CIDER-FM: FOUNDATION MODELS FOR CAUSAL INFERENCE FROM DIVERSE EXPERIMENTAL REGIMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal foundation models (CFMs) amortise causal inference over priors of synthetic structural causal models (SCMs), predicting the effect of an experiment on a specific variable. However, observational data alone may leave multiple causal models compatible with available evidence, while experimental data with interventions on exactly the variable of interest might be unavailable. This work studies CFMs as a method to combine finite observational and surrogate-interventional datasets in order to predict a target conditional interventional distribution (CID) more accurately than with observational data alone. We first formalise the conceptual benefits of surrogate experiments. Building on this analysis, we introduce FOUNDATION MODELS FOR CAUSAL INFERENCE FROM DIVERSE EXPERIMENTAL REGIMES (*CIDER-FM*), a causal foundation model that uses an intervention-aware representation and hierarchical three-axis attention to exchange information across variables, samples, and experimental regimes. We evaluate CIDER-FM against a wide range of baselines across diverse synthetic graph and mechanism families, as well as on both simulated and real-world data from Causal Chambers. Our results demonstrate strong CID prediction performance and show that incorporating experimental context can improve predictions over observational data alone.

1 INTRODUCTION

Causal questions are central to scientific inquiry and decision making. When a variable T can be experimentally manipulated and an outcome Y measured, the causal effect of T on Y can be estimated directly. When such a targeted experiment is infeasible and, for instance, only observational data can be collected, multiple causal models may explain the same observational distribution yet lead to different causal effects (Pearl, 2009; Peters et al., 2017). Therefore, traditional causal inference pairs explicit structural assumptions with bespoke estimators tailored to a particular identification strategy. However, explicit structural assumptions, such as knowledge of the entire set of causal relationships between all variables in a system, might not be justified, especially in complex scientific setups. Even worse, having to make an assumption without the necessary grounds can lead to wrong yet confident predictions.

Recently proposed Causal foundation models (CFMs) offer a complementary approach: pretrained on a distribution of synthetic SCMs, a single model takes as input observational data and amortises causal discovery followed by causal inference using a prior over causal structures (Robertson et al., 2026; Dhir et al., 2026). A CFM approximates the posterior predictive conditional interventional distribution (CID)

$$p(y \mid \mathrm{do}(T=t), \mathbf{x}, \mathcal{D}) = \int p\big(y \mid \mathbf{x}, \psi_{\mathrm{do}(T=t)}\big)\, p(\psi \mid \mathcal{D}, \boldsymbol{x})\, \mathrm{d}\psi, \tag{1}$$

where ψ denotes an SCM and $\mathcal{D}$ the available data. This posterior-predictive view allows uncertainty about both the graph and its mechanisms (from the implicit causal-discovery step) to be propagated into the causal prediction, rather than replaced immediately by a single fixed causal model (Dhir et al., 2024; 2026; Robertson et al., 2026). However, the cost of this approach is that, since existing CFMs only use observational data, a broad SCM prior leaves several causal explanations unresolved (Robertson et al., 2026; Dhir et al., 2026). Mathematically, this corresponds to an uncertain $p(\psi|\mathcal{D}_{obs}, \boldsymbol{x})$. Since the final CID integrates over the posterior belief over SCMs, using only an observational dataset

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.

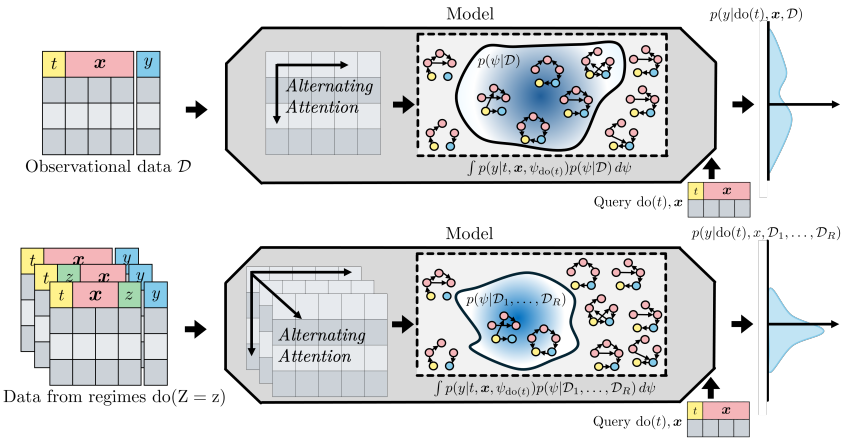


Figure 1: **Conceptual comparison:** observational data (top) and multi-regime experimental data (bottom) in Causal Foundation Models (CFMs). With observational data alone, multiple compatible SCMs may imply different CIDs, producing a broad or multimodal posterior predictive distribution. Adding data from different experimental regimes can exclude or downweight incompatible SCMs and thereby refine the predictive distribution. Compared to existing TFMs, the stacked tables in the lower panel make explicit a third structural axis, namely the different experimental regime, alongside variables and samples along which CIDER-FM performs the attention operation.

in $p(\psi|\mathcal{D}_{obs}, \boldsymbol{x})$ may lead the CID failing to concentrate and instead remaining diffuse or multimodal (the first panel in Fig. 1).

This work studies *interventional data as a complementary source of causal information in CFMs* to overcome this issue by reducing or even fully collapsing the posterior over SCMs compatible with a scenario, over which we need to amortise. More specifically, we replace $\mathcal{D}_{obs}$ by multiple datasets $\mathcal{D}_1, \ldots, \mathcal{D}_R$ that can be observational or interventional data from different experiments. In many scientific settings, observational data coexist with datasets collected under controlled interventions.

Direct experimentation on the treatment variable of interest T may be unavailable, while surrogate experiments $\mathrm{do}(Z = z)$ on other variables are feasible. This motivates exploiting multi-environment data from diverse experimental regimes, including observational data, data from surrogate experiments, and, when available, interventional data on the treatment variable T, to constrain the set of SCMs compatible with the available evidence (Hauser & Bühlmann, 2012; Kivva et al., 2023), as illustrated in the second panel of Fig. 1 and the case study in Fig. 2. Following the prior-data fitted network (PFN) paradigm of synthetic pretraining and in-context learning (Hollmann et al., 2025), we introduce FOUNDATION MODELS FOR CAUSAL INFERENCE FROM DIVERSE EXPERIMENTAL REGIMES (*CIDER-FM*).

Our contributions.

1. **Theoretical analysis.** We formalise how surrogate experiments can enable point identification or narrow identified sets, and provide complementary finite-sample and asymptotic analyses of how they can improve prediction accuracy.
2. **Model architecture.** CIDER-FM uses intervention-aware representations of continuous, row-specific randomised treatment assignments $\mathrm{do}(X = x_i)$ and applies hierarchical three-axis attention across variables, samples within regimes, and experimental regimes to enable in-context causal prediction from multi-environment data.
3. **Synthetic and real-world validation.** We demonstrate CIDER-FM's strong predictive performance across diverse synthetic graph and mechanism families compared with observation-only baselines under matched context budgets. On Causal Chambers, the same model supports simulator-guided experimental design and achieves strong causal prediction performance on the resulting real-world experiments.

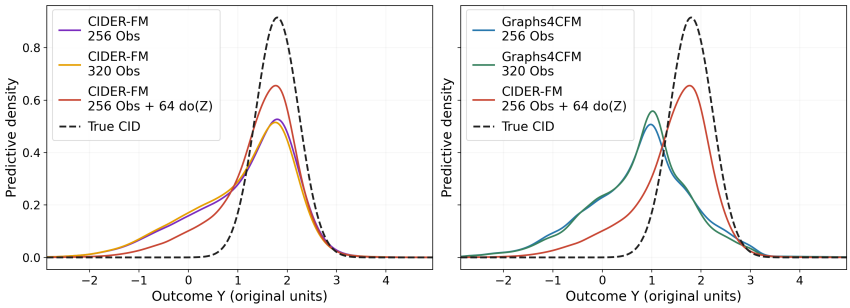


Figure 2: **Case study:** surrogate experiments sharpen conditional interventional predictions. Model predictive densities are visualised using Gaussian KDEs. Left: the same CIDER-FM checkpoint receives 256 observational samples, 320 observational samples, or 256 observational samples plus 64 randomized $\mathrm{do}(Z)$ samples. Right: comparison with observational-only Graphs4CFM (Reuter et al., 2026) without graph input. Details of the task construction and quantitative results are in Section I.

2 Background and Related Work

Structural causal models and interventions A structural causal model (SCM) $\psi = (G, \mathcal{M})$ consists of a causal graph G and mechanisms $\mathcal{M}$ specifying how each variable is generated from its parents and exogenous noise, together with the joint noise distribution (Pearl, 2009; Peters et al., 2017). A hard intervention $\mathrm{do}(T = t)$ replaces the structural assignment for T by $T := t$, removing incoming edges to T in G while leaving the other mechanisms unchanged. Within a specified model class, the available population distributions may be compatible with multiple SCMs. The graph is identifiable if the set of compatible graphs is a singleton. A causal query is identifiable if all these compatible SCMs agree on the target quantity, such as $p_\psi(y \mid \mathrm{do}(T = t), x)$.

Traditional causal methods. Established causal methods span several families. Graphical methods use adjustment criteria and do-calculus to derive causal estimands under an assumed graph (Pearl, 1995; 2009; Shpitser & Pearl, 2006); outcome-regression and propensity-score methods estimate effects through models of the outcome or treatment assignment (Robins, 1986; Rosenbaum & Rubin, 1983). Although these approaches can provide guarantees under their respective conditions, their application requires problem-specific assumptions and typically manual choices of adjustment sets and estimation models.

Tabular Foundation Models. An alternative to fitting a problem-specific estimator is to amortise inference across a distribution of tasks. Tabular foundation models, such as TabPFN (Hollmann et al., 2023; 2025) , are amortised inference methods in the form of neural processes (Garnelo et al., 2018a;b) that have revolutionised the domain of tabular machine learning, setting new benchmark records (Erickson et al., 2026). They work by pretraining on synthetic datasets and adapting to new tasks through in-context learning (ICL), without task-specific parameter updates (Müller et al., 2022; Hollmann et al., 2023; 2025). CIDER-FM, the model we propose, operates in the same synthetic-data training and in-context-learning paradigm as TFMs.

Causal Foundation Models. Causal foundation models (CFMs) extend the ideas behind synthetically trained TFMs to causal queries such as conditional interventional distributions. Existing CFMs broadly follow two strategies. Identifiable-prior CFMs restrict their pretraining priors to structural assumptions in which the target query is identifiable from observational data under specific causal assumptions. For instance, Balazadeh et al. (2026) propose CausalPFN, a CFM that achieves excellent predictive performance on causal effect estimation with binary treatments under unconfoundedness, while Ma et al. (2025) propose priors to train a CFM for varying causal scenarios, such as backdoor-adjustment or front-door adjustment. In contrast, posterior-averaging CFMs learn the posterior predictive distribution induced by a broader prior over SCMs, retaining uncertainty when observationally compatible SCMs imply different interventional distributions (Robertson et al., 2026; Dhir et al., 2026). Both strategies primarily use observational datasets as context, leaving open how to incorporate data from diverse experimental regimes.

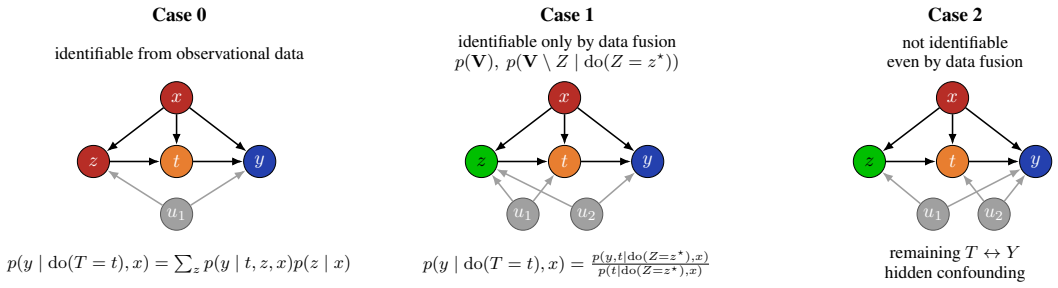


Figure 3: **Case studies:** Three roles of surrogate experiments for $p(y \mid \mathrm{do}(T=t), x)$. In Case 0, the CID is identifiable from observational data alone. In Case 1, it becomes identifiable after incorporating experiments on Z. In Case 2, it remains non-identifiable but with a potentially smaller identified set.

Recently, conditioning CFMs on partial graph information has been proposed as a way to bridge the gap between identifiable-prior and posterior-averaging CFMs (Reuter et al., 2026). Reliable causal graphs, however, are often unavailable in complex scientific setups. Accurately estimating them from data is challenging and treating them as trusted inputs can propagate graph-discovery errors into the final prediction. In this paper, we instead leverage data from multiple environments, not requiring any graphical assumptions.

3 Benefits of Surrogate Experiments

Causal inference from observational data is arguably a well-known and well-studied topic (Pearl, 2009; Rubin, 1974), and having access to experimental data, where the experiment has been performed on precisely the variable T whose effect is to be examined, remains mostly an estimation challenge, not a causal identification problem. In this section, we discuss the case of data from multiple surrogate experiments on different variables, which CIDER-FM takes as input.

Point identification. We first consider the infinite-data limit, in which all available distributions are known exactly. A target CID is point-identifiable if it is the same across all SCMs compatible with the available population distributions. Conditional generalised identification (c-gID) characterises whether such a conditional causal query can be uniquely recovered from an arbitrary collection of observational and interventional distributions under an assumed causal graph (Bareinboim & Pearl, 2012; Lee et al., 2020; Kivva et al., 2023). Fig. 3 illustrates three possible cases.

Partial identification. Failure of point identification does not imply that a surrogate experiment is uninformative. Different SCMs can imply the same observational distribution but different interventional distributions. Requiring agreement with the experimental distribution can therefore exclude some of these observationally compatible SCMs. A graphical illustration arises in DAG models without latent confounding: known perfect interventions can distinguish graphs within an observational Markov equivalence class (MEC), refining it to a potentially smaller interventional MEC. This is because removing incoming edges to the intervention targets can expose differences between otherwise observationally equivalent causal structures (Hauser & Bühlmann, 2012).

To formalise this principle, consider a specified class of SCMs, where ψ encodes both the graph and its mechanisms. Let $[\psi^\star]_O$ denote the SCMs compatible with the population observational distribution, and let $[\psi^\star]_{O+I}$ additionally require compatibility with the available surrogate interventional distribution. For $q_y(\psi) = p_\psi(y \mid \mathrm{do}(T=t), x)$, define the corresponding identified sets as $\mathcal{Q}_S = \{q_y(\psi) : \psi \in [\psi^\star]_S\}$. The additional compatibility requirement implies

$$[\psi^\star]_{O+I} \subseteq [\psi^\star]_O \quad \Longrightarrow \quad \mathcal{Q}_{O+I} \subseteq \mathcal{Q}_O. \tag{2}$$

Thus, at the population level, surrogate experiments can shrink the target's identified set or leave it unchanged. The inclusion may be strict even when $\mathcal{Q}_{O+I}$ is not a singleton.

Finite data. For fixed query values t, x, y, let $q_y(\psi) = p_\psi(y \mid \mathrm{do}(T = t), x)$, where ψ encodes both the graph and its mechanisms. The posterior predictive density (PPD) targeted by a CFM is

$$\widehat{q}_y(\mathcal{D}) = \int p_\psi(y \mid \mathrm{do}(T = t), x)\, p(\psi \mid \mathcal{D})\, d\psi.$$

Suppose that $\psi \sim p(\psi)$ and both datasets are generated from this same SCM under their respective regimes. For $S \in \{O, O + I\}$, define the prior-averaged mean squared error

$$R_S := \mathbb{E}_{\psi, \mathcal{D}_S}\left[\left(q_y(\psi) - \widehat{q}_y(\mathcal{D}_S)\right)^2\right],$$

where $\mathcal{D}_O = \mathcal{D}_{\mathrm{obs}}$ and $\mathcal{D}_{O+I} = (\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{int}})$. If $\mathbb{E}_\psi[q_y(\psi)^2] < \infty$, then

$$R_O - R_{O+I} = \mathbb{E}_{\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{int}}}\left[\left(\widehat{q}_y(\mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{int}}) - \widehat{q}_y(\mathcal{D}_{\mathrm{obs}})\right)^2\right] \geq 0. \tag{3}$$

Thus, incorporating surrogate-interventional data cannot increase the prior-averaged MSE between the exact PPD and the underlying SCM's target density. The reduction is strict whenever the additional data change the PPD at y with positive probability. See Appendix B for the proof. Please note the CFM is trained to approximate $\hat{q}$, which might introduces additional approximation error.

4 Input Representation and Model Architecture

CIDER-FM receives context datasets generated by the same unknown causal system under different experimental regimes. For each query row, it predicts a conditional interventional distribution for the masked outcome. Its architecture represents both how each row was generated and how evidence is exchanged within and across regimes. Appendix D provides details of the input representation, attention layers, and predictive distribution.

Intervention-aware input representation. A task contains R context datasets $\mathcal{D}_1, \ldots, \mathcal{D}_R$, where regime r provides data $\mathcal{D}_r = \{\mathbf{v}_{r,i}\}_{i=1}^{n_r}$ over the same p variables. Each regime has an intervention-target mask $\mathbf{a}_r \in \{0, 1\}^p$, where $a_{r,j} = 1$ indicates that variable j is intervened upon. Each row also has a vector of assigned intervention values $\mathbf{c}_{r,i} \in \mathbb{R}^p$, whose entries are relevant only for intervened variables. Together, $(\mathbf{a}_r, \mathbf{c}_{r,i})$ specifies which structural assignments were externally replaced and the values assigned in row i. Defining $\mathcal{I}_r = \{j : a_{r,j} = 1\}$, the row is generated according to

$$\mathbf{v}_{r,i} \sim p_\psi(\mathbf{V} \mid \mathrm{do}(\mathbf{V}_{\mathcal{I}_r} = \mathbf{c}_{r,i,\mathcal{I}_r})), \tag{4}$$

where ψ denotes the underlying causal system. The observational regime is represented by $\mathbf{a}_r = \mathbf{0}$, allowing observational and interventional data to share the same input format. CIDER-FM accommodates varying numbers of regimes and samples per regime.

CIDER-FM represents every table cell as a d-dimensional token. Each token combines information about the entire row, the individual cell value, the variable's identity and task role, the availability of its value, and its intervention metadata. Writing tildes for preprocessed data values and intervention assignments, the initial token for variable j in context row i of regime r is

$$\mathbf{h}^{C,0}_{r,i,j} = \alpha\, \phi_{\mathrm{row}}(\widetilde{\mathbf{v}}_{r,i}) + \beta\, \phi_{\mathrm{val}}(\widetilde{v}_{r,i,j}) + \mathbf{e}^{\mathrm{id}}_j + \mathbf{e}^{\mathrm{role}}_{\rho_j} + \mathbf{e}^{\mathrm{status}}_{\mathrm{obs}} + \mathbf{e}^{\mathrm{int}}(a_{r,j}, \widetilde{c}_{r,i,j}). \tag{5}$$

The function $\phi_{\mathrm{row}} : \mathbb{R}^p \to \mathbb{R}^d$ encodes the entire row, and its output is shared across all cell tokens in that row. The function $\phi_{\mathrm{val}} : \mathbb{R} \to \mathbb{R}^d$ encodes the individual cell value. The scalar coefficients α and β weight these two contributions.

The indexed $\mathbf{e}$ terms are learned embedding vectors selected by discrete labels. The vector $\mathbf{e}^{\mathrm{id}}_j$ identifies the variable slot, while $\mathbf{e}^{\mathrm{role}}_{\rho_j}$ identifies its task role, with ρ_j denoting treatment, outcome, or conditioning variable. The vector $\mathbf{e}^{\mathrm{status}}_{\mathrm{obs}}$ indicates that the cell value is available to the model.

The intervention encoding $\mathbf{e}^{\mathrm{int}}(a, c)$ is a vector-valued function that combines an embedding of the intervention status a with an encoding of the assigned value c: $\mathbf{e}^{\mathrm{int}}(a, c) = \mathbf{e}^{\text{int-status}}_a + a\, \phi_{\mathrm{int}}(c)$. Here, $\mathbf{e}^{\text{int-status}}_0$ and $\mathbf{e}^{\text{int-status}}_1$ are learned embedding vectors indicating whether the variable is intervened upon, and ϕ_{int} encodes the assigned intervention value. For $a = 0$, the encoding contains only the non-intervened status vector. For $a = 1$, it also includes the encoded assignment $\phi_{\mathrm{int}}(c)$. This distinguishes an intervention assigning the value zero from the absence of an intervention.

Query tokens use the same construction, with the unknown outcome replaced by a masked-target embedding and the requested intervention supplied as metadata. To keep intervention information available throughout the network, the intervention encoding is also injected into every attention layer through a learned gated residual connection.

Hierarchical three-axis attention. Existing table-aware architectures typically exchange information across the variable and sample axes. CIDER-FM introduces the experimental regime as a third axis. Each of its L layers first applies variable-axis attention to the p cell tokens within every context and query row, allowing each token to incorporate information from the other variables in that row. It then applies sample-axis attention independently for each regime and variable, allowing samples within a regime to exchange information while preserving the separation between regimes.

After sample-axis attention, each (regime, variable) pair contains a set of sample tokens. Learned attention pooling compresses this set into K memory tokens, producing a fixed-size summary for each variable in each regime. For each variable, regime-axis self-attention then fuses the corresponding memories from all R regimes. These variable-aligned memories aggregate evidence about the same variable across observational and interventional settings. Finally, separate cross-attention blocks use the fused memories to update the context and query tokens before the next layer.

After the final layer, a linear readout maps the token corresponding to the masked query outcome to the parameters of a full-support bar distribution (Robertson et al., 2026; Hollmann et al., 2025). The model is trained end-to-end by minimising the negative log-likelihood of sampled query outcomes.

5 Experiments

We evaluate CIDER-FM's ability to predict CIDs from data collected across multiple experimental regimes. We begin with general random graphs, considering both linear–Gaussian mechanisms and heterogeneous nonlinear mechanisms with heavy-tailed noise, then use restricted graph families to examine performance stratified by three identification settings in Fig. 3. We then evaluate on real-world data from Causal Chambers (Gamella et al., 2025b). Finally, we conduct architectural ablations to assess the computational efficiency of three-axis attention.

Baselines We compare CIDER-FM with predictive and causal inference methods (Table 5), and CIDER-FM-Obs, a counterpart trained for the same task using only observational context. The general-random-graph and real-world benchmarks include Bayesian linear regression (Bayesian LR), TabPFN v2 (Hollmann et al., 2025), and Graph4CFM (Reuter et al., 2026), with varying levels of ancestral information. The restricted-family benchmark additionally includes linear regression (LR), random forests (RF), partially linear double machine learning (DML) (Chernozhukov et al., 2018), Do-PFN (Robertson et al., 2026), and ArCO-GP (Toth et al., 2025). For the restricted families, we additionally report c-gID-based estimators supplied with the ground-truth ADMG as graph-informed references, excluding them from the main rankings. Given the misspecification affecting external CFMs in our setting (Appendix E), we use CIDER-FM-Obs as our primary matched control.

Metrics For each query, the model predicts $p(y \mid \mathrm{do}(T = t), \mathbf{x}, \mathcal{D})$. We distinguish two evaluation targets: the sampled query outcome Y_q and the generating SCM's conditional interventional mean, $Y_{\mathrm{mean}} = \mathbb{E}_\psi[Y \mid \mathrm{do}(T = t), \mathbf{x}]$. Model predictions are evaluated using negative log-likelihood (NLL), and its predictive mean using MSE and R^2. Where available, comparison with Y_{mean} allows us to ignore the effect of noise added to the mean. We distinguish three context configurations:

- `fusion`: observational and experimental samples from all available regimes.
- `obs_total`: observational samples matching the total number of samples in the fusion context, giving a fair comparison of the contribution of experimental context.
- `obs_fixed`: observational samples matching only the observational portion of the fusion context.

5.1 Synthetic Data

5.1.1 General random graphs

We first sample random acyclic directed mixed graphs with 4–10 observed variables, allowing hidden confounding through bidirected edges. We consider linear–Gaussian mechanisms and ComplexMech,

Table 1: Performance on general random graphs: 560 linear–Gaussian and 512 ComplexMech test SCMs. Entries report the indicated statistic with its SCM-level standard error conditional on fixed checkpoints. Best point estimates are in bold. Infinite NLL arises when the predicted density is numerically zero for a query outcome. ComplexMech Y_{mean} is omitted because reliable ground-truth conditional means are unavailable.

Method		**CIDER-FM**	CIDER-FM-Obs	Bayesian LR	TabPFN v2	Graph4CFM
Data & Info		`fusion`	`obs_total`	`obs_total`	`obs_total`	`obs_total` all ancestry
Linear-Gaussian Y_{mean}	MSE ↓	**95.03** ± 35.10	471.72 ± 167.56	795.25 ± 247.80	386.37 ± 213.81	160.10 ± 60.11
	R^2_{mean} ↑	**0.675** ± 0.045	-0.016 ± 0.030	-21.719 ± 7.873	-3.608 ± 1.736	0.474 ± 0.082
	R^2_{med} ↑	**0.947** ± 0.006	-0.016 ± 0.006	0.843 ± 0.032	0.897 ± 0.016	0.811 ± 0.019
	R^2_{pool} ↑	**0.895** ± 0.067	0.479 ± 0.144	0.122 ± 0.633	0.573 ± 0.347	0.823 ± 0.101
Linear-Gaussian Y_q	NLL ↓	**2.059** ± 0.061	2.445 ± 0.057	13.990 ± 1.961	$+\infty$	3.25e7 ± 2.40e7
	MSE ↓	**138.97** ± 37.70	510.72 ± 168.72	835.55 ± 247.73	425.30 ± 213.14	206.73 ± 63.39
	R^2_{mean} ↑	**0.587** ± 0.019	0.065 ± 0.015	-7.963 ± 2.763	-0.167 ± 0.114	0.498 ± 0.020
	R^2_{med} ↑	**0.741** ± 0.024	-0.009 ± 0.003	0.559 ± 0.060	0.681 ± 0.047	0.599 ± 0.037
	R^2_{pool} ↑	**0.853** ± 0.080	0.461 ± 0.141	0.118 ± 0.583	0.551 ± 0.327	0.782 ± 0.110
ComplexMech Y_q	NLL ↓	-0.395 ± 0.078	**-0.416** ± 0.079	2.081 ± 0.437	$+\infty$	2.12e8 ± 1.28e8
	MSE ↓	**289.26** ± 94.06	299.26 ± 95.40	2588.10 ± 1314.80	771.98 ± 405.36	404.89 ± 124.91
	R^2_{mean} ↑	**0.245** ± 0.022	0.224 ± 0.027	−4.14e4 ± 4.09e4	-0.054 ± 0.119	0.133 ± 0.032
	R^2_{med} ↑	0.110 ± 0.026	**0.112** ± 0.022	-0.061 ± 0.011	0.057 ± 0.020	0.043 ± 0.019
	R^2_{pool} ↑	**0.935** ± 0.052	0.933 ± 0.057	0.420 ± 0.562	0.827 ± 0.153	0.909 ± 0.073

which combines randomly sampled MLP mechanisms, additive or non-additive noise, different noise scales, as well as heavy-tailed noise. Appendix G provides the data construction. Some tasks have very little within-task target variation and therefore exhibit extreme negative R^2. We therefore report the mean and median task-level R^2, together with R^2 computed by pooling query rows across SCMs.

Table 1 shows that CIDER-FM achieves the best point estimates across all reported metrics for both Y_{mean} and Y_q under linear–Gaussian mechanisms. Its advantage over CIDER-FM-Obs and Graph4CFM highlights the benefit of experimental context, while its comparison with TabPFN illustrates the difference between $p(y \mid \text{do}(T = t), \mathbf{x})$ and $p(y \mid T = t, \mathbf{x})$.

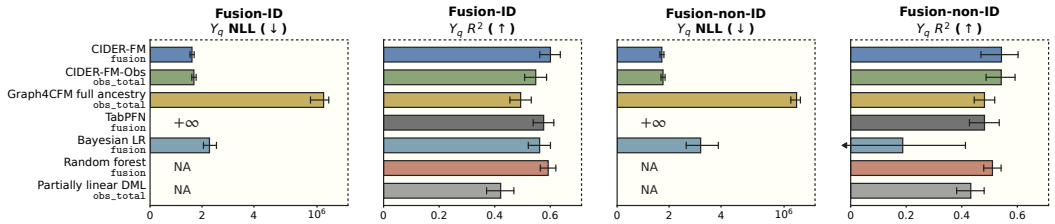
Under ComplexMech, CIDER-FM continues to outperform both predictive baselines across the reported metrics. Compared with TabPFN, query MSE decreases from 771.98 to 289.26, while mean task-level R^2 increases from -0.054 to 0.245. The gains over CIDER-FM-Obs are smaller than under linear–Gaussian mechanisms, with particularly similar NLL, median R^2, and pooled R^2.

An advantage over the observation-only counterpart is not guaranteed when experimental regimes are selected randomly under a fixed total context budget. For example, intervening on an isolated surrogate Z leaves the joint distribution of (T, Y) unchanged and does not resolve any existing confounding between treatment and outcome. To better understand when experimental context is beneficial, we next stratify tasks by the identifiability of the target causal query.

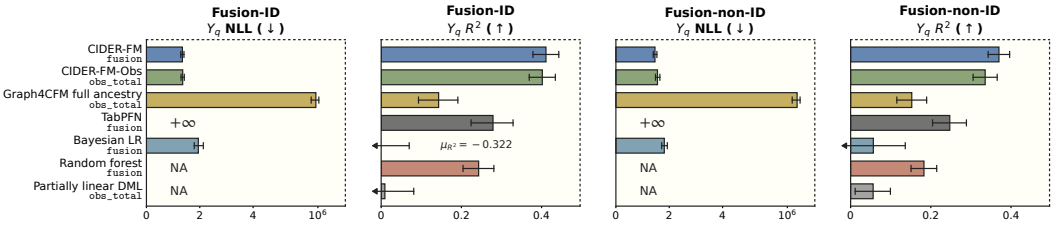
5.1.2 Stratification by identifiability

To examine how predictive performance varies with identifiability, we next consider restricted graph families in which Y is a descendant of both the treatment T and surrogate Z, with hidden confounding represented by bidirected edges. The query conditions on $\mathbf{X}$, comprising all observed variables other than T and Y, including Z. Contexts contain observational data and interventional data on Z, optionally supplemented by an intervention on another observed covariate $X_j \in \mathbf{X} \setminus \{Z\}$. Using c-gID under the generating graph, we distinguish `obs_id`, `fusion_id`, and `fusion_non_id` settings as defined in Fig. 3. Appendix F describes the detailed construction.

The rank comparison contains 19 method configurations for point-prediction metrics and 11 with reported density scores. It includes both observational and pooled-data baselines, together with



(a) Linear–Gaussian



(b) Nonlinear–Gaussian

Figure 4: Selected baseline comparisons on restricted graph families in `fusion_id` and `fusion_non_id`. Error bars indicate 95% confidence intervals.

Graph4CFM variants with and without ancestral information. Graph-informed c-gID estimators are shown as additional references in the full tables (Tables 8 and 9).

The three identification settings and five evaluation metrics (NLL, MSE, and R^2 for Y_q; MSE and R^2 for Y_{mean}) yield 15 setting–metric combinations per mechanism family. CIDER-FM ranks among the top three configurations in all 15 linear–Gaussian combinations and 13 of 15 nonlinear–Gaussian combinations (Table 2). Detailed rankings appear in Tables 6 and 7.

Table 2: Rank-1 and top-3 frequencies of CIDER-FM over five metrics in each of three settings, giving 15 setting–metric combinations per mechanism family.

Mechanism	Rank 1	Top 3
Linear–Gaussian	9/15 (60%)	15/15 (100%)
Nonlinear–Gaussian	7/15 (46.7%)	13/15 (86.7%)

Fig. 4a shows that, under linear–Gaussian mechanisms, CIDER-FM achieves the lowest NLL and highest Y_q R^2 point estimates in both fusion settings. In `fusion_id`, where the available experiments enable identification under the generating graph, CIDER-FM improves over CIDER-FM-Obs on both metrics, highlighting the benefit of experimental context. In `fusion_non_id`, CIDER-FM continues to perform best, despite the query remaining non-identifiable. However, its advantage over CIDER-FM-Obs is smaller in this setting, with nearly identical Y_q R^2.

5.2 Real-world evaluation

We then evaluate CIDER-FM on the light-tunnel task from Causal Chambers (Gamella et al., 2025b). The task concerns how intervening on the red-channel light-source setting $T = \text{red}$ affects the downstream light-intensity measurement $Y = \text{vis}_3$, conditional on the green- and blue-channel settings $\mathbf{X} = (\text{green}, \text{blue})$. We withhold both polarizer variables, leaving the models with incomplete information about the optical system. The goal is to predict the conditional outcome distribution under interventions on the red channel using observational measurements and experiments on other observed variables controlling the light-tunnel experiment.

We select the experimental data sources through *simulation-based experimental design*. Using data generated by the official simulator corresponding to the light-tunnel experiment (Gamella et al., 2025a), we set up different combinations of observational and interventional regimes under a fixed total context budget. We evaluate these candidate designs with the frozen ComplexMech CIDER-FM in Section 5.1.1 and select based on NLL on the simulated data. This procedure selects observational data combined with interventions on green. We then freeze the design and evaluate it using real measurements (without reselecting regimes based on real-data performance). No model retraining is performed, and simulated samples are excluded from the real-data context. (See Appendix H).

Table 3: Real-world performance on Causal Chambers. Values are estimates ± standard errors. Bold indicates the best point estimate in each column. For Graph4CFM, arrows denote supplied ancestral relations; R, G, and B denote the red, green, and blue channels, respectively.

Method	Data & Info	Y_q NLL ↓	Y_q MSE ↓	Y_q R^2_{mean} ↑	Y_q R^2_{med} ↑	Y_q R^2_{pool} ↑
CIDER-FM	`fusion`	5.772 ± 0.031	**6810** ± 399	**0.367** ± 0.026	**0.380** ± 0.038	**0.374** ± 0.026
CIDER-FM-Obs	`obs_total`	**5.770** ± 0.036	6861 ± 340	0.360 ± 0.026	0.351 ± 0.029	0.370 ± 0.026
CIDER-FM-Obs	`obs_fixed`	5.816 ± 0.042	7057 ± 392	0.343 ± 0.028	0.318 ± 0.038	0.352 ± 0.028
CIDER-FM	`obs_total`	5.792 ± 0.029	6956 ± 375	0.352 ± 0.025	0.350 ± 0.026	0.361 ± 0.026
Bayesian LR	`obs_total`	5.850 ± 0.027	7099 ± 400	0.338 ± 0.029	0.352 ± 0.051	0.348 ± 0.029
TabPFN v2	`obs_total`	5.930 ± 0.062	9042 ± 670	0.156 ± 0.062	0.172 ± 0.078	0.169 ± 0.052
Graph4CFM	`obs_total`	5.870 ± 0.044	7156 ± 366	0.334 ± 0.025	0.316 ± 0.034	0.342 ± 0.025
Graph4CFM	`obs_total`; $\mathrm{R} \to Y$	5.970 ± 0.043	7946 ± 445	0.259 ± 0.035	0.256 ± 0.044	0.270 ± 0.035
Graph4CFM	`obs_total`; $\mathrm{R}, \mathrm{G}, \mathrm{B} \to Y$	6.039 ± 0.045	8582 ± 481	0.199 ± 0.041	0.219 ± 0.054	0.211 ± 0.038

Table 3 shows that the selected design reduces MSE by approximately 24.7% relative to the equal-budget TabPFN baseline. Relative to CIDER-FM-Obs, point-prediction scores are slightly better, while NLL is comparable but slightly worse. These results suggest that CIDER-FM could support future experiments by selecting query-specific experimental designs using only simulator-generated data. The selected interventions could then be carried out in the real world, enabling targeted data collection aimed at reducing uncertainty about a causal query, alongside causal prediction from the resulting real world context again with CIDER-FM.

5.3 Ablation studies

Although two-axis architectures can learn from multiple experimental regimes when scaled up and supplied with intervention-aware inputs, our three-axis design aims to improve computational efficiency as multi-regime contexts grow.

Table 4: Throughput ratios (three-axis / two-axis) from short-run profiling on an NVIDIA A100, with 10 variables, $R = 3$, and eight memory tokens per regime and variable. Same width: $d_{3\mathrm{D}} = d_{2\mathrm{D}} = 128$. Parameter matched: $d_{3\mathrm{D}} = 128$, $d_{2\mathrm{D}} = 160$.

Rows per regime	Same width		Parameter matched	
	Inference	Training	Inference	Training
256	0.814	0.865	0.940	**1.002**
512	0.941	**1.001**	**1.112**	**1.167**
1024	**1.108**	**1.199**	**1.297**	**1.383**
2048	**1.357**	**1.484**	**1.561**	**1.684**

Table 4 shows that three-axis attention can be slower for short contexts, but its relative throughput improves as context length increases. At 1024 and 2048 rows per regime, it achieves higher inference and training throughput under both same-width and parameter-matched comparisons. These results support its computational advantage when scaling to more samples per regime in the tested configuration.

6 Discussion, Limitations and Outlook

CIDER-FM demonstrates that CFMs are a promising method for utilising multiple experimental regimes for causal inference. Due to limited available computational resources, our current evaluation is limited to graphs with at most ten variables and does not yet establish how the approach scales to larger systems. More compute would also enable longer pretraining and exploration of more diverse and complex SCM priors. Although the proposed input representation supports simultaneous interventions on multiple variables, our experiments use only one intervention target per regime and assume that all regimes share the same population and non-intervened mechanisms. While in causality experiments on synthetic data are typically considered best practice to establish the fundamental efficacy of any causal inference approach (Poinsot et al., 2025; Reuter et al., 2026), and we validate CIDER-FM on real-world data using Causal Chambers, the next step is to apply CIDER-FM to challenging real-world scientific applications such as single-cell perturbation studies (Gao et al., 2026). We especially believe that the general setup of using simulators together with CIDER-FM for deciding which experiments to carry out is a very promising avenue.

AI USE STATEMENT

We used generative AI tools to improve the clarity of our writing, edit and polish author-written portions of the Appendix, identify potentially relevant literature, assist with the presentation and checking of mathematical arguments, and support the implementation, execution, and monitoring of experiments. We also used these tools as an additional source of critical feedback on the manuscript, including checks of whether our interpretation of the experiments was supported by the reported evidence. The authors reviewed all AI-assisted work.

Research ideas and writing. The research questions, scientific arguments, and central contributions were developed by the authors. Generative AI was used to suggest clearer phrasing, improve readability, and edit author-written Appendix material. It was also used to assist with improving figures and visualisations of the results. The authors reviewed and approved all material included.

Related work. Generative AI was used to help identify publications that had not been found through our conventional literature searches. This supplemented our own literature review. The authors read and independently assessed all publications identified with AI before citing or discussing them.

Mathematical statements and proofs. The theorem, its central proof strategy, and its final statement were developed by the authors. AI assistance was limited to improving the presentation and notation of author-written arguments and providing additional checks of the reasoning. Two authors independently checked the proof, and the authors determined the final mathematical content and presentation.

Implementation and experiments. Coding assistants were used to support the implementation, execution, and monitoring of experiments. The authors reviewed and tested the AI-assisted code and independently checked the reported results. Generative AI was not used to generate the synthetic datasets. The authors remained responsible for the implementations, experimental procedures, and interpretation of the results.

We take responsibility for the complete content of this work, including its text, mathematical statements, experimental results, code, figures, and other artifacts produced with the assistance of AI.

ETHICS STATEMENT

This work is methodological and uses synthetic data and measurements from a physical experimental system; it does not involve human participants or personal data. Causal predictions may be unreliable when the deployment setting differs from the training prior or when relevant assumptions fail, so they should be independently validated before use in high-stakes decision making.

REPRODUCIBILITY STATEMENT

Code for model training and experimental evaluation is available at `https://anonymous.4open.science/r/ciderfm-review-source-A7F2/`. Appendix D specifies the input preprocessing, intervention encoding, hierarchical attention architecture, and predictive distribution, while Appendix J reports the model and training hyperparameters and computational resources. The synthetic graph and mechanism generation procedures, context and query construction, and stratification by identifiability are detailed in Appendices G and F. Appendix E describes the baseline configurations, the data and causal information supplied to each method, relevant prior mismatches, and the evaluation metrics. For Causal Chambers, Appendix H documents the dataset and causal query, simulator-guided context selection, evaluation with frozen checkpoints, and procedures for aggregating results and estimating standard errors. The assumptions and complete proofs for the finite-data and asymptotic theoretical analyses are presented in Appendices B and C, respectively.

REFERENCES

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse Cresswell, and Rahul Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *Advances in Neural Information Processing Systems*, 38:154945–154984, 2026.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. *arXiv preprint arXiv:1210.4842*, 2012.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Anish Dhir, Samuel Power, and Mark Van Der Wilk. Bivariate causal discovery using bayesian model selection. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Anish Dhir, Cristiana Diaconu, Valentinian Lungu, James Requeima, Richard Turner, and Mark van der Wilk. Estimating interventional distributions with uncertain causal graphs through meta-learning. *Advances in Neural Information Processing Systems*, 38:140060–140096, 2026.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems*, 38, 2026.

Juan L Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. *arXiv preprint arXiv:2502.20099*, 2025a.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025b.

Yuche Gao, José Miguel Hernández-Lobato, and Siyuan Guo. Perturbpfn: Probing the limits of synthetic priors in drug perturbation modelling. *arXiv preprint arXiv:2607.23447*, 2026.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pp. 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1): 2409–2464, 2012.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Yaroslav Kivva, Jalal Etesami, and Negar Kiyavash. On identifiability of conditional causal effects. In *Uncertainty in Artificial Intelligence*, pp. 1078–1086. PMLR, 2023.

Sanghack Lee, Juan D Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Uncertainty in artificial intelligence*, pp. 389–398. PMLR, 2020.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. *arXiv preprint arXiv:2506.10914*, 2025.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, pp. 669–688, 1995.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Audrey Poinsot, Panayiotis Panayiotou, Alessandro Leite, Nicolas CHESNEAU, Özgür Şimşek, and Marc Schoenauer. Position: Causal machine learning requires rigorous synthetic experiments for broader adoption. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Arik Reuter, Anish Dhir, Cristiana Diaconu, Jake Robertson, Ole Ossen, Frank Hutter, Adrian Weller, Mark van der Wilk, and Bernhard Schölkopf. Use what you know: Causal foundation models with partial graphs. *arXiv preprint arXiv:2602.14972*, 2026.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *Advances in Neural Information Processing Systems*, 38:174811–174848, 2026.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *AAAI*, pp. 1219–1226, 2006.

Christian Toth, Christian Knoll, Franz Pernkopf, and Robert Peharz. Effective bayesian causal inference via structural marginalisation and autoregressive orders. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4240–4248. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/toth25a.html.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

A CODE REPOSITORY

Our implementation, including code for model training and experimental evaluation, is available at https://anonymous.4open.science/r/ciderfm-review-source-A7F2/.

B FINITE-DATA MSE REDUCTION OF THE PPD

Fix the query values t, x, y. All expectations below are taken under the joint distribution of SCM and data sampled from

$$p(\psi, \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) = p(\psi)\, p(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}} \mid \psi),$$

Recall that the interventional posterior predictive density (PPD) is

$$\widehat{q}_y(\mathcal{D}) = \int p_\psi(y \mid \text{do}(T = t), x)\, p(\psi \mid \mathcal{D})\, d\psi.$$

Integration over ψ includes summation over graphs and integration over their mechanism parameters. We need to assume

$$\mathbb{E}_\psi \left[p_\psi(y \mid \text{do}(T = t), x)^2 \right] < \infty.$$

Note that by Jensen's inequality, this implies that the PPDs also have finite second moments, so all expectations below are finite and well-defined.

Proof of Eq. (3). By the definition of the PPD, for almost every pair of datasets, we have:

$$\int \Big[p_\psi(y \mid \text{do}(T = t), x) - \widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) \Big] p(\psi \mid \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})\, d\psi = 0. \tag{6}$$

Here, the first term integrates to $\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})$ by definition, while the second is constant with respect to ψ and the posterior $p(\psi \mid \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})$ integrates to one.

Now let's decompose the difference using observational data alone:

$$\begin{aligned} &p_\psi(y \mid \text{do}(T = t), x) - \widehat{q}_y(\mathcal{D}_{\text{obs}}) \\ &= \Big[p_\psi(y \mid \text{do}(T = t), x) - \widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) \Big] + \Big[\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - \widehat{q}_y(\mathcal{D}_{\text{obs}}) \Big]. \end{aligned}$$

Squaring and taking expectations gives

$$\begin{aligned} R_O = R_{O+I} &+ \mathbb{E}_{\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}} \left[\big(\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - \widehat{q}_y(\mathcal{D}_{\text{obs}}) \big)^2 \right] \\ &+ 2\mathbb{E}_{\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}} \Bigg[\big(\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - \widehat{q}_y(\mathcal{D}_{\text{obs}}) \big) \\ &\qquad \cdot \int \Big[p_\psi(y \mid \text{do}(T = t), x) - \widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) \Big] p(\psi \mid \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})\, d\psi \Bigg]. \end{aligned}$$

Here, the cross term is written by first averaging over the SCM conditional on both datasets, and then averaging over the datasets. The change between the two PPDs depends only on the datasets, so it can be taken outside the inner integral.

The inner integral is zero by Eq. (6). Consequently,

$$R_O - R_{O+I} = \mathbb{E}_{\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}} \left[\big(\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - \widehat{q}_y(\mathcal{D}_{\text{obs}}) \big)^2 \right] \geq 0.$$

Since the integrand is a square, the inequality is strict exactly when the two PPDs differ at y with positive probability under the marginal distribution of $(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})$. □.

C FINITE-SAMPLE ASYMPTOTIC ESTIMATION ERROR REDUCTION

We also provide theoretical results under the assumptions of a fully-parametric, continuously parametrised SCM. While continuous relaxions exist (Zheng et al., 2018), SCMs, as normally defined, are semi-discrete objects with their DAGs being inherently discrete objects. In this case, the theory in this section only applies *conditional* on a known causal DAG.

Taking those caveats into account, we ask how additional interventional samples affect the estimation error of a posterior predictive CID,

$$\widehat{q}_y(\mathcal{D}) = p(y \mid \mathrm{do}(T=t), x, \mathcal{D}) = \int p_\psi(y \mid \mathrm{do}(T=t), x) p(\psi \mid \mathcal{D})\, d\psi. \tag{7}$$

Assumption 1 (Parametric SCM and sampling scheme)**.** The data-generating process belongs to a correctly specified parametric SCM family

$$\{M_\psi : \psi \in \Psi \subseteq \mathbb{R}^d\}, \tag{8}$$

with true parameter $\psi^\star \in \Psi$. Observational and interventional samples are conditionally independent given $\psi^\star$, and the source label of each sample is known.

Assumption 2 (Bernstein–von Mises regularity)**.** The joint observational–interventional likelihood satisfies the standard regularity conditions for a finite-dimensional parametric Bernstein–von Mises approximation (Van der Vaart, 2000).

Assumption 3 (Smooth target CID)**.** For each fixed outcome value y, the target CID functional

$$q_y(\psi) = p_\psi(y \mid \mathrm{do}(T=t), x) \tag{9}$$

is twice continuously differentiable in a neighbourhood of $\psi^\star$.

Assumption 4 (Local point-identifiability of the target CID)**.** The target CID is locally point-identifiable from the combined observational and interventional distributions.

Let $\mathcal{D}_{\mathrm{obs}} = \{V_i^{\mathrm{obs}}\}_{i=1}^{n_O}$ be observational samples from

$$p^O_{\psi^\star}(\mathbf{v}) = p_{\psi^\star}(\mathbf{V} = \mathbf{v}), \tag{10}$$

and let $\mathcal{D}_{\mathrm{int}} = \{V_j^{\mathrm{int}}\}_{j=1}^{n_I}$ be additional samples from an interventional source, for example

$$p^I_{\psi^\star}(\mathbf{v}) = p_{\psi^\star}(\mathbf{V} \setminus Z = \mathbf{v} \mid \mathrm{do}(Z = z^\star)). \tag{11}$$

The joint log-likelihood is then

$$\ell_{O+I}(\psi) = \sum_{i=1}^{n_O} \log p^O_\psi(V_i^{\mathrm{obs}}) + \sum_{j=1}^{n_I} \log p^I_\psi(V_j^{\mathrm{int}}). \tag{12}$$

Let

$$I_O(\psi^\star) = \mathbb{E}_{\psi^\star}\left[\nabla_\psi \log p^O_\psi(V) \nabla_\psi \log p^O_\psi(V)^\top\right]_{\psi=\psi^\star} \tag{13}$$

and

$$I_I(\psi^\star) = \mathbb{E}_{\psi^\star}\left[\nabla_\psi \log p^I_\psi(V) \nabla_\psi \log p^I_\psi(V)^\top\right]_{\psi=\psi^\star} \tag{14}$$

denote the per-sample Fisher information matrices from the observational and interventional sources, respectively. Under the source independence in Assumption 1, the total Fisher information based on $\mathcal{D}_{\mathrm{obs}}$ and $\mathcal{D}_{\mathrm{int}}$ is

$$I^{(n)}_{O+I} = n_O I_O(\psi^\star) + n_I I_I(\psi^\star) \succeq n_O I_O(\psi^\star) s \tag{15}$$

where $\succeq$ denotes the Loewner order. The inequality follows from $I_I(\psi^\star) \succeq 0$. For compactness, define

$$\Sigma_{O+I} = \left(n_O I_O(\psi^\star) + n_I I_I(\psi^\star)\right)^{-1}. \tag{16}$$

By the Bernstein–von Mises theorem (Van der Vaart, 2000), the posterior over the SCM parameter is asymptotically locally Gaussian around an efficient centre $\widehat{\psi}_{O+I}$, which may be taken to be the maximum likelihood estimator or posterior mode:

$$\psi \mid \mathcal{D}_{\mathrm{obs}}, \mathcal{D}_{\mathrm{int}} \mathrel{\dot\sim} \mathcal{N}\left(\widehat{\psi}_{O+I}, \Sigma_{O+I}\right). \tag{17}$$

Since the target CID is a smooth functional of ψ, the posterior mean of $q_y(\psi)$ is first-order equivalent to the plug-in value at $\widehat{\psi}_{O+I}$:

$$\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) = \int q_y(\psi) p(\psi \mid \mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}})\, d\psi = q_y(\widehat{\psi}_{O+I}) + o_p(n^{-1/2}), \tag{18}$$

where $n = n_O + n_I$. This follows from the local Gaussian posterior approximation and a Taylor expansion of $q_y(\psi)$ around $\widehat{\psi}_{O+I}$. The efficient centre itself satisfies the usual regular parametric asymptotic normality (Van der Vaart, 2000):

$$\widehat{\psi}_{O+I} - \psi^\star \dot{\sim} \mathcal{N}\left(0, \Sigma_{O+I}\right). \tag{19}$$

Thus a first-order Taylor expansion of the CID functional around $\psi^\star$ gives

$$q_y(\widehat{\psi}_{O+I}) - q_y(\psi^\star) = \nabla_\psi q_y(\psi^\star)^\top (\widehat{\psi}_{O+I} - \psi^\star) + o_p(\|\widehat{\psi}_{O+I} - \psi^\star\|). \tag{20}$$

Combining Eq. (18) with the Taylor expansion in Eq. (20) yields

$$\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - q_y(\psi^\star) = \nabla_\psi q_y(\psi^\star)^\top (\widehat{\psi}_{O+I} - \psi^\star) + o_p(n^{-1/2}). \tag{21}$$

Using the asymptotic normality of $\widehat{\psi}_{O+I}$, we obtain the local Gaussian approximation

$$\widehat{q}_y(\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{int}}) - q_y(\psi^\star) \dot{\sim} \mathcal{N}\left(0,\ \nabla_\psi q_y(\psi^\star)^\top \Sigma_{O+I} \nabla_\psi q_y(\psi^\star)\right). \tag{22}$$

The leading asymptotic variance of the posterior predictive CID is therefore

$$\boxed{\nabla_\psi q_y(\psi^\star)^\top \left(n_O I_O(\psi^\star) + n_I I_I(\psi^\star)\right)^{-1} \nabla_\psi q_y(\psi^\star).} \tag{23}$$

If the target CID is already locally identifiable from observational data alone, then the corresponding observational-only approximation is

$$\widehat{q}_y(\mathcal{D}_{\text{obs}}) - q_y(\psi^\star) = O_p\left(\sqrt{g_y^\top \left(n_O I_O(\psi^\star)\right)^{-1} g_y}\right). \tag{24}$$

Because, whenever the inverses are well defined,

$$g_y^\top \left(n_O I_O(\psi^\star) + n_I I_I(\psi^\star)\right)^{-1} g_y \le g_y^\top \left(n_O I_O(\psi^\star)\right)^{-1} g_y. \tag{25}$$

Thus, in a regular parametric setting, adding interventional samples cannot increase the leading asymptotic variance of the posterior predictive CID when the target is point-identifiable from the observational source. It may strictly reduce the asymptotic error constant whenever the interventional source provides information in a direction relevant to the CID functional. One sufficient condition is

$$I_I(\psi^\star)^{1/2} \left(n_O I_O(\psi^\star)\right)^{-1} g_y \neq 0, \tag{26}$$

which states that the interventional likelihood contains information along a direction that affects $q_y(\psi)$.

Geometry of query-relevant information. Eq. (23) shows that finite-sample CID error is controlled not by the full Fisher information matrix alone, but by its inverse projected onto the gradient direction of the target functional, $\nabla_\psi q_y(\psi^\star)$. This projection gives a local geometric view of why interventional data can help.

Consider a small perturbation of the true SCM parameter in a tangent direction v,

$$\psi_\epsilon = \psi^\star + \epsilon v. \tag{27}$$

By the likelihood differentiability conditions underlying Assumption 2, the observational KL divergence between the distributions induced by $\psi^\star$ and ψ_ϵ admits a second-order expansion. To see this, write $\ell_O(\psi; V) = \log p_\psi^O(V)$. Then

$$\ell_O(\psi^\star + \epsilon v; V) = \ell_O(\psi^\star; V) + \epsilon v^\top \nabla_\psi \ell_O(\psi^\star; V) + \frac{1}{2}\epsilon^2 v^\top \nabla_\psi^2 \ell_O(\psi^\star; V) v + o(\epsilon^2). \tag{28}$$

Taking expectation under $p^O_{\psi^\star}$, the first-order term vanishes because the score has mean zero, and the negative expected Hessian equals the Fisher information. Hence

$$\mathrm{KL}\big(p^O_{\psi^\star} \,\|\, p^O_{\psi^\star+\epsilon v}\big) = \frac{1}{2}\epsilon^2 v^\top I_O(\psi^\star)v + o(\epsilon^2). \tag{29}$$

Thus, $v^\top I_O(\psi^\star)v$ measures how distinguishable the local perturbation $\psi^\star + \epsilon v$ is from $\psi^\star$ using observational data. If

$$v^\top I_O(\psi^\star)v = 0, \tag{30}$$

then observational data contain no local second-order information in this direction: perturbations of the SCM along v are locally invisible to the observational likelihood.

The same perturbation may nevertheless change the target CID. By Eq. (20),

$$q_y(\psi^\star + \epsilon v) - q_y(\psi^\star) = \epsilon \nabla_\psi q_y(\psi^\star)^\top v + o(\epsilon). \tag{31}$$

Therefore, if

$$\nabla_\psi q_y(\psi^\star)^\top v \neq 0, \tag{32}$$

then moving along v changes the target CID even though observational data cannot locally distinguish that movement. If the interventional source supplies information along the same direction, for example in case

$$v^\top I_I(\psi^\star)v > 0, \tag{33}$$

the interventional likelihood can locally distinguish perturbations along v. Then the combined information matrix

$$n_O I_O(\psi^\star) + n_I I_I(\psi^\star) \tag{34}$$

can become nonsingular on the query-relevant subspace. In this sense, interventional data can change the problem from one in which the CID cannot be regularly estimated from observational data alone to one in which the CID admits a regular asymptotic approximation.

In conclusion, additional interventional samples can improve finite-sample CID estimation by increasing the Fisher information available in query-relevant directions.

D Details on representation and architecture

D.1 Data fusion: setting and scope

We consider a task generated by a single causal system over a vector of p observed continuous variables $\mathbf{V} \in \mathbb{R}^p$. These include a treatment T, an outcome Y, and conditioning or surrogate intervention variables $\mathbf{X}$, with $Z \in \mathbf{X}$ used where necessary to denote a surrogate intervention target. The model receives a non-empty collection of finite datasets, denoted $\mathcal{D}_1, \ldots, \mathcal{D}_R$, collected from this system under different experimental regimes.

Given this context, the model predicts the conditional interventional distribution (CID)

$$p_\theta(y_i \mid \mathrm{do}(T = t_i)\,, x_i, z_i, \mathcal{D}_1, \ldots, \mathcal{D}_R) \tag{35}$$

for query covariate values x_i and z_i, and a continuous treatment assignment t_i. The schematic task in Fig. 5 illustrates the available source types and the distinction between context data and prediction queries.

Context regimes and query. Let $\mathcal{R}$ denote the set of regimes available in a particular task. Regime $r \in \mathcal{R}$ contributes

$$\mathcal{D}_r = \{\mathbf{v}_{r,i}\}_{i=1}^{n_r}\,, \qquad \mathbf{v}_{r,i} \in \mathbb{R}^p. \tag{36}$$

It is accompanied by an intervention-target mask $\mathbf{a}_r \in \{0,1\}^p$ and row-specific intervention values $\mathbf{c}_{r,i} \in \mathbb{R}^p$. Defining the active target set as $\mathcal{I}_r = \{j : a_{r,j} = 1\}$, row i is generated according to

$$\mathbf{v}_{r,i} \sim p_\psi(\mathbf{V} \mid \mathrm{do}(\mathbf{V}_{\mathcal{I}_r} = \mathbf{c}_{r,i,\mathcal{I}_r}))\,. \tag{37}$$

The observational regime has $\mathcal{I}_r = \varnothing$, equivalently $\mathbf{a}_r = \mathbf{0}$, so observational and interventional datasets share a common representation.

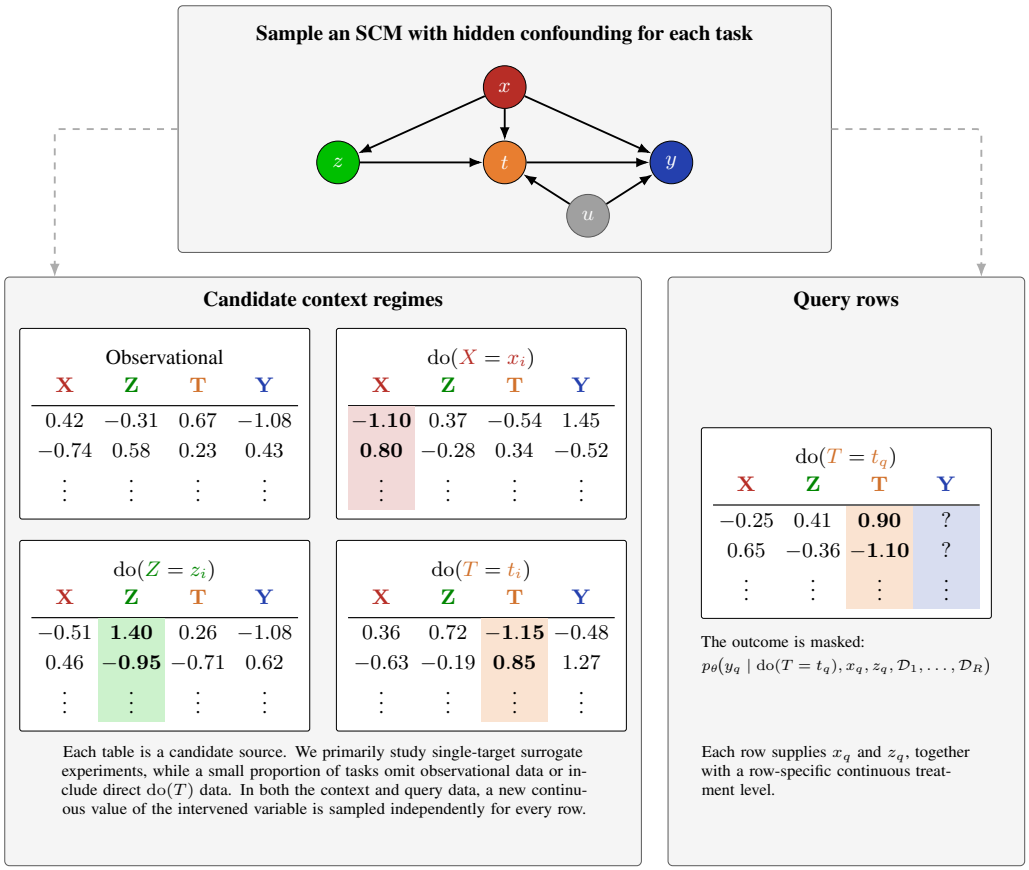


Figure 5: **Data fusion example:** diverse experimental regimes with continuous, row-wise interventions. The upper panel shows one possible latent-confounded data-generating SCM. An actual task contains an observational table and/or single-target interventional tables, while the right table contains CID queries. During pre-training, many SCMs are sampled and used to generate finite context and query data. At inference, only these data are provided; the underlying SCM remains unknown, and CIDER-FM predicts the CID in a single forward pass.

Crucially, an interventional regime is defined by its target rather than by one fixed intervention value. The assigned value is sampled separately for every row. A direct-treatment query table therefore contains rows generated under $\mathrm{do}(T = t_{r,1}), \ldots, \mathrm{do}(T = t_{r,n_r})$, with generally different continuous values. The same construction is used for surrogate experiments $\mathrm{do}(X = x_{r,i})$. Each such table consequently covers a range of intervention levels rather than repeated observations at a single fixed dose.

For each query row i, the model receives the observed conditioning values $\mathbf{x}_i$, the continuous assignment t_i, and query intervention metadata, while y_i is marked as the prediction target.

Availability of regimes. We focus primarily on settings in which causal information is provided by surrogate experiments. We also sample less common cases in which observational data are unavailable, as well as cases in which data from a direct intervention on T are available. The model therefore does not assume a fixed number of regimes or a fixed set of intervention targets.

Set-structured input. The context is treated as a set of regimes, each of which is itself a set of rows. Neither the order of regimes nor the order of rows within a regime carries information. Since the numbers of regimes and rows vary across tasks, inputs are padded for batching and padding entries are excluded using validity masks.

D.2 TOKEN REPRESENTATION

CIDER-FM represents each cell in the table as a token rather than assigning one token to an entire row. This retains the identity and role of each variable while allowing the model to alternate between variable-wise and sample-wise attention. A token combines information about the full row, the scalar cell value, the variable identity, its semantic role (treatment, outcome, conditioning), its cell status (such as target-masked), and the intervention associated with the regime.

D.2.1 TASK-WISE FEATURE NORMALISATION

Before tokenisation, every variable is standardised using all valid context entries for the current task. For context row i in regime r, let $v_{r,i,j}$ denote the value of variable j, and let $m_{r,i,j} \in \{0,1\}$ indicate whether this entry is valid rather than padding. The normalisation statistics are

$$\mu_j = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n_r} m_{r,i,j} v_{r,i,j}}{\sum_{r=1}^{R}\sum_{i=1}^{n_r} m_{r,i,j}}, \qquad \sigma_j^2 = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n_r} m_{r,i,j}(v_{r,i,j}-\mu_j)^2}{\sum_{r=1}^{R}\sum_{i=1}^{n_r} m_{r,i,j}},$$
$$\widetilde{v}_{r,i,j} = \frac{v_{r,i,j}-\mu_j}{\max(\sigma_j, \epsilon)}. \tag{38}$$

The same (μ_j, σ_j) are applied to context intervention values and query values. Thus all regimes use a common task-local coordinate system. The outcome statistics are retained so that the predictive density can later be mapped back to the original outcome scale.

D.2.2 CONTEXT AND QUERY CELL TOKENS

Let d denote the hidden dimension. The tokenisation combines complementary signals so that the model can distinguish cells by their values, variable identities and roles, intervention status, and surrounding row context. Three independent SwiGLU multilayer perceptrons (Shazeer, 2020) define the continuous-input encoders $\phi_{\text{row}} : \mathbb{R}^{p_{\max}} \to \mathbb{R}^d$, $\phi_{\text{val}} : \mathbb{R} \to \mathbb{R}^d$, and $\phi_{\text{int}} : \mathbb{R} \to \mathbb{R}^d$ for a complete standardised row, an individual cell value, and an intervention value, respectively. We use $\mathbf{e}$ instead for learned lookup embeddings of discrete metadata, such as a variable's semantic role, identity, and cell status. For context row i, regime r, and variable j, the initial token is

$$\mathbf{h}^{C,0}_{r,i,j} = \alpha\, \phi_{\text{row}}(\widetilde{\mathbf{v}}_{r,i}) + \beta\, \phi_{\text{val}}(\widetilde{v}_{r,i,j}) + \mathbf{e}^{\text{role}}_{\rho_j} + \mathbf{e}^{\text{id}}_{j} + \mathbf{e}^{\text{status}}_{\text{obs}} + \mathbf{e}^{\text{int}}(a_{r,j}, \widetilde{c}_{r,i,j}), \tag{39}$$

Here, C denotes a context token and the superscript 0 denotes initial representation; Q will analogously denote a query token. The index ρ_j identifies a variable as treatment, outcome, conditioning. The embedding $\mathbf{e}^{\text{status}}_{\text{obs}}$ marks the context cell value as observed by the model rather than missing or masked; it does not denote an observational regime. The learnable scalar coefficients α and β balance the row-level and cell-level encodings. The variable embedding $\mathbf{e}^{\text{id}}_j$ distinguishes variable slots.

The intervention embedding combines a discrete intervention-target indicator with its continuous assignment:

$$\mathbf{e}^{\text{int}}(a, c) = \mathbf{e}^{\text{int-status}}_a + a\, \phi_{\text{int}}(c). \tag{40}$$

Here, $\mathbf{e}^{\text{int-status}}_a$ is the vector selected from a learned two-entry lookup table for $a \in \{0,1\}$. It distinguishes a non-target cell from an intervention-target cell, while $a\, \phi_{\text{int}}(c)$ records the intervention value only when the variable is actively intervened upon. The resulting vector $\mathbf{e}^{\text{int}}(a, c)$ therefore keeps the intervention target and its row-specific assignment within a common d-dimensional representation.

Query tokens follow the same construction. For each query cell, $s_{i,j}$ indicates whether its value is observed or masked as the prediction target. The masked outcome is replaced by zero before standardisation, yielding the model-visible row $\widetilde{\mathbf{v}}^{\text{masked}}_i$; its status embedding marks that entry as a prediction target. For query row i, the initial token is

$$\mathbf{h}^{Q,0}_{i,j} = \alpha\phi_{\text{row}}(\widetilde{\mathbf{v}}^{\text{masked}}_i) + \beta\phi_{\text{val}}(\widetilde{v}^{\text{masked}}_{i,j}) + \mathbf{e}^{\text{role}}_{\rho_j} + \mathbf{e}^{\text{id}}_{j} + \mathbf{e}^{\text{status}}_{s_{i,j}} + \mathbf{e}^{\text{int}}(a^{Q}_{j}, \widetilde{c}^{Q}_{i,j}). \tag{41}$$

Equations equation 39 and equation 41 produce context tokens $\mathbf{H}^{C,0} \in \mathbb{R}^{R_{\max} \times n_{\max} \times p_{\max} \times d}$ and query tokens $\mathbf{H}^{Q,0} \in \mathbb{R}^{n_q \times p_{\max} \times d}$.

D.2.3 INTERVENTION METADATA RESIDUAL CONNECTION

CIDER-FM gives every attention layer direct access to the intervention target and value through a gated residual connection. Let $\mathbf{E}^{C,\mathrm{int}}$ and $\mathbf{E}^{Q,\mathrm{int}}$ denote the tensors obtained by stacking the per-cell vectors $\mathbf{e}^{\mathrm{int}}(a,c)$ across the context and query, respectively. Let L denote the number of attention layers. At layer $\ell \in \{0,\dots,L-1\}$, the context and query representations are

$$\mathbf{H}^{C,\ell}_{\mathrm{meta}} = \mathbf{H}^{C,\ell} + \mathrm{sigmoid}(g^{\ell}_{C}) W^{\ell}_{C}\,\mathrm{LN}\big(\mathbf{E}^{C,\mathrm{int}}\big)\,. \tag{42}$$

$$\mathbf{H}^{Q,\ell}_{\mathrm{meta}} = \mathbf{H}^{Q,\ell} + \mathrm{sigmoid}(g^{\ell}_{Q}) W^{\ell}_{Q}\,\mathrm{LN}\big(\mathbf{E}^{Q,\mathrm{int}}\big)\,. \tag{43}$$

Here, sigmoid is the logistic sigmoid and LN denotes layer normalisation. g^{ℓ}_{C} and g^{ℓ}_{Q} are learned scalar gate logits, and W^{ℓ}_{C} and W^{ℓ}_{Q} are learned linear projections. Equations equation 42 and equation 43 are written for valid cells; padded positions are reset to zero after each residual update. Repeatedly injecting this metadata prevents intervention semantics from being diluted across layers and allows otherwise similar values to be interpreted differently across causal regimes.

Algorithm 1 Hierarchical Three-Axis Attention Stack

Require: Initial tokens $\mathbf{H}^{C,0}$, $\mathbf{H}^{Q,0}$
intervention metadata $\mathbf{E}^{C,\mathrm{int}}$, $\mathbf{E}^{Q,\mathrm{int}}$
Layer count L
Ensure: Updated tokens $\mathbf{H}^{C,L}$, $\mathbf{H}^{Q,L}$

for $\ell = 0,\dots,L-1$ **do**
  $\mathbf{H}^{C,\ell}_{\mathrm{meta}} \leftarrow \mathbf{H}^{C,\ell} + \mathrm{sigmoid}(g^{\ell}_{C}) W^{\ell}_{C}\,\mathrm{LN}(\mathbf{E}^{C,\mathrm{int}})$ ▷ interventional metadata residual connection
  $\mathbf{H}^{Q,\ell}_{\mathrm{meta}} \leftarrow \mathbf{H}^{Q,\ell} + \mathrm{sigmoid}(g^{\ell}_{Q}) W^{\ell}_{Q}\,\mathrm{LN}(\mathbf{E}^{Q,\mathrm{int}})$
  $\mathbf{H}^{C,\ell}_{\mathrm{var}} \leftarrow \mathcal{A}^{\ell}_{V}(\mathbf{H}^{C,\ell}_{\mathrm{meta}})$ ▷ variable axis; row-wise
  $\mathbf{H}^{Q,\ell}_{\mathrm{var}} \leftarrow \mathcal{A}^{\ell}_{V}(\mathbf{H}^{Q,\ell}_{\mathrm{meta}})$ ▷ shared block; row-wise
  for each valid regime–variable pair (r,j) **do**
    $\big(\mathbf{H}^{C,\ell}_{\mathrm{samp}}\big)_{r,:,j,:} \leftarrow \mathcal{A}^{\ell}_{S}\Big(\big(\mathbf{H}^{C,\ell}_{\mathrm{var}}\big)_{r,:,j,:}\Big)$ ▷ sample axis
    $\big(\mathbf{M}^{\ell}_{\mathrm{local}}\big)_{r,:,j,:} \leftarrow \mathrm{PMA}^{\ell}\Big(\mathbf{Q}^{\ell}_{\mathrm{mem}}, \big(\mathbf{H}^{C,\ell}_{\mathrm{samp}}\big)_{r,:,j,:}\Big)$
  end for
  for each valid variable j **do**
    $\big(\mathbf{M}^{\ell}_{\mathrm{fused}}\big)_{:,:,j,:} \leftarrow \mathcal{A}^{\ell}_{R}\Big(\big(\mathbf{M}^{\ell}_{\mathrm{local}}\big)_{:,:,j,:}\Big)$ ▷ regime-axis attention over memories
    $\big(\mathbf{H}^{C,\ell+1}\big)_{:,:,j,:} \leftarrow \mathcal{A}^{\ell}_{C\leftarrow M}\Big(\big(\mathbf{H}^{C,\ell}_{\mathrm{samp}}\big)_{:,:,j,:}, \big(\mathbf{M}^{\ell}_{\mathrm{fused}}\big)_{:,:,j,:}\Big)$ ▷ context feedback
    $\big(\mathbf{H}^{Q,\ell+1}\big)_{:,j,:} \leftarrow \mathcal{A}^{\ell}_{Q\leftarrow M}\Big(\big(\mathbf{H}^{Q,\ell}_{\mathrm{var}}\big)_{:,j,:}, \big(\mathbf{M}^{\ell}_{\mathrm{fused}}\big)_{:,:,j,:}\Big)$ ▷ query reads only
  end for
end for
return $\mathbf{H}^{C,L}, \mathbf{H}^{Q,L}$

D.3 HIERARCHICAL THREE-AXIS ATTENTION

A dense attention block over all cells forms a pairwise attention matrix whose computational and memory costs grow quadratically with the number of tokens. Unlike existing tabular foundation models, CIDER-FM explicitly introduces the experimental regime as a third axis. These operations are repeated for L layers, with intervention metadata supplied at every layer and fused memories returned to both context and query tokens. Algorithm 1 summarises the resulting attention stack; the following subsections describe its components.

We use $\mathbf{H}$ for cell-token states, $\mathbf{M}_{\mathrm{local}}$ and $\mathbf{M}_{\mathrm{fused}}$ for memory tokens, and $\mathbf{Q}^{\ell}_{\mathrm{mem}} \in \mathbb{R}^{K\times d}$ for the K learned pooling queries. The blocks $\mathcal{A}^{\ell}_{V}$, $\mathcal{A}^{\ell}_{S}$, and $\mathcal{A}^{\ell}_{R}$ operate along the three axes, while $\mathcal{A}^{\ell}_{C\leftarrow M}$ and $\mathcal{A}^{\ell}_{Q\leftarrow M}$ perform memory feedback. Batch dimensions, validity masks, and reshaping are omitted for clarity.

D.3.1 VARIABLE-AXIS ATTENTION

The first operation exchanges information across variables independently within each context or query row:

$$\begin{aligned}\left(\mathbf{H}_{\mathrm{var}}^{C,\ell}\right)_{r,i,:,:} &= \mathcal{A}_V^\ell\left(\left(\mathbf{H}_{\mathrm{meta}}^{C,\ell}\right)_{r,i,:,:}\right),\\ \left(\mathbf{H}_{\mathrm{var}}^{Q,\ell}\right)_{i,:,:} &= \mathcal{A}_V^\ell\left(\left(\mathbf{H}_{\mathrm{meta}}^{Q,\ell}\right)_{i,:,:}\right).\end{aligned} \tag{44}$$

This operation allows every variable token to incorporate information from the other variables in the same row. Context and query use the same variable-axis block, but each row forms a separate attention sequence.

D.3.2 WITHIN-REGIME SAMPLE ATTENTION

The second operation exchanges information across context samples separately for each regime and variable:

$$\left(\mathbf{H}_{\mathrm{samp}}^{C,\ell}\right)_{r,:,j,:} = \mathcal{A}_S^\ell\left(\left(\mathbf{H}_{\mathrm{var}}^{C,\ell}\right)_{r,:,j,:}\right), \qquad r = 1,\dots,R, \quad j = 1,\dots,p. \tag{45}$$

As no sample positional embeddings are used and the same attention projections are applied to every row, sample-axis attention is permutation equivariant within each regime. Regimes are processed independently during this step. Query rows bypass this operation so each prediction depends only on its own query inputs and the shared context.

D.3.3 PER-VARIABLE MEMORIES AND REGIME-AXIS ATTENTION

CIDER-FM compresses each (regime, variable) sample set into K memory tokens before exchanging information across regimes. Let

$$\mathbf{Q}_{\mathrm{mem}}^\ell = \begin{bmatrix}\left(\mathbf{q}_{\mathrm{mem},1}^\ell\right)^\top\\ \vdots\\ \left(\mathbf{q}_{\mathrm{mem},K}^\ell\right)^\top\end{bmatrix} \in \mathbb{R}^{K\times d}, \qquad \mathbf{q}_{\mathrm{mem},k}^\ell \in \mathbb{R}^d$$

denote the layer-specific learned pooling queries. They are shared across all regimes and variables within layer ℓ.

Pooling by Multihead Attention (PMA) uses these vectors as queries in a residual cross-attention block, followed by a residual feed-forward update. It gives

$$\left(\mathbf{M}_{\mathrm{local}}^\ell\right)_{r,:,j,:} = \mathrm{PMA}^\ell\left(\mathbf{Q}_{\mathrm{mem}}^\ell, \left(\mathbf{H}_{\mathrm{samp}}^{C,\ell}\right)_{r,:,j,:}\right) \in \mathbb{R}^{K\times d}. \tag{46}$$

Here, the sample tokens $\mathbf{H}_{\mathrm{samp}}^{C,\ell}$ provide the keys and values. Collecting the outputs over all regimes and variables yields $\mathbf{M}_{\mathrm{local}}^\ell \in \mathbb{R}^{R\times K\times p\times d}$, remaining invariance under permutations of samples within a regime.

For a fixed variable j, there are K local memories from each of the R regimes, giving $R\times K$ memory tokens in total. Regime-axis self-attention processes these tokens jointly:

$$\left(\mathbf{M}_{\mathrm{fused}}^\ell\right)_{:,:,j,:} = \mathcal{A}_R^\ell\left(\left(\mathbf{M}_{\mathrm{local}}^\ell\right)_{:,:,j,:}\right) \in \mathbb{R}^{R\times K\times d}. \tag{47}$$

Across all variables, $\mathbf{M}_{\mathrm{fused}}^\ell \in \mathbb{R}^{R\times K\times p\times d}$ and is equivariant to permutations of the regimes. Regime-axis attention is applied separately for each variable, while cross-variable interactions are handled by variable-axis attention.

D.3.4 MEMORY FEEDBACK TO CONTEXT AND QUERY

For each variable j, the fused memories update the context and query tokens through separate cross-attention blocks:

$$\begin{aligned}\left(\mathbf{H}^{C,\ell+1}\right)_{:,:,j,:} &= \mathcal{A}_{C\leftarrow M}^\ell\left(\left(\mathbf{H}_{\mathrm{samp}}^{C,\ell}\right)_{:,:,j,:}, \left(\mathbf{M}_{\mathrm{fused}}^\ell\right)_{:,:,j,:}\right),\\ \left(\mathbf{H}^{Q,\ell+1}\right)_{:,j,:} &= \mathcal{A}_{Q\leftarrow M}^\ell\left(\left(\mathbf{H}_{\mathrm{var}}^{Q,\ell}\right)_{:,j,:}, \left(\mathbf{M}_{\mathrm{fused}}^\ell\right)_{:,:,j,:}\right).\end{aligned} \tag{48}$$

In both blocks, the cell tokens provide the queries and the fused memories provide the keys and values. The updated context tokens enter the next layer, allowing cross-regime information to be refined iteratively. Query tokens read from the memories but do not contribute to their construction.

D.3.5 COMPUTATIONAL COMPLEXITY

Let $N = \sum_{r=1}^{R} n_r$ be the total number of context rows. Ignoring constant factors and the hidden dimension, the principal attention cost per layer is

$$\mathcal{O}\Bigg(\underbrace{(N+n_q)p^2}_{\text{variable axis}} + \underbrace{p\sum_{r=1}^{R} n_r^2}_{\text{sample axis}} + \underbrace{pKN}_{\text{memory pooling}} + \underbrace{p(RK)^2}_{\text{regime axis}} + \underbrace{p(N+n_q)RK}_{\text{memory feedback}}\Bigg). \tag{49}$$

This factorisation preserves the permutation symmetries induced by the set-structured geometry of the input, while avoiding dense attention over all cells. The computational profiling in Section 5.3 shows that the relative efficiency of three-axis attention improves as multi-regime contexts grow. In the tested configuration, it achieves higher inference and training throughput than conventional alternating attention for 2-D tables at 1024 and 2048 rows per regime, under both same-width and parameter-matched comparisons. These results motivate three-axis attention as a structured and computationally attractive approach for scaling to larger multi-regime contexts.

D.4 PREDICTIVE DISTRIBUTION AND LEARNING OBJECTIVE

After the final layer, the model selects the query token corresponding to the designated outcome variable. A linear readout maps this token to the parameters of a full-support bar distribution. With J interior bars, the head emits $J+2$ mixture logits—one for each bar and one for each tail—together with two positive tail-scale parameters. The density is piecewise uniform over fixed intervals in standardised outcome space and uses half-Gaussian tails beyond the outer edges. The experiments in this work use $J = 32$ bars.

Let $z = (y - \mu_Y)/\sigma_Y$ denote the standardised outcome and let $p_\theta^{\text{bar}}(z \mid \cdot)$ be the density represented by the output head. The corresponding density on the original outcome scale is

$$p_\theta(y \mid \cdot) = \frac{1}{\sigma_Y} p_\theta^{\text{bar}}\left(\frac{y-\mu_Y}{\sigma_Y} \,\middle|\, \cdot\right). \tag{50}$$

For a minibatch containing B tasks and n_q queries per task, training minimises the original-scale negative log-likelihood

$$\mathcal{L}(\theta) = -\frac{1}{Bn_q}\sum_{b=1}^{B}\sum_{i=1}^{n_q} \log p_\theta(y_{b,i} \mid \mathcal{C}_b, \mathcal{Q}_{b,i}), \tag{51}$$

where $\mathcal{C}_b$ denotes the multi-regime context and $\mathcal{Q}_{b,i}$ denotes the intervention query. The objective uses only the sampled query outcome: oracle conditional means or graph-derived labels do not enter the loss. Because the network predicts a full density, conditional means, uncertainty intervals, quantiles, and samples can all be obtained from the same output distribution.

E BASELINES AND METRICS

Comparison with causal foundation models. We include Do-PFN (Robertson et al., 2026) as an external causal foundation model using observational context. Its released checkpoint is trained with binary treatments, whereas our benchmarks use continuous treatment values. We pass these values directly to its prediction interface without discretisation or retraining, and report its predictive mean. Consequently, this comparison evaluates Do-PFN under treatment-domain misspecification. We also evaluate Graph4CFM (Reuter et al., 2026) with varying amounts of ancestral information. Its synthetic training prior assumes causal sufficiency, excluding hidden confounders. Graph4CFM is therefore evaluated under prior misspecification on tasks containing hidden confounding, even when supplied with correct directed ancestry. We do not include CausalPFN (Balazadeh et al., 2026) or CausalFM (Ma et al., 2025), whose released implementations address different estimation

Table 5: Methods and context configurations used across the synthetic and real-world experiments.

Full method name	Context policy	Data supplied	Causal information supplied
CIDER-FM and observational controls			
CIDER-FM	`fusion`	N_r obs rows and $(R-1)N_r$ int rows	Intervention targets and values
CIDER-FM	`obs_total`	RN_r obs rows	None
CIDER-FM-Obs	`obs_total`	RN_r obs rows	None
CIDER-FM-Obs	`obs_fixed`	N_r obs rows	None
Predictive methods			
Linear regression	`obs_total`	RN_r obs rows	None
Linear regression	`fusion`	N_r obs rows and $(R-1)N_r$ int rows, pooled	None
Random forest	`obs_total`	RN_r obs rows	None
Random forest	`fusion`	N_r obs rows and $(R-1)N_r$ int rows, pooled	None
TabPFN v2	`obs_total`	RN_r obs rows	None
TabPFN v2	`fusion`	N_r obs rows and $(R-1)N_r$ int rows, pooled	None
TabPFN v2	`obs_fixed`	N_r obs rows	None
Bayesian linear regression	`obs_total`	RN_r obs rows	None
Bayesian linear regression	`fusion`	N_r obs rows and $(R-1)N_r$ int rows, pooled	None
Additional causal methods			
Partially linear DML	`obs_total`	RN_r obs rows	None
Partially linear DML	`obs_fixed`	N_r obs rows	None
Do-PFN	`obs_total`	RN_r obs rows	None
Do-PFN	`obs_fixed`	N_r obs rows	None
ArCO-GP	`fusion`	N_r obs rows and $(R-1)N_r$ int rows	Intervention targets and values
Graph4CFM variants			
Graph4CFM	`obs_total`	RN_r obs rows	None
Graph4CFM	`obs_total`	RN_r obs rows	Oracle ancestry into Y
Graph4CFM	`obs_total`	RN_r obs rows	Oracle all-pairs directed ancestry
Graph4CFM	`obs_total`	RN_r obs rows	$\mathrm{red} \to Y$
Graph4CFM	`obs_total`	RN_r obs rows	$\mathrm{red}, \mathrm{green}, \mathrm{blue} \to Y$
Graph-informed references (excluded from rankings)			
c-gID + linear regression	`obs_total`	RN_r obs rows	Ground-truth ADMG
c-gID + linear regression	`fusion`	N_r obs rows and $(R-1)N_r$ int rows	Ground-truth ADMG
c-gID + Bayesian linear regression	`obs_total`	RN_r obs rows	Ground-truth ADMG
c-gID + Bayesian linear regression	`fusion`	N_r obs rows and $(R-1)N_r$ int rows	Ground-truth ADMG

Protocol. R denotes the total number of regimes, and N_r denotes the number of rows per regime under equal allocation. Obs and int denote observational and interventional data. Predictive baselines pool observational and interventional rows without labels indicating their experimental regimes or intervention targets. The CIDER-FM `obs_total` control uses the fusion-trained checkpoint. Graph4CFM receives ancestral relations rather than a full adjacency graph; arrows denote ancestry. The c-gID references are evaluated only when the query is identifiable from the supplied sources under the ground-truth ADMG.

targets. CausalPFN estimates conditional expected potential outcomes for binary treatments under strong ignorability. CausalFM's released models target binary-treatment CATE estimation under specified back-door, front-door, or instrumental-variable assumptions. These implementations do not directly provide the conditional outcome distribution $p(y \mid \mathrm{do}(T = t), \mathbf{x}, \mathcal{D})$ for continuous t across our graph families with hidden confounding. A comparison would therefore require adapting their estimation targets and training setups, beyond applying the published checkpoints.

Given these prior and task mismatches, we train CIDER-FM-Obs using the same SCM prior and prediction task as CIDER-FM but only observational context, providing our primary matched baseline for assessing the benefit of interventional data.

Metrics. In the main text, we evaluate predictions against sampled query outcomes Y_q, reporting negative log-likelihood (NLL) for the predictive density and MSE and R^2 for its predictive mean. In the appendix, we additionally report MSE and R^2 against the generating SCM's conditional interventional mean, $Y_{\mathrm{mean}} = \mathbb{E}_\psi[Y \mid \mathrm{do}(T = t), \mathbf{x}]$, where this quantity can be computed reliably. These additional metrics assess conditional-mean prediction without the outcome sampling noise present in Y_q.

F Stratified results on restricted graphs

F.1 Experimental protocol

Algorithm 2 Restricted Graph-Family Sampling

Require: Node-count range $[N_{\min}, N_{\max}]$, optional target class $c^\star$
Ensure: ADMG G, available data regimes $\mathcal{S}$, class c

repeat
  Sample $N \sim \mathrm{Uniform}\{N_{\min}, \ldots, N_{\max}\}$ and set $V = \{Z, T, Y, X_0, \ldots, X_{N-4}\}$.
  Uniformly sample one of the five directed cores in which Y is a descendant of both Z and T.
  for $(A, B) \in \{(Z, T), (Z, Y), (T, Y)\}$ **do**
    Add $A \leftrightarrow B$ if an independent $\mathrm{Bernoulli}(0.5)$ draw succeeds. ▷ Hidden confounding
  end for
  Add $X_0 \to Z$, $X_0 \to T$, and $X_0 \to Y$.
  for $i = 1, \ldots, N - 4$ **do**
    for $W \in \{Z, T, Y\}$ **do**
      Add $X_i \to W$ if an independent $\mathrm{Bernoulli}(0.5)$ draw succeeds.
    end for
  end for
  Initialise the available regimes as

$$\mathcal{S} \leftarrow \{P(V), P(V \setminus \{Z\} \mid do(Z))\}.$$

  if an independent $\mathrm{Bernoulli}(0.5)$ draw succeeds **then**
    Uniformly select $X_j \in \{X_0, \ldots, X_{N-4}\}$ and set

$$\mathcal{S} \leftarrow \mathcal{S} \cup \{P(V \setminus \{X_j\} \mid do(X_j))\}.$$

  end if
  Classify $P(Y \mid do(T), V \setminus \{T, Y\})$ using c-gID.
until $c^\star$ is unspecified or the sampled task belongs to $c^\star$
return $G, \mathcal{S}, c$

We sample graphs with 4–10 nodes from five directed cores in which Y is a descendant of both Z and T as shown in Algorithm 2. Bidirected edges between Z, T, and Y introduce hidden confounding, while additional variables act as observed covariates. Each task contains one observational regime and $K \in \{0, 1, 2\}$ interventional regimes. For $K = 0$, the context is observational only; for $K = 1$, it also includes $do(Z)$; for $K = 2$, it additionally includes an experiment on a randomly chosen covariate X_j. Both context and query interventions use continuous assignments, with z_i in $\mathrm{do}(Z = z_i)$ and t_i in $\mathrm{do}(T = t_i)$ varying across rows. We instantiate the graphs using either linear–Gaussian or nonlinear–Gaussian mechanisms and assign each task to one of the three cases in Section 3 using c-gID.

Each regime contains $N_r = 512$ rows. We evaluate the predictive distribution of Y_q using NLL and its posterior mean using MSE and R^2; MSE and R^2 are also reported for the ground-truth conditional causal mean Y_{mean}. Table 5 summarises the comparison methods and the data and causal information supplied to each. The `obs_total` variants match the total row budget of the fusion methods, whereas `obs_fixed` variants use only the observational portion. Graph-informed methods receive the ground-truth ADMG, and the SCM oracle is included only as a reference. All model training and evaluation experiments were conducted on NVIDIA A100 GPUs.

F.2 Stratified results

Tables 6 and 7 report the full results, including means, standard errors, and CIDER-FM's rank for each setting and metric. CIDER-FM is competitive across all three settings and is particularly strong on NLL and R^2. Under nonlinear–Gaussian mechanisms, it achieves the best NLL in every setting and the best R^2 for both targets in the two data-fusion settings. Raw MSE is less uniform, with graph-informed estimators or pooled TabPFN performing best in some cases.

Table 6: Restricted-family linear–Gaussian detailed ranking. Values are mean ± standard error. Rank gives the position of CIDER-FM among 19 methods, excluding all c-gID methods.

Setting	Target	Metric	CIDER-FM mean ± SE	Rank	Best method	Best mean ± SE
obs_id	Y_q	NLL	0.9456 ± 0.0440	2 / 11	CIDER-FM-Obs (obs_total)	0.9338 ± 0.0394
		MSE	0.9097 ± 0.0922	3 / 19	CIDER-FM-Obs (obs_total)	0.8679 ± 0.0766
		R^2	**0.7330 ± 0.0174**	**1 / 19**	**CIDER-FM**	**0.7330 ± 0.0174**
	Y_{mean}	MSE	0.4131 ± 0.0769	3 / 19	CIDER-FM-Obs (obs_total)	0.3711 ± 0.0592
		R^2	**0.8498 ± 0.0241**	**1 / 19**	**CIDER-FM**	**0.8498 ± 0.0241**
fusion_id	Y_q	NLL	**1.6150 ± 0.0403**	**1 / 11**	**CIDER-FM**	**1.6150 ± 0.0403**
		MSE	**1.8332 ± 0.1013**	**1 / 19**	**CIDER-FM**	**1.8332 ± 0.1013**
		R^2	**0.6006 ± 0.0188**	**1 / 19**	**CIDER-FM**	**0.6006 ± 0.0188**
	Y_{mean}	MSE	**0.6746 ± 0.0585**	**1 / 19**	**CIDER-FM**	**0.6746 ± 0.0585**
		R^2	0.7089 ± 0.0398	3 / 19	RF (fusion)	0.7212 ± 0.0186
fusion_non_id	Y_q	NLL	**1.7229 ± 0.0391**	**1 / 11**	**CIDER-FM**	**1.7229 ± 0.0391**
		MSE	2.8267 ± 0.3385	2 / 19	CIDER-FM-Obs (obs_total)	2.7895 ± 0.2367
		R^2	**0.5428 ± 0.0345**	**1 / 19**	**CIDER-FM**	**0.5428 ± 0.0345**
	Y_{mean}	MSE	1.5011 ± 0.3232	2 / 19	CIDER-FM-Obs (obs_total)	1.4586 ± 0.2111
		R^2	**0.6243 ± 0.0551**	**1 / 19**	**CIDER-FM**	**0.6243 ± 0.0551**

Table 7: Restricted-family nonlinear–Gaussian detailed ranking. Values are mean ± standard error. Rank gives the position of CIDER-FM among 19 methods, excluding all c-gID methods.

Setting	Target	Metric	CIDER-FM mean ± SE	Rank	Best method	Best mean ± SE
obs_id	Y_q	NLL	**0.9537 ± 0.0323**	**1 / 11**	**CIDER-FM**	**0.9537 ± 0.0323**
		MSE	1.5096 ± 0.2347	3 / 19	TabPFN v2 (obs_total)	1.3793 ± 0.2732
		R^2	0.3737 ± 0.0154	2 / 19	TabPFN v2 (obs_total)	0.3936 ± 0.0156
	Y_{mean}	MSE	1.1354 ± 0.2335	3 / 19	TabPFN v2 (obs_total)	1.0021 ± 0.2722
		R^2	0.5057 ± 0.0176	2 / 19	TabPFN v2 (obs_total)	0.5299 ± 0.0182
fusion_id	Y_q	NLL	**1.3479 ± 0.0306**	**1 / 11**	**CIDER-FM**	**1.3479 ± 0.0306**
		MSE	3.5265 ± 0.7781	3 / 19	TabPFN v2 (fusion)	3.0792 ± 0.4604
		R^2	**0.4109 ± 0.0169**	**1 / 19**	**CIDER-FM**	**0.4109 ± 0.0169**
	Y_{mean}	MSE	2.7177 ± 0.7732	3 / 19	TabPFN v2 (fusion)	2.2722 ± 0.4526
		R^2	**0.5280 ± 0.0199**	**1 / 19**	**CIDER-FM**	**0.5280 ± 0.0199**
fusion_non_id	Y_q	NLL	**1.4653 ± 0.0341**	**1 / 11**	**CIDER-FM**	**1.4653 ± 0.0341**
		MSE	3.1356 ± 0.4350	4 / 19	TabPFN v2 (fusion)	2.6392 ± 0.2905
		R^2	**0.3699 ± 0.0135**	**1 / 19**	**CIDER-FM**	**0.3699 ± 0.0135**
	Y_{mean}	MSE	2.1801 ± 0.4270	4 / 19	TabPFN v2 (fusion)	1.6919 ± 0.2768
		R^2	**0.5787 ± 0.0166**	**1 / 19**	**CIDER-FM**	**0.5787 ± 0.0166**

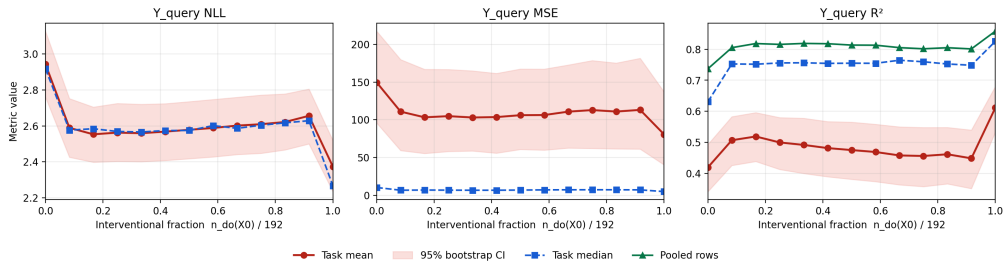


Figure 6: **Case studies:** Effect of observational–interventional context allocation on linear–Gaussian CIDER-FM for a fixed four-node `fusion_id` graph. The total context budget is fixed at 192 rows, while the horizontal axis varies the fraction of randomised interventional samples, $n_{\mathrm{do}(X_0)}/192$, from entirely observational data (0) to entirely interventional data (1).

F.3 Effect of interventional data ratio

To isolate the effect of context allocation, we fix a four-node `fusion_id` graph and vary the fraction of interventional rows while holding the total context budget at 192. All allocations are evaluated on the same held-out SCMs and query rows.

As shown in Figure 6, introducing even a small fraction of interventional rows substantially improves performance over the observation-only allocation, followed by relatively stable performance across mixed allocations.

G Construction of general random graphs and ComplexMech

Algorithm 3 General Random-Graph Sampling

Require: Node-count range $[4, 10]$
Ensure: ADMG $G = (V, E^{\rightarrow}, E^{\leftrightarrow})$
Initialise $E^{\rightarrow} \leftarrow \varnothing$ and $E^{\leftrightarrow} \leftarrow \varnothing$.
Sample $N \sim \mathrm{Uniform}\{4, \ldots, 10\}$ and set $V = \{T, Y, X_0, \ldots, X_{N-3}\}$.
Sample a graph-level directed-edge probability $p_{\rightarrow} \sim \mathrm{Beta}(2, 3)$.
Uniformly sample a permutation $\pi = (\pi_1, \ldots, \pi_N)$ of V.
for $1 \le a < b \le N$ **do**
    Add $\pi_a \rightarrow \pi_b$ if an independent $\mathrm{Bernoulli}(p_{\rightarrow})$ draw succeeds.
end for
Independently sample a graph-level bidirected-edge probability $p_{\leftrightarrow} \sim \mathrm{Beta}(2, 3)$.
for each unordered pair $\{A, B\} \subset V$ **do**
    Add $A \leftrightarrow B$ if an independent $\mathrm{Bernoulli}(p_{\leftrightarrow})$ draw succeeds.
end for
return $G = (V, E^{\rightarrow}, E^{\leftrightarrow})$

General random graphs. For each task, we first sample the number of observed variables as $N \sim \mathrm{Uniform}\{4, \ldots, 10\}$, with node set $\{T, Y, X_0, \ldots, X_{N-3}\}$ as shown in Algorithm 3. We independently draw graph-level edge probabilities $p_{\rightarrow}, p_{\leftrightarrow} \sim \mathrm{Beta}(2, 3)$. After sampling a uniformly random ordering $\pi = (\pi_1, \ldots, \pi_N)$, each directed edge $\pi_a \rightarrow \pi_b$, $a < b$, is included independently with probability $p_{\rightarrow}$. This construction guarantees acyclicity while allowing the directed-edge density to vary across SCMs. Each unordered node pair is independently connected by a bidirected edge with probability $p_{\leftrightarrow}$, representing an independent latent common cause. Bows and graphs without bidirected edges are both allowed. The number K of experimental regimes is sampled from a truncated geometric prior, $p(K = k \mid N) \propto 2^{-(k-1)}$ for $k = 1, \ldots, N - 2$; the resulting tasks contain an empirical mean of 1.76 experimental regimes.

ComplexMech. ComplexMech replaces linear structural equations with randomly sampled continuous MLP mechanisms. Each observed node and each latent root associated with a bidirected edge receives its own MLP, whose width, depth, activation, and noise-injection mode are sampled from the

Table 8: Restricted-family linear–Gaussian validation: all comparison methods. Mean ± SE across 520 SCMs per setting (130 each at 4, 6, 8 and 10 nodes). Bold denotes the best applicable mean in each panel, excluding gray *c-g*ID rows. *N/A*: no density or identification route.

Method	Y_q			Y_{mean}	
	NLL ↓	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑
obs_id					
CIDER-FM	0.9456 ± 0.0440	0.9097 ± 0.0922	**0.7330 ± 0.0174**	0.4131 ± 0.0769	**0.8498 ± 0.0241**
CIDER-FM-Obs (obs_total)	**0.9338 ± 0.0394**	**0.8679 ± 0.0766**	0.7299 ± 0.0190	**0.3711 ± 0.0592**	0.8345 ± 0.0305
LR (obs_total)	*N/A*	1.1995 ± 0.2220	0.6022 ± 0.0768	0.7032 ± 0.2122	0.5799 ± 0.1247
LR (fusion)	*N/A*	1.2662 ± 0.2240	0.5798 ± 0.0787	0.7703 ± 0.2143	0.5556 ± 0.1255
RF (obs_total)	*N/A*	1.5607 ± 0.1293	0.6296 ± 0.0123	1.0665 ± 0.1165	0.7299 ± 0.0151
RF (fusion)	*N/A*	1.5814 ± 0.1392	0.6362 ± 0.0117	1.0902 ± 0.1273	0.7381 ± 0.0139
TabPFN v2 (obs_total)	$+\infty$	0.9016 ± 0.1170	0.7328 ± 0.0229	0.4054 ± 0.0995	0.7977 ± 0.0496
TabPFN v2 (fusion)	$+\infty$	0.9386 ± 0.1222	0.7211 ± 0.0240	0.4422 ± 0.1056	0.7816 ± 0.0511
BLR (obs_total)	1.4017 ± 0.3783	1.1984 ± 0.2216	0.6027 ± 0.0765	0.7021 ± 0.2118	0.5805 ± 0.1244
BLR (fusion)	1.4831 ± 0.3913	1.2653 ± 0.2237	0.5802 ± 0.0785	0.7694 ± 0.2140	0.5561 ± 0.1252
DML (obs_total)	*N/A*	1.6023 ± 0.2267	0.6214 ± 0.0207	1.1095 ± 0.2185	0.6621 ± 0.0378
Do-PFN (obs_total)	*N/A*	1.7719 ± 0.1718	0.6358 ± 0.0152	1.2787 ± 0.1635	0.7282 ± 0.0239
TabPFN v2 (obs_fixed)	$+\infty$	0.9652 ± 0.1202	0.7167 ± 0.0223	0.4700 ± 0.1027	0.7786 ± 0.0464
DML (obs_fixed)	*N/A*	1.9409 ± 0.2622	0.5783 ± 0.0182	1.4473 ± 0.2555	0.6135 ± 0.0337
Do-PFN (obs_fixed)	*N/A*	1.9165 ± 0.1909	0.6089 ± 0.0169	1.4234 ± 0.1832	0.6834 ± 0.0280
ArCO-GP	1.5814 ± 0.0574	3.0684 ± 0.9451	0.5162 ± 0.0193	2.5714 ± 0.9443	0.5553 ± 0.0285
Graph4CFM (no graph)	$(1.23 \pm 0.44) \times 10^7$	3.3129 ± 0.4744	0.4499 ± 0.0181	2.8223 ± 0.4707	0.5159 ± 0.0219
Graph4CFM (ancestors of Y)	$(9.26 \pm 4.03) \times 10^6$	1.6375 ± 0.1809	0.6337 ± 0.0140	1.1479 ± 0.1727	0.7051 ± 0.0202
Graph4CFM (full ancestry)	$(1.66 \pm 0.67) \times 10^7$	1.3903 ± 0.1015	0.6447 ± 0.0141	0.8997 ± 0.0862	0.7137 ± 0.0214
*c-g*ID + LR (obs_total)	*N/A*	1.2001 ± 0.2225	0.6026 ± 0.0768	0.7037 ± 0.2128	0.5810 ± 0.1248
*c-g*ID + LR (fusion)	*N/A*	1.4543 ± 0.2541	0.5216 ± 0.0844	0.9594 ± 0.2444	0.4640 ± 0.1329
*c-g*ID + BLR (obs_total)	0.6731 ± 0.0420	0.7925 ± 0.1470	0.7012 ± 0.0721	0.2997 ± 0.1436	0.8091 ± 0.1050
*c-g*ID + BLR (fusion)	0.7522 ± 0.0434	1.0066 ± 0.1682	0.6267 ± 0.0789	0.5142 ± 0.1642	0.6986 ± 0.1140
fusion_id					
CIDER-FM	**1.6150 ± 0.0403**	**1.8332 ± 0.1013**	**0.6006 ± 0.0188**	**0.6746 ± 0.0585**	0.7089 ± 0.0398
CIDER-FM-Obs (obs_total)	1.6873 ± 0.0424	2.1612 ± 0.1396	0.5477 ± 0.0206	0.9983 ± 0.1029	0.6093 ± 0.0399
LR (obs_total)	*N/A*	4.2547 ± 0.9145	0.1409 ± 0.1760	3.0911 ± 0.9017	-0.3499 ± 0.3784
LR (fusion)	*N/A*	2.0682 ± 0.1213	0.5616 ± 0.0204	0.9135 ± 0.0791	0.6332 ± 0.0331
RF (obs_total)	*N/A*	2.2537 ± 0.1189	0.5497 ± 0.0172	1.0964 ± 0.0804	0.6278 ± 0.0279
RF (fusion)	*N/A*	1.9928 ± 0.0941	0.5924 ± 0.0145	0.8392 ± 0.0508	**0.7212 ± 0.0186**
TabPFN v2 (obs_total)	$+\infty$	2.8170 ± 0.2231	0.4275 ± 0.0327	1.6560 ± 0.1896	0.3209 ± 0.0779
TabPFN v2 (fusion)	$+\infty$	1.9752 ± 0.1103	0.5762 ± 0.0191	0.8193 ± 0.0679	0.6664 ± 0.0297
BLR (obs_total)	4.4714 ± 0.5673	4.2543 ± 0.9143	0.1410 ± 0.1760	3.0908 ± 0.9015	-0.3498 ± 0.3783
BLR (fusion)	2.2876 ± 0.1280	2.0682 ± 0.1213	0.5616 ± 0.0204	0.9135 ± 0.0791	0.6332 ± 0.0331
DML (obs_total)	*N/A*	2.8567 ± 0.1671	0.4214 ± 0.0255	1.6960 ± 0.1274	0.3471 ± 0.0497
Do-PFN (obs_total)	*N/A*	2.1919 ± 0.1199	0.4261 ± 0.1261	1.0304 ± 0.0812	0.0798 ± 0.5483
TabPFN v2 (obs_fixed)	$+\infty$	2.6481 ± 0.2073	0.4610 ± 0.0284	1.4834 ± 0.1759	0.4162 ± 0.0582
DML (obs_fixed)	*N/A*	3.2070 ± 0.1722	0.4034 ± 0.0211	2.0389 ± 0.1358	0.3664 ± 0.0377
Do-PFN (obs_fixed)	*N/A*	2.3191 ± 0.1256	0.2242 ± 0.2896	1.1569 ± 0.0889	-0.7139 ± 1.2678
ArCO-GP	1.7217 ± 0.0425	2.1228 ± 0.1168	0.5351 ± 0.0203	0.9687 ± 0.0730	0.5748 ± 0.0392
Graph4CFM (no graph)	$(5.60 \pm 3.89) \times 10^6$	2.0031 ± 0.1041	0.5899 ± 0.0148	0.8510 ± 0.0624	0.7108 ± 0.0204
Graph4CFM (ancestors of Y)	$(3.62 \pm 2.97) \times 10^6$	2.4361 ± 0.1311	0.4987 ± 0.0197	1.2800 ± 0.0893	0.5025 ± 0.0377
Graph4CFM (full ancestry)	$(3.68 \pm 3.06) \times 10^6$	2.4239 ± 0.1272	0.4933 ± 0.0199	1.2717 ± 0.0851	0.4945 ± 0.0387
*c-g*ID + LR (fusion)	*N/A*	1.2398 ± 0.0610	0.7165 ± 0.0109	0.0807 ± 0.0054	0.9611 ± 0.0031
*c-g*ID + BLR (fusion)	1.2273 ± 0.0279	1.2409 ± 0.0611	0.7163 ± 0.0109	0.0819 ± 0.0055	0.9607 ± 0.0032
fusion_non_id					
CIDER-FM	**1.7229 ± 0.0391**	2.8267 ± 0.3385	**0.5428 ± 0.0345**	1.5011 ± 0.3232	**0.6243 ± 0.0551**
CIDER-FM-Obs (obs_total)	1.7671 ± 0.0395	**2.7895 ± 0.2367**	0.5420 ± 0.0266	**1.4586 ± 0.2111**	0.6103 ± 0.0477
LR (obs_total)	*N/A*	6.2806 ± 1.6874	0.1120 ± 0.1614	4.9374 ± 1.6745	-0.0567 ± 0.2038
LR (fusion)	*N/A*	5.8364 ± 1.6790	0.1872 ± 0.1544	4.4960 ± 1.6662	0.0878 ± 0.1904
RF (obs_total)	*N/A*	3.0442 ± 0.1482	0.5068 ± 0.0164	1.7309 ± 0.1156	0.5857 ± 0.0251
RF (fusion)	*N/A*	2.9690 ± 0.1366	0.5104 ± 0.0159	1.6524 ± 0.0995	0.5952 ± 0.0226
TabPFN v2 (obs_total)	$+\infty$	3.4143 ± 0.2938	0.4447 ± 0.0319	2.0785 ± 0.2677	0.4183 ± 0.0580
TabPFN v2 (fusion)	$+\infty$	3.1366 ± 0.2622	0.4817 ± 0.0274	1.8006 ± 0.2343	0.4915 ± 0.0490
BLR (obs_total)	4.2846 ± 0.4914	6.2784 ± 1.6858	0.1122 ± 0.1613	4.9353 ± 1.6729	-0.0564 ± 0.2036
BLR (fusion)	3.2158 ± 0.3140	5.8342 ± 1.6773	0.1874 ± 0.1543	4.4938 ± 1.6646	0.0880 ± 0.1902
DML (obs_total)	*N/A*	3.4549 ± 0.2172	0.4324 ± 0.0255	2.1251 ± 0.1843	0.3882 ± 0.0495
Do-PFN (obs_total)	*N/A*	3.3871 ± 0.2355	0.5069 ± 0.0200	2.0648 ± 0.2167	0.5800 ± 0.0325
TabPFN v2 (obs_fixed)	$+\infty$	3.2801 ± 0.2779	0.4662 ± 0.0298	1.9469 ± 0.2512	0.4637 ± 0.0524
DML (obs_fixed)	*N/A*	3.5700 ± 0.2166	0.4243 ± 0.0232	2.2489 ± 0.1839	0.3932 ± 0.0428
Do-PFN (obs_fixed)	*N/A*	3.3797 ± 0.2198	0.4956 ± 0.0207	2.0565 ± 0.1995	0.5630 ± 0.0372
ArCO-GP	2.4001 ± 0.0816	4.4181 ± 0.2995	0.3444 ± 0.0249	3.0921 ± 0.2790	0.3353 ± 0.0398
Graph4CFM (no graph)	$(1.76 \pm 0.48) \times 10^7$	5.1302 ± 0.4418	0.3810 ± 0.0169	3.8112 ± 0.4340	0.4570 ± 0.0209
Graph4CFM (ancestors of Y)	$(8.18 \pm 3.01) \times 10^6$	3.3115 ± 0.1773	0.4789 ± 0.0186	1.9872 ± 0.1416	0.4982 ± 0.0304
Graph4CFM (full ancestry)	$(1.17 \pm 0.57) \times 10^7$	3.1786 ± 0.1636	0.4817 ± 0.0191	1.8515 ± 0.1224	0.4969 ± 0.0322

Table 9: Restricted-family nonlinear–Gaussian validation: all comparison methods. Mean ± SE across 520 SCMs per setting (130 each at 4, 6, 8 and 10 nodes). Bold denotes the best applicable mean in each panel, excluding gray *c-g*ID rows. *N/A*: no density or identification route.

Method	Y_q			Y_{mean}	
	NLL ↓	MSE ↓	R^2 ↑	MSE ↓	R^2 ↑
obs_id					
CIDER-FM	**0.9537 ± 0.0323**	1.5096 ± 0.2347	0.3737 ± 0.0154	1.1354 ± 0.2335	0.5057 ± 0.0176
CIDER-FM-Obs (obs_total)	1.0179 ± 0.0337	1.7253 ± 0.3161	0.3430 ± 0.0167	1.3486 ± 0.3153	0.4258 ± 0.0229
LR (obs_total)	*N/A*	4.4135 ± 0.9504	0.2465 ± 0.0185	4.0367 ± 0.9513	0.3103 ± 0.0285
LR (fusion)	*N/A*	4.4630 ± 0.9372	0.2212 ± 0.0171	4.0890 ± 0.9378	0.2764 ± 0.0240
RF (obs_total)	*N/A*	4.3342 ± 0.8977	0.3009 ± 0.0147	3.9576 ± 0.8965	0.3807 ± 0.0194
RF (fusion)	*N/A*	4.1705 ± 0.8432	0.2727 ± 0.0138	3.7958 ± 0.8417	0.3359 ± 0.0180
TabPFN v2 (obs_total)	+∞	**1.3793 ± 0.2732**	**0.3936 ± 0.0156**	**1.0021 ± 0.2722**	**0.5299 ± 0.0182**
TabPFN v2 (fusion)	+∞	1.4054 ± 0.2700	0.3591 ± 0.0151	1.0302 ± 0.2687	0.4754 ± 0.0176
BLR (obs_total)	1.0837 ± 0.0390	4.4135 ± 0.9504	0.2465 ± 0.0185	4.0367 ± 0.9513	0.3103 ± 0.0285
BLR (fusion)	1.1323 ± 0.0379	4.4630 ± 0.9372	0.2212 ± 0.0171	4.0890 ± 0.9378	0.2764 ± 0.0240
DML (obs_total)	*N/A*	4.3299 ± 0.8829	0.1817 ± 0.0190	3.9548 ± 0.8832	0.1295 ± 0.0299
Do-PFN (obs_total)	*N/A*	3.0252 ± 0.5519	0.1326 ± 0.0717	2.6511 ± 0.5511	0.1007 ± 0.1159
TabPFN v2 (obs_fixed)	+∞	1.7490 ± 0.3133	0.3518 ± 0.0159	1.3727 ± 0.3121	0.4565 ± 0.0196
DML (obs_fixed)	*N/A*	5.2764 ± 1.0323	0.1212 ± 0.0198	4.9028 ± 1.0320	0.0448 ± 0.0295
Do-PFN (obs_fixed)	*N/A*	3.4477 ± 0.6594	0.0918 ± 0.0531	3.0729 ± 0.6582	-0.0570 ± 0.1174
ArCO-GP	1.1430 ± 0.0378	3.7139 ± 0.8339	0.2526 ± 0.0163	3.3400 ± 0.8300	0.2804 ± 0.0197
Graph4CFM (no graph)	$(5.78 \pm 1.08) \times 10^6$	4.3147 ± 1.1176	0.2483 ± 0.0162	3.9361 ± 1.1178	0.2755 ± 0.0225
Graph4CFM (ancestors of Y)	$(3.04 \pm 0.58) \times 10^6$	2.7311 ± 0.5113	0.2433 ± 0.0176	2.3546 ± 0.5101	0.2179 ± 0.0260
Graph4CFM (full ancestry)	$(3.65 \pm 0.68) \times 10^6$	2.9989 ± 0.5628	0.2469 ± 0.0173	2.6199 ± 0.5608	0.2316 ± 0.0255
*c-g*ID + LR (obs_total)	*N/A*	4.3975 ± 0.9482	0.2489 ± 0.0184	4.0206 ± 0.9491	0.3161 ± 0.0283
*c-g*ID + LR (fusion)	*N/A*	4.5985 ± 0.9735	0.1966 ± 0.0236	4.2216 ± 0.9742	0.2088 ± 0.0445
*c-g*ID + BLR (obs_total)	1.0793 ± 0.0389	4.3914 ± 0.9500	0.2531 ± 0.0184	4.0147 ± 0.9510	0.3259 ± 0.0287
*c-g*ID + BLR (fusion)	1.1201 ± 0.0396	4.5879 ± 0.9745	0.1992 ± 0.0245	4.2114 ± 0.9752	0.2154 ± 0.0473
fusion_id					
CIDER-FM	**1.3479 ± 0.0306**	3.5265 ± 0.7781	**0.4109 ± 0.0169**	2.7177 ± 0.7732	**0.5280 ± 0.0199**
CIDER-FM-Obs (obs_total)	1.3596 ± 0.0306	3.5618 ± 0.8391	0.4018 ± 0.0170	2.7515 ± 0.8332	0.5083 ± 0.0192
LR (obs_total)	*N/A*	14.6202 ± 2.2011	-0.6410 ± 0.4039	13.7846 ± 2.1971	-1.8777 ± 0.8024
LR (fusion)	*N/A*	13.7198 ± 1.9163	-0.3225 ± 0.2689	12.8941 ± 1.9118	-1.1239 ± 0.5856
RF (obs_total)	*N/A*	7.9823 ± 1.0253	0.2100 ± 0.0209	7.1586 ± 1.0167	0.0497 ± 0.0453
RF (fusion)	*N/A*	7.7285 ± 1.0228	0.2431 ± 0.0198	6.9169 ± 1.0148	0.1285 ± 0.0399
TabPFN v2 (obs_total)	+∞	3.3423 ± 0.4945	0.1852 ± 0.0308	2.5294 ± 0.4851	-0.1412 ± 0.0836
TabPFN v2 (fusion)	+∞	**3.0792 ± 0.4604**	0.2789 ± 0.0270	**2.2722 ± 0.4526**	0.1165 ± 0.0710
BLR (obs_total)	2.1456 ± 0.1137	14.6190 ± 2.2004	-0.6406 ± 0.4036	13.7834 ± 2.1964	-1.8770 ± 0.8020
BLR (fusion)	1.9460 ± 0.0893	13.7190 ± 1.9160	-0.3222 ± 0.2687	12.8933 ± 1.9115	-1.1234 ± 0.5853
DML (obs_total)	*N/A*	7.6543 ± 0.9398	0.0093 ± 0.0391	6.8368 ± 0.9310	-0.5047 ± 0.1010
Do-PFN (obs_total)	*N/A*	8.0483 ± 1.7881	0.1679 ± 0.0440	7.2132 ± 1.7739	0.0605 ± 0.0699
TabPFN v2 (obs_fixed)	+∞	4.8245 ± 0.6867	0.1790 ± 0.0328	4.0117 ± 0.6780	-0.0517 ± 0.0640
DML (obs_fixed)	*N/A*	12.8796 ± 1.6096	-0.0641 ± 0.0434	12.0626 ± 1.6049	-0.6073 ± 0.1409
Do-PFN (obs_fixed)	*N/A*	8.7356 ± 1.3301	-0.3076 ± 0.3601	7.9071 ± 1.3155	-0.6048 ± 0.4368
ArCO-GP	1.8391 ± 0.0749	6.5506 ± 0.9071	0.2190 ± 0.0227	5.7445 ± 0.8990	0.0820 ± 0.0477
Graph4CFM (no graph)	$(1.39 \pm 0.46) \times 10^6$	4.9189 ± 0.7050	0.3303 ± 0.0174	4.1078 ± 0.6983	0.3522 ± 0.0221
Graph4CFM (ancestors of Y)	$(1.12 \pm 0.58) \times 10^6$	6.1153 ± 0.8604	0.1558 ± 0.0234	5.3115 ± 0.8547	-0.1210 ± 0.0485
Graph4CFM (full ancestry)	$(6.78 \pm 2.39) \times 10^5$	6.1298 ± 0.7988	0.1433 ± 0.0255	5.3170 ± 0.7914	-0.1372 ± 0.0505
*c-g*ID + LR (fusion)	*N/A*	12.8379 ± 1.7700	0.2174 ± 0.0149	12.0144 ± 1.7642	0.2480 ± 0.0193
*c-g*ID + BLR (fusion)	1.6651 ± 0.0457	12.8258 ± 1.7674	0.2173 ± 0.0149	12.0021 ± 1.7616	0.2476 ± 0.0193
fusion_non_id					
CIDER-FM	**1.4653 ± 0.0341**	3.1356 ± 0.4350	**0.3699 ± 0.0135**	2.1801 ± 0.4270	**0.5787 ± 0.0166**
CIDER-FM-Obs (obs_total)	1.5593 ± 0.0417	3.1330 ± 0.4179	0.3355 ± 0.0153	2.1793 ± 0.4076	0.4809 ± 0.0241
LR (obs_total)	*N/A*	5.6222 ± 0.7918	0.0712 ± 0.0488	4.6761 ± 0.7868	-0.0468 ± 0.0721
LR (fusion)	*N/A*	5.6966 ± 0.7927	0.0565 ± 0.0482	4.7565 ± 0.7883	-0.0489 ± 0.0682
RF (obs_total)	*N/A*	5.5545 ± 0.8677	0.2272 ± 0.0166	4.6064 ± 0.8600	0.2147 ± 0.0339
RF (fusion)	*N/A*	5.8814 ± 0.9057	0.1829 ± 0.0164	4.9353 ± 0.8978	0.1273 ± 0.0353
TabPFN v2 (obs_total)	+∞	2.6881 ± 0.2951	0.2627 ± 0.0228	1.7397 ± 0.2812	0.2314 ± 0.0541
TabPFN v2 (fusion)	+∞	**2.6392 ± 0.2905**	0.2473 ± 0.0214	**1.6919 ± 0.2768**	0.2204 ± 0.0470
BLR (obs_total)	1.9414 ± 0.0727	5.6221 ± 0.7918	0.0713 ± 0.0488	4.6760 ± 0.7868	-0.0466 ± 0.0720
BLR (fusion)	1.8149 ± 0.0557	5.6965 ± 0.7927	0.0566 ± 0.0481	4.7564 ± 0.7883	-0.0487 ± 0.0682
DML (obs_total)	*N/A*	6.6966 ± 1.0043	0.0561 ± 0.0223	5.7535 ± 1.0000	-0.1741 ± 0.0538
Do-PFN (obs_total)	*N/A*	6.3113 ± 1.1085	0.0890 ± 0.0613	5.3506 ± 1.1013	0.0199 ± 0.1124
TabPFN v2 (obs_fixed)	+∞	3.3522 ± 0.4076	0.2415 ± 0.0213	2.4030 ± 0.3977	0.2131 ± 0.0447
DML (obs_fixed)	*N/A*	7.3199 ± 1.0678	0.0349 ± 0.0207	6.3790 ± 1.0627	-0.1935 ± 0.0445
Do-PFN (obs_fixed)	*N/A*	6.4828 ± 1.0465	0.0206 ± 0.1036	5.5272 ± 1.0423	-0.1177 ± 0.1821
ArCO-GP	1.8616 ± 0.0494	5.7433 ± 0.9454	0.1645 ± 0.0334	4.7852 ± 0.9368	0.0425 ± 0.1217
Graph4CFM (no graph)	$(7.87 \pm 2.32) \times 10^6$	7.1976 ± 1.3303	0.2155 ± 0.0152	6.2658 ± 1.3313	0.2595 ± 0.0232
Graph4CFM (ancestors of Y)	$(4.36 \pm 1.66) \times 10^6$	4.1800 ± 0.4962	0.1516 ± 0.0192	3.2313 ± 0.4857	-0.0012 ± 0.0389
Graph4CFM (full ancestry)	$(6.50 \pm 2.35) \times 10^6$	4.2344 ± 0.5250	0.1525 ± 0.0191	3.2814 ± 0.5137	-0.0021 ± 0.0397

mechanism prior and then held fixed across all regimes of the SCM. Exogenous noise is sampled independently across rows as

$$\epsilon_{j,i} = s_{j,i}\xi_{j,i},$$
$$\xi_{j,i} \sim 0.33\,\mathcal{N}(0,1) + 0.33\ \mathrm{Laplace}\left(0, \frac{1}{\sqrt{2}}\right) + 0.34\,\frac{t_3}{\sqrt{3}}, \qquad s_{j,i} \sim \mathrm{Gamma}(\alpha, \theta). \tag{52}$$

Separate Gamma scale distributions are sampled for root and non-root mechanisms within each SCM. The resulting family combines nonlinear functional relationships, heterogeneous noise scales, heavy-tailed disturbances, and nonlinear hidden confounding.

H Details on causal chamber evaluation

Data and task. We use the light-tunnel dataset `lt_crl_benchmark_v1` (Gamella et al., 2025b;a), following the official collection protocol. It contains 10,000 measurements per environment, with light-source and polarizer commands generated by a software SCM and randomized assignments to individual commands in the intervention environments. The task is $\mathrm{red} \to \mathrm{vis}_3 \mid \mathrm{green}, \mathrm{blue}$, with both polarizers hidden. Query rows come from the red-intervention environment; context sources are observations and interventions on green or blue. The reported case is drawn from an exploratory light-tunnel study.

Simulator-guided design. We use the released deterministic counterparts of these environments, generated by the official `lt.Deterministic` sensor simulator (Gamella et al., 2025a). We enumerate the seven subsets of $\{\mathrm{Obs}, \mathrm{do(green)}, \mathrm{do(blue)}\}$. Each candidate receives 192 context rows, divided equally among its active regimes (Table 10). Using the frozen ComplexMech CIDER-FM, we evaluate each candidate on ten disjoint simulator-data blocks, each with 128 query rows and three context draws. Query rows are shared across candidates. We average NLL over repeats and blocks and select the context with the lowest mean NLL: 96 observational rows and 96 rows from $\mathrm{do(green)}$. This design is then fixed for evaluation on real measurements.

Table 10: Candidate contexts for $\mathrm{red} \to \mathrm{vis}_3 \mid \mathrm{green}, \mathrm{blue}$, each with 192 rows. Bold marks the context selected by simulator NLL.

Context	Obs rows	do(green) rows	do(blue) rows
do(green)	0	192	0
do(blue)	0	0	192
do(green) + do(blue)	0	96	96
Obs + do(green)	**96**	**96**	**0**
Obs + do(blue)	96	0	96
Obs + do(green) + do(blue)	64	64	64

Comparison methods. We use the pretrained ComplexMech CIDER-FM and CIDER-FM-Obs from Section 5.1.1, keeping their weights fixed. All `obs_total` methods receive the same 192 observational rows, including the 96 used by CIDER-FM's fusion context. The `obs_fixed` comparison applies the same CIDER-FM-Obs checkpoint to those 96 rows. We also evaluate the fusion checkpoint with the 192-row observational input. TabPFN uses the local v2 checkpoint and its full predictive distribution for NLL. Graph4CFM receives from no ancestry information, $\mathrm{red} \to Y$, to $\mathrm{red}, \mathrm{green}, \mathrm{blue} \to Y$, with all other off-diagonal relations left unknown. Baseline and metric definitions follow Section E.

Evaluation and aggregation. We divide the real-data confirmation partition into nine disjoint blocks. Each block provides 128 query rows and three context draws sampled without replacement within each draw. The query rows are fixed across methods and repeats, giving 27 episodes with 1,152 distinct query outcomes. We evaluate Y_q in the original sensor units and first average scores over context repeats within each block. NLL, MSE, and mean R^2 give equal weight to the nine blocks; their standard errors are the sample standard deviation of block scores divided by $\sqrt{9}$. Median R^2 is taken across the repeat-averaged block scores, with a block-bootstrap standard error. Pooled

R^2 uses repeat-averaged squared errors and the variation of all distinct query outcomes around their pooled mean; its standard error is estimated by a delete-one-block jackknife.

I CASE STUDY: SURROGATE EXPERIMENTS SHARPEN CONDITIONAL INTERVENTIONAL PREDICTIONS.

This section provides the construction and quantitative evaluation behind Fig. 2. We use two observationally equivalent SCMs whose conditional interventional distributions differ, but which can be distinguished using a surrogate intervention on Z. The models receive neither graph information nor interventions on the target treatment T.

I.1 TASK CONSTRUCTION

Observational equivalence. Consider four observed variables (Z, T, Y, X) and mutually independent standard-normal exogenous noises. With $r = 0.65$ and $s = 0.9$, define

$$\begin{aligned} \psi_{\rightarrow}: \quad & X = \varepsilon_X, \qquad Z = \varepsilon_Z, \\ & T = rZ + \sqrt{1-r^2}\,\varepsilon_T, \qquad Y = sT + \sqrt{1-s^2}\,\varepsilon_Y, \\ \psi_{\leftarrow}: \quad & X = \varepsilon_X, \qquad Y = \varepsilon_Y, \\ & T = sY + \sqrt{1-s^2}\,\varepsilon_T, \qquad Z = rT + \sqrt{1-r^2}\,\varepsilon_Z. \end{aligned} \tag{53}$$

The respective graphs are $Z \to T \to Y$ and $Y \to T \to Z$, with an independent nuisance covariate X and no hidden confounding. Both SCMs induce the same zero-mean Gaussian observational distribution: X is independent of (Z, T, Y), whose covariance matrix is

$$\Sigma_{ZTY} = \begin{pmatrix} 1 & r & rs \\ r & 1 & s \\ rs & s & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.65 & 0.585 \\ 0.65 & 1 & 0.9 \\ 0.585 & 0.9 & 1 \end{pmatrix}. \tag{54}$$

Consequently, no amount of observational data can distinguish these two SCMs by their likelihood. We generate all empirical contexts from $\psi_{\rightarrow}$; $\psi_{\leftarrow}$ explains the ambiguity rather than serving as a second data-generating condition.

Fixed conditional query. Although our main experiments use row-specific randomised query interventions $\mathrm{do}(T = t_i)$, we fix the query here to $p(y \mid \mathrm{do}(T = 2), Z = z_\star, X = 0)$, so the conditioning vector in the main-text notation is $\mathbf{x}_\star = (z_\star, 0)$. We set $z_\star = 0.58335$, the root closest to zero of

$$\varphi(z_\star; 0, 1) = \varphi(z_\star; 1.3, 0.5775), \tag{55}$$

where $\varphi(\cdot; \mu, v)$ denotes a Gaussian density with mean μ and variance v. These are precisely the two densities of Z under $\mathrm{do}(T = 2)$. This choice makes the query covariates equally likely under the two intervened SCMs, so conditioning on them does not by itself favour one SCM. Nevertheless, their CIDs differ:

$$\begin{aligned} q_\star(y) := p_{\psi_{\rightarrow}}(y \mid \mathrm{do}(T = 2), z_\star, 0) = \mathcal{N}(y; 1.8, 0.19), \\ p_{\psi_{\leftarrow}}(y \mid \mathrm{do}(T = 2), z_\star, 0) = \mathcal{N}(y; 0, 1). \end{aligned} \tag{56}$$

In the reverse SCM, intervening on T removes $Y \to T$, leaving Y independent of Z and X. Under an illustrative equal prior over exactly these two SCMs, observational data and the chosen query covariates therefore leave an equal-weight mixture of these two CIDs.

Why a surrogate experiment helps. Under $\mathrm{do}(Z = z)$, the two SCMs instead imply

$$\mathbb{E}_{\psi_{\rightarrow}}[Y \mid \mathrm{do}(Z = z)] = rsz = 0.585z, \qquad \mathbb{E}_{\psi_{\leftarrow}}[Y \mid \mathrm{do}(Z = z)] = 0. \tag{57}$$

Thus a randomised Z experiment supplies information that further observational samples cannot provide about this pair: The ambiguity here comes from the unknown graph direction.

I.2 Paired contexts and frozen models

For each replication we generate 256 observational rows from the forward SCM and 64 independent surrogate-interventional rows. Each experimental row sets $Z_i \sim \mathrm{Uniform}[-4, 4]$, independently of the exogenous noises, and then samples the remaining variables according to the SCM. These rows form one $\mathrm{do}(Z)$ regime with row-specific assignments. No row intervenes on T or Y.

The primary comparison uses an *extra-intervention* protocol: the same 256 observational rows are retained and the 64 experimental rows are added. To distinguish source information from sample-count effects, we also generate a nested 320-row observational context whose first 256 rows are identical to the primary context. This yields an equal-total-budget comparison between 320 observational rows and $256 + 64$ observational and experimental rows.

We evaluate the following frozen models without training or fine-tuning:

- **CIDER-FM:** the same Linear–Gaussian checkpoint (training step 95,000, 128-bar output) receives 256 Obs, 320 Obs, or 256 Obs plus 64 $\mathrm{do}(Z)$ rows. The observational controls are inference-only uses of the same weights, not separately trained CIDER-FM-Obs.
- **Graphs4CFM:** the released fully-conditioned checkpoint of Reuter et al. (2026), with a 1,000-bar output, receives either 256 or 320 observational rows. All graph-relation information is marked unknown.

The 256/320-row observational regimes exceed CIDER-FM's training allocation of 64 rows per regime; its observational-only inputs also constitute source-availability extrapolation. These runs therefore serve as same-checkpoint source-input controls, not as separately trained, in-distribution observational baselines.

We froze the SCM, query, weights and budgets before evaluating 16 new independent context/noise seeds. All replications are retained. The first seed was specified for Fig. 2 before this confirmation; the figure is not an average of densities across seeds or queries.

I.3 Distributional scores and quantitative results

Let $\widehat{q}_d$ and $\widehat{F}_d$ denote the predictive density and CDF for context draw d, and let $F_\star$ be the CDF of $q_\star = \mathcal{N}(1.8, \sigma_\star^2)$, with $\sigma_\star^2 = 0.19$. The expected negative log-likelihood (NLL) and continuous ranked probability score (CRPS) under $q_\star$ are

$$\begin{aligned} \mathrm{NLL}_d &= -\int_{\mathbb{R}} q_\star(y) \log \widehat{q}_d(y)\, \mathrm{d}y, \\ \mathrm{CRPS}_d &= \int_{\mathbb{R}} \left(\widehat{F}_d(u) - F_\star(u)\right)^2 \mathrm{d}u + \frac{\sigma_\star}{\sqrt{\pi}}. \end{aligned} \tag{58}$$

We also report the equal-tailed 90% predictive-interval width and coverage:

$$\begin{aligned} \ell_d &= \widehat{F}_d^{-1}(0.05), \qquad u_d = \widehat{F}_d^{-1}(0.95), \\ W_d &= u_d - \ell_d, \qquad C_d = F_\star(u_d) - F_\star(\ell_d). \end{aligned} \tag{59}$$

Here, W_d measures concentration, while C_d assesses coverage against the nominal level of 0.90.

All scores use the unsmoothed bar distributions in original outcome units, including their tails. We compute NLL analytically under the Gaussian truth and CRPS by numerical quadrature. Table 11 reports mean $\pm$ SE over 16 independent context draws, with $\mathrm{SE} = s_{\mathrm{draw}}/\sqrt{16}$ and s_{draw} the sample standard deviation. In this case, surrogate data improve NLL and CRPS and narrow predictive intervals while maintaining near-nominal coverage. The improvement exceeds that obtained from the same number of additional observational samples.

I.4 Visualisation and scope

For Fig. 2, we smooth each model's bar distribution using a Gaussian KDE fitted to $n = 16{,}384$ stratified inverse-CDF samples, including the tails. All models use Scott's bandwidth rule, $h = s_Y n^{-1/5}$, where s_Y is the sample standard deviation of the predictive draws. The black dashed curve

Table 11: Fixed-query case study (mean ± SE over 16 context draws). O denotes observational rows and I_Z randomised $\mathrm{do}(Z)$ rows. All CIDER-FM rows use the same checkpoint. Nominal coverage is 0.90; the true CID provides an analytic reference.

Model	Context	NLL ↓	CRPS ↓	90% width	Coverage
Graphs4CFM	$256O$	$1.5301 \pm .0619$	$.5864 \pm .0287$	$3.0828 \pm .0702$	$.7903 \pm .0298$
Graphs4CFM	$320O$	$1.3926 \pm .0500$	$.5622 \pm .0245$	$3.1862 \pm .0596$	$.8539 \pm .0223$
CIDER-FM	$256O$	$1.0544 \pm .0186$	$.4404 \pm .0113$	$3.3288 \pm .0233$	$.9260 \pm .0058$
CIDER-FM	$320O$	$1.0383 \pm .0142$	$.4313 \pm .0085$	$3.3268 \pm .0256$	$.9287 \pm .0046$
CIDER-FM	$256O + 64I_Z$	$\mathbf{.7852} \pm .0124$	$\mathbf{.2977} \pm .0055$	$2.5854 \pm .0467$	$.9277 \pm .0049$
True CID	—	.5886	.2459	1.4340	.9000

shows the analytic true CID. KDE smoothing is used only for visualisation. This case illustrates how surrogate experiments can sharpen conditional interventional predictions for a fixed SCM and query, with the results replicated across independent context draws.

J MODEL HYPERPARAMETERS

The reported model was trained using a single NVIDIA A100-SXM4 80GB GPU. A complete 100,000-step training run took approximately 15 GPU-hours.

Table 12: Key hyperparameters of the reported ComplexMech CIDER-FM model.

Hyperparameter	Value
Model configuration	
Transformer layers	4
Hidden dimension	128
Attention heads	4
Latent memory rows per regime–variable pair	8
Predictive distribution	Full-support bar distribution with half-Gaussian tails
Interior bars	128
Trainable parameters	8,150,032
Optimisation	
Training objective	Original-scale negative log-likelihood of Y_q only
Optimizer	AdamW
Base learning rate	3×10^{-4}
Weight decay	10^{-5}
Effective batch size	32 tasks
Warm-up steps	10,000
Learning-rate schedule	Cosine decay to 0.1 times the base learning rate
Total optimisation steps	100,000
Training-prior schedule	10,000 causal-surrogate curriculum steps followed by 90,000 natural-prior steps
Synthetic task prior	
Mechanism family	ComplexMech
Node-count prior	$N \sim \mathrm{Uniform}\{4, \ldots, 10\}$
Graph prior	Random ADMG with bows allowed
Directed-edge probability prior	$p_{\rightarrow} \sim \mathrm{Beta}(2, 3)$
Bidirected-edge probability prior	$p_{\leftrightarrow} \sim \mathrm{Beta}(2, 3)$
Interventional-regime-count prior	$p(K = k \mid N) \propto 2^{-(k-1)}, k = 1, \ldots, N-2$
Observational-regime probability	0.8
Direct-$\mathrm{do}(T)$-regime probability	0.1
Rows per active context regime	64
Query rows	128
Model selection	
Validation suite	560 held-out SCM tasks
Checkpoint criterion	Equal percentile regret over validation Y_q NLL, MSE, and 5%-trimmed task-wise R^2; the test split is never used for selection
Reported checkpoint	Step 96,000