RETHINKING CAUSAL PRIOR-FITTED NETWORKS UNDER UNOBSERVED CONFOUNDING: IDENTIFIABILITY AND RANDOMIZED EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-Fitted Networks (PFNs) amortize the cost of causal effect estimation across tasks, enabling one-pass in-context prediction on new datasets without task-specific optimization. Yet efficient amortized inference does not by itself ensure causal identifiability: accurate effect estimation also depends on whether the observed context contains sufficient information to determine the causal target. Existing identifiability guarantees for causal PFNs rely on unconfounded settings such as strong ignorability, leaving what can be identified from PFN contexts unresolved under unobserved confounding, where the same context distribution may be compatible with different causal effects. We formalize this through *context identifiability*: when all compatible mechanisms agree on the target effect, ideal posterior inference recovers it asymptotically; otherwise, an irreducible bias remains that cannot in general be eliminated by more data or more accurate approximation. This framework further shows that broader causal priors can preserve ambiguity and that instrumental-variable identification still depends on additional structural assumptions. We therefore consider limited randomized evidence alongside confounded observational data and show that retaining source labels can restore identification of the target effect. Guided by this result, we propose *Randomization-Aware PFN (RA-PFN)*, which adapts a pretrained causal PFN through mixed-source prior fitting, source embeddings, and randomization-aware attention while preserving one-pass inference. Across three public benchmarks, RA-PFN achieves the lowest causal effect estimation errors under unobserved confounding with limited randomized evidence, while retaining the inference-efficiency advantage of PFN-based estimation without downstream fitting.

1 INTRODUCTION

Causal effect estimation aims to quantify how an intervention changes an outcome of interest (Pearl, 2009; Hernán & Robins, 2010), which underlies decision making in domains such as healthcare and policy (Corrao et al., 2012; Huynh et al., 2016; Wu et al., 2022). A broad range of machine learning methods have been developed for this task, including tree-based estimators (Hill, 2011; Hahn et al., 2020), meta-learners (Künzel et al., 2019), representation learning approaches (Johansson et al., 2016; Shalit et al., 2017; Shi et al., 2019), and generative models (Yoon et al., 2018; Louizos et al., 2017). Despite their effectiveness, these methods typically fit a dedicated estimator to each new dataset, requiring repeated optimization and model selection.

Prior-Fitted Networks (PFNs) (Müller et al., 2022; Hollmann et al., 2025) remove this per-dataset cost through amortized Bayesian inference: a Transformer trained on synthetic datasets drawn from a prior over mechanisms learns to approximate posterior predictions directly from a dataset placed in its context. Motivated by this paradigm, a bunch of PFN variants emerged tailored for causal effect estimations, by simulating both observations and interventions during prior fitting (Robertson et al., 2026; Balazadeh Meresht et al., 2026; Ma et al., 2026), and thereafter estimate effects on a new dataset in a single forward pass, without any dataset-specific optimization.

Despite the remarkable empirical progress achieved by PFN-based causal effect estimations, the key theoretical guarantee around the *identifiability issue* (Pearl, 2009; Rubin, 1974) remains underex-

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

plored still: in what conditions and which kind of contextual data for input, PFN-based amortized inference achieves accurate estimations of causal effects? A recent study has constructed identifiability results of PFNs by restricting the prior space to unconfounded mechanisms, i.e., the assumption of ignorability (Balazadeh Meresht et al., 2026). However, the realistic data contained unobserved confounders that affecting both treatment and outcome commonly exists across diverse tasks of causal effect estimations (Pearl, 2009). Consequently, the observational distribution may remain compatible with multiple causal effects even if it were known exactly, and existing theory in the unconfounded regime fails. To overcome this issue, existing PFN variants aim to address this ambiguity with different designs, for example propagating uncertainty over a broader causal prior (Robertson et al., 2026), or enriching the context with auxiliary identifying variables such as instruments (Ma et al., 2026). Yet these strategies leave open a basic question: ***when are causal effects theoretically identifiable from the context available to a PFN under unobserved confounding, and how can this theory guide the architecture design of amortized inference by PFNs?***

We address the first question by developing the theory of *context identifiability* towards unobserved confounding, which asks whether the causal target is uniquely determined by the distribution available to the PFN at deployment. Our analysis shows that when two mechanisms induce the same context distribution but imply different effects, at least one incurs irreducible bias for any estimator using that context; increasing model capacity or drawing more samples from the same distribution does not in general resolve the ambiguity. The bottleneck is the identifying information contained in the context rather than model expressiveness. This also shows that placing a valid instrument in context need not identify the population effect without additional structural restrictions.

This motivates the second question: what identifying information can a causal PFN reliably exploit at deployment? We consider the paradigm of a limited randomized controlled trial (RCT) sample accompanying confounded observational data (Kallus et al., 2018; Hatt et al., 2022; Wu & Yang, 2022), which serves as a common strategy towards unobserved confounding as only limited randomized samples offers strong identification evidence without too much cost during the data collection. The challenge is to enable in-context data fusion: a PFN needs to distinguish observational from randomized evidence and learn to use both without task-specific fitting. We address this challenge with *Randomization-Aware PFN (RA-PFN)*, which combines continued prior fitting on mixed-source tasks with source embeddings and randomization-aware attention to estimate causal effects in one forward pass. Our main contributions are summarized as follows:

- We formalize *context identifiability* for causal PFNs under unobserved confounding, connect it to recovery by ideal posterior inference, and characterize the bias that additional context cannot resolve. This framework clarifies the limits of broader priors and instrumental-variable assumptions.
- We use source-labeled randomized evidence to establish an identification setting and guide the design of *Randomization-Aware PFN (RA-PFN)*. Mixed-source prior fitting, source embeddings, and randomization-aware attention let the model learn from both sources while preserving one-pass inference.
- Across three public semi-synthetic benchmarks, RA-PFN achieves the lowest mean errors for conditional and average treatment effects in the main setting with strong unobserved confounding and limited randomized evidence. Further experiments examine component choices, randomized-data availability, confounding strength, and downstream computation.

2 PRELIMINARIES

2.1 CAUSAL ESTIMANDS

Our prediction target is the mean outcome that would result from assigning a treatment to units with given covariates. Let $X \in \mathcal{X}$ denote covariates, $T \in \mathcal{T}$ a discrete treatment, and $Y \in \mathbb{R}$ the observed outcome. Writing $Y(t)$ for the potential outcome under $T = t$, the Conditional Expected Potential Outcome (CEPO) is $\mu_t(x) = \mathbb{E}[Y(t) \mid X = x]$ for $t \in \mathcal{T}$ and $x \in \mathcal{X}$.

For binary $\mathcal{T} = \{0, 1\}$, comparing the two CEPOs gives the Conditional Average Treatment Effect (CATE), $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] = \mu_1(x) - \mu_0(x)$. Averaging this contrast over the target covariate distribution gives the Average Treatment Effect (ATE), $\lambda = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\tau(X)]$.

Throughout, we assume well-defined potential outcomes, consistency $Y = Y(T)$, and finite potential-outcome means. Treatment positivity is specified separately for each data source.

2.2 PFN-BASED CAUSAL EFFECT ESTIMATION

PFNs learn a reusable Bayesian prediction rule from synthetic datasets and apply it to a new context without task-specific optimization (Müller et al., 2022; Hollmann et al., 2025). We use a CEPO-targeted formulation: prior fitting provides the true conditional mean as supervision, while deployment provides only the observed context. This distinction helps explain how unobserved confounding can limit causal prediction even when the network approximates its Bayesian target accurately.

Posterior over the causal target. Let $\pi(\psi)$ denote a prior over structural causal models (SCMs) $\psi \in \Psi$. Each ψ induces a joint distribution $P^{\psi}_{\text{joint}}(X, T, \{Y(t)\}_{t\in\mathcal{T}}, Y)$ and hence an observable context distribution $P^{\psi}_{\text{ctx}}(X, T, Y)$. A context $\mathcal{D}^N_{\text{ctx}} = \{(x_i, t_i, y_i)\}_{i=1}^N$ is sampled from the latter. The PFN observes these factual rows, while the queried mean $\mu_t(x; \psi)$ depends on the underlying SCM. Given the context and query (x, t), Bayes' rule gives the following posterior predictive distribution over the CEPO:

$$\Pi^{\mu_t}\left(B \mid x, \mathcal{D}^N_{\text{ctx}}\right) = \int_{\Psi} \mathbb{I}\left\{\mu_t(x; \psi) \in B\right\} \pi\left(d\psi \mid \mathcal{D}^N_{\text{ctx}}\right), \tag{1}$$

where $\mu_t(x; \psi) = \mathbb{E}_{\psi}[Y(t) \mid X = x]$. This distribution expresses uncertainty about the conditional mean across SCMs; it is not the outcome distribution of an individual unit. The causal PFN is trained to approximate this distribution.

Prior fitting. A causal PFN $Q_\theta(\cdot \mid x, t, \mathcal{D}_{\text{ctx}})$ is trained over synthetic causal tasks. Each task samples a mechanism $\psi \sim \pi$, generates a factual context $\mathcal{D}_{\text{ctx}} \sim (P^{\psi}_{\text{ctx}})^{\otimes N}$, and provides queries (x, t) together with the corresponding ground-truth causal target $\mu_t(x; \psi)$ from the simulator. The model is optimized using the causal data-prior objective

$$\mathcal{L}(\theta) = \mathbb{E}_{\psi, \mathcal{D}_{\text{ctx}}, (x,t)}\left[-\log Q_\theta\left(\mu_t(x; \psi) \mid x, t, \mathcal{D}_{\text{ctx}}\right)\right], \tag{2}$$

where the expectation follows the mechanism, context, and query generation process induced by the causal prior. Minimizing this objective trains Q_θ to approximate the CEPO posterior predictive distribution in Eq. 1 (Robertson et al., 2026; Balazadeh Meresht et al., 2026). The simulator supplies the causal answer during training, but the network receives only factual observations. Prior fitting therefore teaches the model to infer that answer from the information available in a context.

In-context prediction. At deployment, θ is frozen. A new dataset is provided as context, queries are appended as query tokens, and a single forward pass returns $Q_\theta(\cdot \mid x, t, \mathcal{D}_{\text{ctx}})$. We use its mean as the point estimate,

$$f_\theta\left(x, t, \mathcal{D}_{\text{ctx}}\right) = \mathbb{E}_{u\sim Q_\theta(\cdot \mid x, t, \mathcal{D}_{\text{ctx}})}[u], \tag{3}$$

with CATE and ATE obtained by the contrasts in Section 2.1. All network weights remain fixed, and deployment fits no nuisance models or hyperparameters to the target dataset. The resulting efficiency leaves a separate question: does the observed context contain enough information to determine the causal answer under unobserved confounding?

3 CAUSAL PFNS UNDER UNOBSERVED CONFOUNDING

3.1 CONTEXT IDENTIFIABILITY WITH UNOBSERVED CONFOUNDERS

Intuitions of Our Theory. CausalPFN provides a theoretical analysis of causal identifiability, with strong ignorability used as a sufficient condition for identification (Balazadeh Meresht et al., 2026). When strong ignorability is violated by unobserved confounding, however, identification is no longer guaranteed: two SCMs may generate the same context distribution but imply different causal effects. Our theory shows that whenever this occurs, no causal PFN estimator using only this context can guarantee an unbiased estimate of the true causal effect for both SCMs, regardless of context size or network capacity.

Equivalence Class over the Prior SCM Space. Following the equivalence-class viewpoint in CausalPFN (Balazadeh Meresht et al., 2026), we characterize what can be learned from the distribution of the data supplied in context. As the context grows, it can reveal P^{ψ}_{ctx} increasingly well, but cannot distinguish SCMs that induce the same distribution. We therefore define the *context-equivalence class* of ψ as $[\psi]_{\mathrm{ctx}} = \{\psi' \in \Psi : P^{\psi'}_{\mathrm{ctx}} = P^{\psi}_{\mathrm{ctx}}\}$. For a fixed query, let $g(\psi)$ denote the target, such as $\mu_t(x;\psi)$ or $\tau(x;\psi)$. Identifying g does not require distinguishing all SCMs in the same class; it only requires them to agree on g. Definition 3.1 formalizes this requirement.

Definition 3.1 (Context Identifiability)**.** The target g is *context identifiable* on Ψ if it is constant within each context-equivalence class, that is, $P^{\psi}_{\mathrm{ctx}} = P^{\psi'}_{\mathrm{ctx}} \Longrightarrow g(\psi) = g(\psi')$ for all $\psi, \psi' \in \Psi$.

As the context grows, it provides more information about the context distribution, while SCMs within the same class remain indistinguishable. Under unobserved confounding, this matters because the same observed association can arise from different combinations of treatment effects and hidden common causes. Theorem 3.2 shows when learning the context distribution suffices for ideal posterior recovery and when the remaining ambiguity prevents unbiased estimation across all compatible SCMs. Proofs are in Appendices B.1 and B.2.

Theorem 3.2 (Context Identifiability Governs In-Context Estimation)**.** *Let $\widehat{g}_N$ be any estimator computed from an N-row i.i.d. context. For a prior π on Ψ, draw $\widetilde{\psi} \sim \pi$ and let $m_N = \mathbb{E}_\pi[g(\widetilde{\psi}) \mid \mathcal{D}^N_{\mathrm{ctx}}]$ be the idealized posterior mean a causal PFN is trained to approximate.*

(i) Under the regularity conditions stated in Appendix B.1, if g is context identifiable (Definition 3.1), then $m_N \to g(\widetilde{\psi})$ almost surely under the joint prior–data law as $N \to \infty$.

(ii) If instead two mechanisms $\psi_0, \psi_1 \in \Psi$ satisfy $P^{\psi_0}_{\mathrm{ctx}} = P^{\psi_1}_{\mathrm{ctx}}$ while $\Delta = |g(\psi_0) - g(\psi_1)| > 0$, then any $\widehat{g}_N$ with finite expectation under both mechanisms satisfies

$$\max_{j \in \{0,1\}} \left|\mathbb{E}_{\psi_j}[\widehat{g}_N] - g(\psi_j)\right| \geq \frac{\Delta}{2}, \tag{4}$$

and the identical context laws also rule out consistency under both mechanisms.

For PFNs, the theorem highlights why accurate posterior approximation can still leave a biased effect estimate: learning the context distribution does not necessarily determine the causal answer. We make this distinction explicit by separating error in approximating the ideal prediction from the gap between that prediction and the true effect.

Formally, let $R = P^{\widetilde{\psi}}_{\mathrm{ctx}}$ and define $\bar{g}_\pi(R) = \mathbb{E}_\pi[g(\widetilde{\psi}) \mid R]$, the prior-weighted average reached by the ideal posterior mean as the context grows (Appendix B.1). For π-almost every ψ, the estimation error satisfies

$$|\widehat{g}_N - g(\psi)| \;\leq\; \underbrace{\left|\widehat{g}_N - \bar{g}_\pi(P^{\psi}_{\mathrm{ctx}})\right|}_{\text{estimation}} \;+\; \underbrace{\left|\bar{g}_\pi(P^{\psi}_{\mathrm{ctx}}) - g(\psi)\right|}_{\text{non-identifiability bias}}. \tag{5}$$

The first term reflects finite-context and network-approximation errors, which more context and better approximation can reduce. The second measures how far the infinite-context Bayesian prediction can be from the true causal effect, even when the context distribution is known exactly. Under unobserved confounding, SCMs that produce the same context distribution can imply different effects, so their prior-weighted average need not equal the true effect. **More samples from that distribution cannot distinguish these SCMs, and a better network can approximate their average more accurately without identifying which effect is true.** Changing prior weights can change the average, but cannot make it equal different target values across all compatible SCMs. Eliminating this ambiguity therefore requires additional context information or valid structural assumptions.

3.2 Rethinking the Design of PFNs under Unobserved Confounding

We now use Eq. 5 to examine two existing, representative approaches to unobserved confounding: Do-PFN's broader causal prior (Robertson et al., 2026) and CausalFM-IV's instrumental-variable (IV) (Ma et al., 2026) information. **The key question is whether each design eliminates the second term, even when the first is negligible.** In particular, a broader prior can retain conflicting causal answers, whereas IV-based identification depends on additional structural assumptions.

Do-PFN: a broader prior can retain identification bias. Do-PFN includes diverse, potentially confounded SCMs in its prior and represents uncertainty arising from non-identifiability (Robertson et al., 2026). **Representing this uncertainty does not guarantee that the second term in Eq. 5 vanishes.** A simple two-SCM prior illustrates the limitation: the SCMs in Appendix B.3 have the same (T, Y) distribution but ATEs $1/2$ and 0. With equal prior weights, $m_N = \bar{g}_\pi(P^\psi_{\text{ctx}}) = 1/4$ for either SCM and every context size. For the exact posterior-mean estimator, the first term in Eq. 5 is zero, whereas the second is $|1/4 - 1/2| = 1/4$ or $|1/4 - 0| = 1/4$, depending on which SCM is true. Thus, even exact posterior inference can retain bias from unresolved unobserved confounding.

CausalFM-IV: removing bias depends on valid outcome restrictions. CausalFM-IV combines IV information with additional identification assumptions, including an additive separation of treatment and latent contributions to the outcome (Ma et al., 2026). These assumptions make the second term in Eq. 5 vanish within the identified mechanism class. **This guarantee depends on the structural assumptions holding in the downstream task.** Proposition 3.3 gives an additive SCM ψ_0 with effect 1 and a nonadditive SCM ψ_1 with effect 2, both inducing the same IV context law R. If the additive prior yields $\bar{g}_\pi(R) = 1$, the gap in the second term is $|1 - 1| = 0$ under ψ_0 but $|1 - 2| = 1$ under ψ_1, which violates the additive assumption. More context cannot distinguish these SCMs, motivating our study of randomized evidence as additional identifying information.

Proposition 3.3 (Instrument Validity Alone Does Not Identify the Population Effect)**.** *There exist two SCMs with constant X and binary Z, T satisfying instrument independence, exclusion, relevance, treatment positivity given Z, monotonicity, and completeness of the conditional law of T given Z, such that*

$$P^{\psi_0}_{\text{IV}}(Z, T, Y) = P^{\psi_1}_{\text{IV}}(Z, T, Y), \qquad \tau(\psi_0) = 1, \qquad \tau(\psi_1) = 2. \tag{6}$$

Here, ψ_0 is additive in treatment and the latent contribution, whereas ψ_1 has unobserved effect modification. Their identical IV context laws imply worst-case absolute bias at least $1/2$ on this pair for any context-only estimator with finite expectation, and rule out consistency under both SCMs (proof in Appendix B.4).

4 Randomized-Evidence In-Context Learning

4.1 Randomized Evidence as an Identifying Signal

We next ask what additional evidence can resolve the ambiguity identified in Section 3. We consider *causal data fusion*, a common approach to learning under unobserved confounding that combines limited randomized data with confounded observational data (Bareinboim & Pearl, 2016; Pearl & Bareinboim, 2022; Colnet et al., 2024). Existing approaches typically fit a new model for each dataset (Kallus et al., 2018; Hatt et al., 2022; Wu & Yang, 2022). We instead place both sources in a PFN's context and learn the inference procedure across tasks, preserving efficient in-context prediction without downstream fitting.

Formally, each task indexed by ψ induces observational and randomized laws P^ψ_{obs} and P^ψ_{rct}. Let $S \in \{\text{obs}, \text{rct}\}$ denote the source, with $\mathcal{D}^{n_s}_s = \{(x_i, t_i, y_i)\}^{n_s}_{i=1}$ drawn i.i.d. from P^ψ_s for $s \in \{\text{obs}, \text{rct}\}$.

At each covariate value, randomization reveals mean treatment responses without observational assignment bias. These responses identify the target CEPO when they also apply to the target population and the trial covers its covariates and treatments. Theorem 4.1 formalizes these requirements.

Theorem 4.1 (Identification from Source-Labeled Randomized Evidence)**.** *Suppose consistency holds and the observational and randomized sources share the same CEPO: for each t and relevant x, $\mathbb{E}_\psi[Y(t) \mid X = x, S = \text{obs}] = \mathbb{E}_\psi[Y(t) \mid X = x, S = \text{rct}] = \mu_t(x; \psi)$. If $Y(t) \perp\!\!\!\perp T \mid X, S = \text{rct}$, the randomized source covers the target covariate law P^ψ_X ($P^\psi_X \ll P^\psi_{X|S=\text{rct}}$), and $P_\psi(T = t \mid X = x, S = \text{rct}) > 0$ there, then, for target-almost every x,*

$$\mu_t(x; \psi) = \mathbb{E}_\psi[Y \mid X = x, T = t, S = \text{rct}]. \tag{7}$$

Thus the CEPO is context identifiable from the source-aware law, without observational ignorability or an additive outcome restriction. The proof is in Appendix B.5.

Implications for PFN Context Design. For a causal PFN, Theorem 4.1 explains what randomized evidence adds: source-labeled randomized data can rule out causal answers that unobserved confounding leaves compatible with observational data. Under the prior and sampling conditions in Appendix B.5, ideal posterior inference recovers the target as labeled randomized evidence grows. Source labels are crucial because pooling the two sources can hide this identifying information: under i.i.d. mixture sampling, Appendix C.1 constructs two SCMs with the same pooled law $P(X, T, Y)$ and ATEs 0.2 and -0.2, but different labeled laws $P(X, T, Y, S)$. RA-PFN addresses the remaining finite-context problem of learning to exploit this source information effectively.

4.2 Architecture of Randomization-Aware PFN

We propose *Randomization-Aware PFN (RA-PFN)* to learn from mixed observational and randomized contexts while retaining one-pass inference. Source embeddings tell the model how each observation was collected, and attention biases let it adjust the influence of randomized rows (Figure 5, Appendix C.2). Continued prior fitting on mixed-source tasks trains the model to use these signals.

Concretely, RA-PFN represents its source-aware context as

$$\mathcal{D} = \left\{(x_i, t_i, y_i, s_i)\right\}_{i=1}^{N}, \qquad s_i \in \{\text{obs}, \text{rct}\}, \qquad N = n_{\text{obs}} + n_{\text{rct}}. \tag{8}$$

RA-PFN uses s_i both when embedding each context row and when computing attention over context observations.

Prior fitting on mixed-source tasks. We initialize RA-PFN from CausalPFN and continue prior fitting on tasks that contain both confounded observational rows and randomized rows. Within each task, the sources share the covariate and potential-outcome mechanisms and differ in treatment assignment. Varying the randomized proportion exposes the model to different amounts of identifying evidence. Training uses the backbone's soft-histogram implementation of Eq. 2 with the source-aware context in Eq. 8 (Appendix C.2).

Source embedding. The source embedding distinguishes observations collected under different assignment regimes, even when their (x, t, y) values coincide. Let $h_i = \phi_\theta(x_i, t_i, y_i) \in \mathbb{R}^d$ denote the backbone embedding of context row i. We augment it with a learnable embedding indexed by the data source:

$$\widetilde{h}_i = h_i + e_{s_i}, \qquad e_{\text{obs}}, e_{\text{rct}} \in \mathbb{R}^d. \tag{9}$$

These learned embeddings distinguish randomized observations from potentially confounded observational ones. They are added only to context rows, since a query asks about a treatment response and does not belong to either data source.

Randomization-aware attention bias. Source embeddings identify the source of each row. To also let the model adjust the relative attention paid to randomized observations, we add a learned bias to their attention scores. For layer ℓ and attention head h, the attention logit from a receiving row i to a context observation j is modified as

$$\widetilde{a}_{i,j}^{(\ell,h)} = \frac{\left(q_i^{(\ell,h)}\right)^\top k_j^{(\ell,h)}}{\sqrt{d_h}} + \beta_{\ell,h}\mathbb{I}(s_j = \text{rct}), \tag{10}$$

where $\beta_{\ell,h} \in \mathbb{R}$ is a learned scalar. The receiving row i may be a context or query row, while j is a context row. Randomized keys receive the additional bias; observational keys keep their original content-dependent scores.

The bias applies to both *context-to-context* and *query-to-context* attention. Each head learns whether to increase or decrease attention to randomized rows; the architecture does not prescribe that they always receive more weight. Both new modules are zero-initialized, preserving the backbone's initial behavior. These modules expose the source information motivated by the theory, while prior fitting learns how to use it for prediction.

Inference. Given a new dataset with known sources, RA-PFN constructs the context in Eq. 8 and predicts $\widehat{\mu}_{t_q}(x_q) = f_\theta(x_q, t_q, \mathcal{D})$ for each query (x_q, t_q), with CATE and ATE obtained by the contrasts in Section 2.1. As in Eq. 3, inference uses a single forward pass with frozen weights and no task-specific fitting.

Table 1: Performance on public semi-synthetic benchmarks with strong unobserved confounding and 10% RCT data. Results are mean ± standard error over realizations; lower is better. Best and second-best results are bolded and underlined. CPS and PSID PEHE values are scaled by 10^{-3}.

	ACIC		CPS		PSID	
Method	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
S-Learner	3.562±0.508	0.419±0.035	10.911±0.036	0.698±0.007	17.944±0.152	0.588±0.007
T-Learner	4.195±0.326	0.894±0.118	10.184±0.037	0.565±0.011	16.161±0.152	0.398±0.009
X-Learner	4.080±0.395	0.809±0.112	10.193±0.036	0.516±0.011	16.245±0.152	0.366±0.010
DA-Learner	3.319±0.289	0.599±0.080	10.073±0.040	0.504±0.012	15.736±0.157	0.355±0.010
DR-Learner	4.363±0.471	0.912±0.128	10.973±0.036	0.568±0.012	18.047±0.153	0.389±0.010
DML	3.329±0.489	0.468±0.063	10.102±0.041	0.501±0.013	15.925±0.156	0.347±0.011
TARNet	6.099±0.375	0.484±0.095	10.442±0.046	0.519±0.013	16.313±0.170	0.356±0.012
DragonNet	6.166±0.370	0.471±0.096	10.478±0.047	0.514±0.013	16.387±0.171	0.351±0.012
Do-PFN	4.270±0.537	0.646±0.050	12.651±0.048	0.959±0.012	24.257±0.141	1.181±0.006
CausalFM	5.016±0.511	1.027±0.005	12.878±0.016	1.000±0.000	22.701±0.130	1.000±0.000
CausalPFN	3.299±0.377	0.535±0.097	10.086±0.042	0.514±0.012	15.754±0.170	0.354±0.011
Two-Stage	3.574±0.417	0.635±0.143	10.028±0.055	0.415±0.014	16.357±0.170	0.325±0.012
CorNet	4.057±0.446	0.604±0.105	10.315±0.063	0.369±0.016	16.551±0.231	0.317±0.016
R-Learner	3.046±0.553	0.222±0.038	10.264±0.056	0.447±0.017	16.137±0.178	0.335±0.013
RA-PFN (ours)	**2.908±0.380**	**0.204±0.031**	**9.330±0.026**	**0.188±0.008**	**14.821±0.167**	**0.179±0.008**

5 EXPERIMENTS

The experiments test whether RA-PFN turns limited randomized evidence into accurate predictions under unobserved confounding while retaining PFN inference speed. We organize the evaluation around five research questions (RQs): performance against existing estimators (RQ1), source-aware module ablations (RQ2), sensitivity to randomized-data availability (RQ3), robustness to unobserved confounding (RQ4), and the accuracy–efficiency trade-off (RQ5).

5.1 EXPERIMENTAL SETTINGS

Dataset and pre-processing. We evaluate on three public semi-synthetic suites: 10 ACIC 2016 realizations (Dorie et al., 2019) and 100 realizations each from Lalonde-CPS and Lalonde-PSID (LaLonde, 1986). We retain their covariates and potential outcomes and modify treatment assignment to control unobserved confounding. Observational assignment depends on both an observed score $G(X)$ and an outcome-related latent variable U hidden from every estimator: $T^{\text{obs}} \sim \text{Bernoulli}(e(X,U))$, with pre-logit driven by $\sqrt{1-\rho_U}\,G(X) + \sqrt{\rho_U}\,U$. The default $\rho_U = 0.75$ gives strong unobserved confounding. A fraction α of each context is randomized using $T \sim \text{Bernoulli}(0.5)$. The main comparison and ablation use $\alpha = 10\%$; the availability study varies $\alpha \in \{0.01, 0.02, 0.05, 0.10, 0.20, 0.30\}$. All methods share the same 128 context rows, source allocation, realized treatments and outcomes, and held-out queries (Appendix C.3).

Baselines. We compare against three groups: (1) **task-specific causal estimators**: S-, T-, X-, and DA-learners (Künzel et al., 2019), TARNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), DML (Chernozhukov et al., 2018), and DR-Learner (Kennedy, 2023); (2) **PFN-based estimators**: Do-PFN (Robertson et al., 2026), the standard CATE version of CausalFM (Ma et al., 2026), and CausalPFN (Balazadeh Meresht et al., 2026); and (3) **task-specific data-fusion estimators**: Two-Stage (Kallus et al., 2018), CorNet (Hatt et al., 2022), and the integrative R-learner (Wu & Yang, 2022). The fusion estimators and RA-PFN receive source labels; the remaining baselines receive only (X, T, Y). CausalFM uses its standard CATE checkpoint because these contexts contain no instrument. Task-specific methods are tuned on training data only (Appendix C.5).

Training and evaluation. Starting from the official CausalPFN checkpoint, we continue prior fitting with AdamW, updating the prediction head, input and outcome normalization layers, top four Transformer layers, and source-aware modules. Hyperparameters are selected on held-out synthetic validation tasks and fixed for all public evaluations. We measure CATE prediction error using root PEHE $\text{PEHE} = \sqrt{\frac{1}{n_{\text{q}}}\sum_{i=1}^{n_{\text{q}}}(\hat{\tau}(x_i) - \tau(x_i))^2}$, and ATE error using relative absolute error

Table 2: Component ablations under the main setting, with frozen CausalPFN as reference. Results are mean ± standard error; lower is better. Best and second-best results are bolded and underlined. CPS and PSID PEHE values are scaled by 10^{-3}.

	ACIC		CPS		PSID	
Method	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}	PEHE	ϵ_{ATE}
CausalPFN	3.299±0.377	0.535±0.097	10.086±0.042	0.514±0.012	15.754±0.170	0.354±0.011
Source embedding only	<u>2.919±0.389</u>	<u>0.206±0.038</u>	**9.316±0.025**	<u>0.190±0.007</u>	<u>14.846±0.167</u>	<u>0.172±0.008</u>
Attention bias only	2.948±0.385	0.214±0.034	9.339±0.026	0.194±0.007	14.883±0.166	**0.169±0.007**
RA-PFN	**2.908±0.380**	**0.204±0.031**	<u>9.330±0.026</u>	**0.188±0.008**	**14.821±0.167**	0.179±0.008

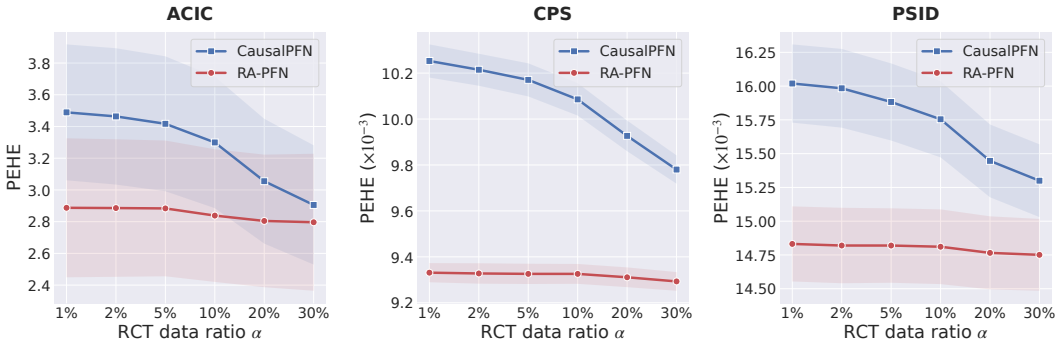

Figure 1: PEHE of RA-PFN and CausalPFN across different RCT proportions under strong unobserved confounding ($\rho_U = 0.75$). Shaded regions denote 70% confidence intervals for ACIC and 90% confidence intervals for CPS and PSID.

$\epsilon_{\text{ATE}} = \frac{|\widehat{\lambda}-\lambda|}{|\lambda|}$, where n_{q} is the number of held-out queries and $\widehat{\lambda} = n_{\text{q}}^{-1}\sum_i \widehat{\tau}(x_i)$, with λ from the true query effects. These metrics summarize prediction error rather than bias alone. Relative ATE error requires nonzero λ and is sensitive to near-zero values. Results are aggregated per realization (Appendix C.4); bold and underlined entries mark the best and second-best means.

5.2 Experimental Results

Performance Comparison (RQ1). RA-PFN achieves the lowest mean error on all six dataset–metric combinations in Table 1, improving both CATE and ATE estimation over the frozen CausalPFN backbone under strong unobserved confounding. This comparison measures the combined effect of mixed-source prior fitting and source-aware modules. Task-specific fusion estimators remain competitive, particularly the integrative R-learner on ACIC, with relative performance varying across datasets and estimands. RA-PFN obtains these results with the same reusable predictor and no downstream fitting.

Ablation Study (RQ2). This ablation asks how the two source-aware modules compare under the same continued prior-fitting protocol. *Source embedding only* removes the attention bias, and *Attention bias only* removes the source embedding; frozen CausalPFN provides a reference. Among the three training-matched variants in Table 2, full RA-PFN has the lowest mean error on four of six metrics, while one-module variants lead on the other two. The small differences support the competitiveness of the combined design, but do not establish that both modules are needed in every setting. Appendix C.1 explains the complementary role of the source-label diagnostic.

Sensitivity to Randomized Evidence Availability (RQ3). To test performance with scarce randomized evidence, we vary α at $\rho_U = 0.75$ while keeping the total context size fixed. In Figure 1, CausalPFN improves as randomized rows replace observational ones, while RA-PFN remains comparatively stable and has its largest advantage at the lowest RCT proportions. This evaluates how the trained predictor uses a small randomized subset together with observational data and its learned prior. It does not establish that a few RCT rows alone suffice for accurate CATE estimation.

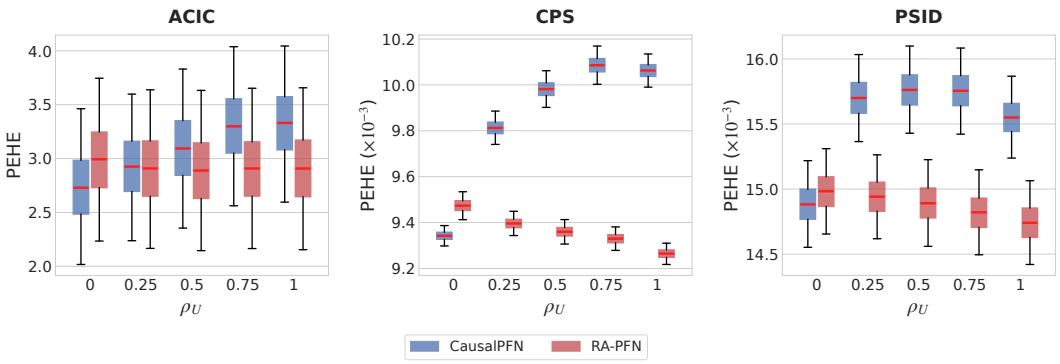

Figure 2: PEHE versus unobserved-confounding strength. Centers show realization means; boxes show interquartile ranges of bootstrap means; whiskers show 95% confidence intervals.

Robustness to Unobserved Confounding (RQ4). We further test whether RA-PFN's advantage increases as hidden causes play a larger role in observational treatment assignment. We increase ρ_U from 0 to 1 while keeping the randomized proportion fixed and the results are summarized in Figure 2. At $\rho_U = 0$, observational treatment assignment depends only on observed covariates, matching the unconfounded setting targeted by CausalPFN; RA-PFN offers little benefit and slightly trails it on some benchmarks. As the latent contribution increases, CausalPFN's error rises substantially, while RA-PFN remains comparatively stable. The growing advantage, largest at $\rho_U = 1$, supports the value of RA-PFN's adaptation in the unobserved-confounding regimes evaluated here.

Accuracy–Efficiency Trade-off (RQ5). We also test whether RA-PFN's gains in effect estimation preserve the low deployment cost of a PFN. Figure 3 compares median PEHE rank and median CPU runtime over all 210 realizations, with 1,000 queries per 128-row context. RA-PFN achieves the best median PEHE rank with runtime comparable to CausalPFN, while task-specific fusion estimators take longer under the evaluated tuning and refitting protocol. RA-PFN incorporates both data sources through in-context prediction, without fitting a new model for each dataset. Timing includes hyperparameter selection and refitting for task-specific estimators but excludes offline prior fitting and checkpoint loading (Appendix C.6). These results support efficient inference on new datasets; they do not compare total training or data-collection costs.

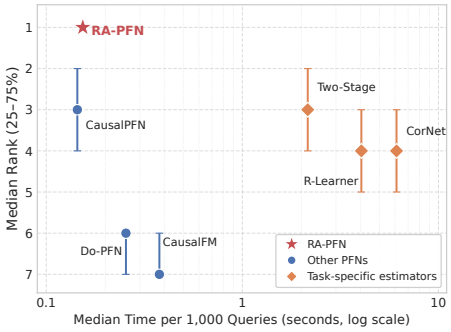

Figure 3: Median PEHE rank versus median runtime across all 210 public realizations. Error bars show rank interquartile ranges.

6 Conclusion

We studied when causal effects can be identified from a PFN's context under unobserved confounding and how this theory can guide architecture design. The central limitation is that SCMs can agree on the observed context distribution while disagreeing on the effect. Their disagreement can persist even when that distribution is known exactly. Context identifiability connects this limitation to ideal posterior inference and clarifies the limits of broader-prior and instrumental-variable designs. Source-labeled randomized evidence resolves the ambiguity when treatment responses are comparable across sources and trial coverage and treatment positivity hold. RA-PFN learns to use this evidence through mixed-source prior fitting, source embeddings, and attention biases. Across three public semi-synthetic benchmarks, it improves finite-context effect estimation while retaining one-pass inference. Extending the learned estimator to broader source shifts and outcome mechanisms remains an important direction.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving the clarity and readability of the manuscript, and suggesting alternative formulations of portions of the text. They were also used as a secondary aid to check the logical consistency and presentation of mathematical arguments. The research ideas, methodology, theoretical claims, assumptions, derivations, proofs, experimental design, and scientific conclusions were developed by the authors. AI tools were not relied upon to establish the correctness of any theoretical result; all mathematical claims and proofs were independently verified by the authors. All AI-assisted text and feedback were carefully reviewed and revised by the authors. We take full responsibility for the final content of this work, including all text, theoretical claims, and experimental conclusions.

REFERENCES

Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

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

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Lucius EJ 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.

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

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 39(1):165, 2024.

Giovanni Corrao, Federica Nicotra, Andrea Parodi, Antonella Zambon, Davide Soranna, Franca Heiman, Luca Merlino, and Giuseppe Mancia. External adjustment for unmeasured confounders improved drug–outcome association estimates based on health care utilization data. *Journal of clinical epidemiology*, 65(11):1190–1199, 2012.

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

Evangelos Dimitriou, Edwin Fong, Jens Magelund Tarp, Karla Diaz-Ordaz, and Brieuc Lehmann. Causal-icm: A data fusion framework for heterogeneous treatment effect estimation with multi-task gaussian processes. *arXiv preprint arXiv:2405.20957*, 2024.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. 2019.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Tobias Hatt, Jeroen Berrevoets, Alicia Curth, Stefan Feuerriegel, and Mihaela van der Schaar. Combining observational and randomized data for estimating heterogeneous treatment effects. *arXiv preprint arXiv:2202.12891*, 2022.

Miguel A Hernán and James M Robins. Causal inference, 2010.

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

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.

Van-Nam Huynh, Vladik Kreinovich, and Songsak Sriboonchitta. *Causal inference in econometrics*. Springer, 2016.

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

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *NeurIPS*, volume 31, 2018.

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

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.

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

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, 2017.

Yuchen Ma, Dennis Frauen, Emil Javurek, and Stefan Feuerriegel. Foundation models for causal inference via prior-data fitted networks. In *International Conference on Learning Representations*, volume 2026, pp. 79065–79098, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022.

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

Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. *Advances in Neural Information Processing Systems*, 36:6862–6901, 2023.

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

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. In *Probabilistic and causal inference: The works of Judea Pearl*, pp. 451–482. 2022.

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

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

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66:688–701, 1974.

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

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Lili Wu and Shu Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pp. 904–926. PMLR, 2022.

Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. *arXiv preprint arXiv:2201.06716*, 2022.

Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of treatment effect and confounding function. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 31(4):2987, 2025.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

Jiaqi Zhang, Joel Jennings, Agrin Hilmkil, Nick Pawlowski, Cheng Zhang, and Chao Ma. Towards causal foundation model: on duality between causal inference and attention. *arXiv preprint arXiv:2310.00809*, 2023.

CONTENTS FOR APPENDIX

A Related Work **14**

B Additional Theoretical Results **14**

B.1 Consistency under Context Identifiability 14

B.2 Proof of the Irreducible Bias Bound 15

B.3 An Observational Equivalence Example 16

B.4 An IV Construction with Unobserved Effect Modification 16

B.5 Identification with Randomized Evidence 18

C Additional Experimental Details **19**

C.1 Controlled Source-Identifiability Diagnostic 19

C.2 RA-PFN Prior-Fitting and Implementation Details 21

C.3 Public Semi-Synthetic Data Generation 22

C.4 Statistical Aggregation and Uncertainty 23

C.5 Additional Baseline Implementation Details 23

C.6 Timing Protocol 24

A Related Work

Causal PFNs and identification. Amortized causal estimators learn across tasks to reduce fitting on each new dataset (Nilforoshan et al., 2023; Zhang et al., 2023; Bynum et al., 2025). Existing theory links CEPO identifiability to posterior consistency in CausalPFN (Balazadeh Meresht et al., 2026), studies inference over possibly non-identifiable mechanisms in Do-PFN (Robertson et al., 2026), and analyzes observational equivalence and valid causal priors in CausalFM (Ma et al., 2026). We use these foundations to ask when a PFN can determine the causal effect under unobserved confounding and how the answer can guide its architecture.

Additional evidence in amortized causal inference. Additional causal evidence can help separate SCMs that observational data alone cannot distinguish. MACE-TNP integrates over uncertain graphs and mechanisms and shows that a few interventional samples can make a controlled setting identifiable (Dhir et al., 2026). Reuter et al. (2026) incorporate ancestral knowledge through learnable attention biases. These studies motivate the broader use of causal information in amortized inference. Our focus is the connection between identification under unobserved confounding and a concrete PFN design trained to combine observational and randomized evidence.

Observational and randomized evidence. Causal data fusion uses randomized evidence to help estimate effects from confounded observational data (Colnet et al., 2024). Approaches include experimental correction of observational estimates (Kallus et al., 2018), shared representation learning (Hatt et al., 2022), joint estimation of effects and confounding functions (Wu & Yang, 2022; Yang et al., 2025), and Bayesian information sharing (Dimitriou et al., 2024). RA-PFN builds on this identification strategy and changes how estimation is performed: it learns across mixed-source tasks, then predicts directly from a new context. This amortizes the per-dataset fitting used by task-specific fusion methods.

B Additional Theoretical Results

B.1 Consistency under Context Identifiability

As an i.i.d. context grows, it reveals the context distribution but cannot distinguish SCMs that share that distribution. The ideal posterior mean therefore converges to a prior-weighted average of their targets; context identifiability makes this average equal the true target. The result below supplies the limit used in Eq. 5 and proves Theorem 3.2(i), following the Bayesian identification perspective of Balazadeh Meresht et al. (2026); Robertson et al. (2026).

Proposition B.1 (Posterior Limit under Context Equivalence)**.** *Suppose the parameter and context spaces are standard Borel, $\psi \mapsto P^{\psi}_{\mathrm{ctx}}$ is measurable, and g is measurable with $\mathbb{E}_{\pi}|g(\widetilde{\psi})| < \infty$. Draw $\widetilde{\psi} \sim \pi$ and, conditionally on $\widetilde{\psi}$, an i.i.d. context sequence. Let $R = P^{\widetilde{\psi}}_{\mathrm{ctx}}$ be the random context law and $m_N = \mathbb{E}_{\pi}[g(\widetilde{\psi}) \mid \mathcal{D}^N_{\mathrm{ctx}}]$. Then*

$$m_N \longrightarrow \mathbb{E}_{\pi}[g(\widetilde{\psi}) \mid R] \quad \textit{almost surely and in } L^1 \tag{11}$$

under the joint prior–data law. This limit equals $g(\widetilde{\psi})$ almost surely if and only if $g(\widetilde{\psi})$ is measurable with respect to R, equivalently, if context identifiability holds up to a π-null set.

Proof of Proposition B.1. We first express the posterior limit as an expectation given the infinite context. We then show that, for inference about the SCM, this context carries exactly the information in its distribution. Let $\mathcal{W}$ be the standard Borel context space and let $\mathcal{P}(\mathcal{W})$ carry its usual evaluation sigma-algebra. The random probability measure $R = P^{\widetilde{\psi}}_{\mathrm{ctx}}$ is measurable by assumption. Write $\mathcal{D}^N_{\mathrm{ctx}} = (W_1, \ldots, W_N)$, $\mathcal{F}_N = \sigma(W_1, \ldots, W_N)$ and $\mathcal{F}_\infty = \sigma(W_1, W_2, \ldots)$. Since $g(\widetilde{\psi})$ is integrable, the upward convergence theorem for conditional expectations gives

$$\mathbb{E}[g(\widetilde{\psi}) \mid \mathcal{F}_N] \longrightarrow \mathbb{E}[g(\widetilde{\psi}) \mid \mathcal{F}_\infty] \quad \text{almost surely and in } L^1. \tag{12}$$

All probabilities and expectations in this proof refer to the joint prior–data law.

We next show that R is measurable with respect to the completed $\mathcal{F}_\infty$. A standard Borel space has a countable probability-determining algebra $\mathcal{A}$. Conditional on $\widetilde{\psi}$, the strong law gives, simultaneously for every $A \in \mathcal{A}$,

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{W_i \in A\} \longrightarrow P_{\mathrm{ctx}}^{\widetilde{\psi}}(A) = R(A) \quad \text{almost surely.} \tag{13}$$

Integrating the conditional probability-one statements over π preserves probability one. The evaluations $(R(A))_{A\in\mathcal{A}}$ determine R, and their inverse on the image is measurable because $\mathcal{P}(\mathcal{W})$ is standard Borel. Hence the infinite sequence determines R almost surely.

The conditional law of the entire sequence given $\widetilde{\psi}$ is $(P_{\mathrm{ctx}}^{\widetilde{\psi}})^{\otimes\mathbb{N}}$, which depends on $\widetilde{\psi}$ only through R. Consequently, $\widetilde{\psi}$ and the sequence are conditionally independent given R. Combining this fact with the measurability of R yields

$$\mathbb{E}[g(\widetilde{\psi}) \mid \mathcal{F}_\infty] = \mathbb{E}[g(\widetilde{\psi}) \mid R, \mathcal{F}_\infty] = \mathbb{E}[g(\widetilde{\psi}) \mid R]. \tag{14}$$

Equations 12 and 14 prove the limit. If $g(\widetilde{\psi}) = \Gamma(R)$ almost surely for a measurable Γ, its conditional expectation given R is itself. Conversely, equality $g(\widetilde{\psi}) = \mathbb{E}[g(\widetilde{\psi}) \mid R]$ almost surely makes $g(\widetilde{\psi})$ measurable with respect to the completed sigma-algebra generated by R, and therefore provides such a measurable version Γ. This proves the equivalence. □

Scope of the consistency statement. The guarantee concerns a mechanism drawn from the specified prior. It gives conditional almost-sure convergence for π-almost every mechanism, not for every point in the prior's support or for an arbitrary mechanism outside it. It provides neither a convergence rate nor uniform consistency. This scope is consistent with the Bayesian analyses of Balazadeh Meresht et al. (2026); Robertson et al. (2026). The argument holds for a fixed query; conditioning on a task-generated query replaces π by its conditional prior, with the same integrability requirement.

From the posterior to a learned PFN. The limit $\bar{g}_\pi(R) = \mathbb{E}_\pi[g(\widetilde{\psi}) \mid R]$ is the Bayesian prediction available once the context distribution is known. A learned PFN estimates it from finitely many rows using a trained network, which accounts for the first term in Eq. 5. The second term, $\left|\bar{g}_\pi(P_{\mathrm{ctx}}^{\psi}) - g(\psi)\right|$, records whether that ideal prediction equals the true effect. Changing the prior can move the average, but cannot make one common prediction equal two different targets: if they are separated by Δ, one incurs error at least $\Delta/2$.

B.2 Proof of the Irreducible Bias Bound

Either of two indistinguishable SCMs could have generated the context, so an estimator has the same expected prediction under both. If their effects differ, that common prediction cannot be unbiased for both possible truths. The proof below quantifies this limitation for any context-based estimator, independently of how it is trained.

Proof of Theorem 3.2(ii). The two context distributions induce the same product law for every sample size N. Applying the same estimator, including any independent randomization with the same law, therefore gives the same output distribution. Let $a_N = \mathbb{E}_{\psi_0}[\widehat{g}_N] = \mathbb{E}_{\psi_1}[\widehat{g}_N]$. The triangle inequality gives

$$\Delta \le |g(\psi_0) - a_N| + |a_N - g(\psi_1)| \le 2 \max_{j\in\{0,1\}} |a_N - g(\psi_j)|,$$

which proves the bound. Consistency under both mechanisms is impossible: identical output laws cannot converge in probability to different constants, because disjoint radius-$\Delta/3$ neighborhoods of the targets cannot both receive probability approaching one under the common law. □

This is a limit on guarantees across compatible mechanisms; an estimator may still be unbiased under a particular one. Source labels, experimental outcomes, or valid causal assumptions can change which mechanisms remain compatible with the estimator's information. Restricting a prior to an identifiable subclass gives a guarantee only within that subclass and does not resolve downstream violations of its assumptions.

B.3 An Observational Equivalence Example

The same treated–untreated association can arise from a treatment effect or from a hidden common cause. The following pair makes this ambiguity explicit and shows why both a restrictive prior and an average over a broader prior can miss the true effect. Let X be constant and consider two binary-outcome mechanisms:

$$\begin{aligned} M_{\mathrm{u}}: \quad & T \sim \mathrm{Bernoulli}(1/2), \quad V \sim \mathrm{Uniform}(0,1), \quad T \perp\!\!\!\perp V, \\ & Y(t) = \mathbb{I}\{V \le 1/4 + t/2\}; \\ M_{\mathrm{c}}: \quad & U \sim \mathrm{Bernoulli}(1/2), \quad T \mid U \sim \mathrm{Bernoulli}(1/4 + U/2), \\ & Y(t) = U. \end{aligned}$$

In both cases $Y = Y(T)$. The first mechanism is unconfounded and has ATE $1/2$. The second is confounded, has ATE 0, and satisfies treatment positivity even conditional on U. Nevertheless, both induce

$$\big(P(T=0,Y=0), P(T=0,Y=1), P(T=1,Y=0), P(T=1,Y=1)\big) = (3/8, 1/8, 1/8, 3/8).$$

An estimator consistent for the observational contrast under M_{u} has the same probability limit under M_{c}, where that contrast is not causal. Including both SCMs in the prior does not resolve the choice: with equal prior weights, their identical likelihoods keep those weights unchanged after every observational context. The exact posterior mean is therefore always $1/4$. This is the optimal average under prior-averaged squared error, but it equals neither true effect. The example illustrates the information limit; it does not assert that a particular checkpoint's finite training prior contains this exact pair.

B.4 An IV Construction with Unobserved Effect Modification

A valid instrument can reveal the effect for units whose treatment it changes while leaving other treatment responses unobserved. We exploit this distinction between local and population effects (Angrist & Imbens, 1995): the construction changes only the unobserved responses of always-takers and never-takers, preserving the IV data distribution while changing the population effect. It proves Proposition 3.3 and illustrates the boundary of the additive identification framework used by CausalFM-IV (Ma et al., 2026; Newey & Powell, 2003). Since $X = x_0$ is constant, $\tau(\psi_j) = \tau(x_0; \psi_j)$ also equals the ATE.

Proof of Proposition 3.3. Set $X = x_0$ almost surely. Let $U \in \{a, c, n\}$ have probabilities $(1/4, 1/2, 1/4)$, and draw $Z \sim \mathrm{Bernoulli}(1/2)$ independently of U. Define $T = T(Z)$ using the potential treatment assignments below. For the two mechanisms M_j indexed by ψ_j, set $Y = Y(T)$ with potential outcomes given in the table.

Type U	Probability	$(T(0), T(1))$	$(Y_0(0), Y_0(1))$	$(Y_1(0), Y_1(1))$
Always-taker a	$1/4$	$(1,1)$	$(2,3)$	$(0,3)$
Complier c	$1/2$	$(0,1)$	$(0,1)$	$(0,1)$
Never-taker n	$1/4$	$(0,0)$	$(1,2)$	$(1,4)$

Here the subscript on $Y_j(t)$ indexes the mechanism, not the treatment.

Instrument conditions. Instrument independence holds because Z is independent of U, which determines all potential treatments and outcomes. Exclusion holds because the outcome depends on Z only through T. All types satisfy $T(1) \ge T(0)$, and

$$P(T = 1 \mid Z = 0) = \frac{1}{4}, \qquad P(T = 1 \mid Z = 1) = \frac{3}{4}.$$

Thus the instrument is relevant and both treatments occur under either instrument value. To check completeness for functions of binary T, suppose $\mathbb{E}[b(T) \mid Z] = 0$. Then

$$\begin{pmatrix} 3/4 & 1/4 \\ 1/4 & 3/4 \end{pmatrix} \begin{pmatrix} b(0) \\ b(1) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The matrix has determinant $1/2$, so $b(0) = b(1) = 0$. Both mechanisms therefore satisfy this completeness condition. Observational ignorability fails: treatment depends on U, and the potential outcomes differ across its types.

Equal observable laws. For either instrument value, an always-taker is observed with $(T, Y) = (1, 3)$ and a never-taker with $(0, 1)$. A complier has $(T, Y) = (0, 0)$ when $Z = 0$ and $(1, 1)$ when $Z = 1$. These values are identical in both mechanisms. The nonzero joint probabilities are

(Z, T, Y)	$(0,0,0)$	$(0,0,1)$	$(0,1,3)$	$(1,0,1)$	$(1,1,1)$	$(1,1,3)$
P_{IV}	$1/4$	$1/8$	$1/8$	$1/8$	$1/4$	$1/8$

Hence the full observable distributions coincide, not merely selected moments.

Different causal targets. In M_0, $Y_0(t) = b_0(U) + t$ with $b_0(a) = 2$, $b_0(c) = 0$, and $b_0(n) = 1$, giving an additive mechanism with effect 1. In M_1, the effects for the three types are $(3, 1, 3)$, so

$$\tau(\psi_0) = 1, \qquad \tau(\psi_1) = \frac{1}{4}\,3 + \frac{1}{2}\,1 + \frac{1}{4}\,3 = 2.$$

Since X is constant, these are also the ATEs. In particular, M_1 cannot have the same additive-in-treatment outcome structure as M_0. The shared observed law gives $\mathbb{E}[Y \mid Z = 0] = 1$ and $\mathbb{E}[Y \mid Z = 1] = 3/2$. Thus the Wald ratio is $(1/2)/(1/2) = 1$, the complier effect in both mechanisms. This completes the construction. □

Implication for IV-specific PFNs. For any integrable estimator based only on (X, Z, T, Y), Theorem 3.2 guarantees worst-case absolute bias of at least $1/2$ on this pair. Consistency for M_0 yields the same limit, 1, under M_1 despite its target 2; with uniform integrability, the asymptotic bias is -1.

This is a boundary of the theoretical additive model class, not a failure within CausalFM-IV's assumptions or a claim about its checkpoint's predictions. Independent mean-zero continuous noise can be added to both potential outcomes, identically across mechanisms, without changing the equality of observable laws or the effect contrast.

What randomized evidence changes. If treatment itself is randomized in a sample from the same potential-outcome population, the two mechanisms no longer induce the same labeled context law. In particular, their randomized outcome means are

$$(\mu_0, \mu_1)_{M_0} = (3/4, 7/4), \qquad (\mu_0, \mu_1)_{M_1} = (1/4, 9/4).$$

Randomization supplies information about the counterfactual outcomes that the observational IV context leaves unresolved. This population distinction does not imply exact estimation from any fixed number of randomized rows.

Identification within the additive class. The additive restriction removes latent variation in the treatment effect, allowing the instrument-induced change in outcomes to determine that effect. When $Y = f(X, T) + h(X, U)$ and $Z \perp\!\!\!\perp U \mid X$, binary treatment gives $f(x, T) = f(x, 0) + T\{f(x, 1) - f(x, 0)\}$. The conditional mean of $h(x, U)$ does not depend on Z. Subtracting the outcome means across instrument values therefore yields the first-stage contrast multiplied by $f(x, 1) - f(x, 0) = \tau(x)$. If both instrument values occur and the contrast is nonzero,

$$\tau(x) = \frac{\mathbb{E}[Y \mid X = x, Z = 1] - \mathbb{E}[Y \mid X = x, Z = 0]}{\mathbb{E}[T \mid X = x, Z = 1] - \mathbb{E}[T \mid X = x, Z = 0]}. \tag{15}$$

A sufficiently expressive PFN trained on the corresponding causal prior may learn to approximate this mapping. The counterexample instead concerns a downstream violation of the additive factorization, while leaving instrument validity unchanged.

B.5 IDENTIFICATION WITH RANDOMIZED EVIDENCE

At a given covariate value, randomization makes the observed mean in each trial treatment group equal its mean potential outcome. The shared-CEPO assumption then transfers that mean to the target population; coverage and positivity specify where it is available.

Proof of Theorem 4.1. For each treatment t and almost every target covariate value x with RCT coverage and treatment positivity,

$$\begin{aligned}\mathbb{E}_\psi[Y \mid X = x, T = t, S = \mathrm{rct}] &= \mathbb{E}_\psi[Y(t) \mid X = x, T = t, S = \mathrm{rct}] \\ &= \mathbb{E}_\psi[Y(t) \mid X = x, S = \mathrm{rct}] \\ &= \mu_t(x; \psi).\end{aligned}$$

The three equalities follow from consistency, randomized assignment conditional on X, and the shared-CEPO assumption. The right-hand side of Eq. 7 is a functional of the source-aware law of (X, S, T, Y), proving identification. CATE follows by subtraction, and ATE additionally integrates over the target covariate law. □

Connection to ideal PFN inference. The identification result determines the causal target from the labeled data distribution. To connect it to PFN inference, consider a prior supported on mechanisms satisfying Theorem 4.1, with the regularity and integrability conditions of Proposition B.1. For an i.i.d. source mixture with a fixed positive RCT probability, that proposition gives convergence of the ideal posterior mean to the target under the joint prior–data law.

For independently sampled sources with fixed counts, the likelihood is the product of the two source-specific likelihoods. If both sample sizes grow without bound, the two infinite sequences identify $(P^\psi_{\mathrm{obs}}, P^\psi_{\mathrm{rct}})$, and the same conditional-expectation argument applies. These results connect identifying information to ideal posterior recovery; they do not provide finite-sample error bounds for the trained RA-PFN.

Scope of identification. RCT coverage means absolute continuity of the target covariate law with respect to the RCT covariate law; topological support alone would not suffice for arbitrary measurable conditional means. Identification does not require treatment effects to be independent of unobserved variables.

A finite-sample statistical benchmark. A finite randomized sample still introduces sampling error even when the target is identifiable. Under stronger assumptions, the RCT-only estimator below has no bias from observational confounding, and its error bound decreases with the number of randomized observations. It provides a statistical benchmark for what randomized evidence can support, separate from the learned RA-PFN estimator.

Proposition B.2 (RCT-Only ATE Estimation)**.** *Assume the conditions of Theorem 4.1 with binary treatment. Suppose additionally that the randomized source has the target covariate law, assigns* $T \sim \mathrm{Bernoulli}(1/2)$ *independently of* $(X, Y(0), Y(1))$*, and satisfies* $|Y| \le B$ *for* $B > 0$*. For* $n_{\mathrm{rct}} \ge 1$ *i.i.d. randomized observations, the Horvitz–Thompson estimator* $\widehat{\lambda}_{n_{\mathrm{rct}}} = n_{\mathrm{rct}}^{-1} \sum_{i:S_i=\mathrm{rct}} 2(2T_i - 1)Y_i$ *is unbiased for the target-population ATE* λ *and satisfies, for every* $\epsilon > 0$*,*

$$\mathbb{E}_\psi\big|\widehat{\lambda}_{n_{\mathrm{rct}}} - \lambda\big| \le \frac{2B}{\sqrt{n_{\mathrm{rct}}}}, \qquad P_\psi\big(|\widehat{\lambda}_{n_{\mathrm{rct}}} - \lambda| > \epsilon\big) \le 2\exp\left(-\frac{n_{\mathrm{rct}}\epsilon^2}{8B^2}\right). \tag{16}$$

Proof of Proposition B.2. Write $W_i = 2(2T_i - 1)Y_i$ for a randomized row. By consistency $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$, so $W_i = 2Y_i(1)$ when $T_i = 1$ and $W_i = -2Y_i(0)$ when $T_i = 0$. Since $T_i \sim \mathrm{Bernoulli}(1/2)$ independently of $(X_i, Y_i(0), Y_i(1))$,

$$\mathbb{E}_\psi[W_i \mid Y_i(0), Y_i(1)] = \tfrac{1}{2}\,2Y_i(1) + \tfrac{1}{2}(-2Y_i(0)) = Y_i(1) - Y_i(0),$$

hence $\mathbb{E}_\psi[W_i \mid S_i = \mathrm{rct}] = \mathbb{E}_\psi[Y_i(1) - Y_i(0) \mid S_i = \mathrm{rct}] = \lambda$, where the last equality uses the shared CEPO and target covariate law. Thus $\widehat{\lambda}_{n_{\mathrm{rct}}}$ is unbiased.

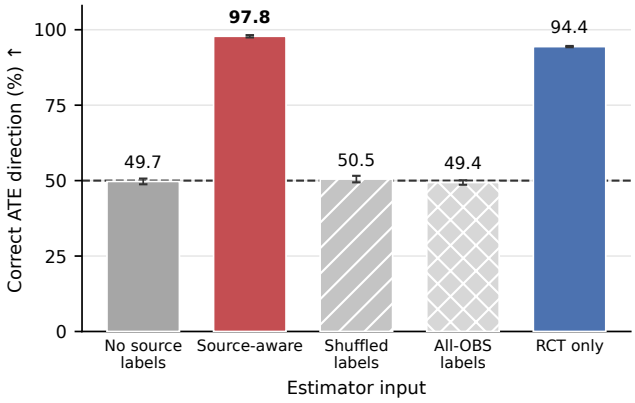

Figure 4: ATE-direction accuracy in the controlled source-label diagnostic. Correct labels yield 97.8% accuracy; removing or corrupting them gives approximately chance performance. Error bars denote standard deviations across runs.

The randomized rows are i.i.d. under ψ, and $|W_i| \le 2B$ because $|2T_i - 1| = 1$ and $|Y_i| \le B$. Jensen and the i.i.d. variance give $\mathbb{E}_\psi|\widehat{\lambda}_{n_{\mathrm{rct}}} - \lambda| \le \sqrt{\mathrm{Var}_\psi(W_1)/n_{\mathrm{rct}}} \le 2B/\sqrt{n_{\mathrm{rct}}}$, using $\mathrm{Var}_\psi(W_1) \le \mathbb{E}_\psi[W_1^2] \le 4B^2$.

As $W_i \in [-2B, 2B]$ has range $4B$, Hoeffding's inequality yields $P_\psi(|\widehat{\lambda}_{n_{\mathrm{rct}}} - \lambda| > \epsilon) \le 2\exp(-2n_{\mathrm{rct}}\epsilon^2/(4B)^2) = 2\exp(-n_{\mathrm{rct}}\epsilon^2/(8B^2))$. Both bounds involve only n_{rct}, so neither the observational rows nor the confounding strength enters. □

Relation to RA-PFN. The benchmark separates two issues: randomization can remove bias from observational confounding, while finite samples still produce estimation error. RA-PFN learns a different prediction rule from mixed-source tasks, so its behavior with limited randomized evidence is evaluated empirically in Section 5. The benchmark alone gives no CATE error rate, no precision gain from observational rows, and no guarantee that adding a fixed number of RCT rows reduces RA-PFN's bias.

C Additional Experimental Details

C.1 Controlled Source-Identifiability Diagnostic

This diagnostic isolates the information carried by source labels. We construct two tasks with opposite treatment effects and identical pooled (X, T, Y) distributions, but different distributions when S is retained. A lightweight estimator tests whether the labels allow it to predict the effect's sign (Figure 4). It is trained from scratch and is separate from RA-PFN.

Each context is generated from an equiprobable task $K \in \{+, -\}$, which is fixed for all observations within the context. Conditional on $K = k$, each observation contains a binary covariate $X_i \in \{0, 1\}$, treatment $T_i \in \{0, 1\}$, and outcome $Y_i \in \{0, 1\}$, with

$$X_i \mid K = k \sim \mathrm{Bernoulli}(0.5).$$

The latent response type is

$$U_i = \big(Y_i(0), Y_i(1)\big) \in \mathcal{U}, \qquad \mathcal{U} = \{(0,0), (0,1), (1,0), (1,1)\},$$

and jointly determines the two potential outcomes. For $k \in \{+, -\}$ and $x \in \{0, 1\}$, define the response-type probability vector

$$\boldsymbol{\pi}_{k,x} = \big(P(U = u \mid X = x, K = k)\big)_{u \in \mathcal{U}},$$

where the entries follow the ordering of $\mathcal{U}$ above. We use the following fixed probabilities as one concrete instantiation of the paired-task construction:

$$\boldsymbol{\pi}_{+,0} = (0.44,\, 0.31,\, 0.11,\, 0.14), \qquad \boldsymbol{\pi}_{+,1} = (0.14,\, 0.31,\, 0.11,\, 0.44),$$
$$\boldsymbol{\pi}_{-,0} = (0.46,\, 0.09,\, 0.29,\, 0.16), \qquad \boldsymbol{\pi}_{-,1} = (0.16,\, 0.09,\, 0.29,\, 0.46).$$

These values are fixed parameters of the controlled construction rather than quantities estimated from the evaluation data.

The source indicator $S_i \in \{\mathrm{obs}, \mathrm{rct}\}$ is sampled independently with equal probability. In the randomized source,

$$T_i \mid S_i = \mathrm{rct} \sim \mathrm{Bernoulli}(0.5).$$

In the observational source, treatment depends on both the observed covariate and latent response type:

$$e_k(x,u) = P(T = 1 \mid X = x, U = u, S = \mathrm{obs}, K = k).$$

Let $\mathbf{e}_{k,x} = (e_k(x,u))_{u\in\mathcal{U}}$. Using the same response-type ordering, the fixed observational propensities are

$$\mathbf{e}_{+,0} = \left(\frac{137}{176}, \frac{10}{31}, 0, \frac{3}{56}\right), \qquad \mathbf{e}_{+,1} = \left(\frac{53}{56}, \frac{21}{31}, 1, \frac{39}{176}\right),$$
$$\mathbf{e}_{-,0} = \left(\frac{19}{92}, \frac{19}{36}, \frac{59}{116}, 1\right), \qquad \mathbf{e}_{-,1} = \left(\frac{17}{64}, 0, \frac{10}{29}, \frac{163}{184}\right).$$

Together with the response-type probabilities above, these propensities constitute a particular construction satisfying the desired causal ambiguity. The observed outcome follows consistency, $Y_i = Y_i(T_i)$. Because U_i affects both observational treatment assignment and the potential outcomes but is withheld from the estimator, the observational source contains unobserved confounding.

By construction, the two tasks have opposite treatment effects,

$$\lambda_+ = \mathbb{E}[Y(1) - Y(0) \mid K = +] = 0.2, \qquad \lambda_- = \mathbb{E}[Y(1) - Y(0) \mid K = -] = -0.2,$$

while inducing the same source-agnostic observable distribution after marginalizing the source indicator:

$$P_+(X,T,Y) = P_-(X,T,Y).$$

For completeness, substituting the specified response types and propensities gives the same joint cell probabilities under either task. In the order $(t,y) = (0,0),(0,1),(1,0),(1,1)$, these are

$$\big(P_k(0,t,y)\big)_{t,y} = \frac{1}{1600}(273, 147, 247, 133),$$
$$\big(P_k(1,t,y)\big)_{t,y} = \frac{1}{1600}(133, 247, 147, 273).$$

The labeled laws differ: at $X = 0$, the randomized treated-outcome means are 0.45 for $K = +$ and 0.25 for $K = -$. Thus marginalization merges mechanisms with different targets, whereas source-aware context distinguishes them. This statement assumes i.i.d. mixture sampling; equal one-row marginals need not imply equal laws under fixed source counts.

The single-precision implementation satisfies the marginal equality up to numerical precision, with

$$\max_{x,t,y} |P_+(x,t,y) - P_-(x,t,y)| = 2.89 \times 10^{-9}.$$

The source-agnostic estimator observes only (X_i, T_i, Y_i), whereas the source-aware variant additionally receives S_i. Both use the same lightweight architecture with a 64-dimensional row representation, four-head learned-query attention, and a scalar prediction head. The model predicts the ATE direction,

$$Z_K = \mathbb{I}\{\lambda_K > 0\} = \mathbb{I}\{K = +\},$$

using binary cross-entropy. We evaluate held-out ATE-direction accuracy across independent training runs. The source-agnostic, shuffled-label, and all-OBS-label controls remove or corrupt S while leaving (X,T,Y) unchanged, whereas the RCT-only condition retains only randomized observations. Error bars in Figure 4 show the across-run standard deviation.

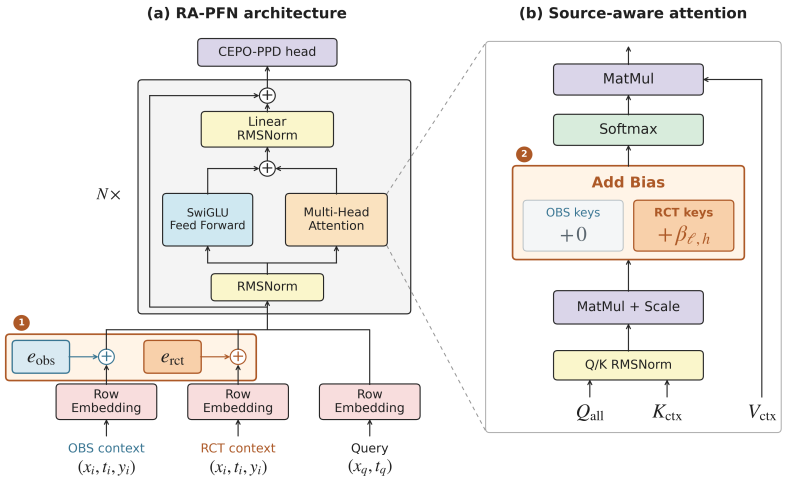

Figure 5: RA-PFN architecture. (1) Source embeddings distinguish observational and randomized context rows. (2) Each attention head adds a learned bias to randomized context keys.

Scope of the diagnostic. Equal pooled context laws make the two tasks indistinguishable to any source-agnostic estimator, regardless of training. The diagnostic tests this information limit with a separate architecture. Attributing RA-PFN's benchmark gains specifically to source labels, beyond continued prior fitting, would require a training-matched variant with both source-aware modules removed. The present diagnostic addresses the value of the information itself, while Table 2 compares ways to incorporate it under matched training.

C.2 RA-PFN Prior-Fitting and Implementation Details

Figure 5 summarizes the two source-aware modifications described in Section 4.2.

Synthetic causal prior. The training tasks vary both unobserved-confounding strength and the amount of randomized evidence. Within a task, the sources share the potential-outcome mechanism but use different treatment assignments, so the model learns from both regimes while predicting the same causal target. RA-PFN is initialized from the official CausalPFN checkpoint and then undergoes continued prior fitting on online-generated causal tasks with observed and unobserved confounding. Each task has $d = 5$ covariates sampled as $X_i \sim \mathcal{N}(0, I_d)$. We generate a latent variable

$$U_i = cX_{i1} + \sqrt{1 - c^2}\,\epsilon_{U,i}, \qquad c \sim \text{Uniform}(-0.65, 0.65), \qquad \epsilon_{U,i} \sim \mathcal{N}(0, 1),$$

and define nonlinear baseline and treatment-effect functions

$$m_0(X_i) = b_0 + 0.75 X_i^\top w_0 + 0.35 \sin(X_{i1}) + 0.15 X_{i2}^2,$$
$$\tau(X_i) = b_\tau + 0.65 \tanh(X_i^\top w_\tau) + 0.20 \sin(X_{i1} X_{i2}),$$

where $w_0, w_\tau, w_A \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d/d)$, $b_0 \sim \text{Uniform}(-0.6, 0.6)$, and $b_\tau \sim \text{Uniform}(-0.7, 0.7)$. Potential outcomes are

$$Y_i(0) = m_0(X_i) + a_U U_i + \epsilon_{Y,i}, \qquad Y_i(1) = Y_i(0) + \tau(X_i),$$

with $a_U \sim \text{Uniform}(0.7, 1.5)$ and $\epsilon_{Y,i} \sim \mathcal{N}(0, 0.2^2)$. The CEPO targets integrate out the latent residual $\epsilon_{U,i}$ and outcome noise $\epsilon_{Y,i}$, giving $\mu_0(X_i) = m_0(X_i) + a_U c X_{i1}$ and $\mu_1(X_i) = \mu_0(X_i) + \tau(X_i)$.

In this generator, the latent variable shifts potential outcomes but does not change the treatment effect $\tau(X)$. Theorem 4.1 also permits latent treatment-effect modification. Its broader identification scope does not by itself establish that this checkpoint generalizes to those additional mechanisms.

For observational rows, treatment is sampled according to

$$T_i^{\text{obs}} \sim \text{Bernoulli}\Big(\sigma\Big[0.45 X_i^\top w_A + \gamma\sqrt{1-c^2}\,\epsilon_{U,i}\Big]\Big),$$

where the pre-logit is clipped to $[-4, 4]$ and $\gamma \sim \text{Uniform}(0, 3)$ controls unobserved-confounding strength. Randomized rows instead use $T_i^{\text{rct}} \sim \text{Bernoulli}(0.5)$ while sharing the same covariate and potential-outcome mechanisms. The task-level randomized proportion is sampled uniformly from $[0.01, 0.30]$, with probability 0.10 of generating a context without randomized rows. Context rows are randomly permuted so that their positions do not reveal the data source.

Optimization and checkpoint protocol. Each optimization step contains eight independently generated tasks, each with 128 context rows and 32 query rows. We run 400 AdamW steps, corresponding to 3,200 online-generated training tasks. Starting from the 20-layer CausalPFN backbone, we update its prediction head, input and outcome normalization layers, top four Transformer layers, source embeddings, and randomization-aware attention parameters. The restriction to the top four layers applies only to the pretrained Transformer weights; the attention-bias scalars are trained in every layer, as specified in Section 4.2.

The learning rates are 3×10^{-5} for the backbone, 10^{-3} for the source embeddings, and 10^{-2} for the attention biases; weight decay is 10^{-5} and the gradient norm is clipped at 1.0. Both new modules are zero-initialized, and training minimizes the original CausalPFN soft-histogram negative log likelihood with smoothing parameter $\sigma = 0.5$.

The source-embedding-only and randomization-aware-attention-only variants in Table 2 use the same initialization, synthetic tasks, seed, optimization steps, and backbone update rule, differing only in which source-aware module is enabled.

C.3 Public Semi-Synthetic Data Generation

The evaluation changes treatment assignment while preserving the benchmark potential outcomes. This lets us vary unobserved confounding and randomized-data availability without changing the causal targets for the fixed query units.

We use ACIC 2016 and the Lalonde-CPS and Lalonde-PSID RealCause benchmarks, which provide covariates and both potential outcomes for each unit. Only the simulator accesses both potential outcomes; an estimator receives factual context tuples (X_i, T_i, Y_i) and, when permitted by the method, the row-level source indicator S_i. The observational and randomized rows share the benchmark covariate and potential-outcome mechanisms and differ only in treatment assignment. For each realization, define the baseline potential-outcome risk

$$B_i = \frac{Y_i(0) + Y_i(1)}{2}.$$

We fit a ridge regression of B_i on the observed covariates using the training units and decompose this risk into a standardized observed component and a standardized residual:

$$G_i = \text{std}\Big\{\widehat{\mathbb{E}}(B_i \mid X_i)\Big\}, \qquad U_i = \text{std}\Big\{B_i - \widehat{\mathbb{E}}(B_i \mid X_i)\Big\}.$$

Here G_i summarizes outcome risk explained by the observed covariates, whereas U_i captures the remaining outcome-related variation used only by the simulator. Because U_i is withheld from the context but also affects observational treatment assignment, it creates the unobserved-confounding channel in our evaluation. We generate the OBS propensity as

$$e_i^{\text{obs}} = \varepsilon + (1 - 2\varepsilon)\,\sigma\Big[b + \kappa\left(\sqrt{1-\rho_U}\,G_i + \sqrt{\rho_U}\,U_i\right)\Big],$$

where b is calibrated so that the mean treatment probability is 0.5 before applying the symmetric overlap floor. Our primary setting uses $\rho_U = 0.75$, $\kappa = 1.5$, and $\varepsilon = 0.05$. Because G_i and U_i are standardized, ρ_U controls the target latent contribution to the pre-logit variance; the realized contribution can differ slightly if the two components are correlated. The overlap floor bounds the propensity in $[0.05, 0.95]$. We then sample

$$T_i^{\text{obs}} \sim \text{Bernoulli}(e_i^{\text{obs}}), \qquad T_i^{\text{rct}} \sim \text{Bernoulli}(0.5),$$

and reveal $Y_i = Y_i(T_i)$ in either data source.

Each realization uses a fixed 128-row context and a fixed source permutation. The main comparison and ablation use $\alpha = 0.10$, corresponding to 13 randomized rows after rounding. The availability study takes nested prefixes of the same permutation to obtain $\alpha \in \{0.01, 0.02, 0.05, 0.10, 0.20, 0.30\}$, corresponding to $\{1, 3, 6, 13, 26, 38\}$ randomized rows; all remaining context rows are observational. The robustness study fixes $\alpha = 0.10$ and varies $\rho_U \in \{0, 0.25, 0.50, 0.75, 1\}$. The endpoints correspond respectively to observational assignment driven entirely by $G(X)$ and entirely by the latent residual U. No covariate columns are omitted in the primary protocol.

Every method receives the same train–test split, context units, source allocation, realized treatments and outcomes, and held-out queries. Source-agnostic methods receive the same (X, T, Y) context with S removed. We repeat every public realization with three fixed source-allocation and treatment seeds. Unlike the Gaussian tasks used for continued prior fitting, this evaluation retains the benchmark covariates and potential outcomes and modifies only the observational and randomized source construction. It therefore tests transfer beyond the training-task generator.

C.4 Statistical Aggregation and Uncertainty

The realization is the statistical unit. We first average each method's PEHE and ATE error over the three fixed source-allocation and treatment repeats within a realization. Tables 1 and 2 report means and standard errors across 10 ACIC and 100 CPS/PSID realization-level observations.

For the availability study, shaded bands are pointwise Student-t intervals of the realization-level mean, computed as $\bar{z} \pm t_{(1+c)/2,n-1} s/\sqrt{n}$. We use $c = 0.70$ for ACIC and $c = 0.90$ for CPS and PSID, as stated in Figure 1.

For the robustness study, we use 20,000 Gaussian parametric-bootstrap draws of the mean based on the realization-level mean, standard deviation, and sample size. Boxes span the 25th–75th percentiles of the bootstrap means, center lines show the realization-level sample mean, and whiskers give the 95% interval.

For the accuracy–efficiency comparison, the three repeats are first averaged within each realization, the seven methods are ranked by PEHE within each of the 210 realizations, and Figure 3 reports the median rank and its 25th–75th percentile range. Runtime is likewise averaged over valid repeats within a realization and then summarized by the median of all 210 realization-level times.

C.5 Additional Baseline Implementation Details

The baselines differ in whether they receive source labels and whether they fit a model to each new context. The details below make these two comparisons explicit.

S-, T-, X-, and DA-Learners use random-forest outcome models, and DR-Learner uses cross-fitted random-forest nuisance models with small-sample stabilization. DML is implemented as Forest DML with random-forest nuisance models, three-fold cross-fitting, and 400 causal trees. TARNet and DragonNet use two-layer, 128-dimensional representations and early stopping on a validation split drawn only from the 128 context observations. These eight estimators are fitted separately to each context.

Do-PFN and CausalPFN use their frozen pretrained checkpoints. Because our public contexts contain no instrument or front-door mediator, CausalFM uses its frozen standard CATE checkpoint rather than its IV or front-door variants. All eleven source-agnostic baselines receive (X, T, Y) without S. The three PFN baselines use frozen weights and perform no downstream optimization; the eight task-specific estimators are fitted to each context as described above.

Two-Stage, CorNet, and the integrative R-Learner additionally use S to distinguish observational from randomized treatment assignment. For each context, we construct a source–treatment-stratified validation split, select among ten configurations using a source-balanced held-out R-loss, and refit the selected configuration on the complete 128-row context. The selection procedure uses factual context observations only and never accesses query outcomes or counterfactuals. All other validation and early-stopping decisions likewise use context data only.

C.6 TIMING PROTOCOL

We measure the computation needed to make predictions for a new dataset once the pretrained models are available. This includes the fitting and selection performed by task-specific methods and the forward prediction performed by PFNs.

We benchmark RA-PFN, CausalPFN, Do-PFN, CausalFM, Two-Stage, CorNet, and R-Learner on an Intel Xeon Gold 6330 CPU. We use all 210 realizations with the same three fixed source-allocation and treatment repeats as in the main comparison. After one untimed warm-up, each method predicts CATE for 1,000 fixed queries from a 128-row context.

To measure per-dataset downstream computation, dataset loading, checkpoint loading, and offline PFN prior fitting are excluded from the timing. For PFN-based methods, the measured time includes context processing and forward prediction. For task-specific estimators, it includes ten sequential validation fits, one full-context refit using the selected configuration, and held-out prediction. Thus, the reported runtime reflects the computation required after receiving a new target dataset under the respective estimation paradigm.

We first average valid repeat timings within each realization and then take the median over all 210 realization-level times. The accuracy coordinate is computed from the same 210 realizations by ranking methods according to realization-level PEHE. Hence, both coordinates in Figure 3 summarize the same set of public causal inference tasks.