THE SYSTEM IS OUT OF CONTROL, BUT CAUSALITY ISN'T: LEVERAGING OPERATIONAL SHIFTS FOR BIVARIATE CAUSAL DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Bivariate causal discovery often relies on subtle asymmetries in a single observational distribution, yet these signals become unreliable when causal effects are weak or masked by noise. Multiple environments can offer stronger evidence through distributional changes and invariances, but many applications provide only two regimes: a reference condition and one alternative. We study causal direction identification in this practically important but theoretically underexplored setting. We introduce *Kernel Residual Displacement* (KRD), which exploits a simple directional asymmetry: under the correct direction, a localized change between the two regimes leaves one residual coordinate invariant, whereas under the reverse direction the change propagates to both residual coordinates. KRD quantifies this asymmetry through kernel distances between residual distributions, requiring neither multiple auxiliary environments nor conditional-independence testing. We establish population identifiability and finite-sample recovery guarantees that explicitly connect successful identification to perturbation magnitude, regression error, sample size, and directional separation. For linear Gaussian systems where causal direction is unidentifiable from a single observational distribution, our analysis further reveals a non-monotone signal-to-noise geometry: identification is strongest when causal signal and downstream noise are balanced and deteriorates toward both near-deterministic and noise-dominated extremes. Experiments across simulations, nonlinear and non-Gaussian mechanisms, and real-world datasets show that even one additional regime can yield substantial causal information beyond conventional single-context discovery.

1 INTRODUCTION

Which of two correlated variables causes the other? This is a fundamental problem in bivariate causal discovery (Peters et al., 2017). From a single observational distribution, causal direction cannot in general be recovered without assumptions that break the statistical symmetry between the two candidate factorizations. In the linear Gaussian case, both directions can be compatible with the same observational distribution (Peters et al., 2017); identifiability can instead arise from structural asymmetries induced by nonlinearity (Hoyer et al., 2008; Rolland et al., 2022), non-Gaussian noise (Shimizu et al., 2006; Hyvärinen & Smith, 2013), or independence between causes and their mechanisms (Daniusis et al., 2012). These principles underlie many successful single-context methods, but they also expose a practical vulnerability: the identifying asymmetry must be strong enough to be statistically visible in the available sample. As the data-generating process approaches a weak-signal regime, the two candidate directions can exhibit increasingly similar residual or distributional behavior and become difficult to distinguish in finite samples; as illustrated in Figure 1, such asymmetries may be weak when causal effects are small or obscured by substantial noise (Kpotufe et al., 2014; Li et al., 2026).

One line of research is to strengthen causal evidence by observing the same system across multiple contexts or environments. Rather than relying on a single observational distribution to reveal direction, these methods use interventions, distribution shifts, or changes in local mechanisms as additional probes: a causal relationship may remain stable, whereas a noncausal one does not, and changes may propagate differently along the causal graph (Peters et al., 2016; Huang et al., 2020;

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.

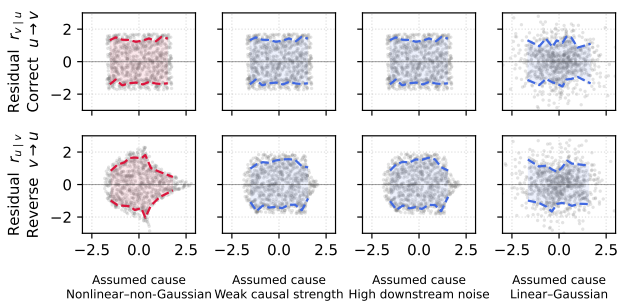

Figure 1: Rows show residuals under the correct (top) and incorrect (bottom) directions. The first column depicts a typical identifiable regime (nonlinear mechanism, non-Gaussian errors), where the two directions differ clearly. As causal strength weakens (second column) or downstream noise increases (third column), this distinction blurs; in the linear Gaussian case (fourth column), the two become nearly indistinguishable.

Mooij et al., 2020; Perry et al., 2022). With sufficiently diverse environments, repeated cross-context contrasts can provide evidence (Montagna, 2025) that is unavailable within any single context; indeed, some identification results explicitly require multiple distinct settings, as in BACKSHIFT (Rothenhäusler et al., 2015). This suggests a natural strategy when within-context asymmetries are weak—collect more contexts. It also leaves a sharper question: how much causal information can be extracted when only one cross-context contrast is available?

This question is not academic: in many applications, two contexts are not a temporary data limitation but the system design itself. Statistical process control (SPC) anchors inference to a stable in-control (IC) regime and asks whether a subsequent regime departs from it (i.e., becomes out-of-control, OOC) (Montgomery & Jennings, 1991); sequential change detection similarly organizes evidence around behavior before and after a change (Page, 1954; Truong et al., 2020); and case–control studies deliberately contrast outcome-defined groups to enable efficient inference without measuring an entire cohort (Flanders, 2020; Prentice & Pyke, 1979). Despite differing objectives, these fields share a pragmatic design principle: a well-characterized reference state provides an anchor, while the scientifically important departure is often sparse, costly, or operationally salient, so a reference-versus-departure representation renders a complex system as an interpretable contrast. The same parsimony imposes a causal limitation: with only two regimes, there is only one contrast to work with. We therefore treat the two-context restriction as a first-class setting and ask whether it suffices for causal direction analysis. Following the reference-versus-departure (or IC-versus-OOC) framework common to these disciplines, we formulate bivariate causal discovery from two regimes and study when their cross-regime changes separate the two candidate directions. The precise structural and perturbation model is deferred to Section 3; here the key point is that the two-regime design enables a focused analysis of population identifiability, finite-sample recovery, and how cross-context change magnitude, causal signal, noise, and sample size govern recoverability.

Our contributions are threefold. First, we propose *Kernel Residual Displacement* (KRD), which compares IC-to-OOC changes in standardized residual-coordinate distributions and selects the direction exhibiting the localized invariance predicted by the causal model, without requiring an explicit independence test. Second, we establish population identifiability and finite-sample sufficient conditions that separate cross-context signal from regression and sampling error. For linear Gaussian systems, the analysis yields an explicit recovery threshold and reveals that directional separation is strongest when causal signal and downstream noise are balanced and deteriorates toward either extreme. Third, experiments across controlled and extended synthetic settings and real-world benchmarks show that KRD is especially effective in regimes where single-context asymmetries are weak while remaining competitive beyond the theoretical setting.

2 Related Works

Single-context methods perform causal discovery by exploiting directional asymmetries within a single observational distribution. Specifically, under suitable functional assumptions, the cause

variable is independent of th noise exclusively in the true causal direction. ANM-HSIC directly tests this independence between the input variable and the residual using the Hilbert-Schmidt Independence Criterion (Hoyer et al., 2008), sharing a similar independence principle with linear non-Gaussian models (LiNGAM) (Shimizu et al., 2006) for full DAG identification. Other variants, such as ANM-Entropy (Kpotufe et al., 2014; Nowzohour & Bühlmann, 2016) and ANM-MML (Stegle et al., 2010), compare entropy and description-length scores of candidate directional models. Beyond noise independence, RECI leverages differences in prediction error (Blöbaum et al., 2018), while IGCI exploits the directional asymmetry between the input distribution and the underlying mechanism (Daniusis et al., 2012). However, these methods rely on model-specific directional asymmetries and can fail when such asymmetries vanish or become too weak to detect from finite samples (Hoyer et al., 2008; Peters et al., 2014; Mooij et al., 2016).

Cross-context methods additionally exploit environmental variation. Two principles particularly relevant to our work are mechanism invariance, often assessed through residual invariance, and sparse mechanism shifts. ICP searches for invariant conditional relationships (Peters et al., 2016), while its nonlinear extensions compare residual distributions under suitable structural assumptions (Heinze-Deml et al., 2018). Broader approaches incorporate context through augmented graphs in JCI (Mooij et al., 2020), conditional independences and changing mechanisms in CD-NOD (Huang et al., 2020), or intervention-aware search in UT-IGSP (Squires et al., 2020). However, invariance alone may not identify a unique model when several candidates remain stable. The sparse mechanism shift principle adds information by favoring structures in which environmental changes affect fewer local mechanisms. MSS formalizes this principle by comparing candidate causal factorizations according to their numbers of changing mechanisms (Perry et al., 2022). Nonparametric implementations that repeatedly compare conditional mechanisms can nevertheless be computationally costly. Our method combines residual invariance with sparse mechanism shifts by evaluating marginal residual distributions rather than computationally intensive joint densities.

3 Problem Formulation

We consider two continuous variables u and v. Without loss of generality, we assume that the true causal direction is $u \to v$, though this direction is unknown to the identification procedure. Their relationship is governed by a bivariate Additive Noise Model (ANM) (Hoyer et al., 2008; Peters et al., 2017), a widely used subclass of Structural Causal Models (SCMs):

$$x_u = \eta_u, \qquad x_v = f(x_u) + \eta_v, \tag{1}$$

where $f : \mathbb{R} \to \mathbb{R}$ represents the causal mechanism, and η_u, η_v are independent exogenous noise terms. To establish theoretical identification, we impose the following standard regularity conditions on the exogenous noises and the mechanism f.

Assumption 1 (Noise and regularity)**.** *The exogenous noises η_u and η_v are independent with zero mean and strictly positive, finite standard deviations σ_u and σ_v. In addition, f is a measurable function satisfying $\mathbb{E}[f(\eta_u)^2] < \infty$.*

The *in-control (IC)* environment corresponds to the unperturbed ANM in equation 1, while the *out-of-control (OOC)* environment describes the same structural mechanism under an external perturbation, as specified in the following assumption.

Assumption 2 (OOC perturbation)**.** *In classic SPC paradigms, the OOC environment exhibits one of two potential shift targets depending on the perturbation location:*

$$\begin{aligned} &\textit{cause shift:} \quad x_u = \eta_u + \delta\sigma_u, && x_v = f(x_u) + \eta_v, \\ &\textit{effect shift:} \quad x_u = \eta_u, && x_v = f(x_u) + \eta_v + \delta\sigma_v. \end{aligned}$$

Here, $\delta > 0$ is the standardized shift magnitude relative to IC standard deviation units (De Magalhães et al., 2009; Qin, 2012; Xue & He, 2021). The cause shift represents an upstream perturbation applied to the cause variable, whereas the effect shift represents a downstream perturbation applied to the effect variable. The resulting OOC variable under each shift is assumed to have a finite second moment.

Given this setup, our objective is to identify the true causal direction,

$$u \to v \qquad \text{or} \qquad v \to u,$$

from the IC and OOC observations. We establish causal identification conditions based on residual changes, first at the population level and then in finite samples.

4 METHOD

4.1 RESIDUAL DISTRIBUTION CHANGES ACROSS IC AND OOC

To identify the true causal direction, we construct the residual coordinates under the two candidate directions:

$$\begin{aligned} u \to v: \quad & r_u = x_u - \mathbb{E}_{\mathrm{IC}}[x_u], && r_{v|u} = x_v - f(x_u), \\ v \to u: \quad & r_v = x_v - \mathbb{E}_{\mathrm{IC}}[x_v], && r_{u|v} = x_u - g(x_v). \end{aligned}$$

Here, $g(x_v) := \mathbb{E}_{\mathrm{IC}}[x_u \mid x_v]$ is the functional form under the assumed direction $v \to u$. Furthermore, for the residual coordinates to be on a comparable scale, we normalize them by $z_u = r_u/s_u$, $z_{v|u} = r_{v|u}/s_{v|u}$, $z_v = r_v/s_v$, and $z_{u|v} = r_{u|v}/s_{u|v}$.

Let P_e^z denote the distribution of the residual coordinate $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ in environment $e \in \{0, 1\}$, corresponding to IC and OOC, respectively. The true direction is identifiable because these residual coordinates behave differently under the two candidate directions. Lemma 1 shows their behavior under the true direction (a clear pattern), while Lemma 2 shows their behavior under the wrong direction (a more entangled response).

Lemma 1 (Residual distribution change under true direction)**.** *Let Assumptions 1–2 hold. Under the true direction $u \to v$, the standardized residual distributions satisfy*

$$\begin{cases} P_1^{z_u} \neq P_0^{z_u} & \text{and} \quad P_1^{z_{v|u}} = P_0^{z_{v|u}}, \quad \textit{cause shift}, \\ P_1^{z_u} = P_0^{z_u} & \text{and} \quad P_1^{z_{v|u}} \neq P_0^{z_{v|u}}, \quad \textit{effect shift}. \end{cases}$$

Lemma 2 (Residual distribution change under reverse direction)**.** *Under the same assumptions, let $\mu_v = \mathbb{E}[f(\eta_u)]$. For the reverse direction $v \to u$, the standardized residual pair is*

$$\left(z_v, z_{u|v}\right) = \begin{cases} \left(\dfrac{f(\eta_u) + \eta_v - \mu_v}{s_v}, \dfrac{\eta_u - g(f(\eta_u) + \eta_v)}{s_{u|v}} \right), & \textit{IC}, \\ \left(\dfrac{f(\eta_u + \delta\sigma_u) + \eta_v - \mu_v}{s_v}, \dfrac{\eta_u + \delta\sigma_u - g(f(\eta_u + \delta\sigma_u) + \eta_v)}{s_{u|v}} \right), & \textit{cause shift}, \\ \left(\dfrac{f(\eta_u) + \eta_v + \delta\sigma_v - \mu_v}{s_v}, \dfrac{\eta_u - g(f(\eta_u) + \eta_v + \delta\sigma_v)}{s_{u|v}} \right), & \textit{effect shift}. \end{cases}$$

In other words, for direct comparison with Lemma 1, except for the degenerate cases listed in Remark 1, both reverse residual distributions change under each shift:

$$\begin{cases} P_1^{z_v} \neq P_0^{z_v} & \text{and} \quad P_1^{z_{u|v}} \neq P_0^{z_{u|v}}, \quad \textit{cause shift}, \\ P_1^{z_v} \neq P_0^{z_v} & \text{and} \quad P_1^{z_{u|v}} \neq P_0^{z_{u|v}}, \quad \textit{effect shift}. \end{cases}$$

Remark 1 (Sources of reverse residual degeneracy)**.** *The three degenerate cases in Lemma 2 are:*

- *Weak causal strength: When the effect of u on v is weak, $g(v) = \mathbb{E}[u \mid v]$ is nearly constant. The effect shift entering g then causes little change in $z_{u|v}$.*
- *Shift invariance: If $f(\eta_u + \delta\sigma_u) = f(\eta_u)$, the cause shift is hidden by a symmetry of f, leaving z_v unchanged in distribution.*
- *Shift equivariance: If $g(f(x + \delta\sigma_u) + \eta_v) - g(f(x) + \eta_v) = \delta\sigma_u$, both the true value of u and its reverse prediction increase by $\delta\sigma_u$ under a cause shift, leaving $z_{u|v}$ unchanged.*

4.2 KERNEL RESIDUAL DISPLACEMENT

With a little abuse of notations, let z and w denote the same residual coordinate under the IC and OOC regimes, with distributions P_0^z and P_1^w, respectively. Lemma 1 and Lemma 2 show that we

can quantify the distribution distance to construct a scoring metric for identifying the true causal direction. We propose a *Kernel Residual Displacement* (KRD) method. Specifically, for P_0^z and P_1^w, their kernel-based displacement $D_k(P_0^z, P_1^w)$ is quantified via Maximum Mean Discrepancy (MMD) (Borgwardt et al., 2006) with a characteristic kernel k:

$$D_k^2(P_0^z, P_1^w) = \mathbb{E}_{Z,Z' \sim P_0^z}[k(Z, Z')] + \mathbb{E}_{W,W' \sim P_1^w}[k(W, W')] - 2\mathbb{E}_{Z \sim P_0^z, W \sim P_1^w}[k(Z, W)], \quad (2)$$

where Z, Z' and W, W' denote independent draws of the given residual coordinate from the IC and OOC distributions P_0^z and P_1^w, respectively. The choice of kernel determines which type of residual changes is captured by D_k. Using the linear kernel $k(z, w) = zw$ yields our **KRD-Linear** variant, for which the distance reduces to the absolute difference in means, $D_k(P_0^z, P_1^w) = \left|\mathbb{E}_{P_0^z}[Z] - \mathbb{E}_{P_1^w}[W]\right|$, offering a simple and computationally efficient measure. To capture more general distributional changes, we adopt the Radial Basis Function (RBF) (Buhmann, 2000) kernel $k_\lambda(Z, Z') = \exp\left(-\frac{(Z-Z')^2}{2\lambda^2}\right)$ $(\lambda > 0)$, which produces the **KRD-RBF** variant.

Once the coordinate-wise MMD metrics are computed, we aggregate them into a directional score to determine the true causal direction. For each candidate direction $i \to j$, KRD defines its directional score as the minimum of the two coordinate displacements:

$$\tau_{\mathrm{KRD}}(i \to j) = \min\left\{D_k(P_0^{z_i}, P_1^{z_i}),\ D_k(P_0^{z_{j|i}}, P_1^{z_{j|i}})\right\}. \quad (3)$$

The minimum follows from Lemma 1. Under the true causal direction, mechanism invariance guarantees that at least one residual coordinate remains unchanged across the two regimes (IC and OOC), yielding a zero population score ($\tau_{\mathrm{KRD}} = 0$). Geometrically, this corresponds to an on-axis residual displacement in Figure 2. Under the reverse direction, the residual changes are often more complex, but specific shifts could accidentally leave one coordinate unchanged. To exclude such cases, Theorem 1 requires both reverse-direction residual displacements to be nonzero, an off-axis displacement in Figure 2, thereby ensuring unique identification of the true direction.

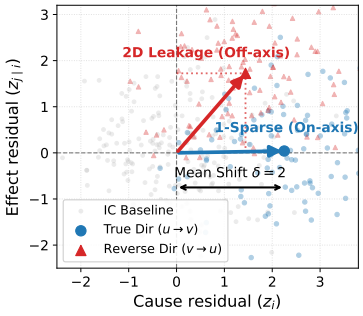

Figure 2: Residual displacement

Theorem 1 (Population identification by KRD)**.** *Under Assumptions 1–2, let k be a characteristic kernel. Excluding the degenerate scenarios in Remark 1, the true causal direction $u \to v$ uniquely satisfies $\tau_{\mathrm{KRD}}(u \to v) < \tau_{\mathrm{KRD}}(v \to u)$. Consequently, the true causal direction is uniquely identified at the population level.*

4.3 Finite-Sample Identification

To estimate $\tau_{\mathrm{KRD}}(u \to v)$ and $\tau_{\mathrm{KRD}}(v \to u)$ with finite samples, let n_0 and n_1 denote the sample sizes of the IC and OOC regimes, respectively. For any residual coordinate $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$, let $\widehat{z}_\ell^{(e)}$ be its fitted value for observation ℓ in environment $e \in \{0, 1\}$ (IC and OOC, respectively). Its empirical KDR is

$$\widehat{D}_k^2(\widehat{z}^{(0)}, \widehat{z}^{(1)}) = \frac{1}{n_0^2}\sum_{\ell,m=1}^{n_0} k(\widehat{z}_\ell^{(0)}, \widehat{z}_m^{(0)}) + \frac{1}{n_1^2}\sum_{\ell,m=1}^{n_1} k(\widehat{z}_\ell^{(1)}, \widehat{z}_m^{(1)}) - \frac{2}{n_0 n_1}\sum_{\ell=1}^{n_0}\sum_{m=1}^{n_1} k(\widehat{z}_\ell^{(0)}, \widehat{z}_m^{(1)}). \quad (4)$$

For the linear kernel, this reduces to $\widehat{D}_k = |\bar{\widehat{z}}^{(1)} - \bar{\widehat{z}}^{(0)}|$, where $\bar{\widehat{z}}^{(e)} = \frac{1}{n_e}\sum_{\ell=1}^{n_e} \widehat{z}_\ell^{(e)}$ is the sample mean of the fitted residual coordinate in environment e. Replacing the population displacements in equation 3 with their empirical estimates gives the finite-sample directional score

$$\widehat{\tau}_{\mathrm{KRD}}(i \to j) = \min\left\{\widehat{D}_k(\widehat{z}_i^{(0)}, \widehat{z}_i^{(1)}), \widehat{D}_k(\widehat{z}_{j|i}^{(0)}, \widehat{z}_{j|i}^{(1)})\right\}.$$

We compute this score for both candidate directions and select the direction with the smaller empirical KRD score:

$$\widehat{d} = \underset{d \in \{u \to v, v \to u\}}{\arg\min}\ \widehat{\tau}_{\mathrm{KRD}}(d). \quad (5)$$

We examine the finite-sample properties of this identification rule in the following assumption.

Assumption 3 (Finite-sample regularity)**.** *We assume that the following conditions hold for the estimated standardized residuals.*

- *Residual estimation accuracy: For any $\gamma \in (0,1)$, the estimated residuals $\widehat{z}^{(e)}$ concentrate around their ideal population counterparts $z^{(e)}$ with probability at least $1-\gamma/2$:*

$$\max_{\substack{z\in\{z_u,z_{v|u},z_v,z_{u|v}\}\\ e\in\{0,1\}}} \frac{1}{n_e}\sum_{a=1}^{n_e}\left|\widehat{z}_a^{(e)} - z_a^{(e)}\right| \le \varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2),$$

  where $\varepsilon_{\mathrm{fit}}(n_0,n_1,\delta)\to 0$ as $\min(n_0,n_1)\to\infty$, capturing the rate of cumulative error from regression, centering, and scale estimation.

- *Residual concentration: For the linear kernel, every ideal standardized residual coordinate $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ is sub-Gaussian (Boucheron et al., 2003) in each environment $e \in \{0,1\}$ with a uniform parameter $\sigma_0 > 0$, satisfying $\mathbb{E}\left[\exp\left(s(z-\mathbb{E}[z])\right)\right] \le \exp\left(\frac{\sigma_0^2 s^2}{2}\right)$ for all $s \in \mathbb{R}$.*

Assumption 3 controls the errors from residual estimation and from approximating population displacements. Theorem 2 provides the finite-sample guarantee based on Assumptions 1–3.

Theorem 2 (Finite-sample identification)**.** *Suppose Assumptions 1, 2, and 3 hold. For either kernel, the rule in equation 5 recovers the true direction with probability at least $1-\gamma$ whenever*

$$\tau_{\mathrm{KRD}}(v\to u) > 2E_{N,\gamma},$$

where $\tau_{\mathrm{KRD}}(v\to u)$ is the population reverse-direction score in equation 3, $N = n_0+n_1$, and the distance error bound is

$$E_{N,\gamma} = \begin{cases} 2\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2) + \sigma_0\sqrt{2\log(16/\gamma)\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}, & \textit{linear},\\ \frac{2}{\lambda}\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2) + \left(1+\sqrt{2\log(16/\gamma)}\right)\left(\frac{1}{\sqrt{n_0}}+\frac{1}{\sqrt{n_1}}\right), & \textit{RBF}. \end{cases}$$

Computational complexity Let $N = n_0 + n_1$ be the total number of IC and OOC observations. With a fixed regression specification and a fixed hyperparameter grid, fitting the candidate regressions and computing their residuals require $\mathcal{O}(N)$ time. The remaining cost depends on the kernel. KRD-Linear compares residual means and therefore has an overall time complexity of $\mathcal{O}(N)$. KRD-RBF evaluates the exact empirical MMD, which requires all within- and cross-context kernel terms and therefore takes $\mathcal{O}(N^2)$ time (Gretton et al., 2012). Both variants use $\mathcal{O}(N)$ memory when the RBF kernel sums are accumulated blockwise without storing the full kernel matrices.

4.4 Explicit Threshold for Linear SCMs

This section provides more explicit answers to several theoretical questions about linear Gaussian systems. We establish population identifiability and finite-sample recovery guarantees that connect successful identification to perturbation magnitude, regression error, sample size, and directional separation. Our analysis further reveals a non-monotone signal-to-noise geometry: identification is strongest when causal signal and downstream noise are balanced, and deteriorates toward both near-deterministic and noise-dominated extremes.

Consider the linear Gaussian model $x_u = \eta_u$ and $x_v = ax_u + \eta_v$, where $\eta_u \sim \mathcal{N}(0,1)$ and $\eta_v \sim \mathcal{N}(0,\sigma_v^2)$ are independent. Under the true direction $u\to v$, one residual coordinate remains unchanged between IC and OOC, giving $\tau_{\mathrm{KRD}}(u\to v) = 0$. Under the reverse direction $v\to u$, the two standardized residual displacements have magnitudes $|\delta|t/\sqrt{1+t^2}$ and $|\delta|/\sqrt{1+t^2}$, where $t = |a|/\sigma_v$ is the signal-to-noise ratio. Therefore,

$$\tau_{\mathrm{KRD}}(v\to u) = |\delta|\phi(t), \qquad \phi(t) := \frac{\min\{1,t\}}{\sqrt{1+t^2}}. \tag{6}$$

Thus, $|\delta|\phi(t)$ gives the population separation between the true and reverse directions. To ensure this population margin yields correct empirical identification under finite-sample estimation error, Corollary 1 quantifies the required sample size and perturbation magnitude.

Corollary 1 (Linear-kernel identification under a linear Gaussian SCM)**.** *Suppose Assumptions 1, 2, and 3 hold. Under the linear Gaussian SCM, linear-kernel KRD correctly identifies the true causal direction $u \to v$ in finite samples with probability at least $1-\gamma$, provided that the OOC perturbation magnitude satisfies*

$$|\delta| > \frac{2(1+\varepsilon_{\mathrm{OLS}})\varepsilon_N}{\phi(t) - 2\varepsilon_{\mathrm{OLS}}}.$$

Furthermore, a sufficient IC sample size n_0 ensuring $\phi(t) > 2\varepsilon_{\mathrm{OLS}}$ is

$$n_0 > 1 + \left(\sqrt{\log(8/\gamma)} + \sqrt{\log(8/\gamma) + 8\log(16/\gamma)/\phi(t)^2}\right)^2.$$

Here, $\varepsilon_{\mathrm{OLS}} = \sqrt{2\log(16/\gamma)/L}$, $\varepsilon_N = \sqrt{2\log(16/\gamma)(n_0^{-1} + n_1^{-1})}$, and $L = n_0 - 1 - 2\sqrt{(n_0-1)\log(8/\gamma)}$.

Signal-to-noise ratio Equation 6 shows how the signal-to-noise ratio t affects causal identifiability. When $t = 1$, the causal signal and downstream noise are balanced, so the two residual displacement coordinates under the reverse direction are equal, and the detection margin reaches its maximum, $\phi(1) = 1/\sqrt{2}$. As $t \to \infty$, the causal relationship becomes nearly deterministic, and the reverse regression can almost perfectly invert the mechanism, causing one reverse-direction residual displacement to vanish. In contrast, as $t \to 0$, the downstream noise dominates the causal signal and the perturbation is transmitted too weakly to be detected, causing the other residual displacement to vanish. Thus, causal identification becomes more difficult at both extremes as $\phi(t) \to 0$.

5 Experiment

We evaluate the proposed method on both synthetic and real-world benchmark datasets against baselines. We measure causal direction identification using the **identification accuracy**, defined as the proportion of pairs for which the predicted direction matches the ground truth.

Datasets For synthetic data, we first use two controlled simulation settings, linear causal relationships with Gaussian noise and nonlinear causal relationships with non-Gaussian noise, to study performance under different signal-to-noise ratios and OOC magnitudes. We then examine 18 extended simulation settings where our theoretical assumptions may not strictly hold, including **(i)** AN, AN-c, LS, LS-c, and MNU (Tagasovska et al., 2020); **(ii)** SIM, SIM-c, SIM-g, and SIM-ln (Mooij et al., 2016); **(iii)** Multi and Net (Guyon et al., 2019); **(iv)** Per, Sig, and Vex (Xi et al., 2025); and **(v)** Qd-V, Sig-V, Rbf-V, and NN-V (Chen et al., 2026). Across these settings, OOC perturbations are randomly introduced to either the cause or the effect variable. For real-world data, we consider datasets without explicit IC/OOC labels, including the **(i)** Tübingen cause–effect pairs (Mooij et al., 2016) and **(ii)** gene network reverse-engineering benchmarks (DREAM) (Marbach et al., 2009), for which OOC samples are first detected from available data and then used in causal discovery, as well as the **(iii)** Tennessee Eastman Process fault-diagnosis (TEP) (Reinartz et al., 2021; Menegozzo et al., 2022), where abnormal operating conditions provide natural OOC episodes. Dataset descriptions and implementation details are deferred to Appendix B.1.

Baselines We compare our method with representative bivariate causal discovery baselines, including **ANM-HSIC** (Hoyer et al., 2008), **ANM-Entropy** (Nowzohour & Bühlmann, 2016), and **ANM-MML** (Stegle et al., 2010) from the additive-noise family, **IGCI** (Daniusis et al., 2012) from the information-geometric family, and the regression-based **RECI** (Blöbaum et al., 2018). For these single-environment methods, we evaluate both a **pooled** version using the combined IC and OOC samples and a **split** version that applies the method separately to IC and OOC and averages the resulting scores. We additionally consider multi-environment methods, including **MSS** (Perry et al., 2022), **UT-IGSP** (Squires et al., 2020), and **CD-NOD** (Huang et al., 2020).

5.1 Synthetic Data Experiments

Results on controlled simulation settings Figure 3 displays performance of all methods across different signal-to-noise ratios t and OOC magnitudes δ. In the linear Gaussian setting, single-context methods remain close to 50%, while pooling IC and OOC data with a large shift ($\delta = 4$) introduces distributional asymmetry that gives ANM-HSIC-Pooled and ANM-Entropy-Pooled some

identification ability. In the nonlinear non-Gaussian setting, single-context methods perform better but decline at large t, where a stronger causal signal can make both directions more explainable and thus weaken the directional asymmetry. Cross-context methods generally show an arch-shaped trend over t, peaking around the balanced case $t = 1$. Larger OOC shifts improve identification in the linear setting, whereas this trend is weaker under nonlinear mechanisms. Overall, KRD performs best in the linear Gaussian setting and is comparable to MSS in the nonlinear non-Gaussian setting.

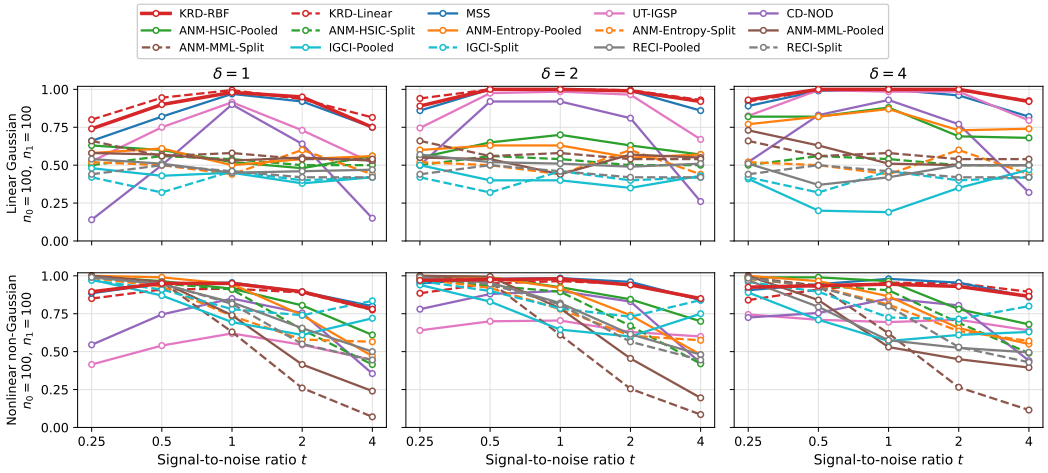

Figure 3: Identification accuracy under controlled simulation settings. The top row shows linear causal relationships with Gaussian errors, and the bottom row shows nonlinear causal relationships with non-Gaussian errors. Columns correspond to different OOC perturbation magnitudes, with the perturbation randomly applied to either the cause or the effect.

Results on extended simulation settings Figure 4 summarizes the results across the 18 extended simulation settings. KRD performs comparably to MSS and consistently outperforms the other baselines, while requiring substantially less computation time. In terms of the two types of OOC perturbation, KRD performs better for effect shifts, while MSS performs better for cause shifts. Detailed results are provided in Appendix B.3, and the computational complexity of KRD and kernel-based independence tests is compared in Appendix B.5.

5.2 Real Data Experiments

Results on unlabeled cases When IC/OOC labels are unavailable, we construct the two environments using anomaly detection. A smooth trend ribbon is fitted to the data, with points inside the ribbon assigned to IC and outlying points assigned to OOC. Cross-context methods are then applied to these environments. The detailed implementation is given in Appendix B.4. Because the Tübingen source groups contain unequal numbers of pairs, we report both overall accuracy across

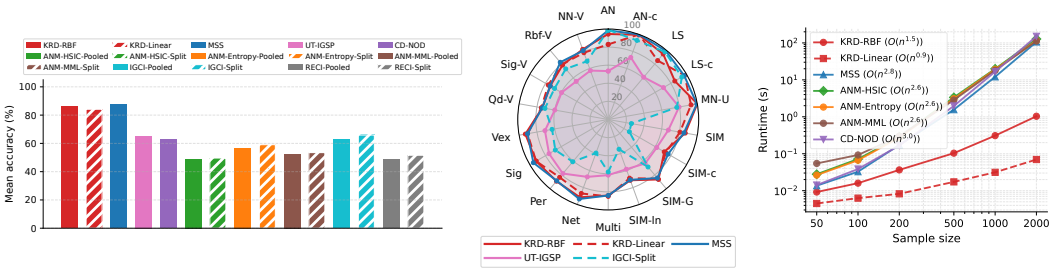

Figure 4: Identification accuracy under OOC perturbations of magnitude $\delta = 2$, randomly applied to either the cause or the effect. Left: average accuracy across all the 18 extended settings. Middle: per-setting accuracy of the top five methods. Right: computation time as sample size increases.

all pairs and group-weighted accuracy that weights each group equally. KRD achieves over 70% under both measures and outperforms all competing methods. Results on DREAM are weaker for all cross-context methods, because the distributional shifts are less clearly separated.

Results on labeled case The TEP for a complex manufacturing system provides explicit labels, where IC and OOC correspond to normal and faulty states, respectively. Faults 1 and 2 introduce step changes in the composition of feed stream 4: Fault 1 changes the A/C ratio while keeping the B composition fixed, and Fault 2 changes the B composition while keeping the A/C ratio fixed. Fault 6 is a sudden loss of A feed in stream 1, representing a feed-supply disruption. Top-performing methods are consistently concentrated among cross-context approaches, whereas single-context baselines struggle. Notably, under the Fault 1 condition where existing methods degrade to near-random guessing, KRD-RBF achieves a 65% accuracy. Across the remaining fault conditions, KRD consistently ranks among the top two performing methods, demonstrating its effectiveness and robustness in complex real-world industrial settings.

Table 1: Identification accuracy (%) on DREAM, Tübingen, and TEP. **Bold** and underlined values indicate the best and second-best distinct available accuracies in each column, respectively.

Method	DREAM				Tübingen		TEP		
	D4S1	D4S2A	D4S2B	D4S2C	Acc	Weig Acc	Fault 1	Fault 2	Fault 6
Cross-context									
Ours (KRD-RBF)	54.17	43.52	52.34	51.06	**69.47**	<u>73.01</u>	**65.00**	<u>54.38</u>	<u>58.75</u>
Ours (KRD-Linear)	54.17	46.91	55.32	51.59	**69.47**	**77.74**	<u>59.38</u>	**55.00**	56.88
MSS (hard)	33.33	30.56	32.98	32.80	41.05	46.13	11.25	10.00	31.88
MSS (soft)	54.17	50.93	50.21	51.85	58.95	64.09	52.50	52.50	**59.38**
CD-NOD	20.83	24.07	28.94	25.13	29.47	36.01	11.25	7.50	26.88
UT-IGSP	54.17	48.77	52.55	52.65	65.26	70.47	52.50	**55.00**	51.25
Single-context									
ANM-HSIC	50.00	**64.20**	<u>58.72</u>	<u>55.03</u>	60.00	63.70	50.63	52.50	55.63
ANM-Entropy	**66.67**	38.89	47.23	40.74	57.89	54.09	38.75	46.25	49.38
ANM-MML	<u>58.33</u>	61.11	**59.15**	**54.50**	<u>67.37</u>	66.72	51.88	52.50	50.00
IGCI	33.33	37.04	39.57	39.15	63.16	66.96	50.63	45.00	51.88
RECI	41.67	<u>61.73</u>	50.64	48.15	58.95	65.82	46.88	41.25	49.38

6 Conclusion

We investigate causal direction identification in the minimal setting of two variables across IC and OOC environments undergoing a mean shift. In the true causal direction, one residual coordinate remains invariant between IC and OOC while the shift is confined to the other, whereas both coordinates typically change in the reverse direction. The proposed KRD method combines residual invariance with the sparse mechanism shift principle by selecting the direction with the smallest residual-coordinate displacement. We establish population and finite-sample identification guarantees; specifically, in linear Gaussian SCMs, identification is governed by sample sizes, shift magnitude, and the signal-to-noise ratio, providing a tractable instance of broader non-linear failure modes related to weak causal strength, mechanism shift invariance, and shift equivariance. By operating on just two environments and replacing expensive conditional independence tests with kernel distances between marginal residual distributions, KRD offers a computationally efficient framework that characterizes the fundamental limits of cross-context identification. While our theoretical analysis assumes additive noise models, empirical gains across various simulated benchmarks and real-world benchmarks demonstrate that residual displacement captures broad structural shifts, opening promising avenues for future extensions to non-additive SCMs, multi-environment systems, and multivariate DAG discovery.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing proofs, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, and implement selected baseline methods. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, design or provide feedback on the research methodology or experiments, implement the proposed method, translate the manuscript, clean or reformat datasets, conduct qualitative or thematic data analysis, or interpret results. We reviewed all AI-assisted work. In particular, we checked the AI-assisted baseline implementations against the original baseline methods and tested the code for correctness, and we independently reviewed and refined all AI-assisted proofs. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies bivariate causal inference through theoretical analysis and experiments on synthetic data and publicly available benchmark datasets. The study does not involve recruiting human participants or collecting new personal data. We have not identified specific ethical concerns arising from the research as presented.

REFERENCES

Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *Journal of machine learning research*, 20(162):1–56, 2019.

Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pp. 208–240. Springer, 2003.

Martin Dietrich Buhmann. Radial basis functions. *Acta numerica*, 9:1–38, 2000.

Yikang Chen, Xingzhe Sun, et al. Causal discovery via quantile partial effect. In *International Conference on Learning Representations*, volume 2026, pp. 148015–148046, 2026.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

Maysa S De Magalhães, AFB Costa, and FD Moura Neto. A hierarchy of adaptive x control charts. *International Journal of Production Economics*, 119(2):271–283, 2009.

W. Dana Flanders. Invited commentary: Two-phase, generalized case-control designs for quantitative longitudinal outcomes and evolution of the case-control study. *American Journal of Epidemiology*, 189(2):91–94, 2020. doi: 10.1093/aje/kwz200.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Isabelle Guyon, Alexander Statnikov, and Berna Bakir Batu. *Cause effect pairs in machine learning*. Springer, 2019.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21(89):1–53, 2020.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *The Journal of Machine Learning Research*, 14(1):111–152, 2013.

Samory Kpotufe, Eleni Sgouritsa, Dominik Janzing, and Bernhard Schölkopf. Consistency of causal inference under the additive noise model. In *International Conference on Machine Learning*, pp. 478–486. PMLR, 2014.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Yiqing Li, Xiaofei Wang, Boyang Sun, Yewei Xia, and Kun Zhang. Independence test for linear non-gaussian data and applications in causal discovery. In *The Fourteenth International Conference on Learning Representations*, 2026.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of computational biology*, 16(2):229–239, 2009.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

Giovanni Menegozzo, Diego Dall'Alba, and Paolo Fiorini. Cipcad-bench: Continuous industrial process datasets for benchmarking causal discovery methods. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pp. 2124–2131. IEEE, 2022.

Francesco Montagna. On the identifiability of causal graphs with the invariance principle. *arXiv preprint arXiv:2510.13583*, 2025.

Douglas C Montgomery and Cheryl L Jennings. *Introduction to statistical quality control*, volume 5. Wiley New York, 1991.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17(32):1–102, 2016.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of machine learning research*, 21(99):1–108, 2020.

Christopher Nowzohour and Peter Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3):471–485, 2016.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979. doi: 10.1093/biomet/66.3.403.

S Joe Qin. Survey on data-driven industrial process monitoring and diagnosis. *Annual reviews in control*, 36(2):220–234, 2012.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Christopher Reinartz, Murat Kulahci, and Ole Ravn. An extended tennessee eastman simulation dataset for fault-detection and decision support systems. *Computers & chemical engineering*, 149:107281, 2021.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pp. 18741–18753. PMLR, 2022.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. *Advances in neural information processing systems*, 28, 2015.

Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

Oliver Stegle, Dominik Janzing, Kun Zhang, Joris M Mooij, and Bernhard Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. *Advances in neural information processing systems*, 23, 2010.

Natasa Tagasovska, Valérie Chavez-Demoulin, and Thibault Vatter. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pp. 9311–9323. PMLR, 2020.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal processing*, 167:107299, 2020.

Johnny Xi, Hugh Dance, Peter Orbanz, and Benjamin Bloem-Reddy. Distinguishing cause from effect with causal velocity models. *arXiv preprint arXiv:2502.05122*, 2025.

Li Xue and Zhen He. Economic design of ewma control charts with variable sampling intervals for monitoring the mean and standard deviation under preventive maintenance and taguchi's loss functions. *Mathematical Problems in Engineering*, 2021(1):6686426, 2021.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

A Proofs of the Main Results

A.1 Proofs for Section 4.1

Lemma 1 (Residual distribution change under true direction)**.** *Let Assumptions 1–2 hold. Under the true direction $u \to v$, the standardized residual distributions satisfy*

$$\begin{cases} P_1^{z_u} \neq P_0^{z_u} & \text{and} \quad P_1^{z_{v|u}} = P_0^{z_{v|u}}, \quad \textit{cause shift}, \\ P_1^{z_u} = P_0^{z_u} & \text{and} \quad P_1^{z_{v|u}} \neq P_0^{z_{v|u}}, \quad \textit{effect shift}. \end{cases}$$

Proof. The zero noise means give $\mathbb{E}_0[x_u] = 0$. Under IC, $r_u = \eta_u$ and $r_{v|u} = \eta_v$, so the fixed IC scales are $s_u = \sigma_u$ and $s_{v|u} = \sigma_v$. Consequently,

$$(z_u, z_{v|u}) = \left(\frac{\eta_u}{\sigma_u}, \frac{\eta_v}{\sigma_v}\right) \qquad \text{under IC.}$$

Under a cause shift, Assumption 2 gives $x_u = \eta_u + \delta\sigma_u$ and $x_v = f(x_u) + \eta_v$. Subtracting the same reference map f cancels its contribution exactly: $r_{v|u} = x_v - f(x_u) = \eta_v$. Therefore,

$$(z_u, z_{v|u}) = \left(\frac{\eta_u}{\sigma_u} + \delta, \frac{\eta_v}{\sigma_v}\right).$$

The second coordinate has the same distribution as under IC. The first has mean δ instead of zero. Since these means are finite and $\delta > 0$, its distribution cannot equal the IC distribution.

Under an effect shift, $x_u = \eta_u$ and $x_v = f(x_u) + \eta_v + \delta\sigma_v$, giving

$$(z_u, z_{v|u}) = \left(\frac{\eta_u}{\sigma_u}, \frac{\eta_v}{\sigma_v} + \delta\right).$$

These two cases prove the stated distributional equalities and inequalities for arbitrary f satisfying the assumptions. □

Lemma 2 (Residual distribution change under reverse direction)**.** *Under the same assumptions, let $\mu_v = \mathbb{E}[f(\eta_u)]$. For the reverse direction $v \to u$, the standardized residual pair is*

$$\big(z_v, z_{u|v}\big) = \begin{cases} \left(\dfrac{f(\eta_u) + \eta_v - \mu_v}{s_v}, \dfrac{\eta_u - g(f(\eta_u) + \eta_v)}{s_{u|v}}\right), & \textit{IC}, \\ \left(\dfrac{f(\eta_u + \delta\sigma_u) + \eta_v - \mu_v}{s_v}, \dfrac{\eta_u + \delta\sigma_u - g(f(\eta_u + \delta\sigma_u) + \eta_v)}{s_{u|v}}\right), & \textit{cause shift}, \\ \left(\dfrac{f(\eta_u) + \eta_v + \delta\sigma_v - \mu_v}{s_v}, \dfrac{\eta_u - g(f(\eta_u) + \eta_v + \delta\sigma_v)}{s_{u|v}}\right), & \textit{effect shift}. \end{cases}$$

In other words, for direct comparison with Lemma 1, except for the degenerate cases listed in Remark 1, both reverse residual distributions change under each shift:

$$\begin{cases} P_1^{z_v} \neq P_0^{z_v} & \text{and} \quad P_1^{z_{u|v}} \neq P_0^{z_{u|v}}, \quad \textit{cause shift}, \\ P_1^{z_v} \neq P_0^{z_v} & \text{and} \quad P_1^{z_{u|v}} \neq P_0^{z_{u|v}}, \quad \textit{effect shift}. \end{cases}$$

Proof. Under IC, $\mathbb{E}_0[x_v] = \mathbb{E}[f(\eta_u)] = \mu_v$ because $\mathbb{E}[\eta_v] = 0$. The reverse residuals are therefore $r_v = x_v - \mu_v$ and $r_{u|v} = x_u - g(x_v)$. The tower property also gives $\mathbb{E}_0[g(x_v)] = \mathbb{E}_0[x_u] = 0$, so both reverse residuals are centered under IC.

Substituting $x_u = \eta_u$ and $x_v = f(\eta_u) + \eta_v$ yields

$$r_v = f(\eta_u) + \eta_v - \mu_v, \qquad r_{u|v} = \eta_u - g(f(\eta_u) + \eta_v).$$

Dividing by s_v and $s_{u|v}$ gives the IC row of the lemma.

For a cause shift, substitution instead yields

$$\begin{aligned} r_v &= f(\eta_u + \delta\sigma_u) + \eta_v - \mu_v, \\ r_{u|v} &= \eta_u + \delta\sigma_u - g\big(f(\eta_u + \delta\sigma_u) + \eta_v\big). \end{aligned}$$

For an effect shift, the corresponding expressions are

$$r_v = f(\eta_u) + \eta_v + \delta\sigma_v - \mu_v,$$
$$r_{u|v} = \eta_u - g\big(f(\eta_u) + \eta_v + \delta\sigma_v\big).$$

Normalizing both cases by the same IC scales proves the other two rows. We now verify that both coordinates change under each shift unless degenerate cases from Remark 1 occur:

1. Cause Shift: For z_v, $P_1^{z_v} \neq P_0^{z_v}$ unless $f(\eta_u + \delta\sigma_u) \overset{d}{=} f(\eta_u)$ (*Shift Invariance*). For $z_{u|v}$, $P_1^{z_{u|v}} \neq P_0^{z_{u|v}}$ unless the shift in x_u is perfectly offset by the shift in the reverse prediction g, i.e., $g\big(f(x + \delta\sigma_u) + \eta_v\big) - g\big(f(x) + \eta_v\big) = \delta\sigma_u$ (*Shift Equivariance*).
2. Effect Shift: For z_v, the additive shift induces a direct mean change $\mathbb{E}_{e=1}[z_v] - \mathbb{E}_{e=0}[z_v] = \delta\sigma_v/s_v \neq 0$, ensuring $P_1^{z_v} \neq P_0^{z_v}$. For $z_{u|v}$, the input argument to g is perturbed by $\delta\sigma_v$, ensuring $P_1^{z_{u|v}} \neq P_0^{z_{u|v}}$ unless g is nearly insensitive to variations in v (*Weak Causal Strength*).

Excluding these conditions, both reverse residual distributions strictly change under both shift types, completing the proof. □

A.2 Proofs for Section 4.2

Theorem 1 (Population identification by KRD)**.** *Under Assumptions 1–2, let k be a characteristic kernel. Excluding the degenerate scenarios in Remark 1, the true causal direction $u \to v$ uniquely satisfies $\tau_{\mathrm{KRD}}(u \to v) < \tau_{\mathrm{KRD}}(v \to u)$. Consequently, the true causal direction is uniquely identified at the population level.*

Proof. Since k is a characteristic kernel, the Maximum Mean Discrepancy $D_k(P, Q)$ is a strict metric on probability distributions satisfying $D_k(P, Q) = 0$ if and only if $P = Q$, and $D_k(P, Q) > 0$ otherwise (Borgwardt et al., 2006). Under the true causal direction $u \to v$, Lemma 1 guarantees that at least one residual coordinate remains invariant across regimes ($P_1^{z_{v|u}} = P_0^{z_{v|u}}$ under a cause shift, or $P_1^{z_u} = P_0^{z_u}$ under an effect shift), yielding at least one zero displacement and thus $\tau_{\mathrm{KRD}}(u \to v) = 0$. Under the reverse direction $v \to u$, Lemma 2 establishes that barring the degenerate conditions in Remark 1, both residual distributions strictly change across environments ($P_1^{z_v} \neq P_0^{z_v}$ and $P_1^{z_{u|v}} \neq P_0^{z_{u|v}}$). The characteristic property of k ensures both coordinate displacements are strictly positive, giving $\tau_{\mathrm{KRD}}(v \to u) > 0$. Consequently, $\tau_{\mathrm{KRD}}(u \to v) = 0 < \tau_{\mathrm{KRD}}(v \to u)$, uniquely identifying $u \to v$ as the true causal direction at the population level. □

A.3 Proofs for Section 4.3

Theorem 2 (Finite-sample identification)**.** *Suppose Assumptions 1, 2, and 3 hold. For either kernel, the rule in equation 5 recovers the true direction with probability at least $1 - \gamma$ whenever*

$$\tau_{\mathrm{KRD}}(v \to u) > 2E_{N,\gamma},$$

where $\tau_{\mathrm{KRD}}(v \to u)$ is the population reverse-direction score in equation 3, $N = n_0 + n_1$, and the distance error bound is

$$E_{N,\gamma} = \begin{cases} 2\varepsilon_{\mathrm{fit}}(n_0, n_1, \gamma/2) + \sigma_0\sqrt{2\log(16/\gamma)\left(\dfrac{1}{n_0} + \dfrac{1}{n_1}\right)}, & \textit{linear}, \\ \dfrac{2}{\lambda}\varepsilon_{\mathrm{fit}}(n_0, n_1, \gamma/2) + \big(1 + \sqrt{2\log(16/\gamma)}\big)\left(\dfrac{1}{\sqrt{n_0}} + \dfrac{1}{\sqrt{n_1}}\right), & \textit{RBF}. \end{cases}$$

Proof. For any candidate direction $d \in \{u \to v, v \to u\}$, the 1-Lipschitz continuity of the minimum operator under the L_∞ norm implies that the score estimation error is bounded by the maximum coordinate-wise displacement error across all four residual coordinates $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$:

$$\big|\widehat{\tau}_{\mathrm{KRD}}(d) - \tau_{\mathrm{KRD}}(d)\big| \leq \max_{z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}} \left|\widehat{D}_k(\hat{z}^{(0)}, \hat{z}^{(1)}) - D_k(P_0^z, P_1^z)\right|.$$

By the triangle inequality, each coordinate-wise error decomposes into a residual fitting error $\mathrm{Err}_{\mathrm{fit}}$ and a finite-sample MMD concentration error $\mathrm{Err}_{\mathrm{sample}}$:

$$\left|\widehat{D}_k(\widehat{z}^{(0)},\widehat{z}^{(1)}) - D_k(P_0^z,P_1^z)\right| \le \underbrace{\left|\widehat{D}_k(\widehat{z}^{(0)},\widehat{z}^{(1)}) - \widehat{D}_k(z^{(0)},z^{(1)})\right|}_{\mathrm{Err}_{\mathrm{fit}}} + \underbrace{\left|\widehat{D}_k(z^{(0)},z^{(1)}) - D_k(P_0^z,P_1^z)\right|}_{\mathrm{Err}_{\mathrm{sample}}}.$$

Under Assumption 3, the average residual estimation accuracy satisfies $\frac{1}{n_e}\sum_{a=1}^{n_e}|\widehat{z}_a^{(e)} - z_a^{(e)}| \le \varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2)$ simultaneously across all coordinates $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ and environments $e \in \{0,1\}$ with probability at least $1-\gamma/2$.

For the linear kernel $k(x,x') = xx'$, $D_k(P_0^z,P_1^z) = |\mathbb{E}[z^{(1)}] - \mathbb{E}[z^{(0)}]|$, and $\widehat{D}_k(A,B) = |\bar{A} - \bar{B}|$. The reverse triangle inequality directly bounds the fitting error by the sum of average residual errors in both environments:

$$\begin{aligned}\mathrm{Err}_{\mathrm{fit}} = \left||\bar{\widehat{z}}^{(1)} - \bar{\widehat{z}}^{(0)}| - |\bar{z}^{(1)} - \bar{z}^{(0)}|\right| &\le \frac{1}{n_1}\sum_{a=1}^{n_1}|\widehat{z}_a^{(1)} - z_a^{(1)}| + \frac{1}{n_0}\sum_{a=1}^{n_0}|\widehat{z}_a^{(0)} - z_a^{(0)}| \\ &\le 2\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2).\end{aligned}$$

For the sampling error, applying the triangle inequality yields $\mathrm{Err}_{\mathrm{sample}} \le \left|(\bar{z}^{(1)} - \bar{z}^{(0)}) - (\mathbb{E}[z^{(1)}] - \mathbb{E}[z^{(0)}])\right|$. By Assumption 3, each ideal coordinate $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ is σ_0-sub-Gaussian, so the sample mean difference $(\bar{z}^{(1)} - \bar{z}^{(0)}) - (\mathbb{E}[z^{(1)}] - \mathbb{E}[z^{(0)}])$ is sub-Gaussian with parameter $\sigma_0\sqrt{\frac{1}{n_0}+\frac{1}{n_1}}$. Standard sub-Gaussian concentration implies that for a single coordinate, $\mathrm{Err}_{\mathrm{sample}} \le \sigma_0\sqrt{2\log(16/\gamma)\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}$ with failure probability at most $\gamma/8$. Applying a union bound across all 4 coordinates $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ guarantees that this sampling bound holds with probability at least $1-\gamma/2$. Combining $\mathrm{Err}_{\mathrm{fit}}$ and $\mathrm{Err}_{\mathrm{sample}}$ gives the total linear kernel error bound:

$$E_{N,\gamma} = 2\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2) + \sigma_0\sqrt{2\log(16/\gamma)\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}.$$

For the RBF kernel $k_\lambda(x,x') = \exp\!\left(-\frac{(x-x')^2}{2\lambda^2}\right)$, let $\Phi(x) = k_\lambda(x,\cdot)$ be the canonical feature map into the RKHS $\mathcal{H}_k$. Using $1-e^{-s} \le s$ for $s \ge 0$, the feature map satisfies $\|\Phi(x) - \Phi(x')\|_{\mathcal{H}_k} = \sqrt{2(1-\exp(-(x-x')^2/2\lambda^2))} \le \frac{|x-x'|}{\lambda}$. The empirical MMD fitting error for $\widehat{D}_k(A,B) = \left\|\frac{1}{|A|}\sum_{a\in A}\Phi(a) - \frac{1}{|B|}\sum_{b\in B}\Phi(b)\right\|_{\mathcal{H}_k}$ is thus bounded via this Lipschitz constant:

$$\mathrm{Err}_{\mathrm{fit}} \le \frac{1}{\lambda n_1}\sum_{a=1}^{n_1}|\widehat{z}_a^{(1)} - z_a^{(1)}| + \frac{1}{\lambda n_0}\sum_{a=1}^{n_0}|\widehat{z}_a^{(0)} - z_a^{(0)}| \le \frac{2}{\lambda}\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2).$$

For the sampling error, bounded kernels ($0 \le k_\lambda \le 1$) satisfy the expectation bound $\mathbb{E}\left[\left|\widehat{D}_k(z^{(0)},z^{(1)}) - D_k(P_0^z,P_1^z)\right|\right] \le \frac{1}{\sqrt{n_0}}+\frac{1}{\sqrt{n_1}}$. Since altering one sample point changes $\widehat{D}_k$ by at most $\frac{2}{n_e}$, McDiarmid's inequality (McDiarmid et al., 1989) ensures that for each coordinate z, $\mathrm{Err}_{\mathrm{sample}} \le \left(1+\sqrt{2\log(16/\gamma)}\right)\left(\frac{1}{\sqrt{n_0}}+\frac{1}{\sqrt{n_1}}\right)$ holds with failure probability at most $\gamma/8$. Taking a union bound over all 4 coordinates $z \in \{z_u, z_{v|u}, z_v, z_{u|v}\}$ limits the total sampling failure probability to $\gamma/2$. Combining $\mathrm{Err}_{\mathrm{fit}}$ and $\mathrm{Err}_{\mathrm{sample}}$ yields the total RBF kernel error bound:

$$E_{N,\gamma} = \frac{2}{\lambda}\varepsilon_{\mathrm{fit}}(n_0,n_1,\gamma/2) + \left(1+\sqrt{2\log(16/\gamma)}\right)\left(\frac{1}{\sqrt{n_0}}+\frac{1}{\sqrt{n_1}}\right).$$

Taking a union bound over the fitting event (holding with probability $\ge 1-\gamma/2$) and the finite-sample sampling event (holding with probability $\ge 1-\gamma/2$), the score error satisfies $\max_d|\widehat{\tau}_{\mathrm{KRD}}(d) - \tau_{\mathrm{KRD}}(d)| \le E_{N,\gamma}$ with overall probability at least $1-\gamma$, where $E_{N,\gamma}$ takes

the explicit expression derived above for the corresponding kernel. By Lemma 1, the population score satisfies $\tau_{\mathrm{KRD}}(u \to v) = 0$, which implies $\widehat{\tau}_{\mathrm{KRD}}(u \to v) \le E_{N,\gamma}$, while for the reverse direction $\widehat{\tau}_{\mathrm{KRD}}(v \to u) \ge \tau_{\mathrm{KRD}}(v \to u) - E_{N,\gamma}$. The decision rule correctly identifies the true causal direction $u \to v$ via $\widehat{\tau}_{\mathrm{KRD}}(u \to v) < \widehat{\tau}_{\mathrm{KRD}}(v \to u)$ whenever:

$$E_{N,\gamma} < \tau_{\mathrm{KRD}}(v \to u) - E_{N,\gamma} \iff \tau_{\mathrm{KRD}}(v \to u) > 2E_{N,\gamma},$$

which completes the proof. □

A.4 Proofs for Section 4.4

Corollary 1 (Linear-kernel identification under a linear Gaussian SCM)**.** *Suppose Assumptions 1, 2, and 3 hold. Under the linear Gaussian SCM, linear-kernel KRD correctly identifies the true causal direction $u \to v$ in finite samples with probability at least $1-\gamma$, provided that the OOC perturbation magnitude satisfies*

$$|\delta| > \frac{2(1+\varepsilon_{\mathrm{OLS}})\varepsilon_N}{\phi(t) - 2\varepsilon_{\mathrm{OLS}}}.$$

Furthermore, a sufficient IC sample size n_0 ensuring $\phi(t) > 2\varepsilon_{\mathrm{OLS}}$ is

$$n_0 > 1 + \left(\sqrt{\log(8/\gamma)} + \sqrt{\log(8/\gamma) + 8\log(16/\gamma)/\phi(t)^2}\right)^2.$$

Here, $\varepsilon_{\mathrm{OLS}} = \sqrt{2\log(16/\gamma)/L}$, $\varepsilon_N = \sqrt{2\log(16/\gamma)(n_0^{-1} + n_1^{-1})}$, *and* $L = n_0 - 1 - 2\sqrt{(n_0-1)\log(8/\gamma)}$.

Proof. By Theorem 2, linear-kernel KRD correctly identifies the true causal direction $u \to v$ with probability at least $1-\gamma$ provided that the population score separation exceeds twice the total error bound: $\tau_{\mathrm{KRD}}(v \to u) > 2E_{N,\gamma}$. For the linear kernel with standardized coordinates ($\sigma_0 = 1$), the total error bound is $E_{N,\gamma} = 2\varepsilon_{\mathrm{fit}}(n_0, n_1, \gamma/2) + \varepsilon_N$, where $\varepsilon_N = \sqrt{2\log(16/\gamma)(n_0^{-1} + n_1^{-1})}$.

To bound the residual fitting error $\varepsilon_{\mathrm{fit}}$, we first analyze the Ordinary Least Squares (OLS) estimator $\widehat{a}$ on n_0 in-distribution samples ($e = 0$), where $x_v = ax_u + \eta_v$ with $x_u \sim \mathcal{N}(0,1)$ and $\eta_v \sim \mathcal{N}(0, \sigma_v^2)$. The coefficient estimation error takes the form $\widehat{a} - a = \frac{\sum_{i=1}^{n_0} x_{u,i}^{(0)} \eta_{v,i}^{(0)}}{\sum_{i=1}^{n_0} (x_{u,i}^{(0)})^2}$. Conditioned on the design vector $\mathbf{x}_u^{(0)} = (x_{u,1}^{(0)}, \ldots, x_{u,n_0}^{(0)})^\top$, the estimation error follows a Gaussian distribution $\mathcal{N}\left(0, \frac{\sigma_v^2}{\|\mathbf{x}_u^{(0)}\|_2^2}\right)$. Since $\|\mathbf{x}_u^{(0)}\|_2^2 \sim \chi^2_{n_0-1}$, Laurent–Massart 1-sided concentration (Laurent & Massart, 2000) for χ^2 random variables guarantees that $\mathbb{P}\left(\|\mathbf{x}_u^{(0)}\|_2^2 \le L\right) \le \frac{\gamma}{8}$, where $L = n_0 - 1 - 2\sqrt{(n_0-1)\log(8/\gamma)}$. Conditioned on $\|\mathbf{x}_u^{(0)}\|_2^2 \ge L$, standard Gaussian tail bounds yield $\mathbb{P}\left(\frac{|\widehat{a}-a|}{\sigma_v} > \sqrt{\frac{2\log(16/\gamma)}{L}}\right) \le \frac{\gamma}{8}$. Taking a union bound over these two events ensures that $\frac{|\widehat{a}-a|}{\sigma_v} \le \varepsilon_{\mathrm{OLS}} := \sqrt{\frac{2\log(16/\gamma)}{L}}$ holds with probability at least $1 - \gamma/4$.

We next translate the coefficient estimation error $\widehat{a} - a$ into the coordinate-wise standardized residual deviation. The ideal residual $z_{v|u} = \frac{x_v - ax_u}{\sigma_v}$ and the estimated residual $\widehat{z}_{v|u} = \frac{x_v - \widehat{a}x_u}{\sigma_v}$ satisfy $|\widehat{z}_{v|u,a}^{(e)} - z_{v|u,a}^{(e)}| = \frac{|\widehat{a}-a|}{\sigma_v}|x_{u,a}^{(e)}| \le \varepsilon_{\mathrm{OLS}}|x_{u,a}^{(e)}|$ for any sample point a in environment $e \in \{0,1\}$. Summing across both environments, the average residual fitting error is bounded by:

$$\begin{aligned} 2\varepsilon_{\mathrm{fit}}(n_0, n_1, \gamma/2) &= \frac{1}{n_0}\sum_{a=1}^{n_0} |\widehat{z}_a^{(0)} - z_a^{(0)}| + \frac{1}{n_1}\sum_{a=1}^{n_1} |\widehat{z}_a^{(1)} - z_a^{(1)}| \\ &\le \varepsilon_{\mathrm{OLS}} \left(\frac{1}{n_0}\sum_{a=1}^{n_0} |x_{u,a}^{(0)}| + \frac{1}{n_1}\sum_{a=1}^{n_1} |x_{u,a}^{(1)}|\right). \end{aligned}$$

In the OOC environment ($e = 1$), the intervention shift magnitude $|\delta|$ implies $|x_{u,a}^{(1)}| \le |\delta| + |x_{u,a}^{(0)}|$ by the triangle inequality. Accounting for the finite-sample deviation of standard Gaussian

variables from their unit second-moment under concentration, the sum of sample averages satisfies $\frac{1}{n_0}\sum_{a=1}^{n_0}|x_{u,a}^{(0)}| + \frac{1}{n_1}\sum_{a=1}^{n_1}|x_{u,a}^{(1)}| \le |\delta| + \varepsilon_N$. Consequently, the average residual fitting error simplifies to $2\varepsilon_{\text{fit}}(n_0, n_1, \gamma/2) \le \varepsilon_{\text{OLS}}|\delta| + \varepsilon_{\text{OLS}}\varepsilon_N$.

Substituting this bound into $2E_{N,\gamma}$ yields:

$$2E_{N,\gamma} = 4\varepsilon_{\text{fit}}(n_0, n_1, \gamma/2) + 2\varepsilon_N \le 2\varepsilon_{\text{OLS}}|\delta| + 2(1+\varepsilon_{\text{OLS}})\varepsilon_N.$$

Combining this total error with the population score separation $\tau_{\text{KRD}}(v \to u) = |\delta|\phi(t)$ from equation 6, the finite-sample identification criterion $\tau_{\text{KRD}}(v \to u) > 2E_{N,\gamma}$ holds whenever:

$$|\delta|\phi(t) > 2\varepsilon_{\text{OLS}}|\delta| + 2(1+\varepsilon_{\text{OLS}})\varepsilon_N \iff |\delta|\left(\phi(t) - 2\varepsilon_{\text{OLS}}\right) > 2(1+\varepsilon_{\text{OLS}})\varepsilon_N.$$

Assuming $\phi(t) > 2\varepsilon_{\text{OLS}}$, dividing both sides by $\phi(t) - 2\varepsilon_{\text{OLS}}$ directly yields the required minimum OOC perturbation magnitude:

$$|\delta| > \frac{2(1+\varepsilon_{\text{OLS}})\varepsilon_N}{\phi(t) - 2\varepsilon_{\text{OLS}}}.$$

To guarantee that the denominator is strictly positive, the in-distribution sample size n_0 must satisfy $\varepsilon_{\text{OLS}} < \frac{1}{2}\phi(t)$. Substituting the definition $\varepsilon_{\text{OLS}} = \sqrt{\frac{2\log(16/\gamma)}{L}}$, this condition becomes:

$$\sqrt{\frac{2\log(16/\gamma)}{L}} < \frac{\phi(t)}{2} \iff L > \frac{8\log(16/\gamma)}{\phi(t)^2}.$$

Expanding $L = (n_0 - 1) - 2\sqrt{n_0 - 1}\sqrt{\log(8/\gamma)}$, we obtain a quadratic inequality in $\sqrt{n_0 - 1}$:

$$(\sqrt{n_0 - 1})^2 - 2\sqrt{\log(8/\gamma)}\sqrt{n_0 - 1} - \frac{8\log(16/\gamma)}{\phi(t)^2} > 0.$$

Solving this quadratic inequality for the positive root $\sqrt{n_0 - 1}$ gives:

$$\sqrt{n_0 - 1} > \sqrt{\log(8/\gamma)} + \sqrt{\log(8/\gamma) + \frac{8\log(16/\gamma)}{\phi(t)^2}}.$$

Squaring both sides and adding 1 completes the lower bound on n_0:

$$n_0 > 1 + \left(\sqrt{\log(8/\gamma)} + \sqrt{\log(8/\gamma) + \frac{8\log(16/\gamma)}{\phi(t)^2}}\right)^2,$$

which completes the proof. □

B Experiments

B.1 Simulation Settings and Baseline Implementations

Baseline implementations The baseline implementations are taken from, or adapted directly from, the public code accompanying the cited studies. We retain the original parameter settings provided by the authors and do not tune them on our evaluation data. We consider **ANM-HSIC** (Hoyer et al., 2008), **ANM-Entropy** (Nowzohour & Bühlmann, 2016), and **ANM-MML** (Stegle et al., 2010) from the additive-noise family, **IGCI** (Daniusis et al., 2012) from the information-geometric family, and the regression-based **RECI** (Blöbaum et al., 2018). To provide a fair comparison when both IC and OOC observations are available, each single-context method is evaluated in two ways. The **pooled** version combines the IC and OOC observations into one sample, giving the method access to all available data while discarding the context labels. The **split** version applies the same method independently to IC and OOC and averages the two directional scores before selecting the causal direction, thereby using both samples without converting the method into a cross-context procedure. We also evaluate methods designed to use environment labels directly, including **MSS** (Perry et al., 2022), **UT-IGSP** (Squires et al., 2020), and **CD-NOD** (Huang et al., 2020). For MSS, the **hard** variant thresholds the mechanism-change tests and may abstain when the two directions have the same number of detected changes, whereas the **soft** variant retains the continuous test evidence

and therefore provides substantially greater decision coverage. Unless explicitly stated otherwise, **MSS** in the main text refers to this soft variant.

Simulation settings Table 2 summarizes the data-generation and evaluation parameters used in the controlled and extended synthetic experiments. The controlled generator supports a broader parameter grid, but the reported figures use only the configurations specified below.

Table 2: Parameter settings for the controlled and extended synthetic experiments.

Component	Setting
Controlled simulations	
Structural settings	Linear mechanisms with Gaussian errors and nonlinear mechanisms with non-Gaussian effect errors.
Reference model	$\sigma_u = 1$ and $a = 1$; each nonlinear mechanism is normalized to have zero mean and unit variance under the IC cause distribution.
Effect-noise scale	$\sigma_v \in \{0.25, 0.5, 1, 2, 4\}$.
OOC location	Cause and effect shifts are evaluated separately.
Shift magnitude	The generator supports $\delta \in \{0.25, 0.5, 1, 2, 4\}$; the reported figures use $\delta \in \{1, 2, 4\}$.
Sample sizes	$n_0, n_1 \in \{100, 500\}$.
Repetitions	100 trials per configuration with master seed 2027.
Extended simulations	
Structural settings	18 settings from five benchmark families, covering additive, heteroscedastic, multiplicative, confounded, neural-network, conditional-flow, and differential mechanisms.
OOC location and magnitude	Cause and effect shifts are evaluated separately at $\delta \in \{1, 2\}$.
Sample sizes	$n_0 = n_1 = 100$.
Repetitions	100 trials for each setting, perturbation location, and magnitude.

B.2 Controlled Synthetic Experiments under Different IC–OOC Sample-Size Allocations

Controlled synthetic data generation For each repetition, we first generate a causal pair following

$$x_u = \eta_u, \qquad x_v = a\, g(x_u) + \eta_v,$$

where $\eta_u \sim \mathcal{N}(0, \sigma_u^2)$ and $\eta_v \sim \mathcal{N}(0, \sigma_v^2)$ are independent. We set $a = 1$ and $\sigma_u = 1$, while varying σ_v to control the noise level. In the nonlinear-non-Gaussian setting, the effect noise is generated through a nonlinear monotone transformation of an auxiliary Gaussian variable. Specifically, we draw a smooth random function F using 256 random Fourier features and define $h'(x) = \exp\{F(x)\}$, so that $h(x) = \int_{-8}^{x} \exp\{F(t)\}\, dt$ is strictly increasing. For $\xi \sim \mathcal{N}(0, 1)$, the transformed variable $h(\xi)$ generally follows a non-Gaussian distribution with trial-specific skewness and kurtosis. We center and scale $h(\xi)$ to unit variance and multiply it by σ_v to obtain the effect noise η_v. A new transformation is sampled for each trial and kept fixed across its corresponding IC and OOC environments.

The IC and OOC samples are drawn independently but share the same function g and baseline noise distributions. IC observations are generated directly from the model above. For each OOC sample, the perturbation location is randomly assigned to either the cause or the effect. A cause shift is generated by

$$x_u = \eta_u + \delta\sigma_u, \qquad x_v = a\, g(x_u) + \eta_v,$$

whereas an effect shift is generated by

$$x_u = \eta_u, \qquad x_v = a\, g(x_u) + \eta_v + \delta\sigma_v.$$

Thus, the OOC environment changes exactly one exogenous component while preserving the causal mechanism and the other noise distribution. Finally, the two observed columns are randomly permuted with equal probability, making the ground-truth direction $x_u \to x_v$ or $x_v \to x_u$ equally likely in the observed data. Under this parameterization, the nominal signal-to-noise ratio in the linear setting is $|a|\sigma_u/\sigma_v = 1/\sigma_v$.

Results under different sample sizes Figure 3 reports the main-text setting $(n_{\mathrm{IC}}, n_{\mathrm{OOC}}) = (100, 100)$. Figures 5, 6, and 7 show $(100, 500)$, $(500, 100)$, and $(500, 500)$, respectively. In each figure, rows distinguish the two mechanism/noise settings, columns distinguish perturbation magnitudes, and the horizontal axis shows the signal-to-noise ratio.

The IC and OOC sample sizes affect different parts of KRD estimation. IC observations determine the candidate regressions, residual centers, scales, and reference distributions, whereas OOC observations determine how accurately the shifted residual distributions are measured. In the linear-Gaussian setting, Figures 5 and 6 show that increasing either sample size to 500 yields consistently high identification accuracy across $t \in [0.25, 4]$ when $\delta = 2$ or 4. When both sample sizes are increased to 500, as shown in Figure 7, reliable identification extends to $\delta = 1$. The nonlinear-non-Gaussian setting shows a similar improvement. These results are consistent with Theorem 2: larger samples reduce the error bound $E_{N,\gamma}$, while stronger perturbations generally increase the population separation $\tau_{\mathrm{KRD}}(v \to u)$ in the examined settings.

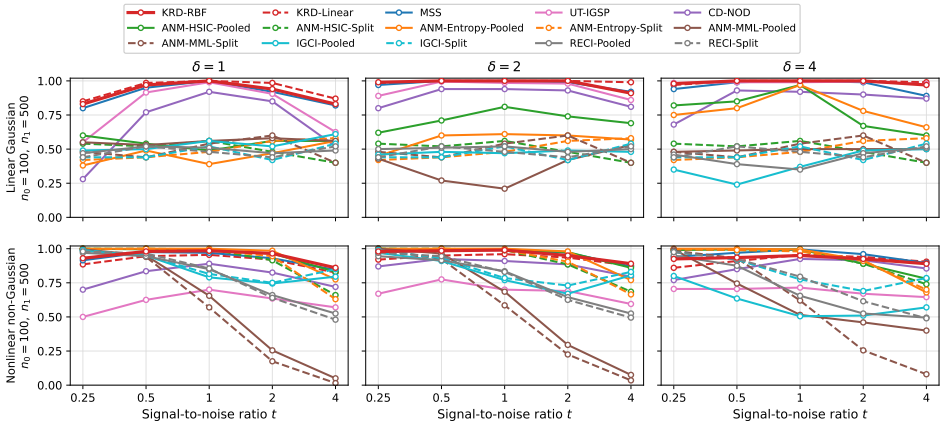

Figure 5: Identification accuracy with $n_0 = 100$ and $n_1 = 500$ under controlled simulation settings. The top row shows linear causal relationships with Gaussian errors, while the bottom row shows nonlinear causal relationships with non-Gaussian errors. Columns correspond to different OOC perturbation magnitudes, with results summarized across cause and effect perturbations.

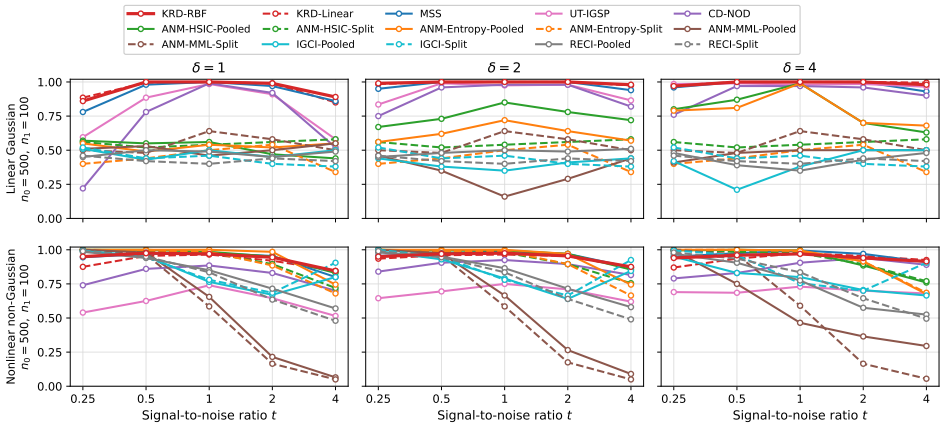

Figure 6: Identification accuracy with $n_0 = 500$ and $n_1 = 100$ under controlled simulation settings. The top row shows linear causal relationships with Gaussian errors, while the bottom row shows nonlinear causal relationships with non-Gaussian errors. Columns correspond to different OOC perturbation magnitudes, with results summarized across cause and effect perturbations.

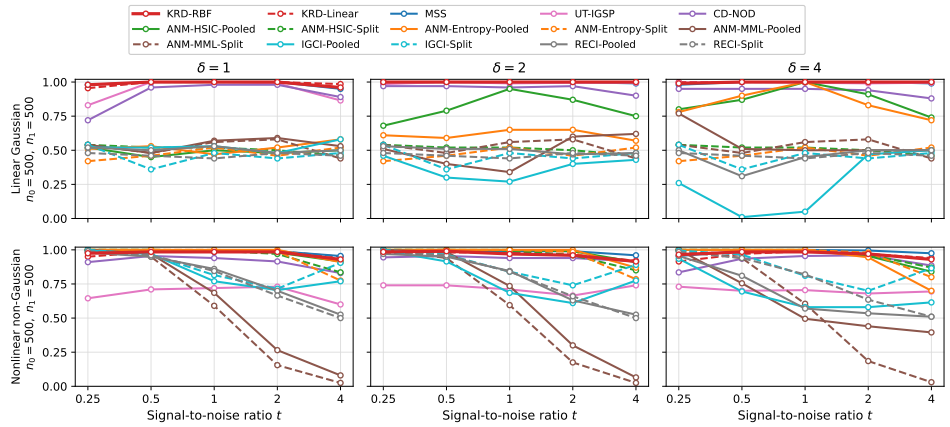

Figure 7: Identification accuracy with $n_0 = 500$ and $n_1 = 500$ under controlled simulation settings. The top row shows linear causal relationships with Gaussian errors, while the bottom row shows nonlinear causal relationships with non-Gaussian errors. Columns correspond to different OOC perturbation magnitudes, with results summarized across cause and effect perturbations.

B.3 Extended Synthetic Experiments across OOC Shift Locations and Magnitudes

Benchmarks We further evaluate KRD on 18 simulation cases adapted from five benchmark families summarized in Table 3. The first family examines different structural-noise relationships (Tagasovska et al., 2020). AN and AN-c use additive noise with Gaussian-process and sigmoid mechanisms, respectively; LS and LS-c introduce input-dependent noise scales and are therefore heteroscedastic; and MN-U uses multiplicative uniform noise. The second family consists of the nonlinear cause–effect models of Mooij et al. (2016), which combine randomly generated nonlinear mechanisms, non-Gaussian exogenous variables, and measurement noise. SIM-c additionally includes a latent common cause, SIM-G uses a Gaussian-like source distribution, and SIM-ln represents a low-measurement-noise setting. The Multi and Net environments (Guyon et al., 2019) introduce further variation in the causal mechanism and noise placement. Multi combines linear or quadratic mechanisms with additive or multiplicative noise introduced before or after the mechanism, whereas Net uses a neural-network mechanism with non-Gaussian inputs. The Per, Sig, and Vex environments (Xi et al., 2025) generate the effect through periodic, sigmoid, and velocity-field-based conditional flows. Finally, Qd-V, Sig-V, Rbf-V, and NN-V (Chen et al., 2026) define nonlinear causal mechanisms by integrating state-dependent velocity fields constructed from quadratic, sigmoid, radial-basis, and neural-network components, respectively, together with multimodal exogenous noise.

Across these environments, the IC and OOC data retain the same structural mechanism, while the OOC perturbation shifts one exogenous variable. Only AN and AN-c are closest to the additive noise model considered in our theory. The remaining settings introduce heteroscedastic or multiplicative noise, latent confounding, non-additive noise placement, conditional flows, or state-dependent differential mechanisms. They therefore evaluate whether KRD remains effective beyond its theoretical setting.

OOC shift locations and magnitudes We evaluate cause and effect shifts separately at two perturbation magnitudes, $\delta \in \{1, 2\}$, resulting in four experimental conditions: cause shift with $\delta = 1$, cause shift with $\delta = 2$, effect shift with $\delta = 1$, and effect shift with $\delta = 2$. This design allows us to examine how the shift target and magnitude affect identification accuracy without averaging over different perturbation locations. Table 2 reports the IC and OOC sample sizes and the number of repetitions used in each condition.

Results At $\delta = 1$, the cause-shift results are reported in Figure 8 and Table 4, while the effect-shift results are reported in Figure 9 and Table 5. The corresponding results for $\delta = 2$ appear in Figure 10 and Table 6 for cause shifts, and in Figure 11 and Table 7 for effect shifts. Each figure presents

Table 3: The 18 extended simulation cases from five benchmark families.

Source	Cases	System characteristics
Tagasovska et al. (2020)	AN, AN-c LS, LS-c MN-U	Additive-noise models with GP or sigmoid mechanisms; heteroscedastic models with input-dependent noise scales; a multiplicative-noise model.
Mooij et al. (2016)	SIM, SIM-c, SIM-G, SIM-ln	Random nonlinear mechanisms with non-Gaussian variables; variants with latent confounding, a Gaussian-like source, or low measurement noise.
Guyon et al. (2019)	Multi, Net	Linear or nonlinear mechanisms with different noise locations and additive or multiplicative noise; a neural-network mechanism with non-Gaussian variables.
Xi et al. (2025)	Per, Sig, Vex	Periodic, sigmoid, and velocity-field-based conditional mechanisms.
Chen et al. (2026)	Qd-V, Sig-V, Rbf-V, NN-V	State-dependent differential mechanisms constructed from quadratic, sigmoid, RBF, or neural-network components, with multimodal noise.

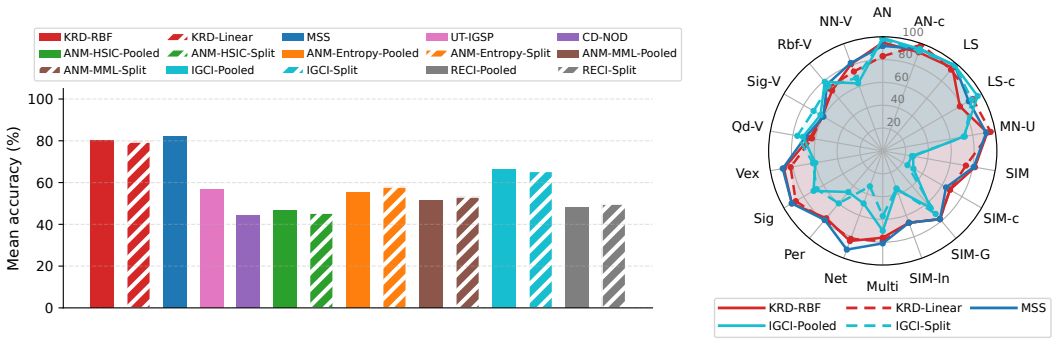

Figure 8: Identification accuracy under OOC perturbations of magnitude $\delta = 1$, applied only to the *cause*. Left: average accuracy across all the 18 extended settings. Right: per-setting accuracy of the top five methods.

the mean accuracy together with selected setting-specific results, whereas the accompanying table includes all evaluated methods.

Across the four conditions, KRD-RBF and MSS (soft) generally perform better than the other cross-context and single-context baselines, whose accuracy varies considerably across simulation settings. Although KRD-RBF and MSS achieve comparable overall performance, their relative advantages depend on the location of the OOC shift. For effect shifts, KRD-RBF achieves higher mean accuracy than MSS, with 79.78% versus 78.22% at $\delta = 1$ and 88.00% versus 84.67% at $\delta = 2$. For cause shifts, MSS performs better, with 81.89% versus 80.06% at $\delta = 1$ and 90.28% versus 84.00% at $\delta = 2$.

This difference reflects how the two methods use environmental changes. Under a cause shift, the root marginal changes directly while the forward conditional mechanism remains invariant, producing a particularly clear sparse-mechanism pattern for MSS. Under an effect shift, the change appears directly in the conditional residual while the root residual remains invariant, which is well matched to the residual-displacement representation used by KRD. Cause shifts can also be propagated or partially absorbed by nonlinear mechanisms and reverse regression, weakening the reverse residual displacement on which KRD relies. Increasing δ strengthens the observable distributional change and improves the mean accuracy of both methods for both shift locations.

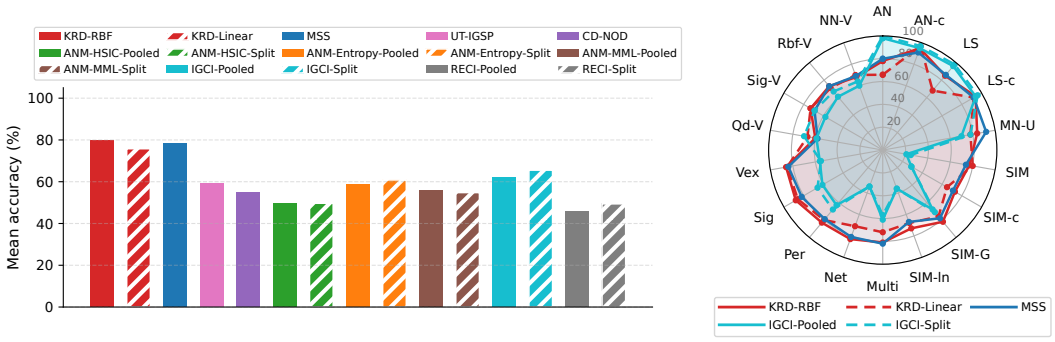

Figure 9: Identification accuracy under OOC perturbations of magnitude $\delta = 1$, applied only to the *effect*. Left: average accuracy across all the 18 extended settings. Right: per-setting accuracy of the top five methods.

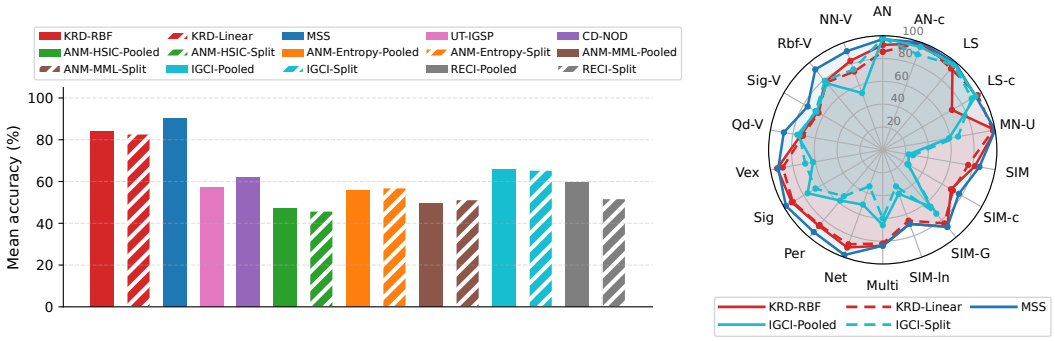

Figure 10: Identification accuracy under OOC perturbations of magnitude $\delta = 2$, applied only to the *cause*. Left: average accuracy across all the 18 extended settings. Right: per-setting accuracy of the top five methods.

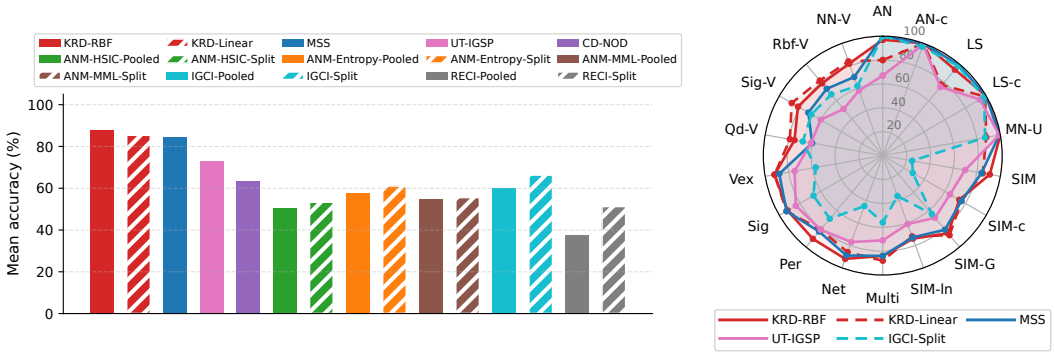

Figure 11: Identification accuracy under OOC perturbations of magnitude $\delta = 2$, applied only to the *effect*. Left: average accuracy across all the 18 extended settings. Right: per-setting accuracy of the top five methods.

Table 4: Causal direction identification accuracy (%) over 100 simulations under cause OOC shifts with shift magnitude $\delta = 1$ (IC standard deviation of the upstream exogenous noise) and sample sizes $n_0 = n_1 = 100$. Avg. denotes mean accuracy across the 18 settings, and Time is the mean runtime (ms) per pair. For each accuracy column, the best and second-best distinct results are bolded and underlined; ties receive the same formatting.

Context	Setting	Method	Avg.	Time/pair (ms)	AN	AN-c	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln	Multi	Net	Per	Sig	Vex	Qd-V	Sig-V	Rbf-V	NN-V
Cross-context	–	Ours (KRD-RBF)	80.06	19.51	95	92	95	78	93	**82**	68	**78**	67	76	84	77	**92**	88	65	61	69	81
		Ours (KRD-Linear)	79.44	7.45	83	99	93	91	**96**	74	68	**78**	67	79	82	**79**	88	82	63	61	73	74
		MSS (hard)	42.00	47.94	60	68	59	45	64	44	28	34	22	43	60	41	58	55	9	16	21	29
		MSS (soft)	**81.89**	47.94	92	95	98	87	92	81	64	**78**	67	**81**	**92**	**79**	**92**	**89**	69	60	76	**82**
		CD-NOD	44.33	46.24	63	72	70	47	62	47	27	41	24	45	63	45	62	53	8	17	23	29
		UT-IGSP	56.56	0.91	44	55	45	51	71	60	54	61	53	50	58	66	71	67	47	63	53	49
Single-context	Pooled	ANM-HSIC	46.61	79.49	99	98	50	17	12	74	74	62	72	41	43	47	15	16	32	47	20	20
		ANM-Entropy	55.28	72.47	**100**	94	97	17	47	78	**80**	57	78	43	48	45	22	28	43	39	34	45
		ANM-MML	51.44	107.12	**100**	**100**	**99**	36	86	66	67	40	**89**	36	49	20	23	25	29	28	16	17
		IGCI	66.17	0.51	99	95	98	**96**	72	27	31	65	36	70	49	47	70	60	71	63	**79**	63
		RECI	48.28	1.57	63	54	55	25	45	62	67	54	66	69	11	0	33	34	63	54	59	55
	Split	ANM-HSIC	45.39	62.17	**100**	96	44	11	6	75	67	52	72	36	48	37	14	20	42	41	33	23
		ANM-Entropy	57.89	55.26	92	86	91	7	55	73	69	59	77	46	56	47	31	43	52	48	54	56
		ANM-MML	53.11	116.03	**100**	**100**	**99**	32	84	66	68	40	86	40	58	24	24	18	37	36	23	21
		IGCI	65.39	0.56	98	93	98	90	73	26	25	72	35	57	33	60	67	62	**76**	**70**	74	68
		RECI	49.72	2.57	83	70	70	25	41	52	69	54	69	71	12	2	36	25	55	54	50	57

Table 5: Causal direction identification accuracy (%) over 100 simulations under effect OOC shifts with shift magnitude $\delta = 1$ (IC standard deviation of the downstream exogenous noise) and sample sizes $n_0 = n_1 = 100$. Avg. denotes mean accuracy across the 18 settings, and Time is the mean runtime (ms) per pair. For each accuracy column, the best and second-best distinct results are bolded and underlined; ties receive the same formatting.

Context	Setting	Method	Avg.	Time/pair (ms)	AN	AN-c	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln	Multi	Net	Per	Sig	Vex	Qd-V	Sig-V	Rbf-V	NN-V
Cross-context	–	Ours (KRD-RBF)	**79.78**	16.86	78	95	85	94	84	77	73	**82**	73	81	**83**	**83**	**88**	**86**	60	**73**	**73**	68
		Ours (KRD-Linear)	75.83	8.57	66	97	68	92	78	80	65	76	68	72	71	80	86	85	67	**73**	71	**70**
		MSS (hard)	31.72	49.77	28	59	33	36	65	30	23	30	18	27	32	31	47	43	10	19	20	20
		MSS (soft)	78.22	49.77	80	91	86	92	92	74	71	78	67	**82**	81	79	82	84	58	69	**73**	69
		CD-NOD	54.78	27.57	66	75	58	53	93	53	38	51	23	43	66	82	75	69	21	31	35	54
		UT-IGSP	59.22	0.88	49	63	49	52	72	60	60	68	55	58	63	63	70	67	48	62	54	53
Single-context	Pooled	ANM-HSIC	49.78	72.53	**100**	96	50	13	17	80	74	46	81	39	58	52	24	21	35	49	37	24
		ANM-Entropy	58.61	69.39	**100**	95	96	15	60	**83**	**81**	50	81	52	55	50	29	30	39	54	44	41
		ANM-MML	55.94	97.29	**100**	**100**	97	41	97	73	72	46	**86**	43	60	9	38	21	39	31	29	25
		IGCI	62.22	0.45	98	95	96	94	70	21	29	68	36	61	34	63	61	56	59	58	61	60
		RECI	45.61	1.42	72	45	57	14	56	59	69	51	66	73	11	1	30	27	56	43	40	51
	Split	ANM-HSIC	49.72	54.53	99	97	53	21	24	74	68	51	77	34	54	48	25	27	37	44	34	28
		ANM-Entropy	60.89	53.02	96	92	95	23	75	79	72	52	80	47	54	56	28	47	55	50	45	50
		ANM-MML	54.83	101.84	**100**	**100**	**99**	46	**99**	66	69	42	82	37	59	15	40	27	29	34	19	24
		IGCI	65.56	0.49	**100**	97	98	**96**	78	25	29	71	36	57	34	68	66	55	**70**	69	67	64
		RECI	49.44	2.31	78	62	63	25	69	56	67	51	71	70	12	2	30	24	54	54	46	56

B.4 Data-derived IC/OOC Partition for Real-data Experiments

When environment labels are unavailable, we assume that IC observations form the dominant continuous relationship between the two variables. We first standardize both variables and fit a robust smooth curve $\widehat{h}$ after ordering the observations by the first supplied coordinate. The vertical residual

$$r_i = \widetilde{y}_i - \widehat{h}(\widetilde{x}_i)$$

measures how far each observation lies from this relationship. Huber regression is used to reduce the influence of large deviations on the fitted curve (Huber, 1992).

Because the residual variation may differ across the input range, we use RBF-kernel change-point detection to divide the ordered observations into regions with similar residual dispersion (Arlot et al., 2019; Truong et al., 2020). A robust local band is then constructed around the fitted curve in each region. Observations inside the band are treated as preliminary IC candidates, while those far from the curve are treated as potential OOC observations.

The preliminary IC candidates may still form several disconnected groups along the first coordinate. We therefore split them at large gaps and retain the largest continuous group. Let L and U denote the smallest and largest standardized first-coordinate values in this group. The final partition is

$$\mathcal{I}_{\mathrm{IC}} = \{i : \text{observation } i \text{ lies inside its local band and } L \le \widetilde{x}_i \le U\}, \qquad \mathcal{I}_{\mathrm{OOC}} = \{1, \dots, n\} \backslash \mathcal{I}_{\mathrm{IC}}.$$

Thus, observations are assigned to OOC when they lie far from the local reference curve or outside the main continuous operating range.

In all experiments, the reference curve is a cubic B-spline with four knots fitted by Huber regression with threshold 1.35. Kernel segmentation uses a minimum segment size of 20 and a penalty of

Table 6: Causal direction identification accuracy (%) over 100 simulations under cause OOC shifts with shift magnitude $\delta = 2$ (2× IC standard deviation of the upstream exogenous noise) and sample sizes $n_0 = n_1 = 100$. Avg. denotes mean accuracy across the 18 settings, and Time is the mean runtime (ms) per pair. For each accuracy column, the best and second-best distinct results are bolded and underlined; ties receive the same formatting.

Context	Setting	Method	Avg.	Time	AN	AN-c	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln	Multi	Net	Per	Sig	Vex	Qd-V	Sig-V	Rbf-V	NN-V
Cross-context	–	Ours (KRD-RBF)	84.00	19.46	92	99	95	70	**100**	82	69	**88**	69	**84**	91	87	92	**94**	73	66	78	83
		Ours (KRD-Linear)	82.89	8.35	86	**100**	93	**96**	98	76	71	84	66	82	88	86	91	89	71	65	77	73
		MSS (hard)	65.67	46.11	86	94	87	66	94	55	36	61	22	61	85	74	85	80	41	40	50	65
		MSS (soft)	**90.28**	46.11	97	**100**	99	94	**100**	**86**	77	**88**	69	**84**	**98**	**94**	**98**	93	**88**	**76**	**92**	**92**
		CD-NOD	62.00	51.50	86	95	87	72	94	53	40	63	27	57	80	70	78	64	23	30	45	52
		UT-IGSP	57.39	0.90	40	47	47	47	60	65	59	59	56	52	58	75	67	67	59	59	59	57
Single-context	Pooled	ANM-HSIC	47.11	82.35	99	**100**	61	24	16	72	78	56	72	38	37	40	15	14	35	48	20	23
		ANM-Entropy	55.61	76.14	99	97	99	25	57	83	**81**	56	79	47	48	40	13	21	43	51	32	30
		ANM-MML	49.78	108.52	**100**	**100**	97	32	88	70	69	43	80	30	49	19	24	17	20	34	10	14
		IGCI	65.83	0.50	98	96	98	94	59	23	26	65	41	66	51	58	76	62	76	67	76	53
		RECI	59.61	1.53	73	63	65	35	53	74	77	61	72	76	18	5	55	45	87	74	72	68
	Split	ANM-HSIC	46.00	61.40	**100**	93	48	7	11	77	67	53	73	35	50	33	17	23	37	46	36	22
		ANM-Entropy	57.11	55.41	95	73	91	14	48	74	69	59	78	54	50	47	31	42	51	48	52	52
		ANM-MML	51.44	113.14	**100**	**100**	**100**	23	79	66	67	43	**81**	38	61	19	29	20	27	39	17	17
		IGCI	65.56	0.54	98	89	97	90	67	27	25	73	34	60	34	53	68	69	74	68	79	75
		RECI	51.94	2.49	77	90	75	29	28	59	70	54	75	70	12	2	46	31	57	53	52	55

Table 7: Causal direction identification accuracy (%) over 100 simulations under effect OOC shifts with shift magnitude $\delta = 2$ (2× IC standard deviation of the downstream exogenous noise) and sample sizes $n_0 = n_1 = 100$. Avg. denotes mean accuracy across the 18 settings, and Time is the mean runtime (ms) per pair. For each accuracy column, the best and second-best distinct results are bolded and underlined; ties receive the same formatting.

Context	Setting	Method	Avg.	Time	AN	AN-c	LS	LS-s	MN-U	SIM	SIM-c	SIM-G	SIM-ln	Multi	Net	Per	Sig	Vex	Qd-V	Sig-V	Rbf-V	NN-V
Cross-context	–	Ours (KRD-RBF)	**88.00**	19.60	97	**100**	94	**100**	**100**	**91**	74	85	74	84	**92**	**91**	92	**92**	75	82	79	82
		Ours (KRD-Linear)	85.39	8.20	80	**100**	77	99	88	85	76	**87**	72	**88**	86	81	**93**	**92**	**79**	**88**	**82**	**84**
		MSS (hard)	57.06	50.29	89	99	82	86	95	55	44	45	31	56	73	46	75	63	14	28	19	27
		MSS (soft)	84.67	50.29	**100**	**100**	**99**	**100**	99	84	76	81	73	84	89	83	**93**	88	60	72	73	70
		CD-NOD	63.56	32.02	93	93	87	87	94	62	53	57	34	65	82	86	78	62	24	25	23	39
		UT-IGSP	73.11	0.90	67	99	75	94	99	70	65	68	61	71	77	81	84	75	61	60	51	58
Single-context	Pooled	ANM-HSIC (GP)	50.72	82.30	**100**	95	47	6	18	77	**79**	48	83	35	58	65	25	24	43	47	36	27
		ANM-Entropy (GP)	57.89	73.84	99	88	93	6	71	84	**79**	48	82	46	58	54	34	34	44	48	34	40
		ANM-MML (GP)	55.06	109.36	**100**	**100**	98	35	96	70	76	46	**87**	40	52	4	28	26	36	42	23	32
		IGCI	59.94	0.51	92	85	93	92	70	21	26	70	29	57	37	65	56	54	55	64	61	52
		RECI	37.44	1.57	72	36	52	11	59	57	70	39	63	68	11	1	16	13	29	24	24	29
	Split	ANM-HSIC	53.33	60.80	99	97	53	32	42	73	69	52	79	41	53	54	32	24	40	46	38	36
		ANM-Entropy	61.11	53.49	96	92	95	28	88	75	69	49	80	46	54	55	34	45	48	49	49	48
		ANM-MML	55.56	114.39	**100**	**100**	**99**	49	**100**	68	69	43	84	41	61	10	47	23	25	36	23	22
		IGCI	66.33	0.56	**100**	97	98	98	87	25	29	64	36	56	45	69	67	57	68	69	67	62
		RECI	51.28	2.55	78	62	65	28	74	60	68	48	71	70	18	2	34	28	53	55	53	56

3. Within each segment, the local residual scale is estimated as 1.4826 times the median absolute residual, using the standard Gaussian calibration of a robust scale estimate (Rousseeuw & Croux, 1993). The band half-width is 1.8 times this scale, with a minimum width of 0.12, and preliminary IC groups are separated when their gap exceeds 15% of the full standardized input range. These parameters are fixed across all datasets and are not selected using causal labels. Figure 12 illustrates the resulting partitions for representative pairs from the Tübingen benchmark.

B.5 Computational Efficiency

Computational-complexity comparison Let $N = n_0 + n_1$, and treat the numbers of variables, environments, and hyperparameter candidates as fixed. In our nonlinear ANM baselines, both candidate directions are fitted using exact Gaussian process regression, whose kernel factorization requires $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory (Rasmussen, 2003). ANM-HSIC additionally constructs and centers two $N \times N$ Gram matrices, adding $\mathcal{O}(N^2)$ time and memory for the dependence score (Gretton et al., 2005); its overall complexity therefore remains $\mathcal{O}(N^3)$. ANM-Entropy applies a one-dimensional spacing-based entropy estimator to the input and fitted residuals, requiring $\mathcal{O}(N \log N)$ time after regression. ANM-MML evaluates the Gaussian process marginal likelihood and fits a fixed collection of univariate mixture models. In both cases, exact Gaussian process fitting remains the dominant operation, giving an overall complexity of $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory in our implementation. Splitting the observations by context replaces these costs by sums over IC and OOC, such as $\mathcal{O}(n_0^3 + n_1^3)$, without changing the worst-case order.

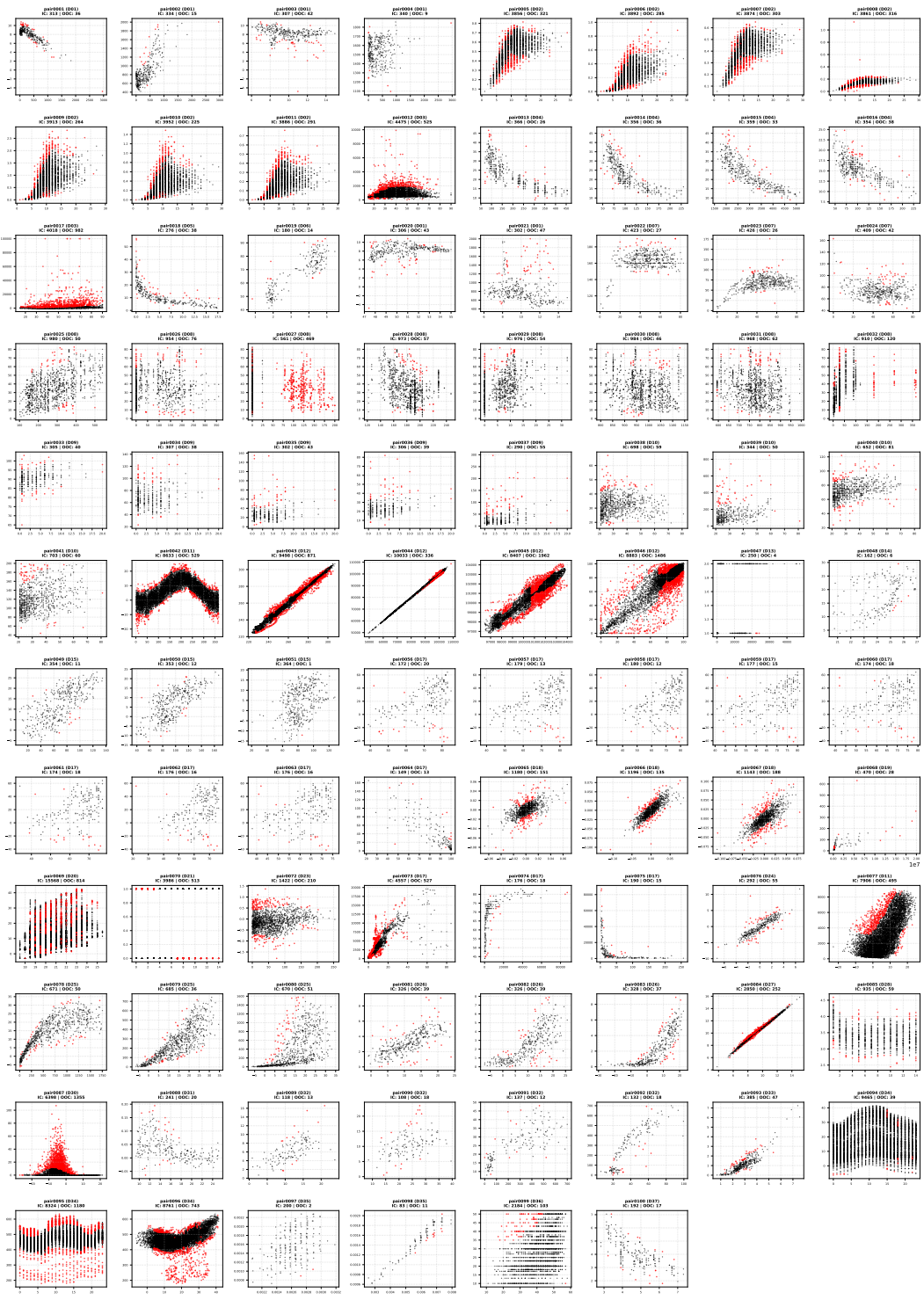

Figure 12: Examples of the data-derived IC/OOC partition on representative Tübingen cause–effect pairs. Black points denote observations assigned to the IC environment, while red points denote observations assigned to the OOC environment.

In our evaluation, MSS and CD-NOD use kernel conditional independence tests. Each KCI test operates on $N \times N$ kernel matrices and requires matrix factorization or eigendecomposition, resulting in $\mathcal{O}(N^3)$ time and $\mathcal{O}(N^2)$ memory (Zhang et al., 2012). Both methods invoke the test multiple times to compare candidate mechanisms or conditional independences. Because the number of such tests is fixed in the present bivariate setting, their overall sample-size complexity remains $\mathcal{O}(N^3)$, although the number of tests can increase in larger graphs.

KRD avoids conditional independence testing and its associated matrix decompositions. With the fixed-size regression specification used here, residual estimation requires $\mathcal{O}(N)$ time. KRD-Linear then compares residual means in $\mathcal{O}(N)$ time, whereas KRD-RBF evaluates the exact empirical MMD in $\mathcal{O}(N^2)$ time (Gretton et al., 2012). Blockwise kernel summation reduces the memory requirement of KRD-RBF from $\mathcal{O}(N^2)$ to $\mathcal{O}(N+B^2)$ for block size B. Consequently, KRD-Linear and KRD-RBF are respectively linear and quadratic in sample size, compared with the cubic-time kernel factorizations used by the evaluated Gaussian-process and KCI-based baselines.

Computation cost in the extended simulation settings The runtime panel in the main text examines how computation grows with sample size, while the extended-simulation tables 4, 5, 6, 7 report mean runtime per pair. At $\delta = 2$, KRD-RBF, KRD-Linear, and MSS (soft) require 19.60, 8.20, and 50.29 ms per pair under effect shifts, respectively. The corresponding cause runtimes are 19.46, 8.35, and 46.11 ms. Thus, both KRD variants run faster than MSS (soft) in these measurements, with the accuracy tradeoff depending on perturbation location.