage estimators

Table 26: Low-sample second-stage alternatives at 200 Shapes3D samples per sign. Values are mean F1 for chain/fork/collider.

Estimator	F1 (chain/fork/collider)
Original MLP gradient	0.889 / 0.667 / 0.889
Standardized linear coefficients	0.444 / 0.417 / 0.667
Marginal permutation loss	0.222 / 0.250 / 0.222
Conditional leave-one-coordinate-out AUC	0.889 / 0.583 / 0.778

Five-model gradient ensembles and diagonal anchoring also fail to improve the aggregate; independent permutation is especially inappropriate because it moves correlated coordinates off their data manifold.

A shared CNN backbone across four targets gives coordinate Spearman 0.896 but reduces the initial chain graph F1 from 0.667 to 0.333. Three mechanism-shift contexts at the same total sample budget do not materially improve coordinate recovery over one context (approximately 0.967 in both cases), although some downstream graphs stabilize. Neither branch improved the prespecified aggregate criterion, so neither entered the main analysis.