FOUNDATION MODELS FOR CAUSAL INFERENCE VIA PRIOR-DATA FITTED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-data fitted networks (PFNs) have recently been proposed as a promising way to train tabular foundation models. PFNs are transformers that are pre-trained on synthetic data generated from a prespecified prior distribution and that enable Bayesian inference through in-context learning. In this paper, we introduce *CausalFM*, a comprehensive framework for training PFN-based foundation models in various causal inference settings. First, we formalize the construction of Bayesian priors for causal inference based on structural causal models (SCMs) in a principled way and derive necessary criteria for the validity of such priors. Building on this, we propose a novel family of prior distributions using causality-inspired Bayesian neural networks that enable CausalFM to perform Bayesian causal inference in various settings, including for back-door, front-door, and instrumental variable adjustment. Finally, we instantiate CausalFM and explicitly train models to perform in-context learning in these settings. We show that CausalFM achieves competitive in-context learning performance even when compared to baselines that are specifically trained for the task at hand. In sum, our framework can be used as a general recipe to train foundation models for various causal inference settings. In contrast to the current state-of-the-art in causal inference, CausalFM offers a novel paradigm with the potential to fundamentally change how practitioners perform causal inference in medicine, economics, and other disciplines.

1 INTRODUCTION

Causal inference is a cornerstone of empirical research in disciplines such as economics (Angrist, 1990; Imbens & Angrist, 1994), medicine (Feuerriegel et al., 2024; Weberpals et al., 2025), and marketing (Varian, 2016). It enables the estimation of causal effects from observational and randomized data, which is essential for reliable decision-making (Kern et al., 2025). In personalized medicine, for instance, it supports identifying the most effective treatment by predicting patient outcomes under different therapeutic options.

In recent years, machine learning, and especially deep learning methods, have gained significant traction in causal inference (Curth & van der Schaar, 2021; Ma et al., 2025; 2024; Melnychuk et al., 2022; Schweisthal et al., 2023; Shalit et al., 2017a; Shi et al., 2019)). These methods offer several advantages for causal effect estimation in practice, including the ability to handle large-scale and high-dimensional datasets with complex confounding structures and to model heterogeneity of causal effects (Feuerriegel et al., 2025). However, most existing approaches require retraining a model for each new dataset. To this end, existing approaches lack the flexibility to perform inference for new datasets without additional retraining, which limits their practicality in real-world settings.

Meanwhile, foundation models have emerged as a transformative paradigm in machine learning (Devlin, 2018; Lahat et al., 2024; Touvron et al., 2023b;a), which offer a key advantage in that they allow for flexible, test-time inference without retraining. These models are pre-trained on large datasets and can generalize across tasks and domains. Examples include large language models (LLMs) in natural language processing and vision transformers in computer vision. However, this paradigm shift toward test-time inference has not yet had a comparable impact on causal inference. Most current approaches in causal machine learning still rely on specialized models tailored to specific tasks, requiring practitioners to manually select, train, and validate appropriate estimation methods for each new dataset.

In this paper, we propose a change to the paradigm for causal inference based on the idea of foundation models trained for tabular causal inference. For this, we build on the recently proposed prior-data fitted networks (PFNs) (Müller et al., 2022; Hollmann et al., 2023), which are transformers pre-trained on purely synthetic datasets generated from a prespecified prior distribution. PFNs enable Bayesian inference purely through in-context learning, allowing for flexible and efficient predictions without requiring additional training for new tasks (Nagler, 2023). While recent works have demonstrated the effectiveness of tabular foundation models based on PFNs for various tasks, only two concurrent works have proposed PFNs tailored for causal inference (Balazadeh et al., 2025; Robertson et al., 2025). However, these are either restricted to specific causal inference settings (namely, **only** back-door adjustment) or do **not** offer identifiability guarantees.

We introduce CausalFM, a comprehensive framework for training PFN-based foundation models for various causal inference settings. For this purpose, we introduce CausalFM priors: a novel family of prior distributions based on structural causal models respecting the underlying causal inference problem at hand. We first formalize and derive necessary criteria on how to construct such SCM-based priors for causal inference in principle. Then, we propose a concrete instantiation using Bayesian neural networks and provide a learning algorithm that leverages the SCM's ability to simulate interventional data to perform Bayesian inference in various causal inference settings.

Compared to classical causal inference methods, models trained based on our CausalFM offer the following advantages: (i) There is **no** need for additional training for new datasets as our CausalFM performs inference *entirely through in-context learning*, enabling fast and flexible deployment across new datasets. (ii) The Bayesian nature of our CausalFM provides *principled uncertainty quantification*, which is critical for downstream decision-making and for detecting situations with poor treatment overlap. (iii) The model *automatically* learns to "select" an identifiability formula based on the data distribution and task at hand. (iv) Our CausalFM builds upon rigorous identifiability guarantees to ensure valid causal inference.

Our **contributions**[1] are: (1) We formalize the constructions of priors based on structural causal models (SCMs) for Bayesian causal inference and derive necessary conditions for their validity. (2) We propose an explicit *CausalFM* prior based on Bayesian neural networks that are compatible with the structure of the causal inference problem at hand. We also propose a learning algorithm to train PFNs for causal inference problems that leverages our CausalFM prior to simulate counterfactuals to mitigate the fundamental problem of causal inference. (3) We propose concrete instantiations of our framework by training PFNs for estimating conditional average treatment effects (CATEs) in different causal inference settings. We show empirically that CausalFM performs competitively and outperforms current state-of-the-art CATE estimators on a variety of benchmarks.

2 Related Work

We provide an overview of related literature streams. Additional related work is in Appendix A.

Amortized causal inference. Several recent papers pre-train large neural networks on synthetic data so that they can solve causal tasks via in-context learning. Examples include causal discovery (Mahajan et al., 2025), ATE estimation under unconfoundedness (Zhang et al., 2024), zero-shot- and few-shot learning (Nilforoshan et al., 2023; Iwata & Chikahara, 2023), and reinforcement-learning (Lee et al., 2023). These methods validate the feasibility of treating causal inference as an in-context learning problem but remain restricted to specific causal inference settings, which typically do **not** allow accommodating unobserved confounding.

Black-box causal inference (BBCI) (Bynum et al., 2025) proposes synthetically-pretrained models to perform causal inference in a variety of settings. However, their approach is different: (i) BBCI does *not* build upon a Bayesian framework. In contrast, building upon PFNs allows us to perform approximate Bayesian causal inference and thus provide rigorous uncertainty quantification. (ii) The proposed data-generating processes in BBCI are *not* tailored for high-dimensional causal inference settings (as the authors mention in their Sec. 7). In contrast, our CausalFM prior leverages Bayesian neural networks inspired by TabPFN (Hollmann et al., 2023) to create SCM-based prior distributions. (iii) Beyond proposing a new method, we provide novel formalizations and theoretical results of constructing valid SCM-based priors for Bayesian causal inference.

[1] Code is available at `https://anonymous.4open.science/r/causal_foundation_model`.

PFNs for causal inference: We are aware of only two concurrent works that propose PFN-based models for causal inference, but each with clear limitations (see Figure 1): (i) (Balazadeh et al., 2025) proposes a PFN similar to ours, but it is restricted to **only** back-door adjustment, i.e., imposes the unconfoundedness assumption throughout their paper. In contrast, we propose a framework for constructing PFN-based foundation models for a *large* class of causal inference problems, *including both front-door adjustment and instrumental variable settings with unobserved confounding*. (ii) Robertson et al. (2025) proposes to train a *single* PFN on various different causal inference settings *without* providing identifiability assumptions to the model. We will show later that the approach of Robertson et al. (2025) has a crucial drawback: because the causal quantity of interest is **not** identified, the PFN learns a posterior that *may never concentrate around the true causal quantity, thus leading to asymptotically non-informative estimators*. In contrast, we propose to infuse our PFNs with identifiability assumptions required for informative causal inference. As such, we follow established philosophy in causal inference that separates identifiability and estimation steps (Kern et al., 2025; Pearl, 2009): the identifiability step should be established by the practitioner using domain knowledge (e.g., establishing whether a certain variable is a valid instrument), while the estimation step can be treated as a purely statistical learning problem.

Table 1: **Overview of identifiability of PFN-based frameworks for causal inference**.

Framework	Backdoor	Frontdoor	IV
CausalPFN Balazadeh et al. (2025)	✓	✗	✗
Do-PFN Robertson et al. (2025)	✓	✗	✗
Ours (CausalFM)	✓	✓	✓

3 Problem Setup

3.1 Background on PFNs

In tabular prediction problems, one considers a population $(X, Y) \sim \mathbb{P} \in \mathcal{P}$. Given a finite sample $\mathcal{D}_n \sim \mathbb{P}$ of size n, the goal is to estimate the conditional distribution $\mathbb{P}(Y = y \mid X = x)$. PFNs formulate this task in a Bayesian non-parametric way by placing a prior distribution Π on $\mathcal{P}$, i.e., a *prior over data-generating distributions* (Müller et al., 2022; Nagler, 2023). Sampling proceeds hierarchically via $\mathbb{P} \sim \Pi$ and i.i.d. data $(X_i, Y_i) \sim \mathbb{P}$. Then, Bayes' rule yields the posterior distribution $\Pi(\mathbb{P} \mid \mathcal{D}_n) \propto \Pi(\mathcal{D}_n \mid \mathbb{P})\,\Pi(\mathbb{P})$, where $\Pi(\mathcal{D}_n \mid \mathbb{P})$ is the likelihood of the sample $\mathcal{D}_n$ under $\mathbb{P}$ and $\propto$ denotes proportionality up to a multiplicative constant. The corresponding *posterior-predictive distribution* is the probability of Y given test point x and observed data $\mathcal{D}_n$, i.e.,

$$\Pi(Y \mid \mathcal{D}_n, x) = \int \mathbb{P}(Y \mid X = x)\,\Pi(\mathbb{P} \mid \mathcal{D}_n)\;\mathrm{d}\mathbb{P}. \tag{1}$$

PFNs are neural networks $q_\theta(Y \mid \mathcal{D}_n, x)$ that parameterize the family of predictive posterior distributions with trainable parameters θ. That is, PFNs map the entire dataset $\mathcal{D}_n$ and a query x to a distribution over $\mathcal{Y}$. In terms of architecture, PFNs are permutation-equivariant transformers (Ashish Vaswani et al., 2017) as they allow for scalable training and leverage the attention mechanism to effectively extract information from $\mathcal{D}_n$. PFNs are trained by minimizing the negative log-likelihood loss $\mathcal{L}(\theta) = \mathbb{E}_{N\sim\Pi_N}\big[\mathbb{E}_{\mathbb{P}\sim\Pi}\big[-\log q_\theta(Y \mid X, \mathcal{D}_N)\big]\big]$, where Π_N is a prior on the sample sizes. In practice, we sample a sample size $N_j \sim \Pi_N$, a probability distribution $\mathbb{P}_j \sim \Pi$, a dataset $\mathcal{D}^j_{N_j} \sim \mathbb{P}_j$, and test points $(x_j, y_j) \sim \mathbb{P}_j$ and then approximate the PFN loss via

$$\hat{\mathcal{L}}(\theta) = \sum_j \big[-\log q_\theta(y_j \mid \mathcal{D}^j_{N_j}, x_j)\big], \tag{2}$$

which is consistent for the exact posterior-predictive under regularity conditions (Nagler, 2023). Note that *all* training data are synthetic, i.e., sampled from the prior Π. Furthermore, the trained PFN can be deployed on arbitrary real datasets *without* further training.

3.2 Task: causal inference

In this paper, we aim to extend PFNs to causal inference. Here, the main challenge is that the object of interest is an *interventional*[2] distribution $\mathbb{P}_{\text{int}}$, yet we only observe data $\mathcal{D}_n \sim \mathbb{P}_{\text{obs}}$ from a potentially different *observational* distribution (Pearl, 2009).

[2]Causal literature often distinguishes between interventional and counterfactual distributions. This is not relevant for the methods of our paper, and we thus use interventional distribution as an umbrella term.

Motivation. As an illustrative example, we consider a standard causal inference setting, called *backdoor adjustment*, where the data comprise $(X, A, Y) \sim \mathbb{P}_{\text{obs}}$, where X are patient covariates, A is a treatment, and Y is an outcome of interest (van der Laan & Rubin, 2006). For example, in medicine, X may contain treatment history or demographic attributes, A may be a medical treatment, and Y a health outcome. Following the potential outcome framework (Rubin, 1974), let $Y(a)$ denote the outcome that would be realized under the treatment $A = a$. The interventional distribution is thus over $(X, A, Y(1) - Y(0)) \sim \mathbb{P}_{\text{int}}$, and a common target functional is the *conditional average treatment effect (CATE)* $Q(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big]$ (Wager & Athey, 2018). The CATE quantifies the expected benefit of providing treatment given the patient's covariates.

Identifiability. To estimate CATE from observational data, we need to impose *identifiability assumptions*, which link the observational and the interventional distributions and allow us to express Q as a functional of the observational distribution (Rosenbaum & Rubin, 1983). These are (i) consistency: $Y(A) = Y$, (ii) positivity: $\mathbb{P}_{\text{obs}}(A = 1 \mid X = x) > 0$, and (iii) Unconfoundedness: $Y(1), Y(0) \perp\!\!\!\perp A \mid X$ in $\mathbb{P}_{\text{int}}$.

Generalized causal inference setting. In the following, we provide a generalized definition of a causal inference setting, that allows us to reason about arbitrary causal inference settings and provide generalized statements beyond the standard example above.

Definition 3.1. We define a *causal inference setting* is a tuple $\mathcal{C} = \big(O, \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}, Q\big)$, where O collects the observed variables (and contains at least A and Y); $(\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}}) \in \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ are paired observational/interventional distributions over O that correspond to an intervention on A; and $Q(\mathbb{P}_{\text{int}})$ is a causal query that is *identifiable*, i.e. there exists a measurable functional $\bar{Q}$ such that $Q(\mathbb{P}_{\text{int}}) = \bar{Q}(\mathbb{P}_{\text{obs}})$ for all $\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}}) \in \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$.

3.2.1 Running examples

■ **Example 1 (back-door adjustment).** Here, we continue the example from above and define $O = (X, A, Y) \sim \mathbb{P}_{\text{obs}}$ with binary A as above. $\mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ contains all observational and interventional distributions that satisfy consistency, positivity, and unconfoundedness. The causal query is the CATE $Q(\mathbb{P}_{\text{int}})(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, which is identified as

$$\bar{Q}(\mathbb{P}_{\text{obs}})(x) = \mathbb{E}_{\mathbb{P}_{\text{obs}}}[Y \mid A = 1, X = x] - \mathbb{E}_{\mathbb{P}_{\text{obs}}}[Y \mid A = 0, X = x]. \tag{3}$$

■ **Example 2 (front-door adjustment).** Let $O = (X, A, M, Y) \sim \mathbb{P}_{\text{obs}}$, where X, A, and Y are defined as above and M is a mediator between A and Y. Interventional distributions are defined using potential outcomes, i.e., $\big(X,\ A,\ M(1), M(0),\ Y(1, M(1)), Y(0, M(0))\big) \sim \mathbb{P}_{\text{int}}$, and the causal query of interest again the CATE $Q(\mathbb{P}_{\text{int}})(x) = \mathbb{E}_{\mathbb{P}_{\text{int}}}\big[Y(1, M(1)) - Y(0, M(0)) \mid X = x\big]$.

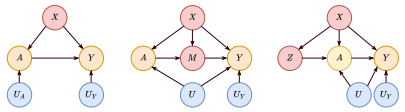

Figure 1: **C-DAGs compatible with the three example causal inference settings.** Yellow variables are observed, blue variables are unobserved, and red variables are clusters of variables.

Identifiability assumptions. We restrict to pairs $(\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}})$ that satisfy (i) consistency: $Y = Y(A, M)$ and $M = M(A)$; (ii) positivity: $\mathbb{P}_{\text{obs}}(A = a \mid X = x) > 0$ and $\mathbb{P}_{\text{obs}}(M = m \mid A = a, X = x) > 0$; and (iii) front-door criterion $M(a) \perp\!\!\!\perp A \mid X = x$, and $Y(a', m) \perp\!\!\!\perp M \mid A = a', X = x$. Under these assumptions, the CATE is identified and $\widetilde{Q}$ is given via the conditional version of Pearl's front-door adjustment formula (Pearl, 2009).

■ **Example 3 (Instrumental variables).** Let $O = (X, Z, A, Y) \sim \mathbb{P}_{\text{obs}}$, where Z is an instrumental variable that causes the treatment A but does not directly cause the outcome Y. The interventional distribution is defined on $\big(X, Z, A, Y(1), Y(0)\big) \sim \mathbb{P}_{\text{int}}$ for a fixed treatment intervention $A = a$. We are again interested in the CATE $Q(\mathbb{P}_{\text{int}})(x) = \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big]$.

Identifiability assumptions. We restrict to pairs $(\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}})$ that satisfy the following conditions (Newey & Powell, 2003): (i): Additive structural equation: $Y = f(X, A) + g(X, U)$, with (unknown) functions f and g and unobserved confounder U, implying that Y does not directly depend on Z; (ii) Independence: $U \perp\!\!\!\perp Z \mid X$; (iii) Relevance: $\mathbb{P}_{\text{obs}}(A \mid X = x, Z = z) > 0$ is non-constant in z; and (iv) Completeness: For every measurable g, if $\mathbb{E}_{\mathbb{P}_{\text{obs}}}[f(x, A) \mid X = x, Z = z] = 0$ for all z, then $f(x, A) = 0$ almost surely in $\mathbb{P}_{\text{obs}}$. Then, the CATE can be shown to be identified via an integral equation (Newey & Powell, 2003; Hartford et al., 2017).

Research question: PFNs have shown to be an effective way to construct tabular foundation models. However, a causal inference setting $\mathcal{C}$ comes with additional challenges, such as the distinction between observational and interventional distribution as well as identifiability assumptions.

Research question

How can we train PFNs for a causal inference setting $\mathcal{C}$ that provides a Bayesian estimator of $Q(\mathbb{P}_{\text{int}})$ given an observational dataset $\mathcal{D}_n \sim \mathbb{P}_{\text{obs}}$ and some context (e.g., values x or a)?

In the following, we introduce CausalFM consisting of (i) appropriate prior distributions that allow for approximating *interventional* predictive posterior distributions as in Eq. (2)(Section 4) and (ii) a training algorithm for the underlying PNFS (see Section 5).

4 CausalFM: Priors

In this section, we construct prior distributions for CausalFM which are based upon identifiable structural causal models (SCMs). We motivate and formalize our approach (Sec. 4.1, provide necessary criteria for valid causal inference (Sec. 4.2), and finally provide a method for constructing such priors in practice (Sec. 4.3). We also provide a complete toy example in Appendix B.

4.1 Introducing SCM-based priors

Naïve approach. A naïve approach for causal inference would construct a prior Π directly for the observational distribution $\mathbb{P}_{\text{obs}}$. If the posterior $\Pi(\mathbb{P}_{\text{obs}} \mid \mathcal{D}_n) \to \mathbb{P}^*_{\text{obs}}$ converges to the ground-truth observational distribution $\mathbb{P}^*_{\text{obs}}$ (i.e., satisfying a Bernstein-von-Mises theorem), we can obtain a consistent Bayesian estimator of our causal query via $\bar{Q}(\Pi(\mathbb{P}_{\text{obs}} \mid \mathcal{D}_n))$. Accordingly, we *could* train a PFN $q_\theta(Y \mid \mathcal{D}_n, x)$ with the loss in Eq. (2) and estimate the CAPO via $\bar{Q}(q_\theta(Y \mid \mathcal{D}_n, x))$.

However, the above approach has *drawbacks*: (i) It requires knowledge of the identification formula $\bar{Q}$, which must be determined on a case-by-case basis depending on the causal inference setting $\mathcal{C}$ at hand. This can be tedious or even hard to compute in practice. For example, the IV setting from Example 3 requires solving integral equations to compute $\bar{Q}$ (Newey & Powell, 2003). (ii) Constructing a prior for $\mathbb{P}_{\text{obs}}$ makes it harder control the distribution of the causal query Q directly. It has been shown in the literature that this can lead to prior misspecification for Bayesian causal inference or slowly converging posterior distributions (Linero & Antonelli, 2022).

Modeling the interventional distribution. Motivated by these drawbacks of constructing priors for only $\mathbb{P}_{\text{obs}}$, we propose to construct priors for *observational-interventional distribution pairs* $(\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}})$, resulting in priors defined on $\mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$. This addresses both drawbacks by (i) inducing an *interventional posterior distribution*, thus only requiring knowledge of Q (not $\bar{Q}$); and (ii) we will see that priors on $\mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ often allow to specify the prior distribution of $Q(\mathbb{P}_{\text{int}}))$ directly. A natural way to define distributions on $\mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ is via structural causal models (SCMs).

Definition 4.1 (SCMs (Pearl, 2009))**.** A (semi-Markovian) *structural causal model (SCM)* $\mathcal{S}$ is a tuple $\big(Z, U, f, \mathbb{P}\big)$, where $\mathbf{Z} = (Z_1, \dots, Z_k)$ are observable **endogenous** variables, U collects latent **exogenous** variables, $f = \{f_{Z_1}, \dots, f_{Z_k}\}$ contains structural assignments $Z_i = f_{Z_i}(pa(Z_i))$ with parents $pa(Z_i) \subseteq Z \cup U$, and $\mathbb{P}$ is a joint distribution on U.

Every SCM induces a unique directed acyclic graph (DAG), $\mathcal{G}^{\mathcal{S}}$ by defining mapping of the parents $pa(Z_i)$ to Z_i with directed edges. We distinguish two types of latent variables U_i in $\mathcal{G}^{\mathcal{S}}$: U_i is an *unobserved confounder* if it is the parent of both A and Y, otherwise, we call it a *noise variable*. Intuitively, an SCM is a simulator: we can draw latent variables $U \sim \mathbb{P}$ and pass them through structural functions f, resulting in an induced observational distribution $\mathbb{P}^{\mathcal{S}}_{\text{obs}}$ over Z. At the same time, we can modify the SCM by intervening on a variable via $do(A = a)$, i.e., fixing the variable and then sampling from the SCM mechanism. This induces a corresponding interventional distribution $\mathbb{P}^{\mathcal{S}}_{\text{int}}$. We call an SCM $\mathcal{S}$ *compatible* with a causal inference setting $\mathcal{C}$, if $(\mathbb{P}^{\mathcal{S}}_{\text{obs}}, \mathbb{P}^{\mathcal{S}}_{\text{int}}) \in \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$.

Definition 4.2 ($\mathcal{C}$-SCM-Priors)**.** A *$\mathcal{C}$-SCM-Prior* is any probability measure $\Pi(\mathcal{S})$ that puts all its mass on SCMs compatible with $\mathcal{C}$. Via the map $\mathcal{S} \mapsto (\mathbb{P}^{\mathcal{S}}_{\text{obs}}, \mathbb{P}^{\mathcal{S}}_{\text{int}})$ every such prior induces a distribution $\Pi\big((\mathbb{P}_{\text{obs}}, \mathbb{P}_{\text{int}})\big)$ on $\mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$.

Sampling from Π therefore amounts to sampling a random latent distribution $\mathbb{P}$ over U as well as random functional assignments f. These can then be used to internally sample an observational dataset $\mathcal{D}_n$, i.e., there is a well-defined likelihood $\Pi(\mathcal{D}_n \mid \mathcal{S})$ induced by $\mathcal{S}$. As a consequence, we can define the posterior distribution over SCMs via $\Pi(\mathcal{S} \mid \mathcal{D}_n) \propto \Pi(\mathcal{D}_n \mid \mathcal{S})\Pi(\mathcal{S})$, where $\propto$ denotes proportionality up to a normalization constant.

Cluster-DAGs. Because an SCM-prior induces a distribution over possibly many DAGs, we compress them into a shared structure. Given variables (Z, U), a Cluster-DAG (C-DAG) (Anand et al., 2023)] is a DAG on clusters $C_1, \ldots, C_k$ which are disjoint subsets of (Z, U). Each $\mathcal{C}$-SCM-Prior induces a unique C-DAG via the following algorithm: (i) draw an edge $C_i \to C_j$ whenever any SCMs $\mathcal{S}$ with $\Pi(\mathcal{S}) > 0$ contains some arrow from any node of C_i to any node of C_j *and* no SCM $\mathcal{S}$ with $\Pi(\mathcal{S}) > 0$ contains some arrow from any node of C_j to any node of C_i; (ii) merge C_i and C_j whenever both directions occur across SCMs $\mathcal{S}$ with $\Pi(\mathcal{S}) > 0$.

4.2 Well-specified priors

The question is now how we should design our prior Π such that the induced posterior $\Pi(\mathcal{S} \mid \mathcal{D}_n)$ allows for valid Bayesian causal inference. We now define a key desirable property of such priors. For this, we call a prior $\Pi(\mathcal{S})$ *well-specified* for $\mathcal{C}$ if, for any true pair $(\mathbb{P}^*_{\text{obs}}, \mathbb{P}^*_{\text{int}})$ and every sample $\mathcal{D}_n \sim \mathbb{P}^*_{\text{obs}}$, it holds that

$$Q\Big(\int \mathbb{P}^{\mathcal{S}}_{\text{int}} \, \Pi(\mathcal{S} \mid \mathcal{D}_n) \, \mathrm{d}\mathcal{S}\Big) \;\longrightarrow\; Q(\mathbb{P}^*_{\text{int}}), \quad n \to \infty. \tag{4}$$

In other words, a well-specified prior ensures that the causal query Q evaluated on the *posterior-predictive interventional distribution* (PPID) $\int \mathbb{P}^{\mathcal{S}}_{\text{int}} \, \Pi(\mathcal{S} \mid \mathcal{D}_n) \, \mathrm{d}\mathcal{S}$ is a consistent estimator of the causal target. If we were able to train a PFN to approximate the PPID of a well-specified prior, we are sure that we can apply Q on this distribution and obtain a consistent estimator.

Identifiability. At this point, one may wonder why we only focus on priors for *identifiable* causal inference settings. Indeed, a recently proposed method, called *do-PFN* (Robertson et al., 2025), does not restrict its PFN priors to identifiable settings. The following result shows that, under weak assumptions, such priors *cannot* be well-specified, leading to asymptotic inconsistency.

Theorem 4.3. *Let $\mathcal{Z}$ be the set of all identifiability-violating SCMs $\mathcal{S}_0$ that satisfy $\mathbb{P}^{\mathcal{S}_0}_{\text{obs}} \in \mathcal{P}_{\text{obs}}$ and $Q\big(\mathbb{P}^{\mathcal{S}_0}_{\text{int}}\big) \neq \bar{Q}\big(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}}\big)$. Assume that Q is a linear functional (e.g., the CATE) and that $\int_{\mathcal{Z}} Q(\mathbb{P}^{\mathcal{S}}_{\text{int}}) \, \Pi(\mathcal{S}) \, \mathrm{d}\mathcal{S} \neq \bar{Q}\big(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}}\big)$ (non-identifiability doesn't cancel out). Then, if $\Pi(\mathcal{S})$ is well-specified for $\mathcal{C}$, it follows that $\Pi(\mathcal{S} \in \mathcal{Z}) = 0$.*

Proof. See Appendix C. □

4.3 Constructing SCM-based priors

C-DAG design. Our method for constructing priors assumes the knowledge of a well-specified C-DAG $\mathcal{G}_c$ for $\mathcal{C}$, meaning that $\mathcal{G}_c$ is induced by some well-specified $\mathcal{C}$-SCM-Prior. Such C-DAGs are usually known for most causal inference settings (see Fig. 1 for C-DAGs compatible with the settings in Examples 1–3).

One point of ambiguity is the modeling of noise variables in C-DAGs. Here, we propose a practical design rule: if $\mathcal{G}_c$ contains an unobserved confounder between A and Y, we only add one additional noise variable to *either* A or Y. Conversely, if $\mathcal{G}_c$ is unconfounded, we add noise parents to *both* A and Y (see Fig. 1). The reasoning is as follows: if $\mathcal{G}_c$ is unconfounded, we need to add noise to both A and Y in order to ensure not restrict ourselves to degenerate observational distributions. Conversely, any unobserved confounder U induces noise into both A and Y, thus removing the need to add noise to both. However, it is still necessary to add *one* additional noise variable to either A or Y since, otherwise, any unconfounded SCM compatible with $\mathcal{G}_c$ would need to be degenerate in either A or Y. We provide a concrete toy example in Appendix B to illustrate this.

Prior construction. We now propose a practical algorithm to construct $\mathcal{C}$-SCM-priors. We assume that we have access to a pair $(\mathcal{G}_c, \mathcal{I})$, where $\mathcal{G}_c$ is a well-specified C-DAG for $\mathcal{C}$ and $\mathcal{I}$ is a set of constraints on SCMs $\mathcal{S}$ compatible with $\mathcal{G}_c$ ensuring that $\mathcal{S}$ is also compatible with $\mathcal{C}$.

■ *Example 1: back-door adjustment.* The observable variables are (X, A, Y) together with noise variables. A compatible C-DAG is in Fig. 1 (left). The constraint set is $\mathcal{I}(\mathcal{S}) = \{\mathbb{P}^{\mathcal{S}}_{\text{obs}}(A = a \mid X = x) > 0\}$, ensuring that all SCMs satisfy the positivity assumption.

■ *Example 2: Front-door adjustment.* Here, the observed variables are (X, A, M, Y) with noise variables and an unobserved confounder U between A and Y. A compatible C-DAG is in Fig. 1 (middle). The constraint set is $\mathcal{I}(\mathcal{S}) = \{\mathbb{P}^{\mathcal{S}}_{\text{obs}}(A = a \mid X = x) > 0, \mathbb{P}^{\mathcal{S}}_{\text{obs}}(M = m \mid X = x, A = a) > 0\}$, ensuring positivity for both treatments and mediators.

■ *Example 3: Instrumental variables.* The observed variables are (X, Z, A, Y), augmented by noise variables and an unobserved confounder U that is a joint parent of A and Y has no edge to the instrument Z; see the compatible C-DAG in Fig. 1 (right). The constraint set is $\mathcal{I}(\mathcal{S}) = \{\mathbb{P}^{\mathcal{S}}_{\text{obs}}(Z = z \mid X = x) > 0, f^{\mathcal{S}}_Y(X, A, U) = f^{\mathcal{S}}(X, A) + g^{\mathcal{S}}(X, U)\}$.

Overall algorithm. Given $(\mathcal{G}_c, \mathcal{I})$, we propose to construct a prior distribution Π over SCMs as follows: First, we order the clusters $(C_1, \dots, C_k)$ according to their hierarchy in the DAG (i.e., C_1 has no parents). Then, we iterate over each cluster C_i as follows: if C_i only contains latent variables, fix their distribution to a standard normal distribution via $U^{(i)} \sim \mathcal{N}(0, \mathbf{I})$. If C_i is a cluster of observed and latent variables, we assign a clustered Bayesian neural network (BNN) prior to C_i (see below). If C_i only contains observed variables, we assign an observational BNN prior.

Clustered BNN prior. For clusters that contain both observed and latent variables, we leverage a BNN-based prior inspired by TabPFN (Hollmann et al., 2023). This prior allows us to effectively sample potentially high-dimensional clusters of variables for which the internal causal structure is irrelevant to infer the causal query of interest. The prior is defined via

$$g^{(i)}_\theta : \text{pa}(C_i) \longrightarrow \mathbb{R}^r, \qquad \theta \sim \Pi_{C_i} \quad \text{s.t. } g^{(i)}_\theta \text{ satisfying } \mathcal{I}(\mathcal{S}_\theta). \tag{5}$$

We then sample random nodes from $g^{(i)}_\theta$ that coincide with observed nodes in C_i, while the remaining nodes serve as latent noise within the cluster. This corresponds to applying the approach taking in TabPFN (Hollmann et al., 2023) to clusters C_i in the C-DAG in which the causal structure does not matter for estimating our causal query.

Observational BNN prior. If C_i contains only observed nodes, we define another BNN via

$$f^{(i)}_\theta : \text{pa}(C_i) \longrightarrow \mathbb{R}^{|C_i|}, \qquad \theta \sim \Pi_{C_i} \quad \text{subject to } f^{(i)}_\theta \text{ satisfying } \mathcal{I}(\mathcal{S}_\theta), \tag{6}$$

and set $C_i = f^{(i)}_\theta\big(\text{pa}(C_i)\big)$. The observed nodes within C_i thus correspond to the output of the neural network and are *not* randomly subsampled neurons.

Example 1: back-door adjustment. Here, the data distribution $\mathbb{P}$ can be separated as follows: $(X, U_X) \sim \mathbb{P}_X$ with U_X denoting noise variables withing the cluster X, $U_A \sim \mathbb{P}_{U_A}$, $U_Y \sim \mathbb{P}_{U_Y}$, $A = f_A(X, U_A)$, and $Y = f_A(X, A, U_Y)$. Our algorithm proceeds as follows: $\mathbb{P}_{U_A}$ and $\mathbb{P}_{U_Y}$ are noise variables and are set to standard normal distributions. The cluster (X, U_X) contains both noise and observed variables, meaning that $\mathbb{P}_X$ is sampled from an clustered BNN prior. Finally, A and Y are observed variables meaning that f_A and f_Y are sampled from observational BNN priors. We refer to Appendix D.1 for full implementation details, including for Example 2 and 3.

Notes on identifiability. Our framework follows established causal inference philosophy and separates identifiability from estimation (Pearl, 2009): the identifiability step (=choosing the causal setting) requires careful modeling and usage of domain knowledge, while the estimation step can be handed over to our CausalFM. If practitioners suspect identifiability assumptions may be violated, we recommend performing causal sensitivity analysis (Dorn & Guo, 2022; Frauen et al., 2023) to assess the extent of potential violations.

5 CausalFM: Training

5.1 Training algorithm

We look at the case where the causal query $Q(\mathbb{P}_{\text{int}}(Y \mid X))$ is a function of the conditional interventional distribution $\mathbb{P}_{\text{int}}(Y \mid X)$ for some contextual observed variables X. This includes, e.g., the CATE $\mathbb{E}[Y(1) - Y(0) \mid X]$ and CAPO $\mathbb{E}[Y(a) \mid X]$ from our running examples.

Our goal is to train a PFN $q_\theta(Y \mid x)$ to approximate the conditional PPID (posterior predictive interventional distribution) $\Pi_{\text{int}}(Y \mid \mathcal{D}_n, X = x) = \int \mathbb{P}^{\mathcal{S}}_{\text{int}}(Y \mid X = x)\, \Pi(\mathcal{S} \mid \mathcal{D}_n)\, \mathrm{d}\mathcal{S}$. Given an SCM prior Π and a prior Π_N over sample sizes, we propose the following modified PFN loss

$$\mathcal{L}(\theta) = \mathbb{E}_{N\sim\Pi_N}\Big[\mathbb{E}_{\mathcal{S}\sim\Pi}\Big[\mathbb{E}_{(X,Y)\sim\mathbb{P}^{\mathcal{S}}_{\text{int}}}\Big[\mathbb{E}_{\mathcal{D}\sim\mathbb{P}^{\mathcal{S}}_{\text{obs}}}\big[-\log q_\theta(Y \mid X, \mathcal{D}_N)\big]\Big]\Big]\Big]. \tag{7}$$

Importantly, the dataset $\mathcal{D}$ is sampled from the *observational* distribution, while the pair (X, Y) is sampled from the *interventional* distribution induced by a random SCM. This ensures that the PFN will aim to predict the interventional outcome Y based on data following the observational distribution. A similar loss has been proposed by Bynum et al. (2025), which, however, is only based on the mean-squared error instead of the negative log-likelihood and thus does not allow an interpretation for approximating the PPID in a Bayesian setting. In particular, modeling the entire PPID allows us not only to provide point estimators of our causal query, but also to account for uncertainty.

In practice, we sample the sample size $N_j \sim \Pi_N$, an SCM $\mathcal{S}_j \sim \Pi$, and an observational dataset $\mathcal{D}^j_{N_j} \sim \mathbb{P}^{\mathcal{S}_j}_{\text{obs}}$ by sampling from the SCM. Then, we modify the SCM by performing the intervention of interest (e.g., $\mathrm{do}(A = a)$) and sample test points $(x_j, y_j) \sim \mathbb{P}^{\mathcal{S}_j}_{\text{int}}$ from the interventional SCM. The approximated PFN-loss is then

$$\hat{\mathcal{L}}(\theta) = \sum_j \big[-\log q_\theta(y_j \mid \mathcal{D}^j_{N_j}, x_j)\big]. \tag{8}$$

Finally, once $q_\theta(Y \mid x)$ is trained, we can obtain an estimator for the causal query via $Q(q_\theta(Y \mid X)$, i.e., by applying the causal query on the approximated PPID by the PFN.

Example: back-door adjustment. Here, we sample a sample size $N_j \sim \Pi_N$, an SCM $\mathcal{S}_j \sim \Pi$ from our constructed prior distribution Π and an observational dataset $\mathcal{D}^j_{N_j} \sim \mathbb{P}^{\mathcal{S}_j}_{\text{obs}}$. Then, we perform two interventions $do(A = 1)$ and $do(A = 0)$ to obtain test points $(x_j, y_j(1) - y_j(0)) \sim \mathbb{P}^{\mathcal{S}_j}_{\text{int}}$. The PFN loss becomes

$$\hat{\mathcal{L}}(\theta) = \sum_j \big[-\log q_\theta(y_j(1) - y_j(0) \mid \mathcal{D}^j_{N_j}, x_j)\big]. \tag{9}$$

Implementation details. Each observation is tokenized during embedding, with separate encoders applied to observational variables. The resulting tokens are processed by a transformer-based PFN to obtain representations, which are subsequently passed to a Gaussian mixture model (GMM) head. Our implementation of $q_\theta(Y \mid x)$ is based on the TabPFN architecture (Hollmann et al., 2023). We train the model with a learning rate of $1\mathrm{e}{-3}$, weight decay $1\mathrm{e}{-5}$, batch size 16, and sequence length 1024 for up to 150 epochs. Training CausalFM on a single NVIDIA A100 GPU takes about 24 hours. Details on the data prior and generation details are provided in Appendix D.1, while the full implementation is given in Appendix D.2.

6 Experiments

We evaluate our method across three causal inference settings: standard CATE estimation, instrumental variables (IV), and front-door adjustment.

Evaluation metrics. We report the precision in estimating heterogeneous effects (PEHE) (Curth & van der Schaar, 2021; Hill, 2011), defined as the root mean squared deviation between predicted and ground-truth CATE, to evaluate the model performance on the CATE estimation task.

Table 2: **Standard CATE estimation** over 10 synthetic datasets and Jobs dataset.

Method	Synthetic	Jobs
BASELINES (A): STANDARD CATE ESTIMATORS		
S-learner (Künzel et al., 2019)	0.734 ± 0.16	0.697 ± 0.18
T-learner (Künzel et al., 2019)	0.661 ± 0.17	0.822 ± 0.18
TARNet (Shalit et al., 2017b)	0.854 ± 0.23	0.864 ± 0.24
DR-learner (Kennedy, 2023b)	0.765 ± 0.17	0.959 ± 0.18
RA-learner (Curth & van der Schaar, 2021)	0.609 ± 0.13	0.652 ± 0.15
X-learner (Künzel et al., 2019)	0.563 ± 0.15	0.802 ± 0.18
BASELINES (B): FOUNDATION MODELS-BASED METHODS		
CausalPFN (Balazadeh et al., 2025)	0.557 ± 0.18	0.528 ± 0.16
DoPFN (Robertson et al., 2025)	0.586 ± 0.19	0.482 ± 0.20
CausalFM (ours)	0.515 ± 0.20	0.478 ± 0.18

Lower = better. Reported: PEHE (mean ± std). Top-three per column are in blue, purple, orange.

6.1 Evaluation for standard CATE setting

Baselines for standard CATE estimation. We consider a broad range of state-of-the-art methods for the conditional treatment effect estimation from the literature: (1) **S-learner** (Künzel et al., 2019): the S-learner is a model-agnostic learner that trains a single regression model by concatenating the covariate and the treatment as input; (2) **T-learner** (Künzel et al., 2019): the T-learner is a model-agnostic learner that trains separate regression models for treated and control groups; (3) **X-learner** (Künzel et al., 2019): builds upon the T-learner by first imputing individual treatment effects in each group and then fitting models to these pseudo-effects; (4) **TARNet** (Shalit et al., 2017b): using representation learning to extract features of covariates and train separate branches for treated and control groups with regularization; (5) **DR-learner** (Kennedy, 2023b): generates pseudo-outcomes based on the doubly-robust AIPW estimator; (6) **RA-learner** (Curth & van der Schaar, 2021): uses a regression-adjusted pseudo-outcome in the second stage. We also include two PFN-based foundation models for treatment effect estimation: (7) **CausalPFN** (Balazadeh et al., 2025) and (8) **DoPFN** (Robertson et al., 2025). Further implementation details are in Appendix D.3.

Results on standard CATE estimation. We benchmark our model on ten synthetic datasets generated under diverse mechanisms, with implementation details in Appendix D. In addition, we evaluate on a semi-synthetic version of the Jobs dataset (Smith & Todd, 2005), derived from the widely used LaLonde study (LaLonde, 1986). Here, we generate outcomes to create a semi-synthetic dataset and allow for evaluation against ground-truth.

Table 2 reports the averaged PEHE across the synthetic datasets (full results in the Appendix) and the Jobs dataset. Our experiments show that CausalFM achieves competitive CATE estimation performance across all benchmarks, *without requiring model retraining*.

Table 3: **IV setting** for CATE estimation with binary and continuous instrument variables.

Method	Binary IV	Continuous IV
BASELINES (A): STANDARD IV ESTIMATORS		
KIV (Singh et al., 2019)	0.454 ± 0.16	0.577 ± 0.20
DRIV (Syrgkanis et al., 2019)	0.531 ± 0.18	0.693 ± 0.20
DeepIV (Hartford et al., 2017)	0.427 ± 0.15	0.516 ± 0.13
DeepGMM (Bennett et al., 2019)	0.503 ± 0.20	0.588 ± 0.21
DMLIV (Syrgkanis et al., 2019)	0.479 ± 0.23	0.618 ± 0.20
DFIV (Xu et al., 2021)	0.709 ± 0.29	0.583 ± 0.30
MRIV (Frauen & Feuerriegel, 2022)	0.688 ± 0.21	0.641 ± 0.24
BASELINES (B): FOUNDATION MODELS-BASED METHODS		
DoPFN (Robertson et al., 2025)	0.523 ± 0.20	0.675 ± 0.37
CausalFM (ours)	0.422 ± 0.16	0.579 ± 0.21

Lower = better. Reported: PEHE (mean ± standard deviation). Top-three per column are in blue, purple, orange.

6.2 Evaluation for IV setting

Baselines for IV setting. We benchmark against a broad set of state-of-the-art IV methods for treatment effect estimation: (1) **KIV** (Singh et al., 2019): a nonlinear extension of two-stage least squares using kernel ridge regression with feature maps; (2) **DFIV** (Xu et al., 2021): extends KIV by parameterizing feature maps with neural networks trained iteratively; (3) **DeepIV** (Hartford et al., 2017): a two-stage neural approach, first estimating the treatment distribution and then solving a counterfactual prediction task; (4) **DeepGMM** (Bennett et al., 2019): formulates IV estimation as a minimax game based on the generalized method of moments, solved via adversarial training; (5) **DMLIV** (Syrgkanis et al., 2019): a double machine learning framework that estimates nuisance functions and learns the CATE by orthogonalized regression; (6) **DRIV** (Syrgkanis et al., 2019): a meta-learner combining DMLIV with doubly robust pseudo-outcomes for improved stability; and (7) **MRIV** (Frauen & Feuerriegel, 2022): a multiply robust framework for binary IVs that directly estimates CATE via pseudo–outcome regression. For foundation model baselines, as CausalPFN (Balazadeh et al., 2025) is only for back-door adjustment, we include DoPFN (Robertson et al., 2025).

Table 4: **Front-door adjustment setting** for CATE estimation.

Method	PEHE
BASELINES (A): STANDARD FRONT DOOR ADJUSTMENT	
Plug-in front-door learner (Linear) Pearl (2009)	1.124 ± 0.28
Plug-in front-door learner (RF) Pearl (2009)	1.364 ± 0.52
Plug-in front-door learner (NN) Pearl (2009)	0.889 ± 0.38
BASELINES (B): FOUNDATION MODELS-BASED METHODS	
DoPFN (Robertson et al., 2025)	1.274 ± 0.24
CausalFM (ours)	0.847 ± 0.34

Lower = better. Reported: PEHE (mean ± standard deviation). Top-three per column are in blue, purple, orange.

Results on IV setting. We evaluate our models on datasets with varying confounding strengths. Table 3 reports the averaged PEHE for binary and continuous IVs. Note that CausalPFN is *not* designed for IV settings. In contrast, we find that our CausalFM consistently achieves comparable

performance relative to standard IV estimators and outperforms biased alternatives. Importantly, in contrast the the standard baselines, these results hold *without requiring model retraining*. Hence, this confirms the flexibility of our approach to IV settings.

6.3 Front-door adjustment

We additionally evaluate our model under the front-door adjustment setting in Table 4. Due to space constraints, details are provided in Appendix H.1. The experiments show the flexibility of our method to perform causal inference in the front-door adjustment setting.

6.4 Computational time

We report computation time in this section. For our method and other foundation-model-based approaches, we show inference time since these models do not need fine-tuning after pretraining. For standard baselines, which must be trained for each dataset separately, we report the average total time per dataset, including both training and inference. As shown in Tables 5 and 6, our model is highly efficient.

Table 5: Overall time comparison for standard CATE setting.

Method	Time (s)
BASELINES (A): STANDARD CATE ESTIMATORS	
S-learner (Künzel et al., 2019)	2.76×10^{0}
T-learner (Künzel et al., 2019)	3.21×10^{0}
TARNet (Shalit et al., 2017b)	3.98×10^{0}
DR-learner (Kennedy, 2023b)	1.78×10^{1}
RA-learner (Curth & van der Schaar, 2021)	1.24×10^{1}
X-learner (Künzel et al., 2019)	1.93×10^{1}
BASELINES (B): FOUNDATION MODELS-BASED METHODS	
CausalPFN (Balazadeh et al., 2025)	1.27×10^{0}
DoPFN (Robertson et al., 2025)	2.31×10^{0}
CausalFM (ours)	4.90×10^{-1}

Lower = better. Reported: Time in seconds.

6.5 Additional experiments

Misspecification of causal settings. We conduct experiments to study the sensitivity to using an incorrect identifiability strategy. Specifically, we generate data from (i) an IV SCM and (ii) a front-door SCM (as in our main experiments), and compare a model trained under the correct identifiability design (IV or front-door, respectively) with a model trained under an incorrect back-door design. The results are in Table 7. As expected, using a misspecified identifiability strategy consistently worsens the PEHE. This highlights the importance of including the correct identifiability assumption into the prior specification, which we propose for CausalFM.

Table 6: Overall time comparison for IV setting.

Method	Time (s)
BASELINES (A): STANDARD IV ESTIMATORS	
KIV (Singh et al., 2019)	3.24×10^{-1}
DRIV (Syrgkanis et al., 2019)	3.87×10^{1}
DeepIV (Hartford et al., 2017)	1.27×10^{1}
DeepGMM (Bennett et al., 2019)	1.85×10^{1}
DMLIV (Syrgkanis et al., 2019)	1.85×10^{1}
DFIV (Xu et al., 2021)	1.74×10^{1}
MRIV (Frauen & Feuerriegel, 2022)	1.56×10^{1}
BASELINES (B): FOUNDATION MODELS-BASED METHODS	
DoPFN (Robertson et al., 2025)	6.53×10^{0}
CausalFM (ours)	4.72×10^{-1}

Lower = better. Reported: Time in seconds.

Prior design choices. We also analyze the robustness of our model to the choice of the prior. For this, we vary the strength of unobserved confounding in the data-generating process, controlled by a parameter $\alpha \in [0, 1]$. Due to space constraints, the results in shown in Fig. 2 in Appendix H.5. The experiments show that our model remains robust as α increases.

Table 7: Analysis of the effect of misspecified identifiability strategies.

Data Generating SCM	Strategy	Identifiability Used	PEHE
IV SCM	Correct	IV	0.422
	Incorrect	Back-door	0.489
Front-door SCM	Correct	Front-door	0.847
	Incorrect	Back-door	0.876

6.6 Discussion

Limitations and future work. The current evaluation is limited to synthetic and semi-synthetic data due to the fundamental problem of missing potential outcomes on real-world data. For future work, it will be interesting to investigate the performance of CausalFM in applied A/B experimental setups to assess its empirical performance and robustness under real-world conditions. Additionally, an important research direction will be to incorporate interpretability or fairness constraints into CausalFM, which is crucial for reliable deployment in practice.

Ethics statement.

Human subjects and IRB. This work does not involve experiments with human subjects. Our training data are *synthetically* generated from prespecified SCM-based priors. For empirical evaluation, we additionally use publicly available benchmark data (e.g., Jobs) where outcomes are generated in a semi-synthetic manner following common practice; no identifiable personal information is introduced by us. Accordingly, no IRB review was required for this study.

Data, privacy, and security. We do not collect, store, or release sensitive personal data. Public datasets are used under their respective licenses, and our semi-synthetic outcome generation avoids re-identification risks. We will document preprocessing and generation steps to support reproducibility.

Bias and potential harms. Causal estimators can be misused if applied outside the assumed identification regime (e.g., back-door, front-door, IV) or under severe violations (e.g., weak instruments, lack of overlap). To mitigate harm: (i) we make assumptions explicit and provide uncertainty quantification; (ii) we advocate domain-expert validation and sensitivity checks before deployment; (iii) we discourage high-stakes automated decision-making without human oversight.

Use of large language models (LLMs). We used LLM-based tools to assist with writing (clarity, grammar) and for literature research. All claims were authored and verified by the authors; citations were cross-checked against primary sources. No sensitive data were provided to LLM tools.

Reproducibility statement. We ensure reproducibility of our results by providing the full implementation and training scripts through an anonymous GitHub repository `https://anonymous.4open.science/r/causal_foundation_model`. The repository contains the necessary code to reproduce our experiments, along with instructions for dataset preparation, model training and evaluation procedures. This setup allows independent researchers to replicate the reported results and extend our work with minimal effort.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task Gaussian processes. In *NeurIPS*, 2017.

Tara V. Anand, Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. Causal effect identification in cluster dags. In *AAAI*, 2023.

Joshua D. Angrist. Lifetime earnings and the vietnam era draft lotter: Evidence from social security administrative records. *The American Economic Review*, 80(3):313–336, 1990.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Benson Li, Junwei Ma, Jesse C. Cresswell, and Rahul G. Krishnan. Causalpfn: Amortized causal effect estimation via in-context learning. *arXiv preprint*, arXiv:2506.07918, 2025. URL http://arxiv.org/pdf/2506.07918v1.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *NeurIPS*, 2019. URL http://arxiv.org/pdf/1905.12495v2.

Lucius E. J. Bynum, Aahlad Manas Puli, Diego Herrero-Quevedo, Nhi Nguyen, Carlos Fernandez-Granda, Kyunghyun Cho, and Rajesh Ranganath. Black box causal inference: Effect estimation via meta prediction. *arXiv preprint*, arXiv:2503.05985, 2025. URL http://arxiv.org/pdf/2503.05985v1.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *AISTATS*, 2021.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 2022. URL http://arxiv.org/pdf/2102.04543v2.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 2024.

Stefan Feuerriegel, Yash Raj Shrestha, and Georg Von Krogh. A new machine learning approach answers what-if questions. *MIT Sloan Management Review*, 66(3):65–69, 2025.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 53(3):879–908, 2023. ISSN 0090-5364. URL http://arxiv.org/pdf/1901.09036v3.

Dennis Frauen and Stefan Feuerriegel. Estimating individual treatment effects under unobserved confounding using binary instruments. In *ICLR*, 2022.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Sharp bounds for generalized causal sensitivity analysis. In *NeurIPS*, 2023.

P. Richard Hahn, Jared S. Murray, and Carlos Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020. URL http://arxiv.org/pdf/1706.09523v4.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *ICML*, 2017.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. In *ICLR*, 2023. URL http://arxiv.org/pdf/2207.01848v6.

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. doi: 10.1038/s41586-024-08328-6.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model tabpfn outperforms specialized time series forecasting models based on simple features. *arXiv preprint*, arXiv:2501.02945, 2025. URL http://arxiv.org/pdf/2501.02945v2.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Tomoharu Iwata and Yoichi Chikahara. Meta-learning for heterogeneous treatment effect estimation with closed-form solvers. *arXiv preprint*, arXiv:2305.11353, 2023. URL http://arxiv.org/pdf/2305.11353v1.

Fredrik D. Johansson, Uri Shalit, and David Sonntag. Learning representations for counterfactual inference. In *ICML*, 2016.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023a.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023b.

Christoph Kern, Unai Fischer-Abaigar, Jonas Schweisthal, Dennis Frauen, Rayid Ghani, Stefan Feuerriegel, Mihaela van der Schaar, and Frauke Kreuter. Algorithms for reliable decision-making need causal reasoning. *Nature Computational Science*, 2025.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Adi Lahat, Kassem Sharif, Narmin Zoabi, Yonatan Shneor Patt, Yousra Sharif, Lior Fisher, Uria Shani, Mohamad Arow, Roni Levin, and Eyal Klang. Assessing generative pretrained transformers (gpt) in clinical decision-making: comparative analysis of gpt-3.5 and gpt-4. *Journal of medical Internet research*, 26:e54571, 2024.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.

Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsesa Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in context reinforcement learning. In *NeurIPS*, 2023.

Antonio R. Linero and Joseph L. Antonelli. The how and why of bayesian nonparametric causal inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1):e1583, 2022. URL http://arxiv.org/pdf/2111.03897v2.

Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. Diffpo: A causal diffusion model for learning distributions of potential outcomes. In *NeurIPS*, 2024.

Yuchen Ma, Jonas Schweisthal, Hengrui Zhang, and Stefan Feuerriegel. A diffusion-based method for learning the multi-outcome distribution of medical treatments. In *KDD*, 2025.

Divyat Mahajan, Jannes Gladrow, Agrin Hilmkil, Cheng Zhang, and Meyer Scetbon. Zero-shot learning of causal models. *arXiv preprint*, arXiv:2410.06128, 2025. URL http://arxiv.org/pdf/2410.06128v2.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *ICML*, 2022. URL http://arxiv.org/pdf/2204.07258v2.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *ICLR*, 2022. URL http://arxiv.org/pdf/2112.10510v7.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *ICML*, 2023.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021. ISSN 0006-3444.

Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Surina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. In *NeurIPS*, 2023.

Judea Pearl. *Causality*. Cambridge University Press, New York City, 2009. ISBN 9780521895606.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-pfn: In-context learning for causal effect estimation. *arXiv preprint*, arXiv:2506.06039, 2025. URL http://arxiv.org/pdf/2506.06039v1.

James M. 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: 1393–1512, 1986.

James M. Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412, 1994. ISSN 0361-0926. doi: 10.1080/03610929408831393.

James M. Robins. Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association on Bayesian Statistical Science*, pp. 6–10, 1999.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of reression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–688, 1994.

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. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663. doi: 10.1037/h0037350.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. In *NeurIPS*, 2023.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 2017a.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017b.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*, 2019.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *NeurIPS*, 2019.

Jeffrey A Smith and Petra E Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of econometrics*, 125(1-2):305–353, 2005.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In *NeurIPS*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Boris van Breugel and Mihaela van der Schaar. Why tabular foundation models should be a research priority. In *ICML*, 2024. URL http://arxiv.org/pdf/2405.01147v2.

Mark J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1):1–31, 2006.

Mark J. van der Laan and Donald B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

Hal R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences (PNAS)*, 113(27):7310–7315, 2016. doi: 10.1073/pnas.1510479113.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Janick Weberpals, Stefan Feuerriegel, Mihaela van der Schaar, and Kenneth L. Kehl. Opportunities for causal machine learning in precision oncology. *NEJM AI*, 2025.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *ICLR*, 2021. URL http://arxiv.org/pdf/2010.07154v3.

Jiaqi Zhang, Joel Jennings, Agril Hilmkril, Nick Pawlowski, Cheng Zhang, and Chao Ma. Towards causal foundation model: On duality between optimal balancing and attention. *arXiv preprint*, arXiv:2310.00809, 2024.

A Extended related work

Prior-data-fitted networks (PFNs) as tabular foundation models. Foundation models are an emerging paradigm that has revolutionized machine learning for various data modalities, particularly language and vision tasks (Devlin, 2018; Lahat et al., 2024; Touvron et al., 2023b;a). The same paradigm is now being explored for tabular data – a modality that underpins the large majority of analyses in science and business (van Breugel & van der Schaar, 2024). Prior-data-fitted networks (PFNs) (Müller et al., 2022) constitute a powerful approach to training tabular foundation models. PFNs are large transformers trained on synthetic data to perform Bayesian inference through in-context learning. TabPFN (Hollmann et al., 2023; 2025) scaled this idea by pairing the transformer with a Bayesian neural network prior over structural causal models (SCMs) and demonstrating state-of-the-art performance on various tabular benchmarks. Subsequent work extended PFNs to time–series forecasting (Hoo et al., 2025) and analyzed their in-context learning abilities theoretically (Nagler, 2023). Critically, all existing PFNs are trained *only* for *predictive* tasks and do **not** target causal estimands; they therefore are **not** designed for causal inference of treatment effects, which is the goal of our paper.

Treatment effect estimation. Causal inference, such as the estimation of average treatment effects, originates from fields like econometrics (Imbens & Angrist, 1994; Angrist, 1990), statistics (van der Laan & Rubin, 2006), and epidemiology (Robins, 1986; 1994). Machine learning methods have been proposed to estimate *heterogeneous* effects to support personalized decision-making. One line of work are frequentist methods, which often build on semiparametric theory (Robins et al., 1994; Robins, 1999), yielding model-agnostic estimators that are doubly robust and Neyman-orthogonal (van der Laan, 2006; Chernozhukov et al., 2018; Nie & Wager, 2021; Foster & Syrgkanis, 2023; Kennedy, 2023a). Another line of work builds upon specific machine learning methods/ architectures such as regression trees (Wager & Athey, 2018) or neural networks (Johansson et al., 2016; Shalit et al., 2017a; Shi et al., 2019) and adopts them to causal inference. Bayesian alternatives include Bayesian additive regression trees (Hahn et al., 2020) or Gaussian-process counterfactual regression (Alaa & van der Schaar, 2017). However, all of the existing estimators above must be *retrained* for every new dataset. In contrast, our CausalFM allows for pre-trained models to approximate Bayesian causal inference.

A.1 Differences between CausalFM, CausalPFN, and Do-PFN

Identifiability. CausalFM separates identifiability from estimation. The central motivation of CausalFM is that causal identification (choosing an identifiable setting such as back-door, IV, or front-door) must be handled before estimation. This mirrors classical causal inference practice and ensures that the PFN *only* learns within an identifiable causal setting. Our reasoning for identifiability is as follows:

(1) Asymptotically unbiased causal inference: As we show in Theorem 4.3, incorporating identifiability assumptions into the prior is *necessary* for asymptotically unbiased causal inference. As a consequence, methods that ignore identifiability assumptions yield biased causal effect estimates, even if we collect large amounts of data. This is highly undesirable in practice.

(2) Informative predictive-posterior distributions: If we do not impose any assumption on the DGP, it is well known that causal effect estimation is not just fundamentally biased, but also that this bias can be of arbitrary size. For example, the backdoor-adjustment bias due to omitted unobserved confounding can be written in closed form depending on confounding strength. Thus, if the PFN-prior assigns positive probability mass for DGPs with arbitrary confounding strength, the predictive-posterior must respect the possibility of arbitrarily biased treatment effects, thus rendering PFN-based inference completely noninformative.

(3) Clear separation between domain knowledge and statistical inference: One might argue that a possible remedy would be to restrict the PFN prior only to DGPs with somewhat "weak" identifiability violations (e.g., weak unobserved confounding). However, we argue that this would correspond to assumptions/ domain knowledge on the DGP, similar to those in our paper, that must be made transparent for practitioners and could also possibly be violated.

In short, our paper follows established causal inference philosophy and separates identifiability from estimation: the identifiability step (choosing the causal setting) requires careful modeling and usage of domain knowledge, while the estimation step can be handed over to our CausalFM. If practitioners suspect identifiability assumptions may be violated, we recommend performing causal sensitivity analysis to assess the extent of potential violations.

In contrast, CausalPFN implicitly assumes *only* back-door adjustment and therefore *cannot* handle IV or front-door, resulting in bias under unobserved confounding. Do-PFN mixes many causal graphs in a single prior without conditioning on which setting is identifiable, which, as our Theorem 4.3 shows, can lead to asymptotically *biased* estimates and non-informative posteriors.

Prior construction. This philosophy requires a fundamentally different prior construction. Because identifiability is encoded at the level of causal structure, CausalFM introduces C-SCM priors and C-DAGs that enforce the assumptions required by each identifiability strategy. Do-PFN does *not* encode identifiability constraints into the prior family for their prior construction. CausalPFN is *restricted* solely to back-door adjustment.

Theoretical guarantees. Beyond this framework, we contribute *new* theoretical results showing that identifiability must be incorporated into PFN priors. Theorem 4.3 proves that if a PFN prior places nonzero mass on SCMs that violate the identifiability conditions of the chosen setting, then the resulting posterior predictive interventional distribution is necessarily misspecified and cannot yield consistent causal effect estimates, even with infinite data. This explains the empirical behavior of Do-PFN, which may return non-informative posteriors when its prior includes SCMs with strong unobserved confounding and no valid instruments. Our theoretical results show that this issue is structural, not merely an implementation detail.

Empirical performance. Empirically, our model outperforms others in different settings. Besides, we also have experiments showing the necessity to have the correct identifiability assumption in the prior specification.

B EXAMPLE FOR SCM-PRIORS

Here, we consider the IV setting from Example 3 with additional normality assumption and empty $X = \emptyset$, i.e., observational distribution $(Z, A, Y) \sim \mathbb{P}_{\text{obs}}$. Let us consider the following class of SCMs:

$$U \sim \mathcal{N}(0,1),\ \epsilon_Z \sim \mathcal{N}(0,1),\ \epsilon_A \sim \mathcal{N}(0,1),\ \epsilon_Y \sim \mathcal{N}(0,1), \tag{10}$$

$$Z = \alpha\epsilon_Z, +\kappa U\ A = \beta Z + \delta\epsilon_A + \gamma U,\ Y = \zeta A + \eta U + \theta\epsilon_Y, \tag{11}$$

where U is an unobserved confounder between A and Y, and ϵ_Z, ϵ_A, ϵ_Y are noise variables, and α, β, γ, δ, ζ, η, and θ are scalars describing the functional dependences between observed and noise variables. Our causal query is $Q(\mathbb{P}_{\text{int}}) = \mathbb{E}[Y(1)] = \zeta$.

General approach. The class of SCMs above is compatible with the linear IV setting whenever it holds that $\kappa = 0$ (independence assumption from Example 3). Hence, we can specify a prior distribution over this class of SCMs by specifying a distribution Π over $(\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta)$ and setting $\kappa = 0$. Note that this automatically specifies a distribution over $\mathbb{P}_{\text{obs}}$ (by sampling from the SCM) *and* $\mathbb{P}_{\text{int}}$ (by intervening and setting $A = 1$ in the SCM). Interestingly, this addresses the two drawbacks of observational priors from above as follows: (i) During the PFN training we can sample $\mathcal{D}_n \sim \mathbb{P}_{\text{obs}}$ and $y(1) \sim \mathbb{P}_{\text{int}}$ and thus fit $q_\theta(y(1) \mid \mathcal{D}_n)$ for the interventional outcome (see Sec. 5 for details). For estimating the causal query we can thus use $Q(q_\theta(y(1) \mid \mathcal{D}_n))$ and do not need access to the potentially unknown $\bar{Q}$. (ii) We can directly control the marginal prior distribution of ζ, thus remedying the above drawbacks and allowing us more control to incorporate prior information of our causal query.

Adding identifiability assumptions to the prior. A key question is whether we should actually impose the identifiability assumption $\kappa = 0$ when constructing a prior. A different approach would be to also put a prior on κ, thus taking account the possibility of identifiability violations in the prior. Such an approach has been proposed by (Robertson et al., 2025), where the authors construct a prior over many possible causal inference settings simultaneously. *However*, as we show in the following, this would make consistent Bayesian estimation of the causal query of interest *impossible*, confirming Theorem 4.3.

Lemma B.1. *Let $S^* = (\alpha^*, \beta^*, \delta^*, \gamma^*, \zeta^*, \eta^*, \theta^*, \kappa^* = 0)$ be an identified ground-truth SCM. Then for any causal target $\zeta \neq \zeta^*$ there exists another SCM $S = (\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta, \kappa)$ with $\kappa \neq 0$ that induces the same observational distribution as S^*.*

Proof. See Appendix C. □

Lemma B.1 has an important consequence: if our prior Π puts positive probability mass on all possible combinations of $(\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta, \kappa)$, the corresponding posterior $\Pi(\cdot \mid \mathcal{D}_n)$ will even for $n \to \infty$ put positive probability mass on any $\zeta \in \mathbb{R}$, thus being completely non-informative about the causal target quantity. As a consequence, any Bayesian point estimator using such a prior (e.g., as in under the approach (Robertson et al., 2025)) will be asymptotically biased.

In contrast, we present a different approach to circumvent the above problems: namely, we propose to *construct PFN-priors that incorporate assumptions that allow for identifiability of the causal target quantity* (e.g., setting $\kappa = 0$ in the above example). As such, we follow established philosophy in causal inference that separates identifiability and estimation steps (Pearl, 2009): the identifiability step should be established by the practitioner using domain knowledge (e.g., establishing whether a certain variable is a valid instrument). Once identifiability has been established, we can use Bayesian modeling and PFN-based models for the *estimation step*.

Which noise variables to model? A key question that remains is what classes of SCMs we can use to specify priors for the causal inference setting $\mathcal{C}$ at hand. Indeed, the class of SCMs is non-unique: as suggested in the main paper, it is not necessary to specify both noise variables ϵ_A and ϵ_Y.

Lemma B.2. *Let $\mathcal{S}^* = (\alpha^*, \beta^*, \gamma^*, \delta^*, \zeta^*, \eta^*, \theta^*)$ be a fixed SCM from the above class with $\text{Var}^*(A \mid z) > 0$ and $\text{Var}^*(Y \mid a) > 0$ for all z, a. Then, there exist unique SCMs $\mathcal{S}_1 = (\alpha_1, \beta_1, \gamma_1, \delta_1 = 0, \zeta_1, \eta_1, \theta_1)$ and $\mathcal{S}_2 = (\alpha_2, \beta_2, \gamma_2, \delta_2, \zeta_2, \eta_2, \theta_2 = 0)$ that induce the same observational distribution as $\mathcal{S}^*$ and thus the same causal query $\zeta_1 = \zeta_2 = \zeta^*$. However, whenever it holds that both $\delta = 0$ and $\theta = 0$, there exists an SCM $\mathcal{S}^*$ for which $\zeta \neq \zeta^*$.*

Proof. See Appendix C. □

Lemma B.2 implies that it *suffices to specify priors over SCMs without either treatment noise ϵ_A or outcome noise ϵ_Y*. However, if we remove both, there exist interventional distributions for which the prior will never put probability mass on the ground-truth causal query, rendering Bayesian inference inconsistent. In the following, we generalize this result to arbitrary SCMs and causal inference settings.

C PROOFS

C.1 PROOF OF THEOREM 4.3

Proof of Theorem 4.3. Assume that $\Pi(\mathcal{S}_0 \in \mathcal{Z}) = w_0 > 0$. Let

$$\mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}}) := \{\mathcal{S} : \mathbb{P}^{\mathcal{Z}}_{\text{obs}} = \mathbb{P}^{\mathcal{S}_0}_{\text{obs}}\}$$

denote the observational equivalence class of $\mathcal{Z}$. In other words, $\mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}})$ contains exactly those SCMs that give rise to the same observational distribution as the models in $\mathcal{Z}$.

Choose any pair of distributions $(\mathbb{P}^{\mathcal{W}}_{\text{obs}}, \mathbb{P}^{\mathcal{W}}_{\text{int}}) \in \mathcal{P}_{\text{obs}} \times \mathcal{P}_{\text{int}}$ induced by an SCM $\mathcal{W}$ with $\Pi(\mathcal{W}) > 0$ and satisfying $\mathbb{P}^{\mathcal{W}}_{\text{obs}} = \mathbb{P}^{\mathcal{Z}}_{\text{obs}}$. Such an SCM $\mathcal{W}$ exists by definition of the equivalence class and has positive prior mass by assumption. By identifiability of the causal inference setting $\mathcal{C}$, it then holds for any $\mathcal{S}_0 \in \mathcal{Z}$ that

$$Q(\mathbb{P}^{\mathcal{W}}_{\text{int}}) \;=\; \bar{Q}(\mathbb{P}^{\mathcal{W}}_{\text{obs}}) \;=\; \bar{Q}(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}}) \neq Q(\mathbb{P}^{\mathcal{S}_0}_{\text{int}}). \tag{12}$$

The key point here is that identifiability fixes the target functional Q uniquely from the observational distribution, and therefore the value obtained under $\mathcal{W}$ must differ from the one induced by $\mathcal{S}_0$ whenever the latter does not agree with the identified functional.

Now draw data $\mathcal{D}_n \sim \mathbb{P}^{\mathcal{W}}_{\text{obs}}$. For every $\mathcal{S} \in \mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}})$, the observational likelihoods coincide for all n because all such models produce the same observational distribution. Hence, the Bayes factors between any two models in $\mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}})$ are always equal to 1, regardless of the sample size. Consequently,

$$\Pi\big(\mathcal{S} \mid \mathcal{D}_n\big) \propto \Pi(\mathcal{S}) \quad \text{for all } \mathcal{S} \in \mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}}), \text{ all } n.$$

Thus, within the equivalence class, the posterior simply mirrors the prior. Outside the equivalence class, however, the likelihood is misspecified, and therefore we know that $\Pi(\mathcal{S}|\mathcal{D}_n) \to 0$ for $\mathcal{S} \notin \mathcal{E}(\mathbb{P}^{\mathcal{Z}}_{\text{obs}})$ as $n \to \infty$.

We now examine the posterior predictive functional. By the above concentration behavior, we have

$$Q\left(\int \mathbb{P}^{\mathcal{S}}_{\text{int}}\, \Pi(\mathcal{S} \mid \mathcal{D}_n)\, d\mathcal{S}\right) \;\to\; \int_{\mathcal{E}(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}})} Q(\mathbb{P}^{\mathcal{S}}_{\text{int}})\, \Pi(\mathcal{S})\, d\mathcal{S}, \tag{13}$$

because only models in the equivalence class retain non-vanishing posterior mass.

Within that class, a fraction w_0 of the prior mass lies on $\mathcal{Z}$, so we can decompose the above limit as

$$\int_{\mathcal{E}(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}})} Q(\mathbb{P}^{\mathcal{S}}_{\text{int}})\, \Pi(\mathcal{S})\, d\mathcal{S} = w_0 \int_{\mathcal{Z}} Q(\mathbb{P}^{\mathcal{S}}_{\text{int}})\, \Pi(\mathcal{S})\, d\mathcal{S} \;+\; (1 - w_0)\, Q\big(\mathbb{P}^{\mathcal{W}}_{\text{int}}\big), \tag{14}$$

where the remaining mass $(1 - w_0)$ is assigned to models observationally equivalent to $\mathcal{Z}$ but not in $\mathcal{Z}$ itself. By (12), the resulting limit cannot equal $Q(\mathbb{P}^{\mathcal{W}}_{\text{int}})$. Hence,

$$\int_{\mathcal{E}(\mathbb{P}^{\mathcal{S}_0}_{\text{obs}})} Q(\mathbb{P}^{\mathcal{S}}_{\text{int}})\, \Pi(\mathcal{S})\, d\mathcal{S} \;\neq\; Q\big(\mathbb{P}^{\mathcal{W}}_{\text{int}}\big),$$

which shows that Π is not well-specified for $\mathcal{C}$. □

C.2 PROOF OF LEMMA B.1 (LINEAR IV)

Proof of Lemma B.1. We prove that for any $\zeta \neq \zeta^*$, there exists an SCM $S = (\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta, \kappa \neq 0)$ that induces the same observational distribution as S^*.

Step 1: Observational distribution of S^*

The observational distribution is characterized by the covariance matrix Σ^* of (Z^*, A^*, Y^*):

$$\text{Var}(Z^*) = (\alpha^*)^2 \tag{15}$$

$$\text{Cov}(Z^*, A^*) = (\alpha^*)^2\beta^* \tag{16}$$

$$\text{Var}(A^*) = (\alpha^*\beta^*)^2 + (\delta^*)^2 + (\gamma^*)^2 \tag{17}$$

$$\text{Cov}(Z^*, Y^*) = \zeta^*(\alpha^*)^2\beta^* \tag{18}$$

$$\text{Cov}(A^*, Y^*) = \zeta^*[(\alpha^*\beta^*)^2 + (\delta^*)^2 + (\gamma^*)^2] + \eta^*\gamma^* \tag{19}$$

$$\text{Var}(Y^*) = \zeta^{*2}[(\alpha^*\beta^*)^2 + (\delta^*)^2 + (\gamma^*)^2] + 2\zeta^*\eta^*\gamma^* + (\eta^*)^2 + (\theta^*)^2 \tag{20}$$

Step 2: Construction of alternative SCM S

The covariance matrix Σ has elements:

$$\mathrm{Var}(Z) = \alpha^2 + \kappa^2 \tag{21}$$

$$\mathrm{Cov}(Z, A) = \alpha^2\beta + \kappa(\kappa\beta + \gamma) \tag{22}$$

$$\mathrm{Var}(A) = (\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2 \tag{23}$$

$$\mathrm{Cov}(Z, Y) = \zeta(\alpha^2\beta + \kappa(\kappa\beta + \gamma)) + \eta\kappa \tag{24}$$

$$\mathrm{Cov}(A, Y) = \zeta[(\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2] + \eta(\kappa\beta + \gamma) \tag{25}$$

$$\mathrm{Var}(Y) = \zeta^2[(\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2] + 2\zeta\eta(\kappa\beta + \gamma) + \eta^2 + \theta^2 \tag{26}$$

Step 3: Parameter matching

To achieve $\Sigma = \Sigma^*$, we need:

$$\alpha^2 + \kappa^2 = (\alpha^*)^2 \tag{27}$$

$$\alpha^2\beta + \kappa(\kappa\beta + \gamma) = (\alpha^*)^2\beta^* \tag{28}$$

$$(\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2 = (\alpha^*\beta^*)^2 + (\delta^*)^2 + (\gamma^*)^2 \tag{29}$$

$$\zeta(\alpha^2\beta + \kappa(\kappa\beta + \gamma)) + \eta\kappa = \zeta^*(\alpha^*)^2\beta^* \tag{30}$$

$$\zeta[(\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2] + \eta(\kappa\beta + \gamma) = \zeta^*[(\alpha^*\beta^*)^2 + (\delta^*)^2 + (\gamma^*)^2] + \eta^*\gamma^* \tag{31}$$

$$\zeta^2[(\alpha\beta)^2 + (\kappa\beta + \gamma)^2 + \delta^2] + 2\zeta\eta(\kappa\beta + \gamma) + \eta^2 + \theta^2 = (\mathrm{Var}(Y^*)) \tag{32}$$

Step 4: Solution construction

We choose $\kappa \neq 0$ such that $|\kappa| < |\alpha^*|$. We set

$$\alpha = \sqrt{(\alpha^*)^2 - \kappa^2}, \tag{33}$$

$$\beta = \frac{(\alpha^*)^2\beta^*}{\alpha^2 + \kappa^2} = \beta^* \quad \text{(from (27) and (28))}, \tag{34}$$

$$\delta = \delta^*, \tag{35}$$

$$\kappa\beta + \gamma = \pm\sqrt{(\gamma^*)^2 - (\alpha^*\beta^*)^2 + (\alpha\beta)^2} \quad \text{(from (29))}. \tag{36}$$

Since $\alpha\beta = \alpha\beta^* = \frac{\alpha}{\alpha^*}\alpha^*\beta^*$, we have .

$$\alpha\beta)^2 = \frac{\alpha^2}{(\alpha^*)^2}(\alpha^*\beta^*)^2 = \frac{(\alpha^*)^2 - \kappa^2}{(\alpha^*)^2}(\alpha^*\beta^*)^2. \tag{37}$$

Therefore, we have

$$\kappa\beta + \gamma = \pm\sqrt{(\gamma^*)^2 + \frac{\kappa^2}{(\alpha^*)^2}(\alpha^*\beta^*)^2}. \tag{38}$$

From Eq. (30) and Eq. (31), we can solve for η via

$$\eta = \frac{\zeta^*(\alpha^*)^2\beta^* - \zeta(\alpha^2\beta + \kappa(\kappa\beta + \gamma))}{\kappa}. \tag{39}$$

Finally, θ is determined from Eq. (32).

Step 5: Existence Verification

The system has 8 parameters $(\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta, \kappa)$ and 6 constraints (the 6 unique entries of the covariance matrix). Since $\zeta \neq \zeta^*$ is fixed and $\kappa \neq 0$ is chosen, we have 6 remaining parameters for 6 constraints. The key observation is that the introduction of confounding ($\kappa \neq 0$) creates additional correlation structures that can compensate for the change in the causal effect ζ, allowing the observational distribution to remain unchanged.

□

Proof of Lemma B.2. Let $\mathcal{S} = (\alpha, \beta, \gamma, \delta, \zeta, \eta, \theta)$ be any SCM from the linear IV class in Eq. (10). Then, the following coefficients are identified via observational data alone:

$$\alpha = \sqrt{\mathrm{Var}(Z)}, \tag{40a}$$

$$\beta = \mathbb{E}[Y \mid Z = 1], \tag{40b}$$

$$\zeta = \frac{\zeta\beta}{\beta} = \frac{\mathbb{E}[\zeta(\beta + \delta\epsilon_A)] - \mathbb{E}[\zeta(\delta\epsilon_A)]}{\beta} = \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{\mathbb{E}[A|Z = 1] - \mathbb{E}[A|Z = 0]} \tag{40c}$$

as well as the combination of coefficients

$$\delta^2 + \gamma^2 = \mathrm{Var}(A \mid Z), \qquad \text{and} \qquad \eta^2 + \theta^2 = \mathrm{Var}(Y \mid A) \tag{41}$$

and the back-door adjustment

$$\mathbb{E}[Y \mid A = 1] - E[Y \mid A = 0] = \zeta + \eta(\mathbb{E}[U \mid A = 1] - \mathbb{E}[U \mid A = 0]) \tag{42}$$

$$= \zeta + \frac{\eta\gamma}{\gamma^2 + \delta^2 + (\beta\alpha)^2}. \tag{43}$$

Note that the back-door adjustment is biased for ζ due to unobserved confounding.

Noiseless treatment case. Let now $\mathcal{S}^* = (\alpha^*, \beta^*, \gamma^*, \delta^*, \zeta^*, \eta^*, \theta^*)$ denote an arbitrary fixed SCM from the linear IV class. We start by constructing $\mathcal{S}_1 = (\alpha_1, \beta_1, \gamma_1, \delta_1 = 0, \zeta_1, \eta_1, \theta_1)$ such that $\mathbb{P}^{\mathcal{S}_1}_{\text{obs}} = \mathbb{P}^{\mathcal{S}^*}_{\text{obs}}$. Because of Eq. (40), we set

$$\alpha_1 = \alpha^*,\ \beta_1 = \beta^*,\ \zeta_1 = \zeta^*. \tag{44}$$

Furthermore, setting $\delta_1 = 0$ implies due to Eq. (41) that

$$\gamma_1^2 = {\delta^*}^2 + {\gamma^*}^2. \tag{45}$$

Due to Eq. (42), it must holds that

$$\frac{\eta_1\gamma_1}{\gamma_1^2 + (\beta_1\alpha_1)^2} = \frac{\eta^*\gamma^*}{{\delta^*}^2 + {\gamma^*}^2 + (\beta^*\alpha^*)^2}, \tag{46}$$

which implies that

$$\eta_1^2 = \frac{{\eta^*}^2{\gamma^*}^2}{{\delta^*}^2 + {\gamma^*}^2}. \tag{47}$$

Finally, due to Eq. (41), we yield

$$\theta_1^2 = {\eta^*}^2 + {\theta^*}^2 - \frac{{\eta^*}^2{\gamma^*}^2}{{\delta^*}^2 + {\gamma^*}^2}, \tag{48}$$

which means that every parameter of $\mathcal{S}_1$ has a unique solution in terms of parameters of $\mathcal{S}^*$ under the constraints of preserving the observational distribution.

Noiseless outcome case. We now construct $\mathcal{S}_2 = (\alpha_2, \beta_2, \gamma_2, \delta_2, \zeta_2, \eta_2, \theta_2 = 0)$ such that $\mathbb{P}^{\mathcal{S}_2}_{\text{obs}} = \mathbb{P}^{\mathcal{S}^*}_{\text{obs}}$. Again, Eq. (40) implies that

$$\alpha_2 = \alpha^*,\ \beta_2 = \beta^*,\ \zeta_2 = \zeta^*, \tag{49}$$

and setting $\theta_2 = 0$ implies due to Eq. (41) that

$$\eta_2^2 = {\eta^*}^2 + {\theta^*}^2. \tag{50}$$

Due to Eq. (42), it must holds that $\frac{\eta\gamma}{\gamma^2+\delta^2+(\beta\alpha)^2} = \frac{\eta\gamma}{{\delta^*}^2+{\gamma^*}^2+(\beta^*\alpha^*)^2}$ which implies that

$$\gamma_2^2 = \frac{{\eta^*}^2{\gamma^*}^2}{{\eta^*}^2 + {\theta^*}^2}. \tag{51}$$

Finally, due to Eq. (41), we have

$$\delta^2 = {\eta^*}^2 + {\gamma^*}^2 - \frac{{\eta^*}^2{\gamma^*}^2}{{\eta^*}^2 + {\theta^*}^2}, \tag{52}$$

which means that every parameter of $\mathcal{S}_2$ has a unique solution in terms of parameters of $\mathcal{S}^*$ under the constraints of preserving the observational distribution. □

D IMPLEMENTATION DETAILS

D.1 IMPLEMENTATION DETAILS OF DATA PRIOR

Data Prior. (i) For each covariate cluster C_i containing latent nodes, we sample a random MLP-style graph over $\mathrm{pa}(C_i)$ by drawing biases and edge-weights from Π_{C_i} and then pruning edges at random to ensure acyclicity. We evaluate this graph with tanh activations and noise (from normal, uniform, Laplace, or logistic distribution) to produce continuous features, then apply randomized thresholds to discretize or binarize a subset, yielding mixed-type covariates via our unstructured BNN prior. (ii) For treatment (and outcome) clusters C_j of purely observed nodes, we instantiate a second BNN $f_\theta^{(j)}$ over $\mathrm{pa}(C_j)$ (with $\theta \sim \Pi_{C_j}$ and the same acyclicity constraint). We forward-propagate the covariates through $f_\theta^{(j)}$ with injected noise to compute a scalar propensity score, then threshold to assign a binary treatment. We forward-propagate both covariates and treatment to obtain potential outcomes. The resulting treatment (and outcome) are sampled from our structured BNN prior.

We sample covariates from a DAG-structured SCM by drawing a random MLP-like directed graph and assigning each node a bias, edge weights sampled from prior distributions. The resulting MLP-like graph is transformed into a DAG by randomly dropping edges, and structural equations with tanh activations and heterogeneous noise distributions (normal, uniform, Laplace, or logistic) generate continuous features. Then we apply a randomized feature transformation that discretizes some features and binarizes others, yielding mixed-type covariates. Next, we assign binary treatments via a separate randomly instantiated MLP and forward-propagate each covariate with injected noise to compute a propensity score.

Input format. CausalFM operates as an in-context learner like other foundation models, meaning it approximates Bayesian inference by conditioning on a dataset provided in its context window. Therefore, it requires an entire dataset to make predictions for specific query samples. (1) Input structure: The model accepts a dataset $\mathcal{D}_n = \{(x_i, a_i, y_i)\}_{i=1}^n$ acting as the context (or support set) and a query point x_{query} (or a batch of query points). (2) Mechanism: The transformer processes the entire sequence of observed data $\mathcal{D}_n$ using self-attention to extract context-dependent representations. It then outputs the posterior predictive distribution for the causal quantity (e.g., CATE given the context $\mathcal{D}_n$ and the specific query x_{query}. (3) Comparison to fine-tuning baselines: In practice, standard baselines require an explicit training phase on a training set before evaluation. In contrast, CausalFM takes the "training" data as the input context (support set) and directly generates predictions for the test data (query set) in a single forward pass.

D.2 IMPLEMENTATION DETAILS OF OUR METHOD

We encode observational data as tokens, and the embedded tokens are then processed through a transformer where attention is applied between the observations. We use transformer-based PFN as an encoder to extract a task- or context-dependent representation from input data. This representation is then passed to a Gaussian mixture model (GMM) head, which predicts the parameters of a GMM, including mixture weights, means, and standard deviations. The model outputs a mixture distribution over the target variable, and is trained end-to-end using the negative log-likelihood (NLL) of the observed targets under the predicted GMM. This enables uncertainty-aware and multi-modal predictions while leveraging the few-shot generalization capabilities of our model.

We instantiate a per-feature transformer tailored to CATE estimation. For a mini-batch with sequence length $S = S_{\text{supp}} + S_{\text{query}}$ (query set followed by support set). Confounders $X \in \mathbb{R}^{S\times B\times F_x}$, treatment $A \in \mathbb{R}^{S\times B\times F_a}$, and factual outcomes $Y \in \mathbb{R}^{S\times B\times F_y}$ are encoded as tokens. To prevent label leakage, we split at $S_{\text{supp}} = \lfloor 0.8\, S \rfloor$ and set $A_t = \mathrm{NaN}$ and $Y_t = \mathrm{NaN}$ for $t \geq S_{\text{supp}}$ (on the query set). The model thus observes (X, A, Y) on support steps and learn to infer CATEs for the query set from X only.

The X stream uses a feature encoder, while A and Y pass through a NaN-indicator handler followed by feature projections. We concatenate the three streams along the token axis to obtain $H_0 \in \mathbb{R}^{B\times S\times (F_g+2)\times E}$, add a feature-token positional embedding, and process H_0 with L transformer encoder blocks (self-attention only). We pool over tokens to produce feature $Z \in \mathbb{R}^{B\times S\times E}$. After lightweight MLP maps Z to a scalar, a 1D K-component GMM head outputs mixture parameters (π, μ, σ) via $\pi = \mathrm{softmax}(W_\pi z/T), \mu = W_\mu z, \sigma = \mathrm{softplus}(W_\sigma z) + \varepsilon$, for each $z \in \mathbb{R}^E$. Our

training loss is the Gaussian-mixture negative log-likelihood (GMM-NLL). We thus obtain the distribution

$$p(\tau \mid x) = \sum_{k=1}^{K} \pi_k(x) \mathcal{N}\left(\mu_k(x), \sigma_k^2(x)\right) \tag{53}$$

And the CATE can be computed through

$$\hat{\tau}(x) = \mathbb{E}[\tau \mid x] = \sum_{k=1}^{K} \pi_k(x) \mu_k(x) \tag{54}$$

We use embedding size $E = 128$, $n_{\text{heads}} = 4$, feed-forward dimension $4E$, $L = 10$ encoder layers, GELU activations, and feature grouping size $= 1$ (per-feature tokens). For the GMM head we set $K = 5$, temperature $T = 1.0$, and variance floor $\varepsilon = 10^{-3}$. We train with Adam (learning rate 10^{-3}, weight decay 10^{-5}), batch size 16, and up to 150 epochs. We use early stopping on validation loss. Empirically, the total training time for causalFM model is about 24 hours on an NVIDIA A100 GPU.

We implement our CausalFM using PyTorch. Our model implementation builds upon the TabPFN architecture (Hollmann et al., 2023) from `https://github.com/PriorLabs/TabPFN/tree/main`.

D.3 Implementation details of baselines

For the standard CATE setting baselines, we follow the implementation from `https://github.com/AliciaCurth/CATENets/tree/main` for most of the CATE estimators, including S-learner (Künzel et al., 2019), T-learner (Künzel et al., 2019), TARNet (Shalit et al., 2017b), X-leaner (Künzel et al., 2019), DR-learner (Kennedy, 2023b), RA-learner (Curth & van der Schaar, 2021). For the foundation model baselines, we follow the author implementation from `https://github.com/vdblm/CausalPFN/tree/main` for CausalPFN (Balazadeh et al., 2025); we follow the author implementation from `https://github.com/jr2021/Do-PFN` for DoPFN (Robertson et al., 2025).

For the IV setting, we follow the implementation from `https://github.com/DennisFrauen/MRIV-Net/tree/main/models` for the most of the IV methods, including KIV (Singh et al., 2019), DFIV (Xu et al., 2021), DeepIV (Hartford et al., 2017), DeepGMM (Bennett et al., 2019), DMLIV (Syrgkanis et al., 2019). For each dataset and method, we evaluated 5 repetitions, each with a different random seed. All methods used the same train-test split.

E SYNTHETIC DATA GENERATION FOR THE STANDARD CATE ESTIMATION SETTING

We construct the standard CATE estimation datasets by sampling covariates X, treatment A, and continuous outcomes Y. The design induces rich nonlinearity while preserving strong ignorability $(A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X)$.

E.1 COVARIATES VIA A DAG-STRUCTURED SCM

We first sample a layered directed graph (an MLP-like DAG), then evaluate a structural causal model (SCM) on its nodes and expose a random subset as observed features.

Graph. Sample number of layers L_X and hidden size H_X from simple discrete priors (see "Hyperparameters" below). Build a layered graph with H_X nodes per layer and fully connect layer ℓ to $\ell+1$. Randomly drop a fraction p^X_{drop} of inter-layer edges to sparsify while keeping acyclicity.

Node equations and noise. For each node j, sample weights $\{w_{jk}\}_{k\in \mathrm{pa}(j)}$, bias b_j, and an exogenous noise distribution $\varepsilon_j \sim \mathcal{D}_j$, where $\mathcal{D}_j$ is drawn from a meta-prior over {Normal, Uniform, Laplace, Logistic} with a random scale. Nodes are evaluated in topological order:

$$s_j \;=\; \sum_{k\in \mathrm{pa}(j)} w_{jk}\, x_k \;+\; b_j \;+\; \varepsilon_j, \qquad x_j \;=\; \tanh(s_j), \tag{55}$$

with the convention $\sum_{k\in\varnothing}(\cdot) = 0$ for roots. Let $U_X = \{\varepsilon_j\}$ denote the collection of all node noises.

Observed features. Sample a feature index set $\mathcal{F} \subseteq V$ with $|\mathcal{F}| = d$ uniformly from all graph nodes. A single observation $X \in \mathbb{R}^d$ is obtained by re-sampling U_X, evaluating (55) over the DAG, and reading out $X = (x_j)_{j\in\mathcal{F}}$. Each sample uses independent U_X.

Feature typing and transformations (Optional). Each selected feature x_j is assigned a random type from {continuous, binary, categorical}. Continuous features are kept in their raw form $x_j \in (-1, 1)$. Binary features are obtained by mapping x_j through a logistic function and drawing a Bernoulli sample. For categorical features, we first sample a base distribution $\pi^0 \in \Delta^{K-1}$ over K categories from a Dirichlet prior. To make the distribution depend on the DAG value x_j, we introduce a fixed direction vector $v \in \mathbb{R}^K$ (normalized) and scale $\alpha > 0$, and form

$$\pi(x_j) \;=\; \mathrm{softmax}\big(\log \pi^0 \;+\; \alpha\, x_j\, v\big). \tag{56}$$

The observed categorical feature is then sampled as $X_i \sim \mathrm{Categorical}(\pi(x_j))$.

E.2 TREATMENT ASSIGNMENT

Given X, we compute a stochastic logit via a feed-forward network with layer-wise exogenous noise and then sample a Bernoulli treatment $A \sim f_A(X, U_A)$.

Network. Sample depth $L_A \geq 3$ and hidden width H_A. Let $h^{(0)} = X \in \mathbb{R}^d$ be the input layer. For hidden layers $\ell = 1, \ldots, L_A - 1$,

$$s^{(\ell)} \;=\; W^{(\ell)} h^{(\ell-1)} \;+\; b^{(\ell)} \;+\; \varepsilon^{(\ell)}, \qquad h^{(\ell)} \;=\; \tanh\big(s^{(\ell)}\big), \tag{57}$$

and the (scalar) output logit

$$s_A \;=\; w^\top h^{(L_A-1)} \;+\; b \;+\; \varepsilon^{(L_A)}. \tag{58}$$

We define the propensity $p = \sigma(s_A)$ and sample

$$A \sim \mathrm{Bernoulli}(p). \tag{59}$$

Let $U_A = \big({\boldsymbol{\varepsilon}^{(\ell)}}_{\ell=1}^{L_A}, , U_B\big)$ collect all exogenous noises of the network and the random variable U_B used for the Bernoulli sampling.

E.3 CONTINUOUS OUTCOME

For each unit, we compute the potential outcomes $Y(0)$ and $Y(1)$ using the *same* exogenous noise U_Y.

Network. Sample depth $L_Y \geq 3$ and width H_Y; optionally drop a fraction p^Y_{drop} of hidden edges to induce sparsity. For a given treatment level $a \in \{0, 1\}$, the input is X and A, then for hidden layers

$$t^{(\ell)}(a) \;=\; V^{(\ell)} h^{(\ell-1)}(a) \;+\; c^{(\ell)} \;+\; \xi^{(\ell)}, \qquad h^{(\ell)}(a) \;=\; \tanh\big(t^{(\ell)}(a)\big), \tag{60}$$

with $h^{(0)}(a) = [X, a]$, and the scalar output logit

$$Y(a) \;=\; v^\top h^{(L_Y-1)}(a) \;+\; c \;+\; \xi^{(L_Y)}. \tag{61}$$

The factual outcome is

$$Y = AY(1) + (1 - A)\, Y(0) \tag{62}$$

Let $U_Y = \{\xi^{(\ell)}\}_{\ell=1}^{L_Y}$ denote outcome-network noises; *the same U_Y is reused when constructing $Y(0)$ and $Y(1)$ for the same unit*.

E.4 INDEPENDENCE AND IDENTIFICATION

All exogenous noises are sampled independently across mechanisms and samples: $U_X \perp U_A \perp U_Y$ and i.i.d. across units. Hence strong ignorability holds:

$$A \perp\!\!\!\perp \{Y(0), Y(1)\} \mid X, \qquad 0 < \Pr(A{=}1 \mid X) < 1, \tag{63}$$

with overlap ensured by the sigmoid in (58) to (59).

E.5 HYPERPARAMETERS AND PRIORS (AS USED IN OUR CODE)

We use simple, reproducible priors for architecture, weights, and noises:

- **Covariate DAG:** $L_X \sim \text{Unif}\{3, 4, 5, 6\}$, $H_X \sim \text{Unif}\{15, \ldots, 40\}$, edge-drop p^X_{drop}=0.5.
- **Treatment net:** $L_A \sim \text{Unif}\{3, 4\}$, $H_A \sim \text{Unif}\{8, \ldots, 20\}$.
- **Outcome net:** $L_Y \sim \text{Unif}\{3, 4, 5\}$, $H_Y \sim \text{Unif}\{10, \ldots, 25\}$, edge-drop p^Y_{drop}=0.4.
- **Weights/biases:** i.i.d. $w, b \sim \mathcal{N}(0, \sigma_w^2)$ with task-specific σ_w.
- **Node noises:** for each node, draw a type in {Normal, Uniform, Laplace, Logistic} and a scale from a wide range; sample fresh noises per unit and layer as in (55), (57)–(58), (60)–(61).
- **Activation:** $\tanh$ for all hidden layers; output layers are linear (logits).
- **Features observed:** choose $\mathcal{F}$ uniformly at random from all DAG nodes, $|\mathcal{F}| = d$.

E.6 GENERATION PIPELINE

For each dataset:

1. Sample the covariate DAG, parameters, and noises; for each unit, evaluate the DAG in topological order to obtain X by reading nodes in $\mathcal{F}$.
2. Given X, construct the treatment network with U_A to get p and sample $A \sim \text{Bernoulli}(p)$.
3. For outcomes, sample U_Y once per unit and use it to compute $Y(0)$ and $Y(1)$ via (61).

E.7 SYNTHETIC DATASETS SIZE

We sample 10000 synthetic training datasets from data prior with different data generation mechanism. Each training datasets contain 1024 data samples. The feature dimensions are also different across the datasets, ranging from 10 to 100. The features are mixed data type with continuous, binary and categorical.

F SYNTHETIC DATA GENERATION FOR THE INSTRUMENTAL VARIABLES (IV) SETTING

We aim at estimating CATEs from observational data under unobserved confounding using IVs. In contrast to the standard CATE setting, where strong unconfoundedness holds, our IV datasets intentionally violate unconfoundedness by introducing an unobserved confounder U that affects both treatment A and outcome Y. Identification is instead driven by an instrument Z that (i) is relevant for A, (ii) has no direct path to Y beyond A (exclusion), and (iii) is conditionally independent of U given X.

Key differences vs. standard CATE. (i) *Ignorability is broken*: $A \not\perp \{Y(0), Y(1)\} \mid X$ due to $U \to A$ and $U \to Y$. (ii) We introduce an *instrument* Z with $Z \perp U \mid X$, $Z \not\perp A \mid X$, and no $Z \to Y$ edge (exclusion). (iii) Outcomes are generated via an *additive* structural form $Y = f(X, A) + g(X, U) + \varepsilon_Y$ with f and g deterministic neural networks; the same ε_Y is reused across $Y(0)$ and $Y(1)$ for a unit to ensure counterfactual consistency.

F.1 COVARIATES AND LATENT CONFOUNDERS VIA A DAG-STRUCTURED SCM

We reuse the DAG-SCM from the standard setting to produce a wide set of base variables W, then split it into observed covariates X and unobserved confounders U. Thus we have different strength of the unobserved confounders from weak to sufficiently strong.

Graph and node equations. Sample number of layers L_X and hidden size H_X, build a layered DAG (fully connect layer ℓ to $\ell{+}1$), and drop a fraction p^X_{drop} of inter-layer edges to sparsify. For each node j, sample weights $\{w_{jk}\}_{k \in \text{pa}(j)}$, bias b_j, and a node-specific exogenous noise $\varepsilon_j \sim \mathcal{D}_j$ (type and scale drawn once per node). Evaluate in topological order

$$s_j \;=\; \sum_{k \in \text{pa}(j)} w_{jk}\, v_k \;+\; b_j \;+\; \varepsilon_j, \qquad v_j \;=\; \tanh(s_j). \tag{64}$$

Draw a feature index set for $W = (v_j)$ with $|W| = d_X + d_U^{\max}$, and then sample the actual confounder dimension $d_U \in \{2, \ldots, 5\}$ uniformly. Split $U \in \mathbb{R}^{d_U}$ from the first d_U coordinates of $W, X \in \mathbb{R}^{d_X}$ from the next d_X coordinates. Node noises $\{\varepsilon_j\}$ are drawn independently per unit.

F.2 INSTRUMENT VARIABLE

We generate Z from X only, ensuring $Z \perp U \mid X$ by construction and precluding any direct $U \to Z$ path. Let ϕ_Z be a feed-forward network with input X and no layer-wise exogenous noise; the network parameters are sampled once per dataset and then fixed. For a unit with covariates X,

$$s_Z \;=\; \phi_Z(X), \qquad Z \;=\; \begin{cases} \text{Bernoulli}\big(\sigma(s_Z)\big), & \text{binary instrument,} \\ s_Z, & \text{continuous instrument.} \end{cases} \tag{65}$$

We randomly choose between the binary and continuous variants when creating datasets. Relevance is induced via the $Z \to A$ path in the treatment mechanism below.

Note that the instrument variable Z has a direct influence on the treatment A, but does not have a direct effect on the outcome Y.

F.3 TREATMENT VARIABLE

Given (X, Z, U), treatment is generated via a deterministic network ϕ_A followed by a Bernoulli draw. There is *no* layer-wise noise inside ϕ_A; the only randomness is the terminal Bernoulli. For a unit,

$$s_A \;=\; \phi_A\big([X; Z; U]\big), \qquad p \;=\; \sigma(s_A), \qquad A \sim \text{Bernoulli}(p). \tag{66}$$

This introduces $U \to A$ and hence breaks ignorability, while maintaining $Z \perp U \mid X$ and $Z \to A$ relevance.

F.4 Outcome variables

The instrument variable Z has no direct effect on the outcomes. Outcomes are generated additively from a treatment channel f and a confounding channel g, both deterministic MLPs with inputs $[X; A]$ and $[X; U]$, respectively. Let $\varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ be an i.i.d. scalar noise drawn once per unit,

$$Y(a) \;=\; f\big(X, a\big) \;+\; g(X, U) \;+\; \varepsilon_Y, \qquad a \in \{0, 1\}, \tag{67}$$

$$Y \;=\; A\,Y(1) + (1 - A)\,Y(0). \tag{68}$$

By construction there is no $Z \to Y$ edge (exclusion), since Z influences Y only through A.

F.5 Independence and identification (IV)

All exogenous noises are sampled independently across units and mechanisms. The IV conditions hold by construction,

$$\text{(Independence)} \quad Z \perp U \mid X, \tag{69}$$

$$\text{(Exclusion)} \quad Y(a) \text{ depends on } X, a, U \text{ and } \varepsilon_Y \text{ only (no } Z), \tag{70}$$

$$\text{(Relevance)} \quad Z \not\perp A \mid X. \tag{71}$$

F.6 Hyperparameters and priors

We use simple priors mirroring our implementation:

- **DAG-SCM for (X, U):** $L_X \sim \mathrm{Unif}\{2, 3, 4, 5\}$, $H_X \sim \mathrm{Unif}\{10, \ldots, 50\}$, edge-drop p_{drop}^X=0.4; node noises ε_j draw a type in {Normal, Uniform, Laplace, Logistic} with random scale.
- **Instrument net ϕ_Z:** depth $L_Z \geq 3$, width $H_Z \sim \mathrm{Unif}\{8, \ldots, 30\}$; output is either Bernoulli with $\sigma(s_Z)$ (binary Z) or real-valued s_Z (continuous Z); no layer-wise noise.
- **Treatment net ϕ_A:** depth $L_A \geq 3$, width $H_A \sim \mathrm{Unif}\{8, \ldots, 30\}$; *no* layer-wise noise; $A \sim \mathrm{Bernoulli}(\sigma(s_A))$.
- **Outcome nets f, g:** depths $L_f, L_g \sim \mathrm{Unif}\{3, \ldots, 6\}$, widths $H_f, H_g \sim \mathrm{Unif}\{10, \ldots, 25\}$; $\varepsilon_Y \sim \mathcal{N}(0, \sigma_Y^2)$ with $\sigma_Y = 0.5$ by default.
- **Weights/biases:** i.i.d. $w, b \sim \mathcal{N}(0, 1)$ sampled once per dataset; $\tanh$ activations.
- **Strength:** $d_U \sim \mathrm{Unif}\{2, \ldots, 5\}$.

F.7 Generation pipeline (IV)

For each dataset, we execute:

1. Sample the covariate DAG and parameters; for each unit, evaluate (64) to obtain a wide matrix then split it into (U, X).
2. Given X, compute the instrument Z via (65) (binary or continuous and mixed).
3. Given (X, Z, U), compute the treatment propensity $p = \sigma(s_A)$ via (66) and sample $A \sim \mathrm{Bernoulli}(p)$.
4. Draw a single ε_Y per unit and compute $Y(0), Y(1)$ using (67); set the factual outcome by (68).

This yields datasets matching the classical IV graph and enabling evaluation of IV estimators.

G SYNTHETIC DATA GENERATION FOR THE FRONT-DOOR–ADJUSTED SETTING

G.1 FRONT-DOOR ADJUSTMENT DATASETS

We next construct datasets satisfying the *front-door* criterion. Besides covariates X, treatment A, and continuous outcomes Y, we introduce a mediator M. The design ensures that A affects Y only through M (no direct $A \to Y$ path), U (unobserved) confounds A and Y but does not affect M.

Covariates via a DAG-structured SCM. Identical to the standard setting: we sample a layered DAG, draw node-wise weights/biases/noise, evaluate in topological order as in (55), and expose d node values as observed features $X \in \mathbb{R}^d$. Independent exogenous noises $U_X = \{\varepsilon_j\}$ are re-sampled per unit.

Latent confounders. From the same SCM evaluation we also retain q additional node values as unobserved confounders $U \in \mathbb{R}^q$ (not revealed to learners). These induce confounding between A and Y.

G.1.1 TREATMENT ASSIGNMENT WITH LATENT CONFOUNDING

Given (X, U), we sample a feed-forward network and generate treatment. Let $L_A \geq 3$ and H_A be the depth and width, respectively. With $h^{(0)} = [X, U]$,

$$s_A^{(\ell)} \;=\; W_A^{(\ell)} h^{(\ell-1)} \;+\; b_A^{(\ell)}, \quad h^{(\ell)} \;=\; \tanh\big(s_A^{(\ell)}\big), \;\; \ell = 1, \ldots, L_A - 1, \tag{72}$$

and scalar logit

$$\tilde{s}_A \;=\; w_A^\top h^{(L_A-1)} \;+\; b_A, \qquad p \;=\; \sigma(\tilde{s}_A), \qquad A \sim \text{Bernoulli}(p). \tag{73}$$

G.1.2 MEDIATOR MECHANISM

The mediator is generated from (X, A) only, thereby enforcing the front-door exclusion $U \nrightarrow M$. Let $L_M \geq 3$, H_M be depth and width, with input $g^{(0)} = [X, A]$,

$$r^{(\ell)} \;=\; W_M^{(\ell)} g^{(\ell-1)} \;+\; b_M^{(\ell)} \;+\; \varepsilon_M^{(\ell)}, \qquad g^{(\ell)} \;=\; \tanh\big(r^{(\ell)}\big), \;\; \ell = 1, \ldots, L_M - 1, \tag{74}$$

and scalar output

$$M \;=\; w_M^\top g^{(L_M-1)} \;+\; b_M \;+\; \varepsilon_M^{(L_M)}. \tag{75}$$

We denote $U_M = \{\varepsilon_M^{(\ell)}\}_{\ell=1}^{L_M}$.

G.1.3 OUTCOME VARIABLE

Outcomes are constructed to satisfy $A \to M \to Y$ as the *only* causal path from A to Y, while allowing $U \to Y$ and $X \to Y$. We decompose Y into an M-path component and a confounding component:

$$\begin{aligned} &\textit{Mediator path:} \quad r_Y^{(\ell)} \;=\; V^{(\ell)}[h^{(\ell-1)}] \;+\; c^{(\ell)} \;+\; \xi^{(\ell)}, \;\; h^{(0)} = [X, M], \;\; h^{(\ell)} = \tanh(r_Y^{(\ell)}), \\ &\qquad\qquad\quad R(X, M) \;=\; v^\top h^{(L_Y-1)} \;+\; c \;+\; \xi^{(L_Y)}, \end{aligned} \tag{76}$$

$$\textit{Confounding path:} \quad G(X, U) \;=\; \tilde{v}^\top \tilde{h}^{(L_G-1)} \;+\; \tilde{c} \;+\; \tilde{\xi}^{(L_G)}, \quad \tilde{h}^{(0)} = [X, U], \; \tilde{h}^{(\ell)} = \tanh(\cdot), \tag{77}$$

and define the potential outcomes

$$Y(a) \;=\; R\big(X,\, M(a)\big) \;+\; G(X, U) \;+\; \epsilon_Y, \qquad M(a) \text{ computed from (74)–(75) with } A{=}a. \tag{78}$$

The factual outcome is $Y = A\,Y(1) + (1-A)\,Y(0)$. By construction there is no direct $A \to Y$ edge; A influences Y solely via M.

G.1.4 HYPERPARAMETERS AND PRIORS

- **Covariate DAG:** $L_X \sim \mathrm{Unif}\{3,4,5,6\}$, $H_X \sim \mathrm{Unif}\{15,\ldots,40\}$, edge-drop p^X_{drop}=0.5; node noises drawn per-node from {Normal, Uniform, Laplace, Logistic} with random scale.
- **Treatment net** (Eq. (72)–(73)): $L_A \sim \mathrm{Unif}\{3,4\}$, $H_A \sim \mathrm{Unif}\{8,\ldots,20\}$.
- **Mediator net** (Eq. (74)–(75)): $L_M \sim \mathrm{Unif}\{3,4\}$, $H_M \sim \mathrm{Unif}\{8,\ldots,20\}$.
- **Outcome nets** (Eq. (76)–(78)): $L_Y, L_G \sim \mathrm{Unif}\{3,4,5\}$, widths $\sim \mathrm{Unif}\{10,\ldots,25\}$; additive Gaussian ϵ_Y with task-specific scale.
- **Weights/biases:** i.i.d. $\mathcal{N}(0,\sigma_w^2)$; $\tanh$ nonlinearity.

G.1.5 GENERATION PIPELINE

For each dataset:

1. Sample the covariate DAG and evaluate to obtain (X,U) (observed X, hidden U).
2. Compute $p(A{=}1 \mid X,U)$ via (72)–(73) and sample A.
3. Evaluate the mediator M from (X,A) using (74)–(75).
4. Sample U_Y once per unit and compute $Y(0)$ and $Y(1)$ via (78) by first obtaining $M(0)$ and $M(1)$ from the mediator net; set $Y = A\,Y(1)+(1-A)\,Y(0)$.

H ADDITIONAL EXPERIMENTS

H.1 EVALUATION IN THE FRONT-DOOR ADJUSTMENT SETTING

H.1.1 BASELINES FOR FRONT-DOOR ADJUSTMENT SETTING

In contrast to the standard CATE or IV settings, there are few established baselines for the front-door case. Identification in this setting is enabled through Pearl's front-door formula (Pearl, 2009). The natural baseline is therefore the **plug-in front-door learner**, which estimates the necessary nuisance components, i.e., $P(M \mid A, X)$, $P(A \mid X)$, and $\mathbb{E}[Y \mid M, X]$ and substitutes them into the identification formula to recover causal quantities. To assess the role of model flexibility in estimating these nuisance functions, we implement the plug-in learner with different regression methods, including linear regression, Random Forests, and neural networks.

H.1.2 RESULTS FOR FRONT-DOOR ADJUSTMENT SETTING

Table 4 reports the averaged PEHE across datasets. We observe that CausalFM achieves competitive CATE estimation. Importantly, these results hold *without requiring model retraining* for our model, demonstrating the adaptability of our approach to the front-door setting.

H.2 ADDITIONAL RESULTS ON THE STANDARD CATE ESTIMATION

We report the detailed standard CATE estimation on 10 synthetic datasets in Table 8. We show our method gives the best estimation on most of the datasets.

Table 8: Standard CATE estimation on 10 synthetic datasets.

Method	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
BASELINES (A): STANDARD CATE ESTIMATORS										
S-learner (Künzel et al., 2019)	0.725	0.583	0.752	0.829	0.614	0.892	0.858	0.421	0.680	0.985
T-learner (Künzel et al., 2019)	0.652	0.496	0.666	0.746	0.552	0.849	0.761	0.357	0.608	0.931
TARNet (Shalit et al., 2017b)	0.769	0.779	0.817	0.984	0.640	0.938	1.405	0.505	0.736	0.968
RA-learner (Curth & van der Schaar, 2021)	0.620	0.421	0.644	0.706	0.523	0.808	0.646	0.353	0.613	0.759
X-learner (Künzel et al., 2019)	0.574	0.400	0.614	0.634	0.381	0.713	0.686	0.302	0.549	0.779
DR-learner (Kennedy, 2023b)	0.783	0.533	0.767	0.947	0.867	0.882	0.791	0.4230	0.653	0.998
BASELINES (B): FOUNDATION MODELS-BASED METHODS										
CausalPFN (Balazadeh et al., 2025)	0.493	0.489	0.585	0.743	0.413	0.615	0.950	0.288	0.453	0.544
DoPFN (Robertson et al., 2025)	0.417	0.313	0.228	0.679	0.591	0.475	0.497	0.551	0.610	0.827
CausalFM (ours)	0.454	0.487	0.515	0.677	0.204	0.618	0.950	0.278	0.442	0.532

Reported: PEHE (Lower = better, best in bold).

H.3 RESULTS ON OTHER DATASETS

In the following, we present detailed results of the experiments with ACIC 2016 datasets. We follow CausalPFN Balazadeh et al. (2025) obtaining data from `https://github.com/BiomedSciAI/causallib/tree/master/causallib/datasets/data/acic_challenge_2016` to evaluate on 10 different datasets with various data generation mechanism. The treatment and outcome were simulated from real-world data corresponding to 4802 individuals and 58 covariates. Table 9 shows the results of the CATE estimation.

Table 9: Standard CATE estimation on ACIC2016 datasets. Reported: PEHE (mean $\pm$ std.)

Method	PEHE
BASELINES (A): STANDARD CATE ESTIMATORS	
S-learner (Künzel et al., 2019)	1.191 ± 0.15
T-learner (Künzel et al., 2019)	1.143 ± 0.14
TARNet (Shalit et al., 2017b)	0.934 ± 0.15
RA-learner (Curth & van der Schaar, 2021)	0.762 ± 0.14
X-learner (Künzel et al., 2019)	0.519 ± 0.16
DR-learner (Kennedy, 2023b)	1.485 ± 0.18
BASELINES (B): FOUNDATION MODELS-BASED METHODS	
CausalPFN (Balazadeh et al., 2025)	0.239 ± 0.11
DoPFN (Robertson et al., 2025)	0.857 ± 0.36
CausalFM (ours)	0.638 ± 0.32

Lower = better (best in bold)

H.4 ADDITIONAL RESULTS FOR THE IV SETTING

Table 10: IV setting for CATE estimation with binary instrument variable reported with PEHE. Results for benchmarking model performance across 10 different datasets under various confounding strength.

Method	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
BASELINES (A): STANDARD CATE ESTIMATORS										
TARNet (Shalit et al., 2017b)	0.789	0.790	0.799	0.789	0.831	0.673	0.582	0.978	0.735	0.642
DR-learner (Kennedy, 2023b)	1.517	1.071	1.022	0.901	0.754	0.676	0.646	1.009	0.664	0.781
BASELINES (B): STANDARD IV ESTIMATORS										
KIV (Singh et al., 2019)	0.660	0.344	0.340	0.394	0.544	0.460	0.299	0.731	0.532	0.241
DFIV (Xu et al., 2021)	0.654	0.245	1.022	0.459	1.145	0.770	0.741	0.366	0.971	0.717
DeepIV (Hartford et al., 2017)	0.614	0.300	0.310	0.372	0.514	0.404	0.309	0.706	0.510	0.235
DeepGMM (Bennett et al., 2019)	0.704	0.403	0.440	0.599	0.569	0.486	0.292	0.737	0.566	0.232
DMLIV (Syrgkanis et al., 2019)	0.712	0.379	0.361	0.433	0.548	0.450	0.293	0.722	0.549	0.344
DRIV (Syrgkanis et al., 2019)	0.869	0.470	0.353	0.368	0.565	0.448	0.272	0.715	0.587	0.667
MRIV (Frauen & Feuerriegel, 2022)	0.759	0.632	0.698	1.011	0.348	0.860	0.929	0.707	0.562	0.380
BASELINES (C): FOUNDATION MODEL-BASED										
DoPFN (Robertson et al., 2025)	0.776	0.265	0.370	0.382	0.552	0.819	0.499	0.794	0.534	0.242
CausalFM (ours)	0.586	0.224	0.374	0.310	0.543	0.464	0.250	0.701	0.553	0.217

Reported: PEHE (mean $\pm$ standard deviation.) Lower = better (best in bold).

Table 11: IV setting for CATE estimation with continuous instrument variable reported with PEHE. Results for benchmarking model performance across 10 different datasets under various confounding strength.

Method	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
BASELINES (A): STANDARD CATE ESTIMATORS										
TARNet (Shalit et al., 2017b)	0.943	0.825	1.025	0.458	1.007	1.316	0.848	1.004	0.825	0.884
DR-learner (Kennedy, 2023b)	1.038	1.055	0.946	0.533	0.955	1.071	1.109	1.502	1.258	0.888
BASELINES (B): STANDARD IV ESTIMATORS										
KIV (Singh et al., 2019)	0.509	0.567	0.699	0.178	0.533	0.948	0.420	0.811	0.602	0.506
DFIV (Xu et al., 2021)	0.526	0.574	0.691	0.171	0.532	0.991	0.428	0.800	0.609	0.506
DeepIV (Hartford et al., 2017)	0.484	0.539	0.664	0.169	0.506	0.901	0.399	0.770	0.572	0.481
DeepGMM (Bennett et al., 2019)	0.543	0.581	0.682	0.165	0.532	1.035	0.437	0.789	0.615	0.505
DMLIV (Syrgkanis et al., 2019)	0.518	0.642	0.701	0.181	0.600	1.009	0.574	0.813	0.611	0.537
DRIV (Syrgkanis et al., 2019)	0.633	0.705	0.870	0.279	0.663	0.873	0.523	1.009	0.749	0.630
MRIV (Frauen & Feuerriegel, 2022)	0.579	0.631	0.760	0.189	0.586	1.091	0.471	0.880	0.669	0.556
BASELINES (C): FOUNDATION MODEL-BASED										
DoPFN (Robertson et al., 2025)	0.471	0.528	0.787	0.322	0.649	1.723	0.416	0.588	0.722	0.545
CausalFM (ours)	0.515	0.600	0.704	0.152	0.538	0.934	0.414	0.826	0.600	0.509

Reported: PEHE (mean $\pm$ standard deviation.) Lower = better (best in bold).

H.5 CHOICE OF PRIOR

We also analyze the robustness of our model to different choices of the prior. In particular, we vary the strength of unobserved confounding in the data-generating process, controlled by a parameter $\alpha \in [0, 1]$. The results in Fig. 2 show that our model remains robust as α increases. This again confirms the strong performance of CausalFM.

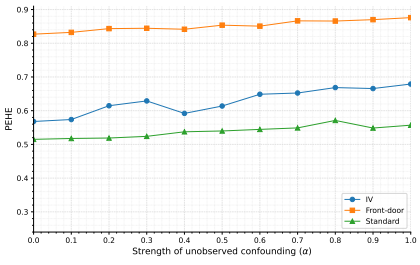

Figure 2: Robustness of our model to difference choices of the prior.