SC-PFN: A CAUSAL FOUNDATION MODEL WITH AMORTIZED STATISTICAL CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior-data fitted networks (PFNs) amortize causal effect estimation, but prediction loss averaged across tasks does not ensure unbiased estimation within each task. Neyman orthogonality motivates residual corrections that cancel first-order nuisance sensitivity in population estimating moments. We introduce SC-PFN (Statistical Correction PFN), a causal foundation model that jointly learns effect prediction and context-conditioned residual weighting. The model predicts observed variables to form residuals and assigns weights to context rows. Weighted residual averages give regularized coefficients, which query weights combine into an update of the base effect estimate. A single model is pretrained across back-door, front-door, and instrumental-variable settings under stated identifying assumptions and applied without dataset-specific parameter updates. We characterize the coefficients as constrained ridge solutions and bound the correction's response to residual perturbations with row and query weights held fixed. Experiments on synthetic and semi-synthetic causal benchmarks show lower mean effect-estimation error than evaluated baselines. Component comparisons and score-shuffling controls support the contribution of the correction and its observation-specific weights.

1 INTRODUCTION

Estimating causal effects from observational data requires identifying assumptions and a statistical estimator (Pearl, 2009; Imbens & Rubin, 2015). Prior-data fitted networks (PFNs) learn inference procedures by pretraining on simulated datasets (Müller et al., 2022). CausalPFN, Do-PFN, and CausalFM extend this approach to causal quantities, using an observational dataset as context without updating network parameters (Balazadeh et al., 2025; Robertson et al., 2025; Ma et al., 2026). We study how statistical correction can be learned as part of this inference procedure.

Minimizing prediction loss across tasks does not require an estimator to be unbiased on each task. For PFN-based Bayesian estimation of average treatment effects, Melnychuk et al. (2026) identify prior-induced confounding bias and apply one-step posterior correction. The effect of an estimation error also depends on how it enters the causal target. For a contrast of two outcome means, equal errors in the same direction cancel, while equal errors in opposite directions add. Direct causal supervision constrains the final prediction but does not prescribe how observational residuals should contribute to its correction. We make this computation explicit and learn it alongside effect prediction.

Orthogonal estimation provides a statistical basis for residual correction (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023; Van der Laan et al., 2011). Neyman orthogonality requires a population estimating moment to be insensitive, to first order, to nuisance perturbations at the true target and nuisance functions. These functions include outcome regressions and treatment probabilities. For suitable target functionals, Riesz representations provide residual weights used to construct such corrections (Chernozhukov et al., 2021; 2022). The required functions and weights differ across back-door (BD), front-door (FD), and instrumental-variable (IV) settings (Ma et al., 2026). This motivates learning a shared correction procedure whose weights depend on the observational context and query.

We introduce **SC-PFN** (Statistical Correction PFN), a causal foundation model that jointly learns a base effect predictor and a residual correction. The model predicts observed variables to form residuals and assigns weights to context rows. Averaged residual–weight products give correction

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.

coefficients, which query weights combine into an effect update. One model is pretrained across BD, FD, and IV tasks under their stated identifying assumptions. Our contributions are as follows.

- We introduce a causal PFN that jointly learns effect prediction, residual weighting, and query-level correction across BD, FD, and IV tasks.
- We characterize the correction coefficients as constrained ridge solutions and bound the update's response to residual perturbations with row and query weights held fixed.
- Experiments on synthetic and semi-synthetic benchmarks show lower mean effect-estimation error than the evaluated PFN baselines. Component comparisons and score-shuffling controls examine the contribution of the correction.

2 Related Work

Causal foundation models. CausalPFN learns posterior distributions over conditional expected potential outcomes under ignorability (Balazadeh et al., 2025), while Do-PFN predicts conditional interventional distributions (Robertson et al., 2025). CausalFM constructs priors compatible with BD, FD, and IV identification settings (Ma et al., 2026). Extensions address continuous treatments (**?**), interventional time series (**?**), and joint structure and outcome prediction (**?**). For the predictor classes studied, Nagler (2023) relates PFN bias and variance to query localization and sensitivity to individual observations.

Statistical correction. Double/debiased machine learning combines orthogonal scores with cross-fitting (Chernozhukov et al., 2018). Orthogonal statistical learning gives excess-risk bounds with estimated nuisance functions (Foster & Syrgkanis, 2023). Riesz regression learns debiasing weights, and RieszNet jointly fits regression functions and Riesz representers (Chernozhukov et al., 2021; 2022). For Bayesian ATE estimation from PFNs, Melnychuk et al. (2026) apply one-step posterior correction using influence functions and obtain nuisance-function posteriors through martingale posteriors. SC-PFN instead learns residual weights and query-level updates jointly with the effect predictor during pretraining. The correction is part of the learned effect estimator, rather than a separate procedure applied to a derived ATE posterior.

3 Background

3.1 Causal Effect Estimation

Let $X \in \mathbb{R}^{d_X}$ denote pretreatment covariates, $A \in \{0, 1\}$ treatment, and $Y \in \mathbb{R}$ the outcome. The potential outcome $Y(a)$ is the outcome under an intervention assigning treatment a. Consistency gives $Y = Y(A)$, so only one potential outcome is observed for each unit (Pearl, 2009; Imbens & Rubin, 2015). A structural causal model (SCM) S specifies how variables are generated from their causes and noise. An intervention $\mathrm{do}(A = a)$ fixes treatment at a by replacing its assignment mechanism, whereas conditioning on $A = a$ selects units who received that treatment.

The conditional average treatment effect (CATE) is

$$\tau_S(x) = \mathbb{E}_S[Y(1) - Y(0) \mid X = x]. \tag{1}$$

The average treatment effect (ATE), $\psi_S = \mathbb{E}_S[\tau_S(X)]$, averages this contrast over the population. Observed treatment-group differences can also reflect confounding. Identification asks whether the observational law determines the effect under stated assumptions; estimation uses a finite sample to approximate it.

Back-door (BD) adjustment accounts for confounding through X. Under consistency, conditional unconfoundedness, and positivity, $\tau_S(x) = \mu_1(x) - \mu_0(x)$, where $\mu_a(x) = \mathbb{E}_S[Y \mid A = a, X = x]$. Unconfoundedness means that treatment is independent of potential outcomes given X; positivity requires both treatment values to have positive probability given X (Imbens & Rubin, 2015; Ma et al., 2026). Front-door (FD) adjustment instead uses an observed mediator on all directed treatment–outcome paths, under the front-door conditions. An instrumental variable (IV) affects treatment, has no direct effect on the outcome, and is independent of unobserved confounders given X. Additional assumptions determine which effect it identifies (Ma et al., 2026).

3.2 PRIOR-DATA FITTED NETWORKS

A prior Π over SCMs describes possible data-generating processes. Let P_S be the observational law of S and $D_{\text{obs}} = \{O_i\}_{i=1}^n \sim P_S^{\otimes n}$ an observational context, where O_i is an observed row. Given $C = (D_{\text{obs}}, x)$, Bayesian inference yields $\Pi(dS \mid C)$, which induces a distribution over $\tau_S(x)$ with mean $\overline{\tau}_\Pi(C) = \int \tau_S(x)\, \Pi(dS \mid C)$. This distribution describes uncertainty about the effect, not variation in individual outcomes (Balazadeh et al., 2025).

A prior-data fitted network (PFN) learns to approximate inference from a dataset and query using simulated tasks (Müller et al., 2022). For direct effect prediction with simulator-provided effect labels, it represents a density $q_\theta(t \mid C)$ over effect values t and trains with

$$\mathcal{L}_{\text{pred}}(\theta) = \mathbb{E}_{S, D_{\text{obs}}, x}\left[-\log q_\theta\big(\tau_S(x) \mid C\big)\right]. \tag{2}$$

The expectation follows the prior and its context and query sampling process. Training uses prior simulations rather than posterior samples; at deployment, predictions depend on the new context while network parameters remain fixed (Müller et al., 2022; Balazadeh et al., 2025).

3.3 STATISTICAL CORRECTION

Weighted residuals. Consider ATE estimation under BD adjustment (Chernozhukov et al., 2018; Kennedy, 2024). Fix the observational law P_0 and write $P_0 f = \mathbb{E}_{P_0}[f(O)]$ and $P_n f = n^{-1}\sum_{i=1}^n f(O_i)$. The outcome regressions $\mu_a(x)$ and propensity $e(x) = P_0(A = 1 \mid X = x)$ are nuisance functions: quantities used for estimation but not themselves the target. The plug-in estimator $\widehat{\psi}_{\text{PI}} = P_n[\widehat{\mu}_1(X) - \widehat{\mu}_0(X)]$ can inherit errors from the fitted regressions. The augmented inverse probability of treatment weighted (A-IPTW) estimator adds a residual correction:

$$\begin{aligned} \widehat{\psi}_{\text{A-IPTW}} &= \widehat{\psi}_{\text{PI}} + P_n\Big[\widehat{H}(A, X)\{Y - \widehat{\mu}_A(X)\}\Big], \\ \widehat{H}(A, X) &= \frac{A}{\widehat{e}(X)} - \frac{1 - A}{1 - \widehat{e}(X)}. \end{aligned} \tag{3}$$

The weights give treated and control residuals opposite signs and account for assignment probabilities. Here, true and fitted propensities are bounded away from zero and one, and nuisance estimates use data separate from the sample averaged by P_n. Cross-fitting exchanges fitting and evaluation roles across folds (Chernozhukov et al., 2018). Observed residuals contain outcome noise as well as regression error; they are not nuisance-function errors.

Neyman orthogonality. For a target ϑ_0 and true nuisance functions η_0, a valid estimating score satisfies $P_0\varphi(O; \vartheta_0, \eta_0) = 0$. It is Neyman-orthogonal if

$$\frac{d}{dt} P_0\varphi(O; \vartheta_0, \eta_0 + tv)\Big|_{t=0} = 0 \quad \text{for every admissible } v. \tag{4}$$

The derivative varies the nuisance functions while holding the target and observational law fixed. The condition removes their first-order influence on the population moment (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023). For A-IPTW, the score is the outcome contrast plus the weighted residual in Eq. (3), minus the ATE. It satisfies Eq. (4) at the true outcome regressions and propensity.

Riesz representations. Let $\mathcal{A}(v)$ be the first-order change of a target functional along a nuisance perturbation v. If $\mathcal{A}$ is continuous and linear on a specified closed subspace of $L^2(P_0)^R$, it has a Riesz representative $H^\star$ satisfying

$$\mathcal{A}(v) = P_0\big[H^\star(O)^\top v(O)\big], \tag{5}$$

where R counts nuisance components and $L^2(P_0)$ denotes square-integrable functions. For the ATE functional of the outcome regression, the weight in Eq. (3) evaluated at the true propensity gives this representation. Riesz-based estimators use such weights to construct residual corrections that cancel the leading nuisance-error term (Chernozhukov et al., 2021; 2022).

Table 1: Identification settings. Treatment is binary; instruments may be binary or continuous.

Setting	Observed row O	Identification conditions
BD	(X, A, Y)	Consistency, treatment positivity, and conditional unconfoundedness.
FD	(X, A, M, Y)	Consistency, treatment and mediator positivity, and conditional front-door criterion.
IV	(X, Z, A, Y)	Additive structural model, exclusion, $U \perp Z \mid X$, relevance, and completeness.

4 Problem Formulation

Task specification. Let $G \in \mathcal{G} = \{\mathrm{BD}, \mathrm{FD}, \mathrm{IV}\}$ index the settings in Table 1 (Ma et al., 2026). Each prior Π_G is supported on SCMs satisfying the corresponding conditions. Pretraining samples G from a fixed mixture, then $S \sim \Pi_G$, an observational context $D_{\mathrm{obs}} \sim P_S^{\otimes n}$, and a query x. The same SCM supplies the label $\tau_S(x) = \Psi_{G,x}(P_S)$ for the effect in Eq. (1). For IV, $Y = f_S(X, A) + g_S(X, U)$ with latent confounder U, so the target is $f_S(x, 1) - f_S(x, 0)$. Full assumptions and identifying formulas are given in Appendix A. The input $C = (D_{\mathrm{obs}}, x)$ includes variable roles and the observed modality pattern, with mediator M in FD and instrument Z in IV. Unavailable modalities are marked as missing, and query rows contain covariates only. Neither G nor S is an explicit model input; effect labels are used only for supervision.

Estimation objective. We seek a PFN that learns statistical correction across settings and estimates $\tau_S(x)$ from C. Its point estimate $\widehat{\tau}_\theta(C)$ is the mean of its predicted effect density. One parameter vector θ is shared across settings and remains fixed on new datasets. For fixed θ, S, x, and n, define the repeated-context bias as

$$b_{\theta,n}(S, x) = \mathbb{E}_{D \sim P_S^{\otimes n}} \left[\widehat{\tau}_\theta(D, x)\right] - \tau_S(x). \tag{6}$$

The expectation varies only the observational sample, not the SCM, query, or pretrained parameters.

Research question. How can a single PFN learn statistical corrections for query-level causal effects across identification settings without dataset-specific parameter updates?

5 Amortized Statistical Correction

SC-PFN jointly learns a base effect predictor and a correction from observational residuals. For each context and query, the network predicts component values, row weights, and query weights. The correction coefficients are computed from these outputs and the observed context rather than predicted by another head. The residual weighting in § 3.3 provides the statistical motivation; Fig. 1 summarizes the effect-estimation procedure.

5.1 Effect Prediction

A transformer encodes $C = (D_{\mathrm{obs}}, x)$ using the observed values, variable roles, and modality information in § 4. Query rows contain covariates only. The representation is shared by the effect and correction outputs. Following the Gaussian-mixture parameterization of Ma et al. (2026), the effect head predicts

$$q_\theta^{\mathrm{base}}(t \mid C) = \sum_{k=1}^{K} \pi_{\theta,k}(C) \mathcal{N}\big(t; \mu_{\theta,k}(C), \sigma^2_{\theta,k}(C)\big)\,, \tag{7}$$

where $\pi_{\theta,k} \geq 0$, $\sum_k \pi_{\theta,k} = 1$, and $\sigma^2_{\theta,k} > 0$. Its mean is $\widehat{\tau}_\theta^{\mathrm{base}}(C) = \sum_k \pi_{\theta,k}(C)\mu_{\theta,k}(C)$. The implementation uses $K = 5$. This base head belongs to the jointly trained estimator; it is not the separately trained predictive-only baseline.

5.2 Residual–Score Construction

The observed modality pattern determines the active component set $\mathcal{A}_C$. The component families include outcome, treatment-balance, mediator-path, first-stage, and instrument-related residuals when

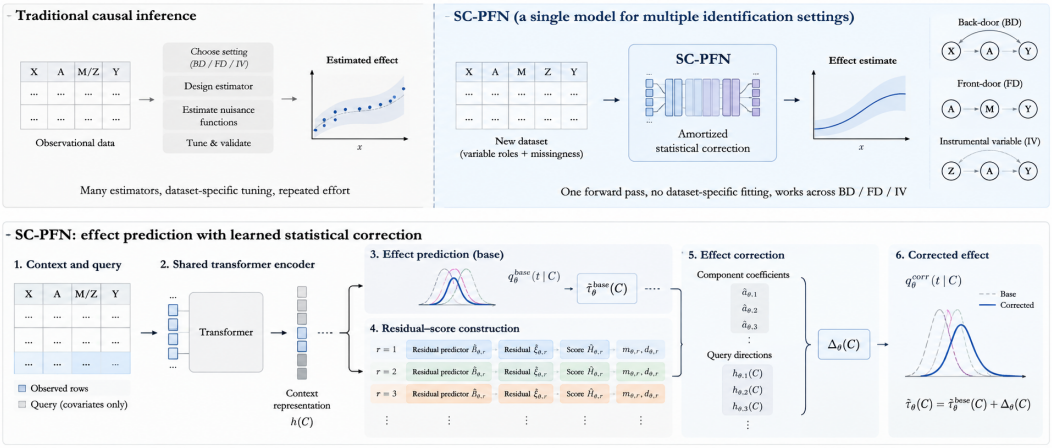


Figure 1: **Overview of the SC-PFN framework.** SC-PFN enables amortized causal effect estimation across back-door, front-door, and instrumental-variable regimes without task-specific retraining (top). A shared encoder outputs a base effect distribution and observational residual scores, which combine into a statistical shift to produce the final debiased effect estimate (bottom).

the required variables are available. The index r denotes a component, not an independently trained estimator for a causal setting. For its scalar observed target $B_r(O_i)$, the model predicts $\widehat{B}_{\theta,r}(O_i; C)$ and forms

$$\widehat{\xi}_{\theta,r}(O_i; C) = B_r(O_i) - \widehat{B}_{\theta,r}(O_i; C). \tag{8}$$

The target is constructed from observed variables, not from the simulator effect label or an unobserved counterfactual. The notation indexes a prediction for context row i; the residual is an observed prediction discrepancy, not the difference between an estimated nuisance function and its true value.

For each component, the network also produces row weights $\widehat{H}_{\theta,r}(O_i; C) \in \mathbb{R}$, $i = 1, \ldots, n$, and one query weight $h_{\theta,r}(C) \in \mathbb{R}$. Row weights determine the contributions of individual residuals. Query weights determine how the resulting component coefficients contribute to the effect at x. Here, "score" refers to a residual weight, not the full estimating score φ in Eq. (4). These outputs depend on the context and query while the network parameters are shared across datasets and identification settings.

5.3 Effect Correction

Component coefficients. Using the empirical average P_n from § 3.3, compute

$$m_{\theta,r}(C) = P_n[\widehat{H}_{\theta,r}\widehat{\xi}_{\theta,r}], \qquad d_{\theta,r}(C) = P_n[\widehat{H}^2_{\theta,r}] + \epsilon, \quad \epsilon > 0. \tag{9}$$

Arguments $(O; C)$ are suppressed inside empirical averages. The numerator retains the signed correspondence between rows and residuals; the denominator accounts for row-weight magnitude. The coefficient is

$$\widetilde{a}_{\theta,r}(C) = \frac{m_{\theta,r}(C)}{d_{\theta,r}(C)}, \qquad \widehat{a}_{\theta,r}(C) = \mathrm{clip}_{a_{\max}}(\widetilde{a}_{\theta,r}(C)), \tag{10}$$

where $a_{\max} > 0$ and $\mathrm{clip}_b(u) = \min\{b, \max\{-b, u\}\}$. For fixed residuals and row weights, this is the unique solution of

$$\widehat{a}_{\theta,r}(C) = \operatorname*{argmin}_{|a| \le a_{\max}} \left\{ P_n\left[(\widehat{\xi}_{\theta,r} - a\widehat{H}_{\theta,r})^2\right] + \epsilon a^2 \right\}. \tag{11}$$

The ridge makes the scalar objective strictly convex, and clipping projects its unconstrained minimizer onto the permitted interval. Appendix B.1 gives the derivation. Each component solves a separate scalar problem, not a joint regression over all row-weight vectors. No iterative optimization is required for these coefficients.

Query-level update. The query weights combine the coefficients into

$$\Delta_\theta(C) = \sum_{r \in \mathcal{A}_C} h_{\theta,r}(C)\widehat{a}_{\theta,r}(C). \tag{12}$$

The update translates the base density and its mean:

$$q_\theta^{\mathrm{corr}}(t \mid C) = q_\theta^{\mathrm{base}}(t - \Delta_\theta(C) \mid C), \tag{13}$$

$$\widehat{\tau}_\theta(C) = \widehat{\tau}_\theta^{\mathrm{base}}(C) + \Delta_\theta(C). \tag{14}$$

Every mixture mean shifts by the same amount, leaving the mixture weights and variance unchanged at fixed C. The residual-regression problem determines a coefficient used in this effect update; the model does not replace its observed residuals by $\widehat{\xi}_{\theta,r} - \widehat{a}_{\theta,r}\widehat{H}_{\theta,r}$.

Effective residual weights. When $|\widetilde{a}_{\theta,r}(C)| \le a_{\max}$ for all active components, clipping leaves the coefficients unchanged and

$$\begin{aligned} W_{\theta,r}(O;C) &= \frac{h_{\theta,r}(C)\widehat{H}_{\theta,r}(O;C)}{d_{\theta,r}(C)}, \\ \Delta_\theta(C) &= \sum_{r \in \mathcal{A}_C} P_n[W_{\theta,r}\widehat{\xi}_{\theta,r}]. \end{aligned} \tag{15}$$

Thus, the effective contribution of an observation depends jointly on its row weight, the component normalization, and the query weight. The model therefore learns both which residuals contribute and how their contributions are combined for a particular query. When a coefficient is clipped, the update remains the one in Eq. (12).

5.4 Joint Pretraining

Auxiliary supervision. The residual–score objective combines four terms:

$$\mathcal{L}_{\mathrm{RS}}(C) = \mathcal{L}_{\mathrm{comp}}(C) + \lambda_{\mathrm{ctr}}\mathcal{L}_{\mathrm{ctr}}(C) + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}}(C) + \lambda_{\mathrm{align}}\mathcal{L}_{\mathrm{align}}(C). \tag{16}$$

The component loss is $\mathcal{L}_{\mathrm{comp}}(C) = \sum_{r \in \mathcal{A}_C} P_n[\widehat{\xi}^2_{\theta,r}]$. The contrast loss is cross-entropy between an observed group label $c_r(O)$ and its prediction $\widehat{p}_{\theta,r}(\cdot \mid O, C)$, with $\lambda_{\mathrm{ctr}} = 1$. The group labels correspond to treatment groups, mediator strata, or instrument groups. The balancing term sums squared differences between score-weighted group representations $\overline{z}_{\theta,r,c}(C)$. The alignment term is $\mathcal{L}_{\mathrm{align}}(C) = |\widehat{\tau}_{\theta,\mathrm{aux}}(C) - \widehat{\tau}_\theta^{\mathrm{base}}(C)|^2$, where $\widehat{\tau}_{\theta,\mathrm{aux}}$ is the auxiliary effect prediction from the residual–score representation. The four expressions are expanded in Appendix D.4. Component and contrast losses use observed targets, whereas alignment compares two network predictions.

Moment regularization. Let $m_\theta(C) = \sum_{r \in \mathcal{A}_C} m_{\theta,r}(C)$. We use

$$\mathcal{L}_{\mathrm{mom}}(C) = \nu\, m_\theta(C) + \frac{\rho}{2} m_\theta(C)^2, \qquad \rho > 0, \tag{17}$$

where ν is the training multiplier. This term acts on the sum of unnormalized component moments. It is distinct from $\Delta_\theta(C)$, which also includes component normalization, clipping, and query weights. Cancellation in $m_\theta(C)$ therefore need not imply a zero effect update.

End-to-end objective. For the task distribution in § 4, the objective is

$$\begin{aligned} \mathcal{L}(\theta;\nu,\rho) = \mathbb{E}_{G,S,D_{\mathrm{obs}},x}\big[&- \log q_\theta^{\mathrm{corr}}(\tau_S(x) \mid C) \\ &+ \lambda_{\mathrm{RS}}\mathcal{L}_{\mathrm{RS}}(C) + \lambda_{\mathrm{mom}}\mathcal{L}_{\mathrm{mom}}(C)\big]. \end{aligned} \tag{18}$$

The expectation averages the per-context–query losses, including the quadratic moment term. The simulator effect label supervises the corrected density, not an isolated base head. Residuals and their coefficients use observed values and network outputs; no oracle row weights or correction labels are supplied. The likelihood and auxiliary objectives train the shared model jointly. Appendix C gives the training procedure; Table D.1 lists the encoder and optimizer settings.

At deployment, the model computes the base density, residuals, weights, and coefficients from the new context and query with θ fixed. The same context supplies the row predictions and empirical moments; the default computation does not partition it for cross-fitting or refit a nuisance model. What is amortized is the construction of the predictors and weights, while the coefficients are calculated from each observed context.

5.5 Properties of the Correction

We now examine how the update responds to residual changes after the weights have been constructed. Fix C, θ, the finite active set, and all row and query weights. Write $H_r = \widehat{H}_{\theta,r}(\cdot; C)$, $h_r = h_{\theta,r}(C)$, and $d_r = P_n[H_r^2] + \epsilon$. For residual arrays $v = (v_r)_{r \in \mathcal{A}_C}$, define

$$\mathcal{K}_C(v) = \sum_{r \in \mathcal{A}_C} h_r \, \mathrm{clip}_{a_{\max}}\left(\frac{P_n[H_r v_r]}{d_r}\right), \qquad \|v\|_{n,\mathcal{A}}^2 = \sum_{r \in \mathcal{A}_C} P_n[v_r^2]. \tag{19}$$

The actual correction is $\Delta_\theta(C) = \mathcal{K}_C(\widehat{\xi}_\theta)$.

Proposition 1 (Sensitivity of the correction)**.** *For finite row and query weights, $n \geq 1$, $\epsilon > 0$, and $a_{\max} > 0$, the map in Eq. (19) satisfies, for all residual arrays v, w,*

$$|\mathcal{K}_C(v) - \mathcal{K}_C(w)| \leq L_C \|v - w\|_{n,\mathcal{A}}, \qquad L_C^2 = \sum_{r \in \mathcal{A}_C} \frac{h_r^2 P_n[H_r^2]}{(P_n[H_r^2] + \epsilon)^2}. \tag{20}$$

Moreover, $|\mathcal{K}_C(v)| \leq a_{\max} \sum_r |h_r|$.

The proof applies the nonexpansiveness of clipping and Cauchy–Schwarz; see Appendix B.2. The factor L_C is the norm of the effective weights in Eq. (15) for the untruncated map. The coefficient constraint also bounds the magnitude of the update, with both bounds depending on the query weights. These results characterize residual amplification at fixed constructed weights, not changes in the weights when the context changes. Holding weights fixed here is an analysis condition, not a stop-gradient operation during training. The population orthogonality condition in Eq. (4) is separate from this bound.

6 Experiments

6.1 Experimental Setup

Benchmarks. We use the CausalFM benchmark (Ma et al., 2026), with ten held-out synthetic datasets for each of BD, FD, binary-IV, and continuous-IV estimation, and its fixed semi-synthetic Jobs dataset for BD. Treatment is binary throughout; the IV variants differ in instrument type. Each synthetic sequence contains 819 context rows and 205 query rows, with only query covariates supplied to the model, following § 4. One SC-PFN model is pretrained on the mixed BD/FD/IV pool and used with fixed parameters in all evaluations. Appendix D gives the data and evaluation details.

Baselines. Table 2 compares task-fitted estimators with pretrained causal models; their references are collected in Appendix D.3. These benchmark reference scores are distinct from the matched training controls in § 6.3. The controls share SC-PFN's backbone, mixed prior, training data, optimizer, schedule, and instantiated parameter budget.

Evaluation metrics. For context D and queries $\mathcal{Q} = \{x_q\}_{q=1}^{Q}$, we report precision in estimation of heterogeneous effects (PEHE):

$$\mathrm{PEHE}(D, \mathcal{Q}) = \left[\frac{1}{Q}\sum_{q=1}^{Q}\{\widehat{\tau}_\theta(D, x_q) - \tau_S(x_q)\}^2\right]^{1/2}. \tag{21}$$

The target follows § 4, including the structural contrast for IV. For SC-PFN and its matched variants, means and standard deviations summarize PEHE across ten synthetic datasets or ten context/query permutations of the fixed Jobs dataset. These repetitions vary the evaluation data, not the pretrained parameters.

6.2 Effect-Estimation Accuracy

SC-PFN has the lowest reported mean PEHE in each of the five comparisons in Table 2. The differences from CausalFM are larger on BD-Synthetic and IV-Continuous than on Jobs and FD. For IV-Continuous, SC-PFN also has lower mean error than DeepIV, the strongest task-fitted reference in that column, with PEHE of 0.456 versus 0.516. The FD difference from CausalFM is 0.015, indicating that the size of the mean improvement varies across settings.

Table 2: Effect-estimation results (PEHE; mean ± standard deviation; lower is better). Task-fitted estimators precede pretrained models in each panel. Bold and underlined means mark the best and second-best reported values, not statistical significance. Binary and continuous refer to the instrument.

(a) Back-door

Method	Synthetic	Jobs
TARNet	0.854±0.23	0.864±0.24
S-learner	0.734±0.16	0.697±0.18
T-learner	0.661±0.17	0.822±0.18
X-learner	0.563±0.15	0.802±0.18
RA-learner	0.609±0.13	0.652±0.15
DR-learner	0.765±0.17	0.959±0.18
CausalPFN	0.557±0.18	0.528±0.16
Do-PFN	0.586±0.19	0.482±0.20
CausalFM	<u>0.515</u>±0.20	<u>0.478</u>±0.18
SC-PFN	**0.440**±0.14	**0.461**±0.06

(b) Instrumental variables

Method	Binary	Continuous
DeepIV	0.427±0.15	<u>0.516</u>±0.13
KIV	0.454±0.16	0.577±0.20
DRIV	0.531±0.18	0.693±0.20
DeepGMM	0.503±0.20	0.588±0.21
DMLIV	0.479±0.23	0.618±0.20
DFIV	0.709±0.29	0.583±0.30
MRIV	0.688±0.21	0.641±0.24
Do-PFN	0.523±0.20	0.675±0.37
CausalFM	<u>0.422</u>±0.16	0.579±0.21
SC-PFN	**0.399**±0.15	**0.456**±0.13

(c) Front-door

Method	PEHE
FD-Linear	1.124±0.28
FD-RF	1.364±0.52
FD-NN	0.889±0.38
Do-PFN	1.274±0.24
CausalFM	<u>0.847</u>±0.20
SC-PFN	**0.832**±0.35

Table 3: Contribution of the correction (PEHE; mean ± standard deviation; lower is better). The first four variants are retrained under the matched setup. The last row shuffles scores in the full model at inference without retraining.

Configuration	BD-Syn.	Jobs	IV-Bin.	IV-Cont.	FD
Predictive only	0.528±0.19	0.489±0.17	0.431±0.17	0.592±0.22	0.861±0.23
Auxiliaries only	0.497±0.17	0.481±0.12	0.421±0.16	0.547±0.19	0.851±0.28
No effect update	0.463±0.16	0.470±0.09	0.409±0.15	0.489±0.16	0.842±0.31
No moment penalty	0.456±0.15	0.467±0.08	0.406±0.15	0.476±0.15	0.838±0.33
SC-PFN	**0.440**±0.14	**0.461**±0.06	**0.399**±0.15	**0.456**±0.13	**0.832**±0.35
Shuffled scores	0.551±0.20	0.520±0.15	0.465±0.18	0.614±0.24	0.891±0.31

6.3 Contribution of the Correction

Matched training controls. We independently train four variants. *Predictive only* removes auxiliary supervision, the moment penalty, and the effect update. *Auxiliaries only* retains the residual–score objective but omits the moment penalty and effect update. The other two variants remove only the effect update or only the moment penalty from the full model. Disabled outputs remain instantiated. These comparisons evaluate training configurations, rather than removing an update from a fixed trained predictor. In particular, the predictive-only control is not the base head within the jointly trained SC-PFN.

In Table 3, auxiliary supervision reduces mean PEHE relative to predictive-only training in all five evaluations, but does not match the full model. Removing the effect update increases mean error even when auxiliary supervision and moment regularization are retained. Removing the moment penalty also increases mean error in every setting. The differences are larger on BD-Synthetic and IV-Continuous than on Jobs and FD.

Row-specific weighting. We shuffle learned row scores across context rows at inference, preserving their values but breaking their correspondence with observations. PEHE increases in every evaluation, from 0.440 to 0.551 on BD-Synthetic and from 0.456 to 0.614 on IV-Continuous. Unlike the retrained controls, this intervention leaves the network parameters fixed. The result supports the role of observation-specific weighting in the trained estimator, rather than the score values alone.

6.4 Sensitivity to Observational Context

Nonlocal-context contamination. Fig. 2(a) increases the fraction α of context rows that follow global outcome patterns but are irrelevant to the query-local effect in the stress test. The delocalization

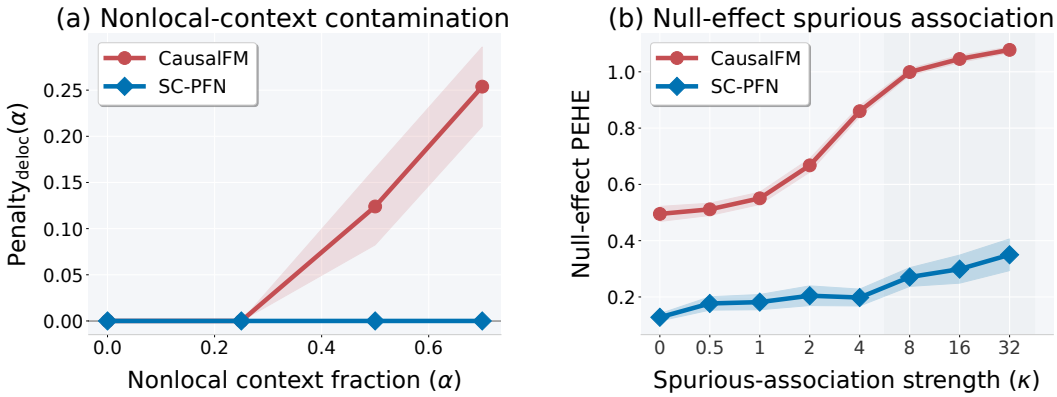


Figure 2: Sensitivity to observational context. (a) Delocalization penalty under nonlocal-context contamination. (b) Null-effect PEHE as spurious association increases. Lower values are better.

penalty increases for CausalFM but remains nearly flat for SC-PFN over the tested range. This comparison examines sensitivity to context composition.

Spurious association under a null effect. Fig. 2(b) fixes the true effect at zero and increases the spurious-association parameter κ. SC-PFN has lower PEHE than CausalFM at every plotted value, although its error also increases as the association strengthens. Under this null, PEHE is the root mean square of predicted effects. It measures false-effect magnitude rather than repeated-context bias alone.

7 Conclusion

We presented SC-PFN, a causal foundation model that learns effect prediction and residual correction jointly. A shared model constructs row weights and query-level updates across identification settings without dataset-specific parameter updates. Our analysis characterizes the coefficients as constrained ridge solutions and bounds residual amplification with the constructed weights held fixed. Benchmark results show lower mean effect-estimation error than the evaluated baselines. Component comparisons show gains beyond predictive training and auxiliary supervision alone, while score shuffling shows the estimator's use of observation-specific weights. These findings support learning statistical correction within pretrained causal inference under the identifying assumptions and task families studied.

AI Use Statement

Large language models were employed exclusively to enhance text clarity and polish academic prose. The core conceptualization, methodological formulations, theoretical derivations, and empirical analyses were conceived and executed entirely by the authors, who accept complete accountability for this work.

Ethics Statement

Our investigation addresses core algorithmic and methodological principles using open-access scientific datasets. This research does not utilize private individual records, involve participants, or foresee plausible pathways for societal harm or ethical compromise.

Reproducibility Statement

The accompanying codebase, including model implementations and experimental routines, is accessible via the anonymous repository at https://anonymous.4open.science/r/SC-PFN-55668. Detailed mathematical proofs, experimental specifications, baseline parameters, and data processing workflows are comprehensively laid out in the manuscript and supplementary material.

REFERENCES

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. *Advances in Neural Information Processing Systems*, 2025.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression. *arXiv preprint arXiv:2104.14737*, 2021.

Victor Chernozhukov, Whitney Newey, Vıctor M Quintas-Martınez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pp. 3901–3914. PMLR, 2022.

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*, pp. 1810–1818. PMLR, 2021.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Dennis Frauen and Stefan Feuerriegel. Estimating individual treatment effects under unobserved confounding using binary instruments. In *International Conference on Learning Representations*, 2023.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International conference on machine learning*, pp. 1414–1423. PMLR, 2017.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

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

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, pp. 207–236, 2024.

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.

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*, 2026.

Valentyn Melnychuk, Vahid Balazadeh, Stefan Feuerriegel, and Rahul G Krishnan. Frequentist consistency of prior-data fitted networks for causal inference. *arXiv preprint arXiv:2603.12037*, 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.

Thomas Nagler. Statistical foundations of prior-data fitted networks. In *International Conference on Machine Learning*, pp. 25660–25676. PMLR, 2023.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

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*, 2025.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems*, 32, 2019.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

Appendices

A Identification Settings **13**

A.1 Back-door Adjustment 13

A.2 Front-door Adjustment 13

A.3 Instrumental Variables 13

B Properties of the Correction **14**

B.1 Constrained Ridge Coefficients 14

B.2 Sensitivity with Constructed Weights Held Fixed 14

B.3 Translation of the Effect Density 15

B.4 Statistical Reference 15

C Training Algorithm **16**

D Implementation Details **16**

D.1 Synthetic Task Generation 16

D.2 Architecture and Optimization 17

D.3 Baselines 17

D.4 Residual–Score Losses 19

D.5 Evaluation Metrics and Matched Controls 19

D.6 Observational-Context Stress Tests 19

D.7 BD Reference-Score Diagnostic 20

E Broader Impacts **20**

A Identification Settings

We state the identifying assumptions and functionals used in § 4, following Pearl (2009); Ma et al. (2026). Fix an SCM S and suppress its subscript where unambiguous. All conditional means are assumed finite on the relevant support. These formulas define the causal targets; they are not additional estimators fitted by SC-PFN at deployment.

A.1 Back-door Adjustment

For $O = (X, A, Y)$, assume consistency $Y = Y(A)$, $(Y(0), Y(1)) \perp A \mid X$, and $0 < e(X) < 1$ almost surely, where $e(x) = P(A = 1 \mid X = x)$. With $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$, exchangeability and consistency give

$$\mathbb{E}[Y(a) \mid X = x] = \mathbb{E}[Y(a) \mid A = a, X = x] = \mu_a(x), \qquad \tau_S(x) = \mu_1(x) - \mu_0(x). \tag{22}$$

Positivity ensures that both conditional means are defined on the covariate support used for evaluation.

A.2 Front-door Adjustment

For $O = (X, A, M, Y)$, assume consistency for treatment and mediator interventions, treatment and mediator positivity on the adjustment support, and the conditional front-door criterion. The mediator intercepts all directed paths from A to Y. Given X, there is no unblocked back-door path from A to M, and conditioning additionally on A blocks the back-door paths from M to Y. The benchmark states the corresponding exchangeability conditions as

$$M(a) \perp A \mid X, \qquad Y(a', m) \perp M \mid A = a', X. \tag{23}$$

These conditions are used together with complete mediation, not as a replacement for it. Let $\mu(x, a, m) = \mathbb{E}[Y \mid X = x, A = a, M = m]$ and $p_a(x) = P(A = a \mid X = x)$. The conditional front-door formula is

$$\psi_a(x) = \int \left\{ \sum_{a'=0}^{1} \mu(x, a', m) p_{a'}(x) \right\} dP(m \mid A = a, X = x), \tag{24}$$

$$\tau_S(x) = \psi_1(x) - \psi_0(x). \tag{25}$$

The integral is a sum for discrete mediators. The required positivity includes support for evaluating $\mu(x, a', m)$ under both treatment values in the inner sum.

A.3 Instrumental Variables

For $O = (X, Z, A, Y)$, use the structural model

$$Y = f_S(X, A) + g_S(X, U), \qquad U \perp Z \mid X. \tag{26}$$

Here U is an unobserved confounder, and exclusion is encoded by the absence of a direct dependence of Y on Z. Treatment remains binary for both instrument types. The benchmark assumes instrument relevance and completeness for the structural function class (Ma et al., 2026). Relevance requires $P(A = 1 \mid X = x, Z = z)$ to vary with z. Completeness requires a candidate difference u satisfying $\mathbb{E}[u(X, A) \mid X, Z] = 0$ to vanish almost surely.

Identifying equation. Let $c_S(x) = \mathbb{E}[g_S(X, U) \mid X = x]$ and $s_S(x, a) = f_S(x, a) + c_S(x)$. Taking conditional expectations in Eq. (26) yields

$$\mathbb{E}[Y \mid X = x, Z = z] = \sum_{a=0}^{1} s_S(x, a) P(A = a \mid X = x, Z = z). \tag{27}$$

Completeness makes the solution unique within the specified class. The additive term cancels in the contrast, giving

$$\tau_S(x) = s_S(x, 1) - s_S(x, 0) = f_S(x, 1) - f_S(x, 0). \tag{28}$$

This is the structural effect specified by the benchmark, not a complier-specific local average treatment effect.

Scope. Variable roles and modality patterns specify the available inputs. They do not establish the validity of an adjustment set, mediator, or instrument. The priors encode the identifying conditions, and the estimator is evaluated within those task families.

B Properties of the Correction

This section proves the properties in § 5.5 for the update defined in the Method section. The calculations concern finite real-valued arrays with $n \geq 1$ and a finite active component set. Network parameters, the context, and constructed weights are held fixed throughout the perturbation analysis.

B.1 Constrained Ridge Coefficients

Fix one component and write $H_i = \widehat{H}_{\theta,r}(O_i; C)$, $\xi_i = \widehat{\xi}_{\theta,r}(O_i; C)$, $m = P_n[H\xi]$, $s = P_n[H^2]$, and $d = s + \epsilon$.

Completing the square. The objective in Eq. (11) satisfies

$$\begin{aligned} J(a) &= P_n[(\xi - aH)^2] + \epsilon a^2 \\ &= P_n[\xi^2] - 2am + da^2 \\ &= P_n[\xi^2] - \frac{m^2}{d} + d\left(a - \frac{m}{d}\right)^2. \end{aligned} \tag{29}$$

Since $d > 0$, the unconstrained minimizer is uniquely m/d. Minimizing over $[-a_{\max}, a_{\max}]$ projects this value onto the interval, so

$$\widehat{a} = \mathrm{clip}_{a_{\max}}(m/d). \tag{30}$$

Strict convexity also gives uniqueness under the constraint. Each component solves its own scalar problem; this is not a joint regression on all component scores.

Residual-regression identity. For the unconstrained coefficient $\widetilde{a} = m/d$,

$$P_n[H(\xi - \widetilde{a}H)] = m - \frac{m}{s+\epsilon}s = \frac{\epsilon}{s+\epsilon}m. \tag{31}$$

With clipping, the corresponding expression is $m - \widehat{a}s$. This identity concerns the hypothetical regression residual $\xi - aH$. The deployed estimator instead shifts the effect density through Eqs. (12) and (13); it does not recompute its observational residuals as $\xi - aH$.

B.2 Sensitivity with Constructed Weights Held Fixed

Proof of Prop. 1.

Clipping. Projection onto a closed interval is nonexpansive: $|\mathrm{clip}_b(u) - \mathrm{clip}_b(v)| \leq |u - v|$. Holding H_r, h_r, and $d_r = P_n[H_r^2] + \epsilon$ fixed gives

$$\begin{aligned} |\mathcal{K}_C(v) - \mathcal{K}_C(w)| &\leq \sum_r \frac{|h_r|}{d_r} |P_n[H_r(v_r - w_r)]| \\ &\leq \sum_r \frac{|h_r|\sqrt{P_n[H_r^2]}}{d_r} \sqrt{P_n[(v_r - w_r)^2]} \\ &\leq \left(\sum_r \frac{h_r^2 P_n[H_r^2]}{d_r^2}\right)^{1/2} \left(\sum_r P_n[(v_r - w_r)^2]\right)^{1/2} \\ &= L_C \|v - w\|_{n,\mathcal{A}}. \end{aligned} \tag{32}$$

The second line applies Cauchy–Schwarz over observations; the third applies it over components.

Magnitude. Each clipped coefficient has magnitude at most $a_{\max}$. Consequently,

$$|\mathcal{K}_C(v)| \le \sum_r |h_r| \left|\mathrm{clip}_{a_{\max}}\left(\frac{P_n[H_r v_r]}{d_r}\right)\right| \le a_{\max} \sum_r |h_r|. \tag{33}$$

□

Role of the ridge. For $s \ge 0$, $(s+\epsilon)^2 \ge 4s\epsilon$. Hence

$$L_C^2 \le \frac{1}{4\epsilon} \sum_r h_r^2. \tag{34}$$

This bound still depends on the query weights and the ridge. It is not context-uniform without additional control of the query weights.

Effective weights. Define the untruncated map separately by

$$\mathcal{K}_C^0(v) = \sum_r P_n[W_r v_r], \qquad W_r = \frac{h_r H_r}{d_r}. \tag{35}$$

It agrees with $\mathcal{K}_C$ whenever clipping is inactive for the residual array under consideration. The norm of $\mathcal{K}_C^0$ under $\|\cdot\|_{n,\mathcal{A}}$ is L_C: Cauchy–Schwarz gives the upper bound, and $v = W/\|W\|_{n,\mathcal{A}}$ attains it for $W \neq 0$. For $W = 0$, the norm is zero. The effective weight includes both the query weight and the normalization, not just the row-level score.

Scope. The proposition describes changes to residual arrays after weight construction, not the complete predictor's response to a changed dataset. Holding weights fixed in this analysis does not introduce a stop-gradient operation during training.

B.3 Translation of the Effect Density

For fixed C, let $\Delta = \Delta_\theta(C)$. The substitution $u = t - \Delta$ in $q_\theta^{\mathrm{corr}}(t \mid C) = q_\theta^{\mathrm{base}}(t - \Delta \mid C)$ gives

$$\begin{aligned} \int q_\theta^{\mathrm{corr}}(t \mid C)\, dt &= 1, \\ \mathbb{E}_{q_\theta^{\mathrm{corr}}}[t \mid C] &= \mathbb{E}_{q_\theta^{\mathrm{base}}}[t \mid C] + \Delta, \\ \mathrm{Var}_{q_\theta^{\mathrm{corr}}}(t \mid C) &= \mathrm{Var}_{q_\theta^{\mathrm{base}}}(t \mid C). \end{aligned} \tag{36}$$

These identities concern the output density at one context. They do not establish frequentist coverage or preserve the variance of point estimates across different contexts.

B.4 Statistical Reference

The Riesz representation in Eq. (5) provides the statistical motivation for residual weighting (Chernozhukov et al., 2021; 2022). For a fixed target, assume its first variation $\mathcal{A}$ is continuous and linear on a closed subspace $\mathcal{T} \subseteq L^2(P_0)^R$ with its $L^2(P_0)$ inner product and representative $H^\star$. For any $g \in \mathcal{T}$,

$$P_0\|g\|_2^2 - 2\mathcal{A}(g) = \|g - H^\star\|_{P_0}^2 - \|H^\star\|_{P_0}^2. \tag{37}$$

This follows by expanding the square and substituting $\mathcal{A}(g) = P_0[(H^\star)^\top g]$. Over a nonempty closed linear subspace, the minimizer is the orthogonal projection of $H^\star$ onto that subspace. This is a population reference identity, not a characterization of the minimizer of SC-PFN's auxiliary loss.

The observed residual $B_r - \widehat{B}_{\theta,r}$ contains noise as well as prediction error; it is not a nuisance perturbation $\widehat{\eta} - \eta_0$. The constrained ridge solution and conditioned sensitivity bound therefore do not establish Neyman orthogonality of the learned effect estimator.

Algorithm 1 SC-PFN pretraining

Require: Task priors $\{\Pi_G\}$ and their mixture; parameters θ; loss weights $\lambda_{\mathrm{RS}}, \lambda_{\mathrm{mom}}$; moment parameters ν, ρ; ridge $\epsilon > 0$; coefficient bound $a_{\max} > 0$
Ensure: Pretrained parameters θ

for each scheduled training step **do**
  1. Task sampling and base prediction
  Sample G, $S \sim \Pi_G$, context $D_{\mathrm{obs}} \sim P_S^{\otimes n}$, and query x.
  Obtain the simulator effect label $\tau_S(x)$.
  Encode $C = (D_{\mathrm{obs}}, x)$ and predict $q_\theta^{\mathrm{base}}(t \mid C)$ using Eq. (7).
  2. Residual–score construction
  for all $r \in \mathcal{A}_C$ **do**
    Predict $\widehat{B}_{\theta,r}(O_i; C)$, $\widehat{H}_{\theta,r}(O_i; C)$, and $h_{\theta,r}(C)$.
    Form $\widehat{\xi}_{\theta,r}(O_i; C) = B_r(O_i) - \widehat{B}_{\theta,r}(O_i; C)$.
    Compute $m_{\theta,r} = P_n[\widehat{H}_{\theta,r}\widehat{\xi}_{\theta,r}]$ and $d_{\theta,r} = P_n[\widehat{H}_{\theta,r}^2] + \epsilon$.
    Compute $\widehat{a}_{\theta,r} = \mathrm{clip}_{a_{\max}}(m_{\theta,r}/d_{\theta,r})$.
  end for
  Form $m_\theta(C) = \sum_r m_{\theta,r}(C)$ and $\mathcal{L}_{\mathrm{RS}}(C)$.
  3. Effect correction
  Compute $\Delta_\theta(C) = \sum_r h_{\theta,r}(C)\widehat{a}_{\theta,r}(C)$.
  Set $q_\theta^{\mathrm{corr}}(t \mid C) = q_\theta^{\mathrm{base}}(t - \Delta_\theta(C) \mid C)$.
  4. Objective and parameter update
  Compute $\mathcal{L}_{\mathrm{pred}} = -\log q_\theta^{\mathrm{corr}}(\tau_S(x) \mid C)$.
  Compute $\mathcal{L}_{\mathrm{mom}} = \nu m_\theta(C) + \frac{\rho}{2} m_\theta(C)^2$.
  Update θ using the batch objective in Eq. (18) and the optimizer settings.
  Update ν and ρ according to the configured moment-penalty schedule.
end for

C Training Algorithm

This appendix expands the pretraining procedure in § 5.4. Each episode samples a causal task, predicts a base effect density, constructs residual–score components, and computes an effect update. The setting G determines the simulator prior but is not an explicit model input; the network receives the observations, variable roles, and modality pattern. Algorithm 1 shows the episode-level computation. Training batches use the settings in Table D.1.

At deployment, θ is fixed. The network constructs residuals and weights from the new context, computes the coefficients and density translation, and returns $\widehat{\tau}_\theta(C) = \widehat{\tau}_\theta^{\mathrm{base}}(C) + \Delta_\theta(C)$. The default calculation uses the same context for weight construction and empirical moments.

D Implementation Details

This appendix specifies the data, model configuration, optimization, baselines, and evaluation protocol used in § 6. SC-PFN is pretrained once and evaluated in context without updating network parameters on an evaluation dataset.

D.1 Synthetic Task Generation

For BD, FD, and IV, we use the CausalFM SCM benchmark pipeline (Ma et al., 2026). Observed rows are (X, A, Y), (X, A, M, Y), and (X, Z, A, Y), respectively. Each episode contains an observational context, query covariates, and effect labels from the same SCM. The index G is used for task generation and evaluation grouping, not as an explicit model input. Unavailable mediator or instrument entries use missing-value tokens. Our model is pretrained on the mixed BD/FD/IV pool.

Each synthetic evaluation sequence contains 1,024 rows, with 819 context rows and 205 query rows. Query inputs contain X only: A, Y are hidden for BD; A, M, Y for FD; and Z, A, Y for IV. Metrics are evaluated only on query rows. This is a context/query construction for inference, not a split for

fitting network parameters. The target is the CATE for BD and FD and the structural contrast in Eq. (28) for IV. The Jobs evaluation uses the fixed semi-synthetic benchmark dataset. Context stress tests are described in Appendix D.6.

D.2 Architecture and Optimization

SC-PFN uses a transformer encoder over the context and query. The shared representation supplies the base effect density, component predictions, row weights, and query weights. The density is a Gaussian mixture with $K = 5$. Active residual components are selected from the observed modality pattern, and their coefficients are computed using Eq. (10). Table D.1 lists the configuration.

D.3 Baselines

We compare SC-PFN with PFN-based causal models, back-door CATE estimators, front-door plug-in estimators, and IV estimators. The evaluated methods and settings are listed in Table 2. Task-fitted methods fit their models on each evaluation dataset; pretrained PFNs make predictions from the observational context. Benchmark reference scores are distinguished from matched SC-PFN training controls.

PFN-based causal estimators.

- **CausalPFN** (Balazadeh et al., 2025) is trained to estimate causal quantities under ignorability. It takes an observational context and predicts effects without fitting new parameters on the target dataset. It is included in the BD comparison.
- **Do-PFN** (Robertson et al., 2025) predicts conditional interventional distributions from observational contexts and intervention queries. It does not require a user-specified DAG. Effect estimates are obtained from its interventional predictions on the benchmark settings where it is evaluated.
- **CausalFM** (Ma et al., 2026) constructs SCM priors for BD, FD, and IV estimation and trains PFNs on causal targets. It provides the prior and benchmark framework used here. SC-PFN adds the learned residual correction described in § 5.

Back-door CATE estimators.

- **S-learner** (Künzel et al., 2019) fits one outcome model using covariates and treatment. It estimates CATE by contrasting predictions at $A = 1$ and $A = 0$.
- **T-learner** (Künzel et al., 2019) fits separate outcome regressions for treated and control units and takes their difference.
- **X-learner** (Künzel et al., 2019) imputes treatment effects using outcome regressions and fits effect models to the resulting pseudo-outcomes.
- **RA-learner** (Curth & Van der Schaar, 2021) uses a regression-adjusted pseudo-outcome to estimate CATE.
- **DR-learner** (Kennedy, 2023) constructs doubly robust pseudo-outcomes from outcome regressions and treatment probabilities, then regresses them on covariates.
- **TARNet** (Shalit et al., 2017) learns a shared covariate representation with separate treated and control outcome heads.

Front-door estimators.

- **FD-Linear** (Pearl, 2009) evaluates the front-door identifying formula using linear nuisance models.
- **FD-RF** (Pearl, 2009) uses random forests for the nuisance components of the front-door formula.
- **FD-NN** (Pearl, 2009) uses neural networks for these nuisance components.

Table D.1: Architecture and optimization hyperparameters. N denotes sequence length and n the number of context rows.

Category	Hyperparameter	Value
Data	Prior structures	BD, FD, IV
	Training pool	Mixed BD/FD/IV
	Sequence split	$n = \lfloor 0.8N \rfloor$ context rows; $N - n$ query rows
	Covariate dimension d_X	Dataset-dependent; maximum 128 feature tokens
	Treatment type	Binary
	Query sampling	Last 20% of each sequence
Input	Row layouts	(X, A, Y), (X, A, M, Y), (X, Z, A, Y)
	Missing modalities	Learned missing-value token / NaN indicator encoder
	Explicit setting label	Not used
Backbone	Encoder	Transformer encoder
	Number of layers	10
	Hidden dimension	128
	Attention heads	4
	Feed-forward dimension	512
	Dropout	0.0
	Normalization	Post-norm LayerNorm, $\epsilon_{\mathrm{LN}} = 10^{-5}$
Heads	Effect distribution	Gaussian mixture
	Mixture components K	5
	Residual components	Outcome, treatment-balance, mediator-path, first-stage, instrument-related
	Weight outputs	Row weights $\widehat{H}_{\theta,r}$ and query weights $h_{\theta,r}$
Optimization	Optimizer	Adam
	Learning rate	10^{-3}
	Weight decay	10^{-5}
	Batch size	32 episodes
	Training duration	150 epochs
	Learning-rate schedule	ReduceLROnPlateau; factor 0.5, patience 5
	Warm-up steps	100
	Gradient clipping	1.0
	Mixed precision	Not used
Loss	GMM likelihood weight	1.0
	Moment-loss setting	$\eta_{\mathrm{orth}} = 0.1$
	Contrast weight λ_{ctr}	1.0
	Balance weights	Representation: 0.1; treatment: 0.05
	Alignment weights	Factual: 0.1; mediator, first-stage, instrument, effect consistency: 0.05 each
	Moment multiplier ν	Initialized at 0; augmented-Lagrangian update; clipped to $[-100, 100]$
	Moment penalty ρ	Initialized at 1; growth 1.02; maximum 50
	Coefficient regularization	Ridge $\epsilon = 10^{-8}$; $a_{\max} = 5$
Reporting	Training seed	42
	Hardware	NVIDIA RTX Pro 6000

Instrumental-variable estimators.

- **DeepIV** (Hartford et al., 2017) first models treatment conditional on instruments and covariates, then uses this distribution to learn the outcome response.
- **KIV** (Singh et al., 2019) uses kernel ridge regression to estimate first-stage and structural relationships.
- **DRIV** (Syrgkanis et al., 2019) combines first-stage, reduced-form, and effect-model components through doubly robust pseudo-outcomes.
- **DeepGMM** (Bennett et al., 2019) formulates IV estimation as a generalized method-of-moments problem and uses adversarial training to enforce the moment restrictions.
- **DMLIV** (Syrgkanis et al., 2019) estimates nuisance functions and then fits treatment effects through an orthogonalized regression objective.
- **DFIV** (Xu et al., 2021) uses neural feature learning for first-stage and structural IV regression.
- **MRIV** (Frauen & Feuerriegel, 2023) estimates heterogeneous effects using multiply robust pseudo-outcomes.

D.4 Residual–Score Losses

For component r, let $c_r(O)$ be its contrast label and $\widehat{p}_{\theta,r}(\cdot \mid O, C)$ the predicted class distribution. Let $\overline{z}_{\theta,r,c}(C)$ be the score-weighted mean representation in group c, and $\widehat{\tau}_{\theta,\mathrm{aux}}(C)$ the auxiliary effect estimate. The terms in Eq. (16) are

$$\mathcal{L}_{\mathrm{comp}}(C) = \sum_{r \in \mathcal{A}_C} P_n[\widehat{\xi}^2_{\theta,r}], \tag{38}$$

$$\mathcal{L}_{\mathrm{ctr}}(C) = \sum_{r \in \mathcal{A}_C} P_n[-\log \widehat{p}_{\theta,r}(c_r(O) \mid O, C)], \tag{39}$$

$$\mathcal{L}_{\mathrm{bal}}(C) = \sum_{r \in \mathcal{A}_C} \sum_{c<c'} \|\overline{z}_{\theta,r,c}(C) - \overline{z}_{\theta,r,c'}(C)\|_2^2, \tag{40}$$

$$\mathcal{L}_{\mathrm{align}}(C) = |\widehat{\tau}_{\theta,\mathrm{aux}}(C) - \widehat{\tau}^{\mathrm{base}}_{\theta}(C)|^2. \tag{41}$$

The first two terms supervise component targets and contrast labels. The balancing and alignment terms regularize group representations and the auxiliary effect prediction. The contrast term has fixed coefficient $\lambda_{\mathrm{ctr}} = 1$. The aggregate moment penalty is given in Eq. (17) and is separate from the effect update.

D.5 Evaluation Metrics and Matched Controls

We use PEHE as defined in Eq. (21). For SC-PFN and its matched controls, synthetic results summarize ten held-out datasets per benchmark. Jobs results summarize ten row permutations of one fixed dataset, using seeds $0, \ldots, 9$. These repetitions do not vary the pretrained parameters. The standard deviations describe variation over these evaluation units, not independent training seeds or confidence intervals.

The training variants in Table 3 share the backbone, mixed prior, data, optimizer, schedule, and instantiated parameter budget, and are trained independently from initialization. Predictive-only training removes auxiliary and moment losses and the effect update. Auxiliaries-only training retains the residual–score loss but omits the moment penalty and effect update. The other two training controls remove only the effect update or only the moment penalty. Disabled outputs remain instantiated. The shuffled-score row instead permutes the full model's row scores at inference, without retraining.

D.6 Observational-Context Stress Tests

The prior-interface comparison evaluates CausalFM with the designated correct and incorrect interfaces and SC-PFN with its shared model. It concerns use of the represented identification setting, not recovery under arbitrary violations of causal assumptions.

In the nonlocal-context test, α changes the fraction of rows that follow global outcome patterns but are irrelevant to the query-local effect. Fig. 2(a) reports the delocalization penalty. The null-effect test fixes $\tau_S(x) = 0$ and increases spurious association through κ. Its PEHE in Fig. 2(b) is the root mean square of predicted effects, measuring false-effect magnitude rather than repeated-context bias alone.

D.7 BD Reference-Score Diagnostic

The diagnostic uses ten BD episodes and 96 perturbations per episode. Its reference propensity is

$$e_{\mathrm{ref}}(x) = \min\{0.925, \max\{0.075,\, 0.5 + 0.5\{e_{\mathrm{sim}}(x) - 0.5\}\}\}\,, \tag{42}$$

where e_{sim} is the simulator propensity. The stabilized BD score H_{ref} uses this transformed propensity. For a residual perturbation v, the reference and model responses are

$$s_{\mathrm{ref}}(v) = P_n[H_{\mathrm{ref}} v], \qquad s_\theta(v) = P_n[\widehat{H}_\theta v]. \tag{43}$$

Scores are standardized and responses are affine-calibrated within each episode. The calibration is fitted and evaluated on the same perturbations, some of which are reference-aligned.

The learned scores yield calibrated $R^2 = 0.939$ and MAE $= 0.073$, compared with 0.113 and 0.292 for row-shuffled scores. These values measure in-sample reference agreement, not a simulator finite-difference test or error of the final query-level correction. The episode indices denote diagnostic tasks, not independently pretrained models.

E Broader Impacts

SC-PFN learns causal effect estimation and residual correction across datasets, reducing the need for dataset-specific model fitting. Incorrect causal assumptions, invalid instruments or mediators, and distribution shift can nevertheless produce misleading effects and harmful decisions. The method should be used with domain knowledge, assumption checks, sensitivity analysis, and human oversight, rather than as an autonomous decision-making system.