ECRL: EXPLAINABLE CAUSAL REPRESENTATION LEARNING VIA SEMANTIC ANCHORING OF EXOGENOUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding dynamical systems requires not only accurate prediction but also representations that reveal which external mechanism drives each state change, enabling mechanism-level explainability. Yet predictively equivalent causal representations may assign different latent blocks to the same named mechanism, leaving the semantic meaning of block-level interventions unresolved. We study this semantic orientation ambiguity using anchor environments that reveal which exogenous block changed, but not its realized value. Under affine residual ambiguity, known block dimensions, and block-spanning mean shifts, we show that off-target response bounds cross-block mixing, with zero response identifying named blocks up to within-block invertible affine transformations. We derive a finite-sample certificate and extend the result to exact temporal innovations. Motivated by this theory, ECRL localizes anchor responses in transition innovations while sharing coordinates with the state, decoder, and predictor. Known-affine calibration validates the certificate, while full-span neural experiments evaluate factor-value-free intervention-effect prediction. Across five controlled SCM families, 40 seeds, and eight held-out mechanism banks, ECRL reduces effect MSE from 1.1442 to 1.0856 versus a capacity-matched CITRIS-style model, a 5.12% reduction with a two-way cluster-bootstrap interval of $[0.0438, 0.0740]$. ECRL thus links conditional semantic orientation to mechanism-specific effect prediction.

1 INTRODUCTION

Causal representation learning seeks representations that recover underlying generative factors from observational and interventional data (Ahuja et al., 2023; Liang et al., 2023). Such representations are particularly important in dynamical systems, where understanding which external mechanism drives a state change is essential for interpretation, diagnosis, and targeted intervention. However, accurate prediction alone does not guarantee that latent coordinates carry stable mechanism-level semantics.

Existing identifiable CRL methods characterize latent recovery under structural assumptions and invariances (Ahuja et al., 2023; Yao et al., 2025), while concept-based approaches emphasize application-relevant semantics (Rajendran et al., 2024). Intervention extrapolation and controllable dynamical models further study how learned representations support interventions and prediction beyond observed regimes (Saengkyongam et al., 2024; Yao et al., 2024), and intervention-centric benchmarks highlight the need for evidence beyond idealized identification (Liu et al., 2023; Gamella et al., 2025). However, these results do not fully resolve correspondence to protocol-named mechanisms under residual affine ambiguity. In particular, predictively equivalent representations may still assign different meanings to the same latent block, leaving mechanism-specific intervention semantics ambiguous. We therefore study a controlled protocol in which the changed exogenous block is known but its realized value is hidden, and ask: *when do named anchor responses determine block correspondence, and how much cross-block mixing can remain?*

Our contributions follow from this question. **Theory:** mean shifts spanning each block yield a response–mixing bound, with finite-sample, moment, and exact-innovation extensions. **Method:** ECRL localizes innovation responses in coordinates shared with states and prediction (Fig. 1). **Evidence:** known-affine calibration tests the certificate, while full-span frozen neural evaluation tests

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

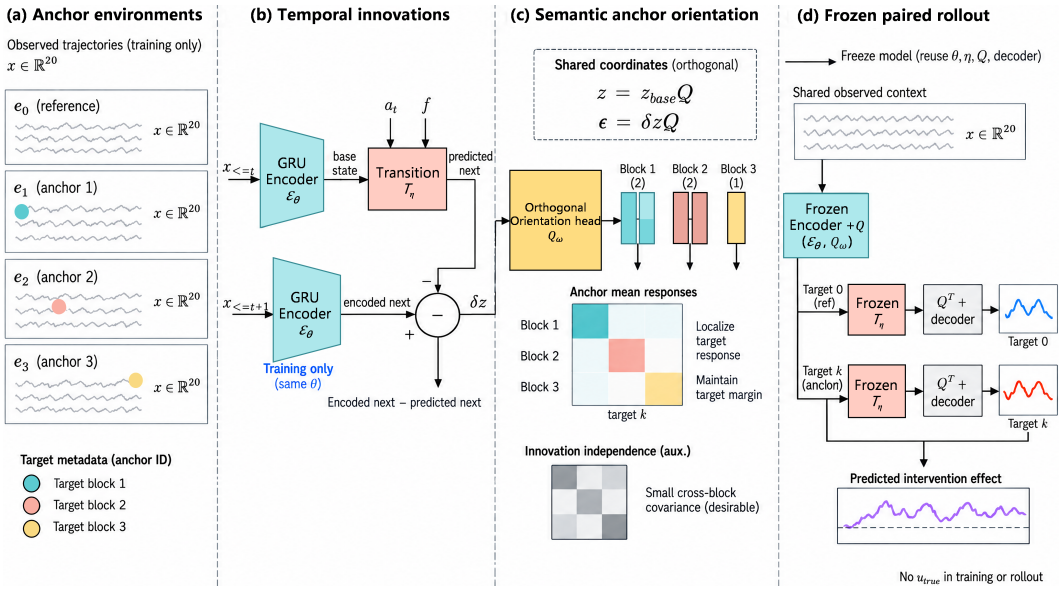


Figure 1: ECRL overview. Named anchors localize innovation responses in shared latent coordinates for mechanism-specific effect prediction. Theoretical certificates require affine residuals and block-spanning coverage; all panels are schematic.

factor-value-free intervention-effect prediction against a matched CITRIS-style comparator. Across five controlled SCM families, 40 seeds, and eight held-out mechanism banks, ECRL achieves a 5.12% reduction in intervention-effect MSE relative to a capacity-matched CITRIS-style model.

2 Related Work

Interventions, identifiability and explainability. We define explainability as a stable correspondence between learned blocks and named external mechanisms. Identification depends on mixing restrictions and intervention structure: linear latent models under nonlinear mixing (Buchholz et al., 2023), nonparametric unknown interventions (von Kügelgen et al., 2023), soft interventions (Zhang et al., 2023), and score-based recovery (Varıcı et al., 2025) yield distinct guarantees. Unpaired domains (Sturma et al., 2023), general environments (Zhang et al., 2024; Ng et al., 2025), multi-node interventions (Varıcı et al., 2024; Bing et al., 2024), and observed-variable grouping (Morioka & Hyvärinen, 2024) supply different identifying information. Intrinsic linear ambiguity (Jin & Syrgkanis, 2024), shifted mechanisms (Chen et al., 2024b), and observational layer-wise recovery (Welch et al., 2024) delimit achievable resolution. These results are fundamental to explainability because semantic meaning cannot be stable when the underlying representation itself is unconstrained; Yet identification up to admissible transformations need not align learned blocks with protocol-named mechanisms. We therefore focus on this remaining ambiguity within a known affine residual class and quantify residual cross-block semantic mixing among named exogenous blocks.

Finite-sample guarantees for explainable representations. Acartürk et al. (2024) analyze the sample complexity of interventional CRL, while Lee et al. (2026) establish recovery guarantees with few unknown intervention environments. For explainability, finite-sample analysis is important because a semantic correspondence should be empirically auditable rather than justified only in the population limit. Our certificate bounds semantic mixing using independent calibration data and known support and coverage bounds.

Temporal causal representation learning. iCITRIS identifies multidimensional factors from temporal interventions with known targets and instantaneous effects (Lippe et al., 2023). NCTRL and CtrlNS exploit unknown nonstationarity and sparse transitions (Song et al., 2023; 2024); CaRiNG uses temporal context under non-invertible generation (Chen et al., 2024a). IDOL leverages sufficient

variability together with sparse delayed and instantaneous influences without target labels (Li et al., 2025c). Hierarchical dynamics (Li et al., 2025b), locally sparse system-parameter models (Baumgartner et al., 2026) and non-Markovian SCMs (Li et al., 2024) further broaden the temporal setting. These approaches make temporal latent structure increasingly identifiable, but explainability additionally requires separating mechanism-specific external variation from variation propagated through the system state. Our exact-innovation identity motivates localizing anchor responses in transition innovations, where newly entering exogenous variation can be associated with named mechanisms.

Causal responses, abstraction, and explainability. Direct-effect analysis learns treatment-responsive outcome directions (Durand et al., 2025), while DECAF uses targets to identify reusable factors (Talon et al., 2024). Causal abstractions characterize the resolution supported by a given intervention family (Li et al., 2025a), and internal auxiliaries yield subspace equivalence under additional mixing and alignment assumptions (Kim et al., 2026). These works move causal representation learning toward task- and intervention-relevant explainability, showing that useful semantics need not require coordinate-wise recovery of every latent variable. Our endpoint is therefore a bound on cross-block semantic mixing in a fixed protocol-named order, while deliberately preserving arbitrary invertible transformations within each block. Taken together, prior work provides increasingly strong guarantees for recovering latent variables, dynamics, subspaces, and intervention-responsive factors; our complementary focus is whether the remaining representation ambiguity can be sufficiently constrained to give learned blocks stable, named mechanism-level semantics without observing realized exogenous values.

3 Problem Setup

Sources, states, and observations. Let $U = (U^{(1)}, \dots, U^{(K)})$ be an exogenous source vector with known block dimensions r_k and $r = \sum_k r_k$. The static identification problem observes $X = \psi(U) \in \mathbb{R}^D$, where ψ has a measurable left inverse on its support. This permits abduction in principle; it does not itself imply that a learned representation is affine in U. In the temporal setting, S_t denotes a latent *state*, ξ_t an external *innovation*, a_t an observed action, and X_t an observation. We reserve Z_t for learned states and ϵ_t for learned residuals. States may be temporally or cross-block dependent even when innovations are independent.

Named anchor environments. For a reference environment e_0 and anchor $e_{k,\ell}$ targeting block k, assume

$$P_U^{e_0} = \bigotimes_{m=1}^{K} P_m^0, \qquad P_U^{e_{k,\ell}} = P_k^{\ell} \otimes \bigotimes_{m \neq k} P_m^0. \tag{1}$$

The structural map is fixed within this family. The learner receives target identity k, not realized $U^{(k)}$; names come from the experimental protocol. Contrasts $\ell \in \mathcal{A}_k$ must expose every direction of the target block for the mean-based guarantee.

Identification target. Within the residual class $Z = AU + b$, $A \in \mathrm{GL}(r)$, write $A_{jk} \in \mathbb{R}^{r_j \times r_k}$ for its block submatrices. Named-block orientation means $A_{jk} = 0$ whenever $j \neq k$. The admissible ambiguity is a block-diagonal invertible affine transformation in the *fixed named order*. A permutation exchanging two named blocks is not admissible; a change of basis or scale within one block is. The result concerns this equivalence class, not the direction of edges in a causal graph.

Available information and distribution shifts. Training uses observed trajectories, actions, family IDs, reference/anchor matching IDs, target masks, and intervention strengths. Prediction receives the observed context, future action sequence, family ID, and a specified target mask and strength. True source values are excluded from training, validation selection, and rollout; they are used only in designated simulation diagnostics. Thus "without factor values" does not mean "without intervention metadata." Anchor shifts alter source distributions within a fixed mechanism. Held-out *mechanism banks* additionally change generator parameters and test empirical transfer; they are not members of the fixed-mechanism family required by Eq. (1).

4 ANCHOR-RESPONSE GUARANTEES FOR SEMANTIC ORIENTATION

A rotation makes the issue explicit. For two scalar sources, consider $Z_1 = \cos\theta\, U_1 + \sin\theta\, U_2$ and $Z_2 = -\sin\theta\, U_1 + \cos\theta\, U_2$. Composing the decoder with the inverse rotation preserves reconstruction. Shifting only the mean of U_1 by d changes the mean of Z_2 by $-d\sin\theta$: its squared off-target response is $d^2\sin^2\theta$, exposing the squared mixing coefficient.

Anchor Coverage and Semantic Response. Let $\Delta\mu_{k,\ell}$ be the true block-k mean contrast and define

$$G_k = \frac{1}{|\mathcal{A}_k|}\sum_{\ell\in\mathcal{A}_k}\Delta\mu_{k,\ell}\Delta\mu_{k,\ell}^\top, \quad \gamma_k = \lambda_{\min}(G_k), \quad R_{jk}^\mu = \frac{1}{|\mathcal{A}_k|}\sum_{\ell}\|\Delta\mu_{j,k,\ell}^Z\|_2^2. \tag{2}$$

Means are assumed finite, and $\Delta\mu_{j,k,\ell}^Z$ is the learned block-j contrast. Mean spanning requires $\gamma_k > 0$; increasing a collinear shift's magnitude cannot expose an orthogonal direction.

Theorem 1 (Anchor-Response Bound for Named-Block Orientation)**.** *Suppose Eq.* (1) *holds, $X = \psi(U)$ has a measurable left inverse on its support, the learned and true block dimensions agree, and $Z = AU + b$ with invertible A. If $\gamma_k > 0$, then*

$$R_{jk}^\mu = \mathrm{tr}(A_{jk}G_kA_{jk}^\top) \;\geq\; \gamma_k\|A_{jk}\|_F^2, \qquad \|A_{jk}\|_F \leq \sqrt{R_{jk}^\mu/\gamma_k}. \tag{3}$$

If this coverage condition holds for every block, zero off-target responses identify the named blocks up to invertible within-block affine transformations.

Proof. Only source block k changes, so linearity gives $\Delta\mu_{j,k,\ell}^Z = A_{jk}\Delta\mu_{k,\ell}$. Averaging the squared norm yields the trace identity in Eq. (3). Since $G_k \succeq \gamma_k I$, the trace is at least $\gamma_k\,\mathrm{tr}(A_{jk}A_{jk}^\top)$. If every off-target response vanishes, all off-diagonal blocks of A vanish. Matching block dimensions and invertibility of A then make each diagonal block invertible. Conversely, a block-diagonal A has zero off-target response under these anchors. □

Certificate scope and interpretation. Mixing is measured in the coordinates defining A and G_k. A positive coverage margin γ_k makes the response bound informative by separating weak mixing from weak anchor excitation. The certificate operates on raw response under explicit coverage, rather than on relative response ratios. Application of the certificate to a learned model additionally requires validating the assumed affine residual relation.

Proposition 2 (Finite-Sample Certificate for Cross-Block Mixing)**.** *Fix the representation independently of calibration data. Suppose every environment supplies n independent samples, each coordinate in learned block j is bounded in absolute value by a known M_j, and L counts all block–environment mean estimates. For $0 < \delta < 1$, set*

$$b_j(n,\delta) = 2M_j\sqrt{r_j}\sqrt{\frac{2\log(4Lr_j/\delta)}{n}}, \qquad \widehat{B}_{jk}^\mu = \frac{\sqrt{\widehat{R}_{jk}^\mu} + b_j(n,\delta)}{\sqrt{\gamma_k}}. \tag{4}$$

Under Theorem 1, all reported $\|A_{jk}\|_F \leq \widehat{B}_{jk}^\mu$ hold simultaneously with probability at least $1-\delta$.

Proof. Hoeffding's inequality and a union bound over coordinates and environments bound each anchor-minus-reference block-mean error by b_j. Stack the contrasts over ℓ and divide by $\sqrt{|\mathcal{A}_k|}$. The triangle inequality bounds the population stacked norm by $\sqrt{\widehat{R}_{jk}^\mu} + b_j$. Apply Eq. (3). Shared reference samples do not invalidate the union bound. □

Independent trajectory summaries can replace independent samples when they are the experimental units and satisfy the same boundedness assumption. A valid lower bound on γ_k may be used in the denominator, but an unvalidated estimate cannot. Our direct calibration knows the source geometry and affine map; it does not provide these quantities automatically for a neural checkpoint or a heavy-tailed real system.

Beyond mean-spanning anchors. Covariance changes can reveal directions invisible to mean shifts. Define the raw moment response $R^{\mathrm{mv}}_{jk} = |\mathcal{A}_k|^{-1}\sum_\ell(\|\Delta\mu^Z_{j,k,\ell}\|_2^2 + \|\Delta\Sigma^Z_{j,k,\ell}\|_F^2)$ and

$$\kappa^2_{k,q} = \inf_{\substack{B\in\mathbb{R}^{q\times r_k}\\ \|B\|_F=1}} \frac{1}{|\mathcal{A}_k|}\sum_\ell \left(\|B\Delta\mu_{k,\ell}\|_2^2 + \|B\Delta\Sigma_{k,\ell}B^\top\|_F^2\right). \tag{5}$$

If $\kappa_{k,r_j} > 0$, then $\|A_{jk}\|_F \le g^{-1}(R^{\mathrm{mv}}_{jk}/\kappa^2_{k,r_j})$, where $g(s) = \min\{s^2, s^4\}$. To see this, write $A_{jk} = sB$, $\|B\|_F = 1$: its mean and covariance energies scale as s^2 and s^4, respectively. Their sum is at least $g(s)\kappa^2_{k,r_j}$. This bound has a fourth-root regime near zero; the square-root rate requires mean span. Mean-only and mean–covariance responses do *not* generally have the same zero set.

Temporal transfer and boundary conditions. For an additive state process $S_{t+1} = F(S_t, a_t) + \xi_{t+1}$, let $Z_t = AS_t + b$ and define an exact conjugate predictor $T_Z(z, a) = AF(A^{-1}(z - b), a) + b$. Then

$$Z_{t+1} - T_Z(Z_t, a_t) = A\xi_{t+1}. \tag{6}$$

Theorem 1 applies to these residuals if the innovations and their anchors meet its conditions, without requiring independent states. For approximate predictors, the response additionally contains a transition-error contrast, which characterizes the gap between the exact identity and the learned neural residual. Anchor coverage determines which mixing directions are observable: directions outside the span of G_k remain unconstrained by mean responses.

5 Semantic Anchoring for Mechanism-Level Explainability

5.1 Shared State–Innovation Coordinates for Semantic Orientation

The neural implementation uses row-vector coordinates. A GRU encoder maps observed history, actions, and family identity f to a base state $\widetilde{z}_t = E_\theta(x_{\le t}, a_{\le t}, f)$. An orthogonal head Q_ω gives $z_t = \widetilde{z}_t Q_\omega$, partitioned into the known dimensions. With a skew-symmetric H_ω, we use $Q_\omega = (I + H_\omega)^{-1}(I - H_\omega)$. This head searches an orthogonal class, not all of $\mathrm{GL}(r)$; any nonorthogonal correction must be learned by the encoder.

Let c summarize the encoded context, $\widetilde{T}_\eta$ be the base transition, and D_ϕ the base decoder. We use distinct notation for their oriented counterparts:

$$T^Q_\eta(z, a, f, c) = \widetilde{T}_\eta(zQ^\top_\omega, a, f, c)Q_\omega, \qquad D^Q_\phi(z, f) = D_\phi(zQ^\top_\omega, f). \tag{7}$$

For a fixed predictor, the shared orientation preserves zero-intervention reconstruction; joint-training effects are evaluated empirically. The learned one-step innovation is

$$\epsilon_t = \left(\widetilde{z}_{t+1} - \widetilde{T}_\eta(\widetilde{z}_t, a_t, f, c_t)\right) Q_\omega. \tag{8}$$

Here a_t denotes the action associated with the transition. Observed next states are used for training residuals, never future rollout. Learned residuals need not satisfy the exact identity in Eq. (6).

5.2 Anchor–Guided Semantic Alignment

We estimate reference and anchor means from trajectory-averaged innovations within each matched family group. Let s_f be a detached pre-intervention innovation RMS scale, floored at 10^{-3}. For a contrast e targeting $k(e)$, define the dimensionless energy

$$E_{j,e} = \left\| \frac{\widehat{\mu}^{\epsilon,\mathrm{anchor}}_{j,e} - \widehat{\mu}^{\epsilon,\mathrm{ref}}_{j,e}}{s_f} \right\|_2^2. \tag{9}$$

This empirical, scale-normalized contrast is distinct from the raw population R^μ in the theorem. Positive rescaling preserves the zero set of the *same* mean response, but changes its quantitative magnitude. Averaging losses over valid matched contrasts $\mathcal{E}$, we minimize

$$\mathcal{L}_{\mathrm{loc}} = -\frac{1}{|\mathcal{E}|}\sum_e \log\frac{E_{k(e),e} + \varepsilon}{\sum_j E_{j,e} + \varepsilon}, \qquad \mathcal{L}_{\mathrm{margin}} = \frac{1}{|\mathcal{E}|}\sum_e \mathrm{softplus}(0.18 - E_{k(e),e}). \tag{10}$$

Localization favors the named block; the margin penalizes uniformly weak responses. These surrogates are not the absolute certificate in Eq. (4).

We standardize pre-intervention innovations within each family and expand coordinates as $h(v) = [v, v^2 - \mathbb{E}(v^2), \tanh(v)]$. The auxiliary $\mathcal{L}^{\epsilon}_{\text{ind}}$ averages squared cross-block entries of $\text{Cov}(h(\epsilon) \mid f)$. This auxiliary penalizes selected cross-block dependencies, not full statistical dependence. The joint system's objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + 0.16\mathcal{L}_{\text{loc}} + 0.06\mathcal{L}_{\text{margin}} + 0.05\mathcal{L}^{\epsilon}_{\text{ind}}. \tag{11}$$

The shared prediction objective combines observation rollout MSE, 0.35 times reconstruction MSE, and 0.12 times latent rollout MSE to detached encoded targets, with variance and off-diagonal covariance stabilization weighted 0.015 and 0.005. Full has no separate off-target hinge: the localization ratio already compares target and off-target energies. Independence remains an auxiliary bias rather than a condition of Theorem 1.

5.3 Training and Mechanism–Specific Effect Prediction

Optimization has three stages: 12 epochs learn the base representation and dynamics; 8 update only the orientation head; 4 calibrate decoder and transition-side parameters with orientation frozen. Comparators receive the same total epochs and validation opportunities.

The intervention specification consists of a block mask $I \in \{0, 1\}^K$ and signed strength s. Let $M(I)$ expand this mask to its block coordinates. In the base gauge, a learned intervention head J_η predicts a bounded direction from state, action, family, and context. The oriented step is

$$z_{t+1} = T^Q_\eta(z_t, a_t, f, c) + 0.24s\, M(I) \odot \left[J_\eta(z_t Q_\omega^\top, a_t, f, c) Q_\omega\right]. \tag{12}$$

The mask acts *after* rotation; this is why the named coordinate assignment matters to the query. The base transition uses a learned increment d_η through $0.68\widetilde{z} + 0.32\tanh(d_\eta) + 0.08 d_\eta$, and J_η has a $\tanh$ output. These architectural choices are shared with the matched comparator, controlling model capacity so that the comparison isolates semantic orientation.

At evaluation, encode a shared 16-step observed context once, retain its latent state and context code, and run Eq. (12) for 20 steps twice: with $(I, s) = (0, 0)$ and with the specified target and strength. Both predictions use the same observed action sequence and frozen parameters, and are decoded by D^Q_ϕ. Their observation-space difference is the predicted effect. The simulator pairs its ground-truth trajectories using identical initial conditions and random draws; the model itself does not read future exogenous noise or future observations. This yields factor-value-free, target-informed paired intervention-effect prediction.

6 Experiments

6.1 Protocols and Evaluation

Known-affine calibration, primary neural comparison, and the independent objective audit remain separate protocols. All three protocols use block-spanning anchor sets with ranks $(2, 2, 1)$; they differ in whether the affine map and source geometry are known and in the learning objective being evaluated.

Controlled neural benchmark. All five SCM families have state blocks $(2, 2, 1)$, observation dimension $D = 20$, and two-dimensional actions. They comprise stable affine dynamics, nonlinear intra-block dynamics, heavy-tailed noisy dynamics, lagged cross-block coupling, and instantaneous-effect stress. Innovations range from Gaussian to Student-t noise; the heavy-tailed case adds sparse outliers. A nonlinear injective coupling and full-column-rank projection define the observation component where applicable, while additive observation noise can destroy exact abduction. The affine generator provides a controlled setting for assessing how closely the learned neural representation approaches the theorem's residual class.

Training uses mechanism IDs 0–2 and strengths 0.5 and 1.0; validation includes strength 0.75. Development mechanisms are 31–32. Eight predeclared, disjoint mechanism banks evaluate frozen

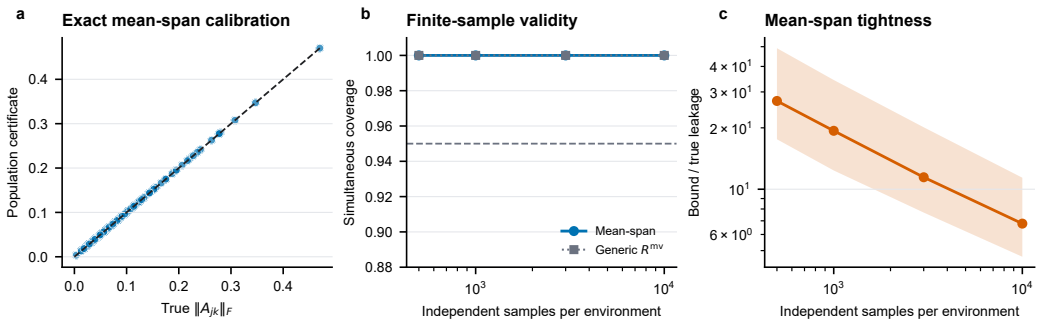


Figure 2: Known-affine certificate calibration: (a) population accuracy, (b) simultaneous finite-sample coverage, and (c) bound tightness across sample sizes. The experiment evaluates certificates, not neural recovery.

models. Queries change strength to 1.5 or 2.0, reverse sign, or compose targets. Within each paired query the mechanism is shared; between banks its parameters change. The primary prediction protocol uses block-spanning anchor mean shifts with ranks $(2, 2, 1)$, matching the block dimensions and satisfying the geometric coverage condition of Theorem 1.

Comparators and selection. The primary comparator is an actual capacity-matched CITRIS-style recurrent model. It uses the same predictor and target information with its own target-localization regularization. Both methods use 40 seeds. Anchor-only removes innovation dependence; State-Independence applies that penalty to states; Full+Off adds the legacy off-target hinge. These diagnostics use 20 matched seeds, not an unmatched comparison against the full 40-seed mean. Architectures have a GRU encoder, two-layer decoder/transition networks, latent dimension five, and approximately 49.6k parameters. Every method has a three-trial tuning budget, 24 epochs, and batch size 192; validation selects learning rate 10^{-3} and weight decay 10^{-4}. Selection uses observations and metadata only. Checkpoints are frozen before bank evaluation, with no bank-dependent fitting, permutation, or early stopping.

Metrics and statistical units. For paired trajectory p and horizon h, let $\delta x_{p,h} = x^I_{p,h} - x^0_{p,h}$ and $\widehat{\delta x}_{p,h} = \widehat{x}^I_{p,h} - \widehat{x}^0_{p,h}$. The primary endpoint is

$$\text{EffectMSE} = \frac{1}{PHD} \sum_{p=1}^{P} \sum_{h=1}^{H} \|\widehat{\delta x}_{p,h} - \delta x_{p,h}\|_2^2. \tag{13}$$

Effect cosine diagnoses direction; ID rollout MSE reveals predictive trade-offs. A direct orientation analysis complements these predictive metrics. Statistical tests use paired *seed-level* differences averaged across banks. The primary interval independently resamples seed rows and bank columns of the complete matrix; 40 seeds evaluated on eight banks do not constitute 320 independent training runs. Component tests use Holm correction over three two-sided tests; horizon and family decompositions are descriptive.

6.2 Explainability Guarantees.

We calibrate the theorem separately using 20 known invertible affine maps, bounded independent source samples, and one mean-shift direction per coordinate. A reference is uniform on $[-1, 1]^5$; each anchor shifts one coordinate by 0.8. Thus $G_k = (0.8^2/r_k)I$, with full ranks $(2, 2, 1)$. Four leakage scales are cycled across the 20 map seeds, and each map is evaluated at $N \in \{500, 1000, 3000, 10000\}$ samples per environment. There are six off-diagonal block comparisons per map. Known maps and support bounds are used for calibration, not given to a trained predictor.

Figure 2 distinguishes validity from tightness. Population mean bounds coincide with the known leakage because the anchor Gram matrices are isotropic; This isolates certificate validity from

Table 1: Frozen comparison over 40 seeds and eight held-out banks. Values are means ± 95% seed-level t-interval half-widths.

Method	Effect MSE ↓	Effect cosine ↑	ID MSE ↓
CITRIS-style	1.1442 ± 0.0157	0.6402 ± 0.0121	$\mathbf{0.8142 \pm 0.0040}$
ECRL	$\mathbf{1.0856 \pm 0.0160}$	$\mathbf{0.6902 \pm 0.0105}$	0.8297 ± 0.0045

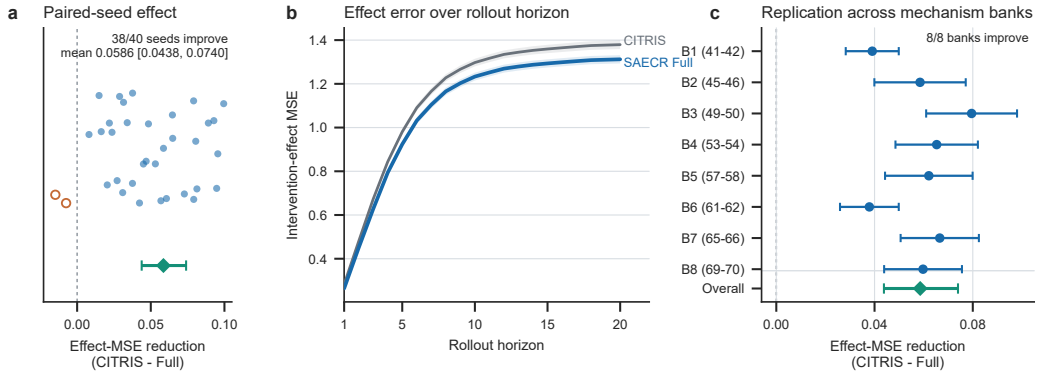


Figure 3: Held-out effect prediction across paired seeds, rollout horizons, and mechanism banks. Intervals are seed-level 95% t-intervals except for the overall two-way cluster-bootstrap interval.

the separate representation-learning problem. Both finite certificates cover all six comparisons simultaneously in all 20×4 map–sample-size settings. The mean-bound median ratio falls from 27.09 at $N = 500$ to 6.78 at $N = 10000$. The broader moment bound is more conservative, while the mean certificate tightens systematically as sample size increases. Coverage here evaluates certificate validity over 20 independent affine-map seeds, with six simultaneous off-diagonal comparisons per map.

6.3 Effectiveness.

Table 1 reports the primary frozen-bank comparison. ECRL lowers effect MSE from 1.1442 to 1.0856, an absolute reduction of 0.05859 and a relative reduction of 5.12%. The two-way cluster-bootstrap 95% interval for the reduction is $[0.0438, 0.0740]$; the paired two-sided Wilcoxon test gives $p = 1.27 \times 10^{-11}$ on 40 bank-averaged seed differences. Effect cosine improves from 0.6402 to 0.6902. The improvement is specifically concentrated on the mechanism-level intervention objective targeted by ECRL, with a modest 1.91% ID-rollout trade-off.

Figure 3 shows the reduction is positive for 38 of 40 seeds and all eight bank averages; individual bank reductions range from 0.0379 to 0.0795. The horizon curve shows an advantage sustained across the evaluated future, while absolute errors grow. These panels demonstrate consistent replication across all eight predeclared held-out mechanism banks and 38 of 40 training seeds. The complete primary and component record contains 140 trained checkpoints and 1120 frozen bank evaluations, with no missing method–seed–bank entries or failed final runs. No evaluator reads true factor values to generate a prediction.

6.4 Ablation Study

Table 2A compares six configurations under matched seeds, mechanism banks, architecture, and training budgets. The two new variants remove only $\mathcal{L}_{\text{loc}}$ or $\mathcal{L}_{\text{margin}}$, respectively, while retaining the remaining objectives and model components.

Localization is the main contributor to semantic orientation. Relative to w/o Localization, Full improves selectivity by 0.3048 (95% CI $[0.2574, 0.3499]$) and reduces OOD effect MSE by 0.1062 (95% CI $[0.0664, 0.1470]$), both with Holm-adjusted $p = 0.0002$. Removing the margin has little effect on selectivity ($p = 0.525$), but increases effect MSE by 0.0451 (95% CI $[0.0258, 0.0645]$;

Table 2: Component studies over 20 paired seeds. A: semantic-alignment ablations and OOD effect prediction; B: an independent mean-versus-moment objective audit.

A. Explainability-oriented V3 ablations			
Method	Selectivity ↑	Effect MSE ↓	On / Off response
ECRL Full	0.8954	1.0853	1.0304 / 0.1190
w/o Localization	0.5906	1.1915	0.3391 / 0.2399
w/o Margin	0.8916	1.1304	1.0565 / 0.1228
Anchor-only	0.8780	1.1008	0.9896 / 0.1326
State-Independence	0.8984	1.0878	1.0109 / 0.1119
Full+Off	0.9023	1.0857	1.0698 / 0.1147
B. Independent mean-versus-moment audit			
Endpoint	Reduction	95% CI	p
OOD effect MSE	0.0261	$[-0.0161, 0.0677]$	0.255
ID rollout MSE	0.0472	$[0.0394, 0.0548]$	$< 10^{-9}$
Off-target R^{mv}	0.00087	$[-0.00218, 0.00427]$	0.608

adjusted $p = 0.0002$). The auxiliary variants yield comparable effect MSE. Localization also reflects an ID–OOD trade-off: removing it improves ID rollout MSE from 0.8308 to 0.7084.

Table 2B evaluates the covariance extension using block-spanning anchors of ranks $(2, 2, 1)$. Moment-Aligned reduces ID rollout MSE from 0.8260 to 0.7788, increases DDR from 10.46 to 18.78, and improves selectivity from 0.698 to 0.764. Its OOD effect-MSE reduction is smaller and statistically inconclusive ($p = 0.255$), so we retain the simpler mean-only objective as the default. Held-out affine-fit R^2 also increases from 0.535 to 0.585, providing an additional diagnostic of the learned representation's affine structure.

6.5 Generalization.

The primary protocol satisfies the geometric coverage requirement: its anchor mean-shift ranks are $(2, 2, 1)$, matching the block dimensions. The remaining limitation is therefore representational rather than geometric. Full-span anchors do not by themselves establish that learned neural residuals obey the affine relation assumed by Theorem 1, and approximation error in the transition model may contaminate anchor responses. Thus, the primary protocol permits a neural comparison under full coverage, but the observed prediction gains should not by themselves be interpreted as proof of exact block identification.

Family averages favor ECRL in all five settings, with absolute reductions of 0.052 in the affine family, 0.035 in nonlinear-invertible dynamics, 0.070 under heavy-tailed noise, 0.091 with lagged coupling, and 0.046 with instantaneous effects. Lagged coupling remains hardest in absolute error. ECRL shows positive mean reductions across all five SCM families; family-wise results are reported descriptively. Robustness to anchor misspecification and transition-model error remains open.

7 Discussion and Conclusion

Named exogenous blocks require more than predictive equivalence. Under affine residuals and mean-spanning anchors, off-target responses bound cross-block mixing, and zero response identifies named blocks up to within-block affine changes. Finite calibration tests this bound on known maps, while the temporal identity motivates innovation-based orientation without assuming independent states.

ECRL consistently improves factor-value-free, target-informed intervention-effect prediction across frozen mechanism banks, with a modest ID trade-off. Although the primary anchors satisfy full-span coverage, imperfect learned affine fit precludes interpreting predictive gains as complete block identification. Future work will extend checkpoint-level certificate calibration to richer nonlinear and real-system interventions. Together, ECRL links semantic anchoring of exogenous blocks to mechanism-level explainability and intervention-effect prediction.

REFERENCES

Emre Acartürk, Burak Varıcı, Karthikeyan Shanmugam, and Ali Tajer. Sample Complexity of Interventional Causal Representation Learning. In *NeurIPS*, 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 *ICML*, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Markus W. Baumgartner, Anson Lei, Joe Watson, and Ingmar Posner. Disentangling Dynamical Systems: Causal Representation Learning Meets Local Sparse Attention. In *CLeaR*, 2026. URL https://proceedings.mlr.press/v323/baumgartner26a.html.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying Linearly-Mixed Causal Representations from Multi-Node Interventions. In *CLeaR*, 2024. URL https://proceedings.mlr.press/v236/bing24a.html.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep K. Ravikumar. Learning Linear Causal Representations from Interventions under General Nonlinear Mixing. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8e5de4cb639ef718f44060dc257cb04f-Abstract-Conference.html.

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 *ICML*, 2024a. URL https://proceedings.mlr.press/v235/chen24ai.html.

Tianyu Chen, Kevin Bello, Francesco Locatello, Bryon Aragam, and Pradeep Ravikumar. Identifying General Mechanism Shifts in Linear Causal Representations. In *NeurIPS*, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Homer Durand, Gherardo Varando, and Gustau Camps-Valls. Learning Causal Response Representations through Direct Effect Analysis. In *UAI*, 2025. URL https://proceedings.mlr.press/v286/durand25a.html.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity Checking Causal Representation Learning on a Simple Real-World System. In *ICML*, 2025. URL https://proceedings.mlr.press/v267/gamella25a.html.

Jikai Jin and Vasilis Syrgkanis. Learning Linear Causal Representations from General Environments: Identifiability and Intrinsic Ambiguity. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/741aab8b41a2987867acc9939ad50383-Abstract-Conference.html.

Kwonho Kim, Heejeong Nam, Inwoo Hwang, and Sanghack Lee. On Causal Representation Learning with Internal Auxiliaries. In *UAI*, 2026. URL https://proceedings.mlr.press/v337/kim26e.html.

Inbeom Lee, Tongtong Jin, and Bryon Aragam. Beyond Identifiability: Learning Causal Representations with Few Environments and Finite Samples. arXiv preprint arXiv:2603.25796, 2026. URL https://arxiv.org/abs/2603.25796.

Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled Representation Learning in Non-Markovian Causal Systems. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bd7ffcbd088b5e12f3b9eecc9c498b27-Abstract-Conference.html.

Xiusi Li, Sékou-Oumar Kaba, and Siamak Ravanbakhsh. On the Identifiability of Causal Abstractions. In *AISTATS*, 2025a. URL https://proceedings.mlr.press/v258/li25g.html.

Zijian Li, Minghao Fu, Junxian Huang, Yifan Shen, Ruichu Cai, Yuewen Sun, Guangyi Chen, and Kun Zhang. Towards Identifiability of Hierarchical Temporal Causal Representation Learning. In *NeurIPS*, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/1c6decac1477fbcc2cbf12d314ce0133-Abstract-Conference.html.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the Identification of Temporal Causal Representation with Instantaneous Dependence. In *ICLR*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5726513facc85f802be4a25e77fb9765-Abstract-Conference.html.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/67089958e98b243d5cc1881ad60418b8-Abstract-Conference.html.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In *ICLR*, 2023. URL https://openreview.net/forum?id=itZ6ggvMnzS.

Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning. In *CLeaR*, 2023. URL https://proceedings.mlr.press/v213/liu23a.html.

Hiroshi Morioka and Aapo Hyvärinen. Causal Representation Learning Made Identifiable by Grouping of Observational Variables. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/morioka24a.html.

Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal Representation Learning from General Environments under Nonparametric Mixing. In *AISTATS*, 2025. URL https://proceedings.mlr.press/v258/ng25a.html.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From Causal to Concept-Based Representation Learning. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b76a9959151d377ddd2c77a275a97475-Abstract-Conference.html.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Ravikumar, Niklas Pfister, and Jonas Peters. Identifying Representations for Intervention Extrapolation. In *ICLR*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aa9508b2577e0d9c824da9e2e27fe9d6-Abstract-Conference.html.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric P. Xing, and Kun Zhang. Temporally Disentangled Representation Learning under Unknown Nonstationarity. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/19a567abaec3990cb40d7a013556fecd-Abstract-Conference.html.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal Temporal Representation Learning with Nonstationary Sparse Transition. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8cef4e4bcb85f7d4a1005a9db018d6b6-Abstract-Conference.html.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired Multi-Domain Causal Representation Learning. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6c8985579293e0209bdaa4f21bb1d237-Abstract-Conference.html.

Davide Talon, Phillip Lippe, Stuart James, Alessio Del Bue, and Sara Magliacane. Towards the Reusability and Compositionality of Causal Representations. In *CLeaR*, 2024. URL https://proceedings.mlr.press/v236/talon24a.html.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear Causal Representation Learning from Unknown Multi-node Interventions. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ca70528fb11dc8086c6a623da9f3fee6-Abstract-Conference.html.

Burak Varıcı, Emre Acartürk, 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://jmlr.org/papers/v26/24-0194.html.

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 *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/97fe251c25b6f99a2a23b330a75b11d4-Abstract-Conference.html.

Ryan Welch, Jiaqi Zhang, and Caroline Uhler. Identifiability Guarantees for Causal Disentanglement from Purely Observational Data. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ba2f8120cd723c63e6d72b5a2fd16803-Abstract-Conference.html.

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying Causal Representation Learning with Dynamical Systems for Science. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/83eb339ed42297658fa24b5cec939285-Abstract-Conference.html.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying Causal Representation Learning with the Invariance Principle. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/85381f4549b5ddf1d48e2e287d7d3d15-Abstract-Conference.html.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability Guarantees for Causal Disentanglement from Soft Interventions. In *NeurIPS*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9d3a4cdf6f70559e8c6fe02170fba568-Abstract-Conference.html.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal Representation Learning from Multiple Distributions: A General Setting. In *ICML*, 2024. URL https://proceedings.mlr.press/v235/zhang24br.html.