WHAT YOU OBSERVE DETERMINES HOW YOU IDENTIFY CAUSAL EFFECTS: EVALUATING CAUSAL MODELS ACROSS OBSERVATIONAL VIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal foundation models (CFMs) pre-trained on data generated from various structural causal models (SCMs) have been proposed for estimating causal effects from observational data. However, differences in pre-training environments and evaluation protocols make it difficult to assess how their performance depends on the information available for causal identification. To enable controlled comparisons, we introduce **CausalIDView**[1], a multi-view benchmark that holds the SCM realization and target estimand fixed while varying only the observational view available to the estimator. Each observational view corresponds to a distinct identification regime under the benchmark's maintained causal assumptions. Across these matched views, no CFM consistently performs best and model rankings vary substantially. Under controlled structural changes, CFMs exhibit model-specific failures to maintain stable estimates when true effects are unchanged and to track genuine effect changes. We also examine whether combining explicit identification with strong predictive estimation is effective. A modular approach that pairs a predictive tabular foundation model with regime-specific identification procedures is competitive with CFMs and outperforms several of them. These findings motivate cross-regime comparisons to assess the empirical value of CFMs.

1 INTRODUCTION

Machine-learning methods for causal effect estimation have become useful estimators by flexibly estimating outcome regressions, propensity scores and other nuisance functions under a specified set of identification assumptions (Athey et al., 2019; Kennedy, 2023; Künzel et al., 2019; Nie and Wager, 2021; Shalit et al., 2017; Shi et al., 2019). Recently, Causal Foundation Models (CFMs) have been proposed to amortize causal inference across datasets (Balazadeh Meresht et al., 2026; Ma et al., 2026; Robertson et al., 2026). At test time, they take observational datasets as context and produce causal predictions through in-context learning without updating pre-trained parameters. This shifts the computational burden from dataset-specific training to shared pre-training, making CFMs attractive for applications requiring causal analyses across multiple datasets. However, current CFMs are trained and evaluated under different observational structures and causal assumptions. It remains unclear how their performance changes when the information available for identification changes.

Existing evaluations do not directly isolate this question. Across existing CFM studies, both pre-training priors and evaluation protocols differ, including the data-generating processes and causal assumptions used for evaluation (Balazadeh Meresht et al., 2026; Ma et al., 2026; Robertson et al., 2026). Robertson et al. (2026) and Ma et al. (2026) consider multiple identification regimes, but their evaluations use separately constructed SCMs rather than the same underlying system. This makes it difficult to determine how much performance depends on the available identifying information. A controlled comparison should therefore hold the causal world and target effect fixed while varying only which variables the estimator observes. A causal world consists of an SCM and one complete data realization, including potential outcomes.

We introduce CAUSALIDVIEW, a multi-view benchmark that places causal effect estimators on a common evaluation. For each instance, we hold the SCM realization, query units, and target con-

[1]Code is available at `https://anonymous.4open.science/r/CausalIDView-04E5`.

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

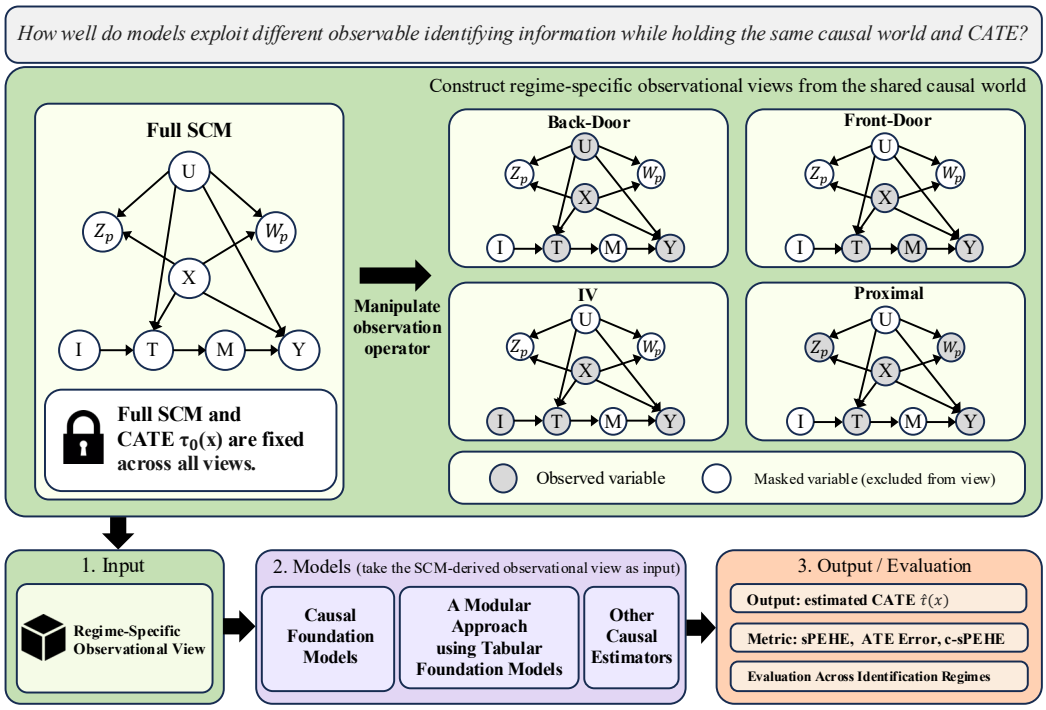


Figure 1: **Overview of CAUSALIDVIEW.** Each realized SCM is mapped to matched observational views while preserving the query units and target CATE. We can also compare the Hidden Confounding (HC) view, described in Appendix B.5.

ditional average treatment effect (CATE) fixed, while varying only what the estimator observes. As illustrated in Figure 1, each causal world is exposed through matched views supporting back-door adjustment (Pearl, 1993), front-door identification (Pearl, 2022), instrumental-variable identification (Angrist et al., 1996), and proximal identification (Miao et al., 2018), together with a hidden-confounding stress view. CAUSALIDVIEW does not assume all views are native to each model since current CFMs were not necessarily pre-trained for every regime. Instead, it characterizes how CFM inference changes under controlled shifts in identifying information and compares CFMs with other causal estimators. We complement CATE accuracy evaluations with semi-synthetic benchmarks and null-effect tests using pre-treatment outcomes from real-world randomized trials.

We further ask whether a strong predictor coupled with explicit identification can rival CFMs. To this end, we use modular estimators equipped with the pre-trained tabular foundation model TabPFN (Grinsztajn et al., 2026; Jäger et al., 2026) to estimate the observable quantities that each regime's identification strategy requires. No CFM leads in every regime, while TabPFN-based modular estimators remain competitive with CFMs. Under partial identification, TabPFN-based bound estimators achieve the lowest endpoint errors. Whether structural perturbations preserve or alter the true CATE, CFMs falter in model-specific ways. On semi-synthetic benchmarks and real-world null-effect tests, the best estimator varies by dataset.

Our main contributions are:

- **A common evaluation ground for causal effect estimators.** We introduce a paired multi-view benchmark that evaluates CFMs and causal inference methods on observational views derived from the same realized SCM, while holding the query units and target fixed.
- **Complementary diagnostics of causal estimation.** We characterize model-specific failures under structural changes, dataset-dependent accuracy and null-effect behavior, and evaluation of partial-identification bounds.
- **Effectiveness of modular identification.** We demonstrate that combining predictive tabular foundation models with explicit causal procedures yields competitive point estimates and lower bound-endpoint errors in the evaluated settings.

2 BACKGROUND

2.1 CATE AND COMMON CAUSAL CONDITIONS

CATE. We use the potential-outcomes framework (Rubin, 1974). We consider a binary treatment $T \in \{0, 1\}$, an outcome Y, and potential outcomes $Y(t)$ under $T = t$. X denotes the observed pre-treatment covariates indexing treatment-effect heterogeneity. The treatment-specific conditional mean and conditional average treatment effect (CATE) are

$$\mu_t(x) := \mathbb{E}[Y(t) \mid X = x], \qquad \tau_0(x) := \mu_1(x) - \mu_0(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{1}$$

Common causal conditions. We maintain consistency together with the standard stable-unit treatment-value conditions (Rubin, 1980). Consistency requires that the observed outcome equal the potential outcome under the received treatment, $Y_i = Y_i(T_i)$ for index i. Well-defined treatment versions require each treatment value to represent a single causally relevant intervention, and no interference requires that one unit's potential outcomes not depend on the treatment assignments of other units. The positivity and support requirements specific to each identification regime are imposed over the target covariate support and stated in Appendix B.

2.2 POINT-IDENTIFICATION REGIMES FOR CATE

We focus on four point-identification strategies, depending on which auxiliary causal variables are available.

Back-Door (BD) identification. When the confounder U is observed and (X, U) satisfies conditional exchangeability and positivity, back-door adjustment identifies the target CATE (Pearl, 2022; Rosenbaum and Rubin, 1983).

$$\tau_{\mathrm{BD}}(x) := \int \Big\{ \mathbb{E}[Y \mid T = 1, X = x, U = u] - \mathbb{E}[Y \mid T = 0, X = x, U = u] \Big\}\, dP(u \mid X = x)$$

Front-Door (FD) identification. When U is unmeasured but a mediator M satisfying the front-door criterion is observed, the total treatment effect remains identifiable (Pearl, 2022). For a continuous mediator, the corresponding conditional front-door functional is

$$\tau_{\mathrm{FD}}(x) := \int \Big\{ p(m \mid T = 1, X = x) - p(m \mid T = 0, X = x) \Big\} \times \left[\sum_{a \in \{0,1\}} \mathbb{E}[Y \mid M = m, T = a, X = x]\, P(T = a \mid X = x) \right] dm$$

Instrumental-Variable (IV) identification. When a valid binary instrument I is observed, the conditional Wald functional is

$$\tau_{\mathrm{IV}}(x) := \frac{\mathbb{E}[Y \mid I = 1, X = x] - \mathbb{E}[Y \mid I = 0, X = x]}{\mathbb{E}[T \mid I = 1, X = x] - \mathbb{E}[T \mid I = 0, X = x]}$$

The conditional Wald ratio generally identifies a conditional local average treatment effect under the standard relevance, exogeneity, exclusion, and monotonicity conditions (Angrist and Imbens, 1995; Angrist et al., 1996).

Proximal Causal (PX) identification. When U is unmeasured but the confounding proxies $(Z_{\mathrm{p}}, W_{\mathrm{p}})$ are observed, let $h_t(w, x)$ be an outcome confounding bridge satisfying $\mathbb{E}[Y \mid Z_{\mathrm{p}} = z, T = t, X = x] = \mathbb{E}[h_t(W_{\mathrm{p}}, x) \mid Z_{\mathrm{p}} = z, T = t, X = x]$. Under the corresponding proxy-independence, bridge-existence, and completeness conditions, the proximal g-formula (Miao et al., 2018; Sverdrup and Cui, 2023; Tchetgen Tchetgen et al., 2024) identifies

$$\tau_{\mathrm{PX}}(x) := \mathbb{E}[h_1(W_{\mathrm{p}}, x) - h_0(W_{\mathrm{p}}, x) \mid X = x]$$

Under the benchmark's maintained regime-specific conditions, with the shared-target construction described in Section 4.1.1, all four identification strategies recover the same target CATE:

$$\boxed{\tau_{\mathrm{BD}}(x) = \tau_{\mathrm{FD}}(x) = \tau_{\mathrm{IV}}(x) = \tau_{\mathrm{PX}}(x) = \tau_0(x)} \tag{2}$$

Appendix B formally provides the assumptions and derivations for each identification regime.

3 Related Work

Identification of Causal Effects Causal identification concerns whether a target causal estimand is uniquely determined by the observed-data distribution under maintained causal assumptions. In graphical causal models, assumptions encoded by the causal structure can render interventional or counterfactual quantities identifiable from observational data. In general, complete identification procedures characterize when causal queries can be uniquely recovered from lower levels of the causal hierarchy and provide graphical certificates when such recovery is impossible (Shpitser, 2020; Shpitser and Pearl, 2008). When multiple causal models remain observationally indistinguishable yet imply different values of the target estimand, point identification fails. Rather than treating such cases as entirely uninformative, partial identification characterizes the set of causal effects compatible with the observed distribution and maintained assumptions, often through lower and upper bounds. This perspective has a long history in econometrics and causal inference, including bounds for treatment effects under imperfect compliance (Balke and Pearl, 1997) and the broader theory of partially identified probability distributions (Manski, 2003).

Causal Foundation Models for Treatment-Effect Estimation Recent work explores amortized causal effect estimation through pre-trained foundation models. Do-PFN (Robertson et al., 2026) pre-trains a Prior-Data Fitted Network (PFN) (Müller et al., 2021) over diverse causal structures using paired observational and interventional samples to predict conditional interventional distributions. CausalPFN (Balazadeh Meresht et al., 2026) focuses on treatment-effect estimation under strong ignorability across heterogeneous data-generating processes, and CausalFM (Ma et al., 2026) explicitly separates identification from estimation through identification-aware SCM priors covering back-door, front-door, and instrumental-variable settings. Causal foundation models have largely targeted point estimates or prior-dependent posterior predictions, with recent work extending to partial identification (Bellot and Dhir, 2026).

Benchmarks for Causal Effect Estimation Existing benchmarks for causal effect estimation rely largely on synthetic or semi-synthetic data and evaluate estimators under controlled variations in data-generating processes. Widely used benchmarks such as IHDP (Hill, 2011) and the ACIC challenges (Dorie et al., 2019; Hahn et al., 2019) vary outcome functions, treatment assignment, overlap, and effect heterogeneity, while RealCause (Neal et al., 2020) improves realism by learning generative models from observational datasets. However, these benchmarks largely vary the DGP within a fixed identification regime rather than evaluating estimators across different identification regimes. This motivates benchmarks spanning diverse SCMs and identification regimes, analogous to broad generalization benchmarks such as TabArena (Erickson et al., 2026).

4 Data Curation

CausalIDView is designed to contrast the performance of estimators targeting the same causal effect as the available identifying information changes. Rather than generating a separate dataset for each identification regime, we generate each complete causal world once and derive multiple observational views from it. Within each world, the SCM realization, query units, and target CATE remain fixed, while only the variables available to the estimator change.

To complement the analysis on our controlled synthetic benchmark, we examine baseline performance on semi-synthetic and real-world datasets that reflect characteristics of empirical data. Using semi-synthetic datasets with more realistic covariate distributions, treatment assignments and outcome variability, we evaluate CATE estimation accuracy against the available ground-truth effects. In the real-world setting, where ground-truth CATE is not observed but is theoretically known to be zero, we investigate whether models introduce treatment-effect bias into their estimates.

4.1 Synthetic Dataset

4.1.1 Generating Shared Causal Worlds

Shared causal structure. Each world is generated by the causal graph in Figure 1. The observational views are projections of a shared generating process, not independently constructed regime-specific datasets. The confounder U affects treatment assignment and outcome levels, while the treatment effect on Y is mediated entirely through M. A randomized binary instrument I affects the outcome only through treatment, and the pre-treatment proxies (Z_p, W_p) provide noisy measurements of U without directly affecting treatment or outcome. All variables are generated jointly before any observational view is constructed.

Shared target CATE. To realize the common-target identity in equation 2, we restrict U to enter both potential outcomes through the same additive term $\gamma_U(X)U$, rather than modifying the treatment gain. The mediator has a fixed treatment-induced shift δ, and its contribution to the outcome has coefficient $\beta(X)$. We use shared unit-level mediator and outcome noises across treatment arms. Within a fixed world, the resulting unit-level effect and its conditional expectation satisfy

$$\begin{aligned} Y_i(1) - Y_i(0) &= \beta(X_i)\{M_i(1) - M_i(0)\} = \delta\beta(X_i), \\ \tau_0(x) &= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] = \delta\beta(x). \end{aligned} \tag{3}$$

Under this construction, the unit-level treatment effect is determined entirely by X:

$$Y_i(1) - Y_i(0) = \tau_0(X_i).$$

Therefore, the conditional average treatment effect in equation 1 is exactly the calibrated effect surface $\tau_0(x)$, and the same target is attached to every observational view of a given causal world. The restriction also removes variation in treatment gains across latent states or compliance types at fixed X, so the conditional IV complier effect coincides with this target CATE.

We then construct the remaining mechanisms of the SCM so that each observational view satisfies the assumptions required to identify this same target. Although the observed variables and identifying functionals differ across views, each point-identified view recovers the same $\tau_0(x)$. Appendix B provides the assumptions and derivations and Appendix D reports construction and identification audits.

Confounding in the observational contrast. The gain restriction does not eliminate confounding from the observed treatment-group contrast. Under this SCM,

$$\begin{aligned} &\mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x] \\ &\quad = \tau_0(x) + \underbrace{\gamma_U(x)\big\{\mathbb{E}[U \mid T = 1, X = x] - \mathbb{E}[U \mid T = 0, X = x]\big\}}_{\text{confounding bias}}. \end{aligned} \tag{4}$$

The common confounding term cancels between the two potential outcomes of the same unit, but not between treatment groups, because treatment selection changes the conditional distribution of U. The construction aligns the causal targets across regimes without eliminating confounding in either the level or shape of the observational contrast.

4.1.2 Observational Views

After generating a complete world, we derive multiple observational views by changing only which auxiliary variables are available to the estimator. The BD, FD, IV, and PX views retain U, M, I, and (Z_p, W_p), respectively, alongside (X, T, Y).

We assign a single disjoint context–query split, denoted by $\mathcal{C}$ and $\mathcal{Q}$, and reuse it across all views. For view r, let $V^{(r)}$ denote its retained auxiliary variables. The observational context is

$$\mathcal{D}_{\text{obs}}^{(r)} = \left\{ (X_i, T_i, Y_i, V_i^{(r)}) : i \in \mathcal{C} \right\}. \tag{5}$$

Excluding a variable from a view changes its availability to the estimator, not its role in generating the data. In particular, we do not resample treatment assignments, outcomes, structural parameters, or unit-level noise when constructing a new view. All views within a world are evaluated on the same query targets $\{\tau_0(X_i) : i \in \mathcal{Q}\}$. Auxiliary variables supply information for estimation without redefining the target conditioning set.

Consequently, within-world performance differences across views are not driven by different realized outcomes, query populations, or oracle effect values. This pairing compares the performance of models on the same realized causal problem while varying the information available for identification. Potential outcomes and oracle effects are reserved for evaluation and data validation.

4.2 Semi-Synthetic Dataset

We evaluate the CATE estimation performance of existing back-door baselines using ACIC 2016 (Dorie et al., 2019) and LaLonde-PSID/CPS datasets (Dehejia and Wahba, 1999; 2002; LaLonde, 1986) based on RealCause (Neal et al., 2020). ACIC 2016 is a benchmark that combines real covariates from the Collaborative Perinatal Project with researcher-designed treatment assignment and outcome-generating mechanisms. By preserving realistic covariate distributions, it enables

the evaluation of estimation performance under diverse data-generating conditions, including nonlinearity, varying degrees of overlap, and treatment-effect heterogeneity. RealCause-based LaLonde-PSID/CPS generates data from real covariates together with conditional treatment-assignment and outcome distributions learned from observational data. The original data concern the effect of participation in a job-training program on subsequent earnings and consist of the treatment group from the National Supported Work program paired with nonexperimental comparison groups drawn from PSID and CPS, respectively. RealCause fits generative models to the observational data under an assumed causal structure, providing ground-truth CATEs defined by the resulting data-generating process. We use back-door baselines as the relevant comparison methods since both benchmarks assume that confounding between treatment and outcome can be sufficiently controlled using the observed covariates. Detailed experimental settings are provided in Appendix E.3.

4.3 Real-World Dataset

Estimating causal effects in real-world settings is an important problem, yet the ground-truth counterfactuals needed to evaluate treatment-effect estimates are generally unavailable. To address this limitation, we construct a real-world evaluation setting with known ground-truth CATEs by using variables measured before treatment assignment in randomized controlled trials (RCTs) as negative control outcomes. A negative control outcome is an outcome that cannot be affected by the treatment of interest (Arnold and Ercumen, 2016). In particular, an outcome realized before treatment assignment cannot be altered by a subsequent treatment. Ashby et al. (2025) formalize this property in the potential-outcome framework as a zero treatment effect for every individual. Thus, for a pre-treatment outcome Y^{pre}, we have $Y_i^{\mathrm{pre}}(1) = Y_i^{\mathrm{pre}}(0)$, implying $\tau_0(x) = \mathbb{E}[Y^{\mathrm{pre}}(1) - Y^{\mathrm{pre}}(0) \mid X = x] = 0$ for all x.

Following this principle, we use three real-world RCT datasets. For ACTG 175 (Hammer et al., 1996), we focus on the treatment arms randomized to zidovudine (AZT) monotherapy and AZT+didanosine (ddI) combination therapy. Baseline CD4 count, measured before treatment assignment, serves as the outcome. The Pennsylvania Reemployment Bonus Demonstration (Corson et al., 1992) compares the control group with a reemployment-bonus treatment arm. We use earnings measured before randomized bonus assignment as the outcome. In the Illinois Unemployment Insurance Incentive Experiment (Woodbury and Spiegelman, 1987), the treatment contrast is between the control group and claimants offered a reemployment bonus. Base-period earnings measured before treatment assignment serve as the outcome. Appendix E.4 provides the conditions and proofs under which the null effect remains valid when treatment arms are selected in multi-treatment RCTs.

5 Experimental Design

5.1 Metrics

For query units $\{x_i\}_{i=1}^{N_q}$, we evaluate CATE estimation using

$$\mathrm{sPEHE} = \sqrt{\frac{1}{N_q}\sum_{i=1}^{N_q}\big(\widehat{\tau}(x_i) - \tau_0(x_i)\big)^2}.$$

Let $\tau_{\mathrm{ATE},\mathcal{Q}} = N_q^{-1}\sum_{i=1}^{N_q}\tau_0(x_i)$ and $\widehat{\tau}_{\mathrm{ATE},\mathcal{Q}} = N_q^{-1}\sum_{i=1}^{N_q}\widehat{\tau}(x_i)$. We additionally report the ATE error $|\widehat{\tau}_{\mathrm{ATE},\mathcal{Q}} - \tau_{\mathrm{ATE},\mathcal{Q}}|$ and centered sPEHE (c-sPEHE), which separate errors in the query-mean effect from errors in the centered CATE function with $\mathrm{sPEHE}^2 = (\widehat{\tau}_{\mathrm{ATE},\mathcal{Q}} - \tau_{\mathrm{ATE},\mathcal{Q}})^2 + \text{c-sPEHE}^2$. Appendix E.1.1 provides the detailed definitions and derivation.

Standard point-estimation metrics do not measure whether an estimator responds appropriately to structural changes. We therefore consider Stay and Move conditions. Stay changes the structural mechanism while preserving the target CATE, $\tau_0^{\mathrm{stay}}(x_i) = \tau_0(x_i)$, so E_{stay} measures spurious changes in the estimated CATE under a CATE-preserving intervention. Move instead changes the target CATE itself and E_{move} measures how accurately the corresponding change in the estimated CATE tracks the true CATE change. For a function f, let $\|f\|_{\mathcal{Q}} := \sqrt{N_q^{-1}\sum_{i=1}^{N_q} f(x_i)^2}$. Both metrics are dimensionless and lower values are better.

$$E_{\mathrm{stay}} = \frac{\|\widehat{\tau}^{\mathrm{stay}} - \widehat{\tau}\|_{\mathcal{Q}}}{\|\tau_0\|_{\mathcal{Q}}}, \qquad E_{\mathrm{move}} = \frac{\|(\widehat{\tau}^{\mathrm{move}} - \widehat{\tau}) - (\tau_0^{\mathrm{move}} - \tau_0)\|_{\mathcal{Q}}}{\|\tau_0^{\mathrm{move}} - \tau_0\|_{\mathcal{Q}}}.$$

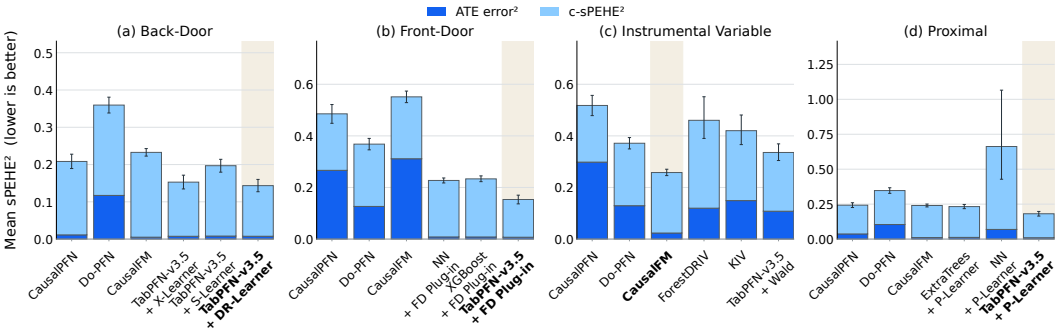


Figure 2: **Point estimation across observational views.** We compare the evaluated model configurations under (a) Back-Door, (b) Front-Door, (c) IV, and (d) Proximal views. Bars show mean sPEHE^2 across 40 SCM worlds, decomposed into mean squared query-sample ATE error (dark blue) and mean $\mathrm{c\text{-}sPEHE}^2$ (light blue). Shading marks the lowest observed mean in each panel. CausalPFN and Do-PFN receive the auxiliary variables available in each view as ordinary covariates, except for the post-treatment mediator.

5.2 Baselines

Causal foundation models. We evaluate CausalPFN, Do-PFN, and CausalFM (Balazadeh Meresht et al., 2026; Ma et al., 2026; Robertson et al., 2026). In these configurations, CausalPFN and Do-PFN use auxiliary variables as ordinary features, without IV/proximal role annotations or specialized wrappers. The front-door *X-only* variants omit M from both context and query, avoiding conditioning total-effect predictions on the factual mediator.

Modular approaches. Our modular approaches pair supervised machine-learning models with explicit causal estimation procedures for the corresponding identification regimes. The predictive models estimate the outcome, propensity, or other nuisance quantities required by each procedure, while the causal estimation step determines how predictions are combined into treatment-effect estimates.

In the back-door setting, we use S- and X-Learners (Künzel et al., 2019) and a DR-Learner (Kennedy, 2023). Unlike the S- and X-Learners, the DR-Learner constructs doubly robust pseudo-outcomes by augmenting outcome contrasts with inverse-propensity-weighted residual corrections. Under front-door identification, we replace the predictive components of the plug-in estimator, including the treatment propensity, mediator, and outcome models, while retaining the front-door averaging procedure. The IV variant estimates the outcome mean and treatment probability under each instrument value and combines their contrasts through the conditional Wald ratio. For the proximal setting, we use a P-Learner (Sverdrup and Cui, 2023), replacing the predictive models used for bridge estimation and final pseudo-outcome regression.

We instantiate these procedures with predictive models including TabPFN-v3.5 (Jäger et al., 2026), XGBoost (Chen and Guestrin, 2016), neural networks, and ExtraTrees (Geurts et al., 2006) as applicable.

Others. We additionally include regime-specific causal estimators that do not belong to either the CFM or modular predictor–procedure categories, including ForestDRIV and KIV for instrumental-variable estimation (Singh et al., 2019; Syrgkanis et al., 2019).

Implementation details of all baselines are provided in Appendix E.2.

6 Results

RQ1: How does performance vary across observational views? Figure 2 compares the evaluated model–input configurations while holding the causal world and target CATE fixed, and shows that the lowest-error configuration differs across observational views. The TabPFN-based DR-Learner and front-door plug-in attain the lowest mean sPEHE^2 in back-door and front-door, respectively, whereas CausalFM leads in IV. In the proximal view, the TabPFN-based P-Learner attains the lowest mean sPEHE^2 among the evaluated configurations. The decomposition of sPEHE^2 distinguishes errors in the query-mean effect from errors in the centered CATE function and shows that their contributions vary across model–view configurations. Centered CATE error dominates most

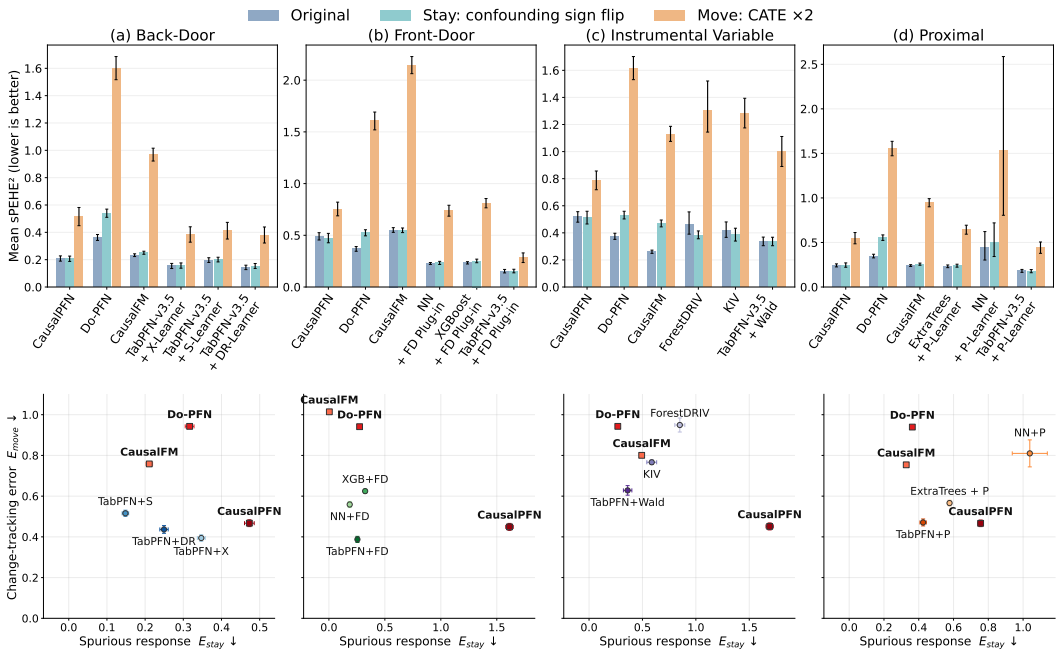


Figure 3: **Structural-response audit across observational views.** We evaluate model responses under (a) Back-Door, (b) Front-Door, (c) IV, and (d) Proximal views. **Top:** Bars show mean sPEHE^2 under the Stay and Move conditions. **Bottom:** Each point represents an evaluated model configuration, positioned by its mean spurious-response error E_{stay} on the horizontal axis and mean change-tracking error E_{move} on the vertical axis.

back-door results and several proximal results, whereas mean-effect error is substantial for particular front-door and IV configurations. Thus, a single total-error score can obscure distinct estimation failures. The back-door and FD comparisons illustrate the practical value of combining a predictive model with an explicit causal procedure, while the IV comparison shows that this advantage does not hold uniformly across regimes.

RQ2: Do estimators respond appropriately to structural changes? Figure 3 examines whether estimators respond to changes in the target causal effect rather than to incidental structural changes. Starting from $Y(t) = \mu(X) + \gamma_U(X)U + \beta(X)M(t) + \epsilon_Y$, we construct a Stay intervention by reversing the confounding contribution, yielding $\tau_0^{\mathrm{stay}}(x) = \tau_0(x)$, and a Move intervention by doubling the mediator–outcome response, yielding $\tau_0^{\mathrm{move}}(x) = 2\tau_0(x)$. An appropriate response therefore requires invariance under Stay and accurate effect-change tracking under Move. The resulting response map reveals qualitatively different CFM failures. CausalPFN shows relatively large Stay errors, indicating that its predictions can change even when the target CATE is preserved. On the other hand, Do-PFN shows consistently large Move errors with poor tracking when the CATE itself changes. CausalFM exhibits a view-dependent mixture of these behaviors. In contrast, the TabPFN-v3.5 modular estimators generally occupy a more balanced region with lower errors on both axes, suggesting that combining a strong predictor with an explicit identification rule better preserves invariance while tracking genuine effect changes. These results motivate evaluating structural responses alongside point-estimation accuracy and developing CFMs that more explicitly distinguish estimand-preserving structural shifts from those that alter the causal effect.

RQ3: What happens when confounders are hidden? As illustrated in Figure 6, synthetic and IHDP comparisons reveal how estimation error changes when a confounder is omitted. Most models show pronounced error increases. Separately, the binary partial-identification companion evaluates recovery of identified-set boundaries. TabPFN-v3.5 based models achieve the lowest endpoint RMSE in the Manski, IV and sensitivity bounds. Appendix B.5 defines the identified sets and Appendix F.1 reports the detailed results.

Table 1: **Back-door evaluation on semi-synthetic benchmarks.** Mean squared errors over 40 repeated realizations. Lower is better. For each dataset–metric column, the lowest value is shown in red and the second-lowest value in orange. Values are reported as mean ± standard deviation. LaLonde-PSID and LaLonde-CPS results are shown in units of $\times 10^6$.

Estimator	ACIC 2016			Lalonde-PSID ($\times 10^6$)			Lalonde-CPS ($\times 10^6$)		
	sPEHE^2	ATE error^2	c-sPEHE^2	sPEHE^2	ATE error^2	c-sPEHE^2	sPEHE^2	ATE error^2	c-sPEHE^2
CausalPFN	3.24 ±3.45	0.1074 ±0.2224	3.13 ±3.34	44.0 ±61.5	27.2 ±41.7	16.7 ±21.6	11.2 ±8.03	7.55 ±8.10	3.65 ±0.782
Do-PFN	31.48 ±20.02	12.19 ±9.29	19.29 ±18.65	373 ±37.2	213 ±34.8	160 ±7.20	80.6 ±4.70	43.6 ±3.81	37.0 ±1.77
CausalFM	23.56 ±17.52	5.49 ±4.84	18.08 ±17.49	423 ±7.50	212 ±6.85	211 ±4.49	85.6 ±1.31	44.4 ±0.711	41.2 ±1.42
Causal Tree	11.43 ±9.75	1.30 ±1.41	10.14 ±9.86	181 ±23.7	2.23 ±4.34	179 ±22.9	61.2 ±8.83	12.4 ±8.83	48.8 ±0.0000
TARNet	9.44 ±8.15	0.2352 ±0.3589	9.20 ±8.09	131 ±212	46.6 ±90.2	84.4 ±130	49.2 ±21.3	23.4 ±6.55	25.7 ±21.2
GRF / Causal Forest	8.30 ±9.66	0.1903 ±0.2633	8.11 ±9.68	439 ±62.8	247 ±58.6	192 ±10.5	96.3 ±19.9	47.7 ±19.6	48.6 ±1.56
CFRNet	11.43 ±9.38	0.5777 ±0.7625	10.85 ±9.16	18.4 ±33.6	3.77 ±6.94	14.7 ±28.3	18.7 ±11.5	11.0 ±8.39	7.64 ±8.62
TabPFN-v3.5 + X-Learner	1.11 ±1.56	0.0258 ±0.0417	1.09 ±1.55	456 ±295	223 ±180	233 ±127	145 ±132	86.7 ±107	58.2 ±26.7
TabPFN-v3.5 + S-Learner	1.25 ±2.03	0.0282 ±0.0290	1.23 ±2.03	415 ±116	200 ±80.2	215 ±44.3	74.6 ±9.64	37.0 ±8.13	37.6 ±4.74
TabPFN-v3.5 + DR-Learner	1.17 ±1.54	0.0222 ±0.0280	1.15 ±1.54	42.8 ±63.8	20.0 ±33.0	22.8 ±32.6	15.3 ±23.5	9.10 ±19.3	6.23 ±5.46

Table 2: **RCT negative-control null-CATE benchmarks.** Mean squared null-CATE error over 40 repeated context/query realizations. Lower is better. For each dataset–metric column, the lowest value is shown in red and the second-lowest distinct value in orange. Tied values receive the same color. Each dataset uses its primary context size n_C.

Estimator	ACTG 175 ($n_C = 768$)			Pennsylvania ($n_C = 1024$)			Illinois ($n_C = 2048$)		
	sPEHE^2	ATE error^2	c-sPEHE^2	sPEHE^2	ATE error^2	c-sPEHE^2	sPEHE^2	ATE error^2	c-sPEHE^2
CausalPFN	0.0146 ±0.0086	0.0028 ±0.0035	0.0118 ±0.0068	0.0166 ±0.0089	0.0030 ±0.0036	0.0136 ±0.0064	0.0051 ±0.0040	0.0011 ±0.0019	0.0039 ±0.0031
Do-PFN	0.0203 ±0.0113	0.0113 ±0.0094	0.0090 ±0.0062	0.0177 ±0.0080	0.0099 ±0.0052	0.0077 ±0.0037	0.1249 ±0.0220	0.0960 ±0.0210	0.0289 ±0.0069
CausalFM	0.0053 ±0.0028	0.0009 ±0.0010	0.0045 ±0.0029	0.0023 ±0.0011	0.0016 ±0.0010	0.0007 ±0.0003	0.0016 ±0.0008	0.0004 ±0.0005	0.0011 ±0.0005
Causal Tree	0.0416 ±0.0485	0.0022 ±0.0028	0.0394 ±0.0484	0.0173 ±0.0352	0.0042 ±0.0058	0.0131 ±0.0308	0.0082 ±0.0197	0.0017 ±0.0026	0.0065 ±0.0193
TARNet	0.0692 ±0.0308	0.0037 ±0.0038	0.0655 ±0.0306	0.0993 ±0.0345	0.0040 ±0.0053	0.0953 ±0.0324	0.0239 ±0.0142	0.0016 ±0.0022	0.0223 ±0.0137
GRF / Causal Forest	0.0159 ±0.0073	0.0024 ±0.0033	0.0135 ±0.0059	0.0110 ±0.0082	0.0024 ±0.0034	0.0086 ±0.0054	0.0231 ±0.0078	0.0014 ±0.0024	0.0217 ±0.0068
CFRNet	0.0660 ±0.0429	0.0030 ±0.0038	0.0630 ±0.0420	0.1094 ±0.0645	0.0044 ±0.0100	0.1049 ±0.0630	0.0167 ±0.0134	0.0012 ±0.0015	0.0155 ±0.0127
TabPFN-v3.5 + X-Learner	0.0232 ±0.0141	0.0025 ±0.0034	0.0207 ±0.0131	0.0237 ±0.0139	0.0030 ±0.0036	0.0207 ±0.0117	0.0213 ±0.0117	0.0011 ±0.0017	0.0202 ±0.0113
TabPFN-v3.5 + S-Learner	0.0008 ±0.0010	0.0005 ±0.0008	0.0003 ±0.0004	0.0015 ±0.0009	0.0005 ±0.0005	0.0010 ±0.0005	0.0062 ±0.0070	0.0004 ±0.0009	0.0058 ±0.0065
TabPFN-v3.5 + DR-Learner	0.0159 ±0.0133	0.0020 ±0.0027	0.0139 ±0.0126	0.0247 ±0.0170	0.0024 ±0.0030	0.0223 ±0.0157	0.0153 ±0.0109	0.0010 ±0.0017	0.0143 ±0.0103

RQ4: Are estimator rankings consistent across semi-synthetic and real-world data? Tables 1 and 2 evaluate CATE accuracy on semi-synthetic data and null-effect behavior on real-world pre-treatment outcomes, respectively. Both evaluations show dataset-dependent performance, with no consistently best estimator. By mean sPEHE^2, the best TabPFN-v3.5 meta-learner is the X-Learner on ACIC 2016, the DR-Learner on both LaLonde benchmarks, and the S-Learner on all three real-world null-effect tests. Among all evaluated methods, TabPFN-based estimators attain the lowest error on ACIC 2016, ACTG 175, and Pennsylvania, whereas CFRNet, CausalPFN, and CausalFM lead on LaLonde-PSID, LaLonde-CPS, and Illinois, respectively. The TabPFN-based DR-Learner ranks second overall on both LaLonde benchmarks. These results support modular estimation as a competitive alternative to CFMs, while highlighting that the choice of causal estimator remains consequential even with the predictive backbone fixed.

7 Conclusion

We introduce CausalIDView to evaluate how causal estimators respond to changes in identifying information. By holding the realized SCM, query units, and target CATE fixed across observational views, the benchmark enables controlled comparisons that are difficult to obtain from separately constructed datasets. Changes in the rankings of the evaluated CFMs highlight the value of these matched comparisons across identification regimes. The evaluation also reveals that TabPFN-based estimators achieve the best point-estimation performance in the back-door, front-door, and proximal views. These matched comparisons further reveal model-specific failure patterns as the available identifying information changes. Such findings motivate CFMs that better account for the identifying roles of observed variables. Across the semi-synthetic and real-world evaluations, no single model consistently performs best, highlighting the need for more reliable generalization across diverse data-generating settings. Future work should extend CausalIDView to broader causal settings and further investigate TabPFN-based estimator designs across identification regimes.

AI USE STATEMENT

We used large language models (LLMs) during manuscript preparation to assist with language editing and to provide feedback on clarity, logical flow, and presentation. We also used LLMs for brainstorming and critical feedback on experimental design and for discussing possible interpretations of experimental results. All research questions, methodological and experimental decisions, analyses, and conclusions were determined and verified by the authors. The authors reviewed and revised all AI-assisted content and take full responsibility for the accuracy and integrity of the work.

REFERENCES

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

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Benjamin F Arnold and Ayse Ercumen. Negative control outcomes: a tool to detect bias in randomized trials. *Jama*, 316(24):2597–2598, 2016.

Ethan Ashby, Bo Zhang, Genevieve G Fouda, Youyi Fong, and Holly Janes. Negative control outcome adjustment in early-phase randomized trials: Estimating vaccine effects on immune responses in hiv exposed uninfected infants. *Statistics in medicine*, 44(13-14):e70142, 2025.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709.

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.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176, 1997.

Alexis Bellot and Anish Dhir. Foundation models for partial causal identification. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026. URL https://openreview.net/forum?id=jCbehzZBsk.

Björn Bokelmann and Stefan Lessmann. Improving uplift model evaluation on randomized controlled trial data. *European Journal of Operational Research*, 313(2):691–707, 2024.

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.

Lucius EJ Bynum, Rajesh Ranganath, and Kyunghyun Cho. Computational identifiability. *arXiv preprint arXiv:2606.19361*, 2026.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Xiaohong Chen, Timothy M Christensen, and Elie Tamer. Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018, 2018.

Walter Corson, Paul Decker, Shari Dunstan, and Stuart Kerachsky. Pennsylvania reemployment bonus demonstration: Final report. Unemployment Insurance Occasional Paper 92-1, U.S. Department of Labor, Employment and Training Administration, Unemployment Insurance Service, Washington, DC, 1992.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela Van Der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

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.

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

Dennis Frauen, Fergus Imrie, Alicia Curth, Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. A neural framework for generalized causal sensitivity analysis. In *International Conference on Learning Representations*, volume 2024, pages 29413–29444, 2024.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

AmirEmad Ghassami, Andrew Ying, Ilya Shpitser, and Eric Tchetgen Tchetgen. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *International conference on artificial intelligence and statistics*, pages 7210–7239. PMLR, 2022.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Anna Guo, David Benkeser, and Razieh Nabi. Flexible nonparametric inference for causal effects under the front-door model. *arXiv preprint arXiv:2312.10234*, 2023.

Shantanu Gupta, Zachary C Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. In *Uncertainty in Artificial Intelligence*, pages 982–991. PMLR, 2021.

P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.

Junha Ham, Deokgyu Kim, Doeun Kim, Serjin Kim, and Sanghack Lee. Causal foundation models perform better without post-treatment variables. In *2nd ICML Workshop on Foundation Models for Structured Data*, 2026.

Scott M Hammer, David A Katzenstein, Michael D Hughes, Holly Gundacker, Robert T Schooley, Richard H Haubrich, W Keith Henry, Michael M Lederman, John P Phair, Manette Niu, et al. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335 (15):1081–1090, 1996.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL `https://doi.org/10.1198/jcgs.2010.08162`.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

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.

Benjamin Jäger, Nick Erickson, Léo Grinsztajn, Felix Birkel, Klemens Flöge, Oscar Key, Kürşat Kaya, Jonas Kübler, Adèle Frankel, Tobias Schröder, et al. Tabpfn-3.5: Technical report. *arXiv preprint arXiv:2609.17895*, 2026.

Emil Javurek, Dennis Frauen, Marie Brockschmidt, Jonas Schweisthal, and Stefan Feuerriegel. Amortizing causal sensitivity analysis via prior data-fitted networks. *arXiv preprint arXiv:2605.10590*, 2026.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021.

Yonghan Jung. Debiased front-door learners for heterogeneous effects. In *International Conference on Learning Representations*, volume 2026, pages 128840–128867, 2026.

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

Stanislav R Kirpichenko, Andrei Vladimirovich Konstantinov, and Lev Vladimirovich Utkin. Focat: foundation model for estimating the conditional average treatment effect. *Doklady Mathematics*, 112(1):279–287, 2025. doi: 10.1134/S1064562425700280.

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.

Sergei O Kurashkin, Vadim S Tynchenko, Aleksei S Borodulin, Vladimir A Nelyub, Nikolay O Kalutsky, and Tee Connie. The current generation of tabular foundation models: A critical review. *Machine Learning and Knowledge Extraction*, 8(8):244, 2026.

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

Adi Lin. *Causal Inference Using Bayesian Deep Learning*. University of Technology Sydney (Australia), 2020.

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, pages 79065–79098, 2026.

Charles F Manski. *Partial identification of probability distributions*. Springer, 2003.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Francisco Mourao, David Hajage, Daria Bystrova, Bertrand Bouvarel, Nathanaël Lapidus, Fabrice Carrat, and Benjamin Glemain. Prior-data fitted networks for causal inference: a simulation study with real-world scenarios. *arXiv preprint arXiv:2603.15928*, 2026.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. Realcause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

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

Miruna Oprescu, Jacob Dorn, Marah Ghoummaid, Andrew Jesson, Nathan Kallus, and Uri Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In *International Conference on Machine Learning*, pages 26599–26618. PMLR, 2023.

Judea Pearl. [bayesian analysis in expert systems]: Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269, 1993.

Judea Pearl. Causal diagrams for empirical research (with discussions). In *Probabilistic and causal inference: The works of Judea Pearl*, pages 255–316. 2022.

Lennart Purucker, Andrej Tschalzev, Nick Erickson, Gioia Blayer, David Holzmüller, Alan Arazi, Alexander Pfefferle, Mustafa Tajjar, Gaël Varoquaux, and Frank Hutter. Beyond iid: How general are tabular foundation models, really? *arXiv preprint arXiv:2606.30410*, 2026.

Hongxiang Qiu, Marco Carone, Ekaterina Sadikova, Maria Petukhova, Ronald C Kessler, and Alex Luedtke. Optimal individualized decision rules using instrumental variable methods. *Journal of the American Statistical Association*, 116(533):174–191, 2021.

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.

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

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

Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.

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

Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

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

Ilya Shpitser. Identification in causal models with hidden variables. *Journal de la société française de statistique*, 161(1):91–119, 2020.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(64):1941–1979, 2008. URL `http://jmlr.org/papers/v9/shpitser08a.html`.

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

Erik Sverdrup and Yifan Cui. Proximal causal learning of conditional average treatment effects. In *International Conference on Machine Learning*, pages 33285–33298. PMLR, 2023.

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.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

Dennis Thumm, Billy Tim Anthony, and Ying Chen. Dotime: A synthetic benchmark generator for interventional and counterfactual time series. *arXiv preprint arXiv:2607.27263*, 2026.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *The annals of mathematical statistics*, 11(3):284–300, 1940.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550, 2018.

Stephen A Woodbury and Robert G Spiegelman. Bonuses to workers and employers to reduce unemployment: Randomized trials in illinois. *The American Economic Review*, pages 513–530, 1987.

A Formal Problem Setup

This appendix clarifies the distinction between the generating causal model, its realized data, and the information available to an estimator. It also specifies the identification terminology and the scope of the synthetic benchmark.

A.1 Identification Terminology

Identified set for a general estimand. Let O denote an observed random vector with population distribution P_{obs}, and let $\mathcal{A}$ denote a collection of maintained causal assumptions. Let $\mathcal{M}(\mathcal{A})$ denote the class of full-data causal laws satisfying $\mathcal{A}$, and consider a generic causal estimand

$$\theta(Q) \in \Theta,$$

where Q denotes a candidate full-data causal law and Θ is the corresponding parameter space. Following the identified-set formulation (Manski, 2003), the identified set is

$$\mathcal{I}_\theta\big(P_{\text{obs}}; \mathcal{A}\big) := \{\theta(Q) \in \Theta : Q \in \mathcal{M}(\mathcal{A}),\ Q_O = P_{\text{obs}}\}\,, \tag{6}$$

where Q_O denotes the observed-data distribution induced by Q. Hence, $\mathcal{I}_\theta(P_{\text{obs}}; \mathcal{A})$ contains all values of the target estimand that are compatible with both the observed-data distribution and the maintained assumptions.

Point, partial, and no informative identification. Let Q_0 denote the true full-data causal law and $\theta_0 := \theta(Q_0)$. Assuming $Q_0 \in \mathcal{M}(\mathcal{A})$, we distinguish

$$\begin{aligned} \mathcal{I}_\theta(P_{\text{obs}}; \mathcal{A}) &= \{\theta_0\}, && \text{point identification,} \\ \{\theta_0\} \subsetneq \mathcal{I}_\theta(P_{\text{obs}}; \mathcal{A}) &\subsetneq \Theta, && \text{partial identification,} \\ \mathcal{I}_\theta(P_{\text{obs}}; \mathcal{A}) &= \Theta, && \text{no informative identification.} \end{aligned}$$

We use non-point identification for the broader case in which the identified set is not a singleton. It therefore includes informative partial identification and should not be equated with the absence of any information about the target. The final case above is the limiting situation in which the observed distribution and maintained assumptions rule out no value in the original parameter space. Identification status is consequently always relative to a specified target, observed-data distribution, and set of maintained assumptions, as summarized in Figure 4.

We now specialize the general definition to the target considered in this work. For a fixed covariate value x, define under a candidate causal law Q

$$\tau_Q(x) := \mathbb{E}_Q[Y(1) - Y(0) \mid X = x]\,.$$

Taking $\theta(Q) = \tau_Q(x)$ in Equation 6, the identified set for the target CATE in regime r is therefore

$$\Theta_I^{(r)}(x) := \mathcal{I}_{\tau(x)}\big(P_{\text{obs}}^{(r)}; \mathcal{A}_r\big) = \Big\{\tau_Q(x) \in \Theta : Q \in \mathcal{M}(\mathcal{A}_r),\ Q_{O^{(r)}} = P_{\text{obs}}^{(r)}\Big\}\,.$$

The true target is $\theta_0 = \tau_0(x)$. Across the observational regimes in CausalIDView, $\tau_0(x)$ is held fixed, while $O^{(r)}$, $P_{\text{obs}}^{(r)}$, and the assumptions $\mathcal{A}_r$ may differ.

When $\Theta_I^{(r)}(x) = \{\tau_0(x)\}$, the target is point identified at x. Equivalently, there exists a regime-specific observed-data functional $\Phi_x^{(r)}$ such that $\tau_0(x) = \Phi_x^{(r)}\Big(P_{\text{obs}}^{(r)}\Big)$ under $\mathcal{A}_r$. Different observational views may identify the same target through different observed-data functionals. Conversely, the absence of a point-identification claim for the HC view does not by itself imply $\Theta_I^{(\text{HC})}(x) = \Theta$. The assumptions and identifying functionals associated with each regime are detailed in Appendix B.

Identification versus statistical estimation. Identification concerns whether the target is uniquely determined by the population observed-data law under maintained causal assumptions (Shpitser, 2020; Shpitser and Pearl, 2008). Statistical estimation instead concerns recovering an identified functional, or an identified set, from finite data. A target may therefore be point identified while remaining statistically difficult to estimate. Conversely, a point prediction may happen to be close to the oracle effect even when point identification has not been established from the available observational view.

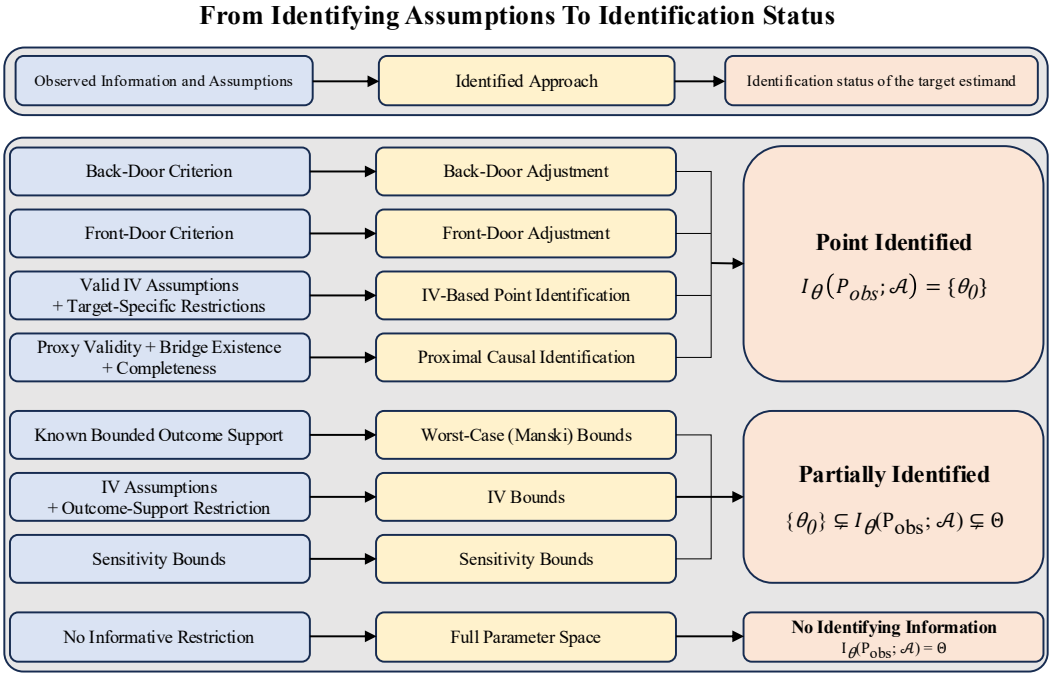


Figure 4: **Taxonomy of identification.** The three identification status categories are defined by the identified set. The assumptions and identification approaches shown are representative examples, including those relevant to our benchmark, rather than an exhaustive enumeration of possible identification regimes.

A.2 Benchmark Scope

The synthetic generator does not include longitudinal or time-varying treatments, dynamic treatment regimes, interference between units, continuous or multivalued treatments, censoring or survival outcomes, or multidimensional latent confounders.

B Identification Regimes

We expand the identification arguments in Section 2 and show how each observational view supports the same target CATE. Structural arguments are distinguished from finite-sample diagnostics, which are reported separately in Appendix D.

B.1 Back-Door

Assumptions. The back-door view observes (X, U, T, Y), with (X, U) measured before treatment. In addition to the common consistency conditions, we require:

1. *Conditional exchangeability:* $\{Y(0), Y(1)\} \perp T \mid X, U$.
2. *Positivity:* $0 < P(T = 1 \mid X = x, U = u) < 1$ for every (x, u) in the target support.

Recovering the CATE. Let $m_t(x, u) := \mathbb{E}[Y \mid T = t, X = x, U = u]$. Exchangeability and consistency allow the potential-outcome mean to be computed by averaging the observed outcome regression over the same confounder distribution in both arms (Pearl, 2022):

$$\mu_t^{\mathrm{BD}}(x) = \mathbb{E}[Y(t) \mid X = x] = \sum_{u \in \{-1,+1\}} m_t(x, u) P(U = u \mid X = x).$$

For the SCM outcome law, $m_t(x, u) = \mu(x) + t\tau_0(x) + \gamma_U(x)u$. The prognostic and confounding terms therefore cancel within each (x, u) stratum, giving

$$\begin{aligned} \tau_{\mathrm{BD}}(x) &= \sum_u \{m_1(x, u) - m_0(x, u)\} P(U = u \mid X = x) \\ &= \sum_u \tau_0(x) P(U = u \mid X = x) = \tau_0(x). \end{aligned}$$

Table 3: Observational contexts constructed from each complete causal world. All views share the same context and query units, treatment assignments, factual outcomes, and oracle effects.

View	Observed context variables
Back-door (BD)	X, U, T, Y
Front-door (FD)	X, M, T, Y
Instrumental variable (IV)	X, I, T, Y
Proximal (PX)	$X, Z_{\mathrm{p}}, W_{\mathrm{p}}, T, Y$
Hidden confounding (HC)	X, T, Y

B.2 Front-Door

Assumptions. The front-door view observes (X, T, M, Y) and hides U. Conditional on X, we require (Pearl, 2022):

1. *Complete mediation:* M intercepts every directed path from T to Y.
2. *Treatment–mediator exchangeability:* no back-door path between T and M remains open given X.
3. *Mediator–outcome adjustment:* all back-door paths between M and Y are blocked by (T, X).
4. *Positivity and support:* both treatment arms have positive probability given X, and mediator values used in the functional are supported in both arms.

Recovering the CATE. Write $Q(m, s, x) := \mathbb{E}[Y \mid M = m, T = s, X = x]$ and $p_s(x) := P(T = s \mid X = x)$. In the SCM, $M \perp U \mid T, X$, so

$$Q(m, s, x) = \mu(x) + \beta(x)m + \gamma_U(x)\mathbb{E}[U \mid T = s, X = x].$$

First, average this regression over the observed treatment distribution:

$$\begin{aligned} G(m, x) &:= \sum_{s \in \{0,1\}} Q(m, s, x)p_s(x) \\ &= \mu(x) + \beta(x)m + \gamma_U(x) \sum_s \mathbb{E}[U \mid T = s, X = x]p_s(x) \\ &= \mu(x) + \beta(x)m, \end{aligned}$$

where the last equality uses $\sum_s \mathbb{E}[U \mid T = s, X = x]p_s(x) = \mathbb{E}[U \mid X = x] = 0$. This step removes the treatment-specific latent composition from the outcome regression. Next, integrate over the mediator distribution under each treatment arm:

$$\begin{aligned} \mu_t^{\mathrm{FD}}(x) &= \int G(m, x)p(m \mid T = t, X = x)\, dm \\ &= \mu(x) + \beta(x)\mathbb{E}[M \mid T = t, X = x] = \mu(x) + \beta(x)\delta t. \end{aligned}$$

Consequently,

$$\tau_{\mathrm{FD}}(x) = \mu_1^{\mathrm{FD}}(x) - \mu_0^{\mathrm{FD}}(x) = \beta(x)\delta = \tau_0(x).$$

The total effect thus requires integrating over the two mediator laws, not holding the query unit's factual mediator fixed.

B.3 Instrumental Variable

Assumptions. The IV view observes (X, I, T, Y). Let $T^I(j)$ denote potential treatment under instrument value j. We use the following conditions:

1. *Exogeneity:* $I \perp \{Y(0), Y(1), T^I(0), T^I(1)\} \mid X$.
2. *Exclusion:* I affects Y only through T.
3. *Instrument positivity and relevance:* $0 < P(I = 1 \mid X = x) < 1$ and $\mathbb{E}[T \mid I = 1, X = x] \neq \mathbb{E}[T \mid I = 0, X = x]$.
4. *Monotonicity:* $T^I(1) \geq T^I(0)$ almost surely.
5. *Within-X gain invariance:* the benchmark additionally imposes $Y(1) - Y(0) = \tau_0(X)$, so gains do not vary with latent state or compliance type at fixed X.

Recovering the CATE. The first four conditions give the conditional Wald ratio a complier-effect interpretation and the final restriction makes that effect equal to the CATE (Angrist et al., 1996). To see this directly, the SCM has $I \perp U \mid X$ and $\mathbb{E}[U \mid X = x] = 0$, hence

$$\begin{aligned}\mathbb{E}[Y \mid I = j, X = x] &= \mu(x) + \tau_0(x)\mathbb{E}[T \mid I = j, X = x] + \gamma_U(x)\mathbb{E}[U \mid I = j, X = x] \\ &= \mu(x) + \tau_0(x)\mathbb{E}[T \mid I = j, X = x].\end{aligned}$$

Define the instrument-induced differences

$$\begin{aligned}\Delta_Y(x) &:= \mathbb{E}[Y \mid I = 1, X = x] - \mathbb{E}[Y \mid I = 0, X = x], \\ \Delta_T(x) &:= \mathbb{E}[T \mid I = 1, X = x] - \mathbb{E}[T \mid I = 0, X = x].\end{aligned}$$

Subtracting the two reduced-form means cancels $\mu(x)$ and yields

$$\Delta_Y(x) = \tau_0(x)\Delta_T(x), \qquad \tau_{\mathrm{IV}}(x) = \frac{\Delta_Y(x)}{\Delta_T(x)} = \tau_0(x).$$

Thus, dividing the instrument-induced outcome change by the instrument-induced treatment change recovers the shared CATE under the additional gain restriction.

B.4 Proximal Identification

Assumptions. The proximal view observes (X, Z_p, W_p, T, Y), where Z_p and W_p are the treatment- and outcome-inducing proxies, respectively. We require (Miao et al., 2018; Tchetgen Tchetgen et al., 2024):

1. *Latent exchangeability:* $Y(t) \perp T \mid U, X$.
2. *Proxy restrictions:* $Y \perp Z_p \mid T, U, X$ and $W_p \perp (T, Z_p) \mid U, X$.
3. *Bridge existence:* an integrable outcome bridge h_t satisfies

$$\mathbb{E}[Y \mid Z_p = z, T = t, X = x] = \mathbb{E}[h_t(W_p, x) \mid Z_p = z, T = t, X = x].$$

4. *Completeness:* for $V = U$ and $V = W_p$, and every square-integrable g,

$$\mathbb{E}[g(V) \mid Z_p, T = t, X = x] = 0 \quad \Longrightarrow \quad g(V) = 0$$

   almost surely under the corresponding conditional law. In this binary construction, these are full-rank conditions.
5. *Support:* both latent strata have positive probability given X, $0 < P(T = 1 \mid X, U) < 1$, and the required proxy cells have positive conditional probability.

Recovering the CATE. For binary proxies, define the observed matrix and mean vector by

$$A_t(x)_{z,w} := P(W_p = w \mid Z_p = z, T = t, X = x), \qquad y_t(x)_z := \mathbb{E}[Y \mid Z_p = z, T = t, X = x]$$

for $z, w \in \{0, 1\}$. The bridge equation is a two-equation linear system,

$$A_t(x)\boldsymbol{h}_t(x) = \boldsymbol{y}_t(x), \qquad \boldsymbol{h}_t(x) = A_t(x)^{-1}\boldsymbol{y}_t(x),$$

where $\boldsymbol{h}_t(x) = (h_t(0, x), h_t(1, x))^\top$. After solving this system for each arm, average the bridge values over $W_p \mid X = x$:

$$\mu_t^{\mathrm{PX}}(x) = \sum_{w \in \{0,1\}} h_t(w, x) P(W_p = w \mid X = x).$$

For the SCM, write $p_W^{\pm}(x) := P(W_p = 1 \mid U = \pm 1, X = x)$. An explicit bridge is

$$h_t(w, x) = \mu(x) + t\tau_0(x) - \gamma_U(x) + \frac{2\gamma_U(x)}{p_W^+(x) - p_W^-(x)}\{w - p_W^-(x)\}.$$

Only the term $t\tau_0(x)$ changes between treatment arms, so $h_1(w, x) - h_0(w, x) = \tau_0(x)$. Therefore,

$$\tau_{\mathrm{PX}}(x) = \sum_w \{h_1(w, x) - h_0(w, x)\} P(W_p = w \mid X = x) = \tau_0(x).$$

B.5 Hidden Confounding and Partial Identification

The Hidden-Confounding (HC) view observes only (X, T, Y). For the main continuous-outcome SCM, its outcome law and the zero conditional mean of its noise give

$$\mathbb{E}[Y \mid T = t, X = x] = \mu(x) + t\tau_0(x) + \gamma_U(x)\mathbb{E}[U \mid T = t, X = x].$$

Subtracting the observed treatment-arm means therefore gives

$$\begin{aligned} &\underbrace{\mathbb{E}[Y \mid T = 1, X = x] - \mathbb{E}[Y \mid T = 0, X = x]}_{\text{observed group contrast}} \\ = \underbrace{\tau_0(x)}_{\text{target CATE}} &+ \underbrace{\gamma_U(x)\big\{\mathbb{E}[U \mid T = 1, X = x] - \mathbb{E}[U \mid T = 0, X = x]\big\}}_{\text{remaining confounding bias } b_{\mathrm{HC}}(x)}. \end{aligned} \tag{7}$$

The term $\gamma_U(X)U$ does not cancel when comparing different treatment groups, because treatment selection changes their conditional distributions of U. Confounding bias remains even though U does not modify the treatment gain. Since this bias can vary with x, it can distort the shape of the observational contrast as well as its mean. Equation 7 describes the observational contrast, not necessarily the output of every evaluated model. We score HC point predictions against the oracle $\tau_0(x)$ as a hidden-confounding stress test, without a point-identification claim.

We distinguish the three partial-identification approaches in Figure 4: Worst-Case (Manski) bounds, IV bounds, and sensitivity bounds. These evaluations use a separate partial-identification companion SCM described in Appendix C.5, distinct from the SCM used for the point-estimation benchmark. The Manski and sensitivity views observe (X, T, Y), whereas the IV view additionally observes I. The three views share the same realized (X, T, Y), context–query split, query units, and target CATE, enabling paired comparisons across the partial-identification settings while varying the available identifying information and assumptions. Following Appendix A.1, each identified set is defined relative to its observed distribution and maintained assumptions.

For finite-sample estimation, we keep the corresponding identifying functional fixed and estimate the required observable quantities from the context data. The Manski and IV estimators first estimate the relevant cell probabilities and then apply the closed-form Manski bounds or the monotone-IV linear program, respectively. If estimated IV cells are infeasible, they are projected onto the monotone-IV response-type polytope before endpoint computation. For sensitivity bounds, we evaluate CSA-PFN (MSM) (Javurek et al., 2026), B-Learner (Oprescu et al., 2023), and NeuralCSA (Frauen et al., 2024) against a common numerical MSM reference. Detailed implementations of the partial-identification estimators are provided in Appendix E.2.4. Throughout, we distinguish the population identified set, its numerical reference, and the estimated interval $[\widehat{L}_m(x), \widehat{U}_m(x)]$. Target containment denotes whether $\tau_0(x)$ lies in an estimated set, rather than confidence-interval coverage.

B.5.1 Worst-case (Manski) bounds

For binary potential outcomes $Y(t) \in \{0, 1\}$, define the four observed cells $q_{ty}(x) := P(T = t, Y = y \mid X = x)$. Under consistency alone, factual outcomes constrain the observed arm while missing counterfactuals remain unrestricted within the binary support. The resulting sharp CATE set is (Manski, 2003)

$$\begin{aligned} \Theta_I^{(\mathrm{M})}(x) &= [L_{\mathrm{M}}(x), U_{\mathrm{M}}(x)], \\ L_{\mathrm{M}}(x) &= -q_{10}(x) - q_{01}(x), \\ U_{\mathrm{M}}(x) &= q_{11}(x) + q_{00}(x). \end{aligned} \tag{8}$$

The width is exactly one because the four cells sum to one. These bounds use the potential-outcome support $\{0, 1\}$.

B.5.2 Monotone-IV bounds

Let $q_{jty}(x) := P(T = t, Y = y \mid I = j, X = x)$ for the matched IV view. We maintain consistency, conditional instrument exogeneity and exclusion, instrument positivity and relevance, and treatment monotonicity $T^I(1) \geq T^I(0)$. Unlike the point-identification argument in Appendix B.3, this identified-set model does not impose within-X gain invariance or equality of mean effects across compliance types. The conditional response-type law is represented by 12 probabilities: three treatment types crossed with all four binary outcome types,

$$r = (d_0^r, d_1^r, y_0^r, y_1^r) \in \mathcal{R} := \{(0, 0), (0, 1), (1, 1)\} \times \{0, 1\}^2.$$

Here d_j^r is the treatment under $I = j$ and y_t^r is the outcome under $T = t$. Let $q^{\mathrm{IV}}(x)$ stack the eight conditional cells and define

$$A_{(j,t,y),r} := \mathbf{1}\{d_j^r = t,\ y_t^r = y\},$$
$$\Pi_x := \{\pi \in \mathbb{R}^{12} :\ \pi \geq 0,\ \mathbf{1}^\top \pi = 1,\ A\pi = q^{\mathrm{IV}}(x)\}.$$

The same response-type probabilities reproduce both instrument arms; this encodes conditional exogeneity, while the mapping through d_j^r encodes exclusion. The sharp set follows from two linear programs (Balke and Pearl, 1997):

$$\begin{aligned} \Theta_I^{(\mathrm{IVb})}(x) &= [L_{\mathrm{IV}}(x), U_{\mathrm{IV}}(x)], \\ L_{\mathrm{IV}}(x) &= \min_{\pi \in \Pi_x} \sum_{r \in \mathcal{R}} (y_1^r - y_0^r)\pi_r, \\ U_{\mathrm{IV}}(x) &= \max_{\pi \in \Pi_x} \sum_{r \in \mathcal{R}} (y_1^r - y_0^r)\pi_r. \end{aligned} \tag{9}$$

No outcome monotonicity or generator-specific gain restriction is added to these programs. For compatible population distributions, $\Theta_I^{(\mathrm{IVb})}(x) \subseteq \Theta_I^{(\mathrm{M})}(x)$ on the matched companion views.

B.5.3 Sensitivity bounds

Let $e(x,u) := P(T = 1 \mid X = x, U = u)$ and $e(x) := P(T = 1 \mid X = x)$. The marginal sensitivity model (MSM) maintains, for a specified $1 \leq \Gamma < \infty$,

$$\Gamma^{-1} \leq \frac{e(x,u)/\{1 - e(x,u)\}}{e(x)/\{1 - e(x)\}} \leq \Gamma. \tag{10}$$

The compatible causal laws satisfy consistency, positivity, and latent exchangeability $\{Y(0), Y(1)\} \perp T \mid X, U$ and reproduce the observed (X, T, Y) distribution. The latent variable in this nonparametric compatible-law class is otherwise unrestricted; the generator's binary latent support and within-X gain invariance are not imposed (Oprescu et al., 2023).

The MSM identified set. We specialize the CATE identified set in Appendix A.1 to these assumptions. For $t \in \{0,1\}$, write $p_t(x) := P(T = t \mid X = x)$ and let $F_t(\cdot \mid x)$ denote the observed law of $Y \mid T = t, X = x$, with a finite absolute first moment. The odds restriction induces the normalized reweighting class

$$\begin{aligned} \ell_t^\Gamma(x) &:= p_t(x) + \{1 - p_t(x)\}/\Gamma, \\ u_t^\Gamma(x) &:= p_t(x) + \Gamma\{1 - p_t(x)\}, \\ \mathcal{W}_t^\Gamma(x) &:= \left\{ w :\ \ell_t^\Gamma(x) \leq w(y) \leq u_t^\Gamma(x) \quad F_t\text{-a.s.}, \quad \int w(y)\, dF_t(y \mid x) = 1 \right\}. \end{aligned}$$

The extremal potential-outcome means are

$$\underline{\mu}_t^\Gamma(x) := \inf_{w \in \mathcal{W}_t^\Gamma(x)} \int y\, w(y)\, dF_t(y \mid x),$$
$$\overline{\mu}_t^\Gamma(x) := \sup_{w \in \mathcal{W}_t^\Gamma(x)} \int y\, w(y)\, dF_t(y \mid x).$$

The sharp CATE set under this model is (Oprescu et al., 2023)

$$\begin{aligned} \Theta_I^{(\mathrm{HC})}(x;\Gamma) &= [L_\Gamma(x), U_\Gamma(x)], \\ L_\Gamma(x) &= \underline{\mu}_1^\Gamma(x) - \overline{\mu}_0^\Gamma(x), \\ U_\Gamma(x) &= \overline{\mu}_1^\Gamma(x) - \underline{\mu}_0^\Gamma(x). \end{aligned} \tag{11}$$

Normalization makes each reweighted outcome law a probability law. At $\Gamma = 1$, all weights equal one and the set reduces to the observed treatment-group mean contrast. Larger Γ allows greater hidden selection and yields nested, nonshrinking population sets. The set contains $\tau_0(x)$ when the true law satisfies the maintained assumptions at the specified Γ.

Algorithm 1 SCM generation and multi-view construction

Input: World seed s_w; settings in Table 4

Output: Full world $\mathcal{W}_w$, paired views $\{\mathcal{O}_w^{(r)}\}_r$, and shared split $(\mathcal{C}_w, \mathcal{Q}_w)$

1: Select the treatment-effect family using s_w and sample its parameters and the remaining SCM parameters.
2: Calibrate $\tau_w(x)$ and the treatment intercept $b_{T,w}$ on an auxiliary covariate sample; set $\beta(x) \leftarrow \tau_w(x)/\delta$.
3: **for** $i = 1, \ldots, N_c + N_q$ **do**
4: Draw root variables (X_i, U_i, I_i) and independent exogenous noises for treatment, proxies, mediator, and outcome.
5: Generate $T_i^I(0), T_i^I(1)$ using the same treatment threshold; set $T_i \leftarrow T_i^I(I_i)$.
6: Generate $(Z_{p,i}, W_{p,i})$ conditionally independently given (X_i, U_i).
7: Construct paired potentials using the same $\epsilon_{M,i}$ and $\epsilon_{Y,i}$ in both arms:
$M_i(t) \leftarrow \delta t + \sigma_M \epsilon_{M,i}$,
$Y_i(t) \leftarrow \mu(X_i) + \gamma_U(X_i)U_i + \beta(X_i)M_i(t) + \sigma_Y \epsilon_{Y,i}, \qquad t \in \{0, 1\}$.
8: Select the factual values: $M_i \leftarrow M_i(T_i)$ and $Y_i \leftarrow Y_i(T_i)$.
9: **end for**
10: Assemble $\mathcal{W}_w$ from the generated variables and potential outcomes.
11: Randomly partition the unit indices into disjoint sets $\mathcal{C}_w$ and $\mathcal{Q}_w$, with $|\mathcal{C}_w| = N_c$ and $|\mathcal{Q}_w| = N_q$.
12: **for** $r \in \{\mathrm{BD}, \mathrm{FD}, \mathrm{IV}, \mathrm{PX}, \mathrm{HC}\}$ **do**
13: Project the same full world onto the variables retained in view r:
$\mathcal{O}_w^{(r)} \leftarrow \left\{(X_i, T_i, Y_i, V_i^{(r)})\right\}_{i=1}^{N_c+N_q}$.
14: Reuse $(\mathcal{C}_w, \mathcal{Q}_w)$ and the query targets $\{\tau_w(X_i) : i \in \mathcal{Q}_w\}$ without resampling.
15: **end for**

C SCM AND CAUSAL WORLDS

This appendix describes the synthetic SCM used in Table 3, the parameter and noise sampling procedure and the construction of paired observational views. The main comparison uses 40 worlds, indexed by $w \in \{0, \ldots, 39\}$, each containing 1,024 context units and 100 query units.

C.1 FULL SCM

Causal world and structure. Each world describes the effect of a binary treatment $T \in \{0, 1\}$ on a continuous outcome Y, mediated entirely through a continuous post-treatment variable M. Pre-treatment covariates $X \in \mathbb{R}^{50}$ and a confounder $U \in \{-1, +1\}$ influence both treatment assignment and the outcome, with U inducing the back-door path $T \leftarrow U \rightarrow Y$. A binary instrument $I \in \{0, 1\}$ also affects treatment assignment but influences the outcome only through the directed path $I \rightarrow T \rightarrow M \rightarrow Y$. X, U, and I form the roots of the SCM. The pre-treatment proxies $(Z_p, W_p) \in \{0, 1\}^2$ are generated from (X, U), providing noisy measurements of the confounder without directly affecting treatment or outcome.
The full world retains these variables together,

$$\mathcal{W} = \{X, U, I, Z_p, W_p, T, M(0), M(1), M, Y(0), Y(1), Y\}.$$

C.2 SCM GENERATION ALGORITHM

Algorithm 1 summarizes how a single causal world is generated and converted into paired observational views. World-level mechanisms are sampled and calibrated first, followed by unit-level realizations and a shared context–query split. The structural equations and parameter settings are detailed in Sections C.3 and C.4.

C.3 STRUCTURAL EQUATIONS AND DESIGN CHOICES

We suppress the world index when describing the structural equations for a single realized world.

Root variables. The root distributions are

$$X_i \sim \mathcal{N}(0, I_{50}), \qquad P(U_i = -1) = P(U_i = +1) = \tfrac{1}{2}, \qquad I_i \sim \mathrm{Bernoulli}(\tfrac{1}{2}).$$

Treatment assignment. For $j \in \{0, 1\}$, define

$$e_j(x, u) = \mathrm{expit}\{b_T + 0.55x^\top v_T + 0.75u + 1.50j\},$$
$$T_i^I(j) = \mathbf{1}\{V_{T,i} \le e_j(X_i, U_i)\}, \qquad V_{T,i} \sim \mathrm{Uniform}(0, 1),$$
$$T_i = T_i^I(I_i).$$

Because the same threshold $V_{T,i}$ is used for both instrument values and the coefficient on j is positive, the construction satisfies $T_i^I(1) \ge T_i^I(0)$.

Proxies. The treatment-inducing proxy and outcome-inducing proxy are generated by

$$Z_{p,i} \mid X_i, U_i \sim \text{Bernoulli}\big[\text{expit}\{0.20X_i^\top v_Z + 2U_i\}\big],$$
$$W_{p,i} \mid X_i, U_i \sim \text{Bernoulli}\big[\text{expit}\{0.20X_i^\top v_W + 2U_i\}\big].$$

They are noisy pre-treatment measurements of the latent confounder, not direct causes of treatment or outcome.

Mediator and outcomes. For $t \in \{0, 1\}$,

$$M_i(t) = t + 0.75\epsilon_{M,i}, \qquad M_i = M_i(T_i), \qquad \epsilon_{M,i} \sim \mathcal{N}(0, 1).$$

The prognostic surface and outcome-confounding coefficient are

$$\mu(x) = 0.7x^\top v_\mu + 0.25\sin\{x^\top \text{roll}(v_\mu, 1)\},$$
$$\gamma_U(x) = 0.8\{1 + 0.75\tanh(x^\top v_\gamma)\}.$$

Let $\tau(x)$ denote the calibrated treatment-effect surface and set $\beta(x) = \tau(x)$.
The potential outcomes are

$$Y_i(t, m) = \mu(X_i) + \gamma_U(X_i)U_i + \beta(X_i)m + 0.75\epsilon_{Y,i}, \qquad \epsilon_{Y,i} \sim \mathcal{N}(0, 1),$$

and

$$Y_i(t) = Y_i\{t, M_i(t)\}, \qquad Y_i = T_iY_i(1) + (1 - T_i)Y_i(0).$$

Why the target is shared across views. Because the same mediator noise and the same outcome noise are shared across the two treatment arms, we obtain

$$Y_i(1) - Y_i(0) = \beta(X_i)\{M_i(1) - M_i(0)\} = \tau(X_i).$$

Hence the benchmark target is identical across sibling observational views of the same realized world. Changing the observational view changes only which auxiliary variables are observed and which identification assumptions are available to the estimator; it does not change the underlying query units or the target effect values.

C.4 Parameterization and Fixed Settings

Sparse directions. Each coefficient direction is sampled as a sparse unit vector in $\mathbb{R}^{50}$. The number of active coordinates is fixed by role: the CATE directions use 10 active coordinates each, v_μ and v_T use 14, v_Z and v_W use 8, and v_γ uses 10.

Treatment-effect families. Let $a = x^\top v_a$ and $b = x^\top v_b$. The uncalibrated effect surface $q_w(x)$ is chosen from one of the following five families: linear, quadratic-interaction, threshold/piecewise, Fourier/nonstationary, and shallow MLP. The world seed determines the family so that the 40 worlds contain eight worlds from each family.

Calibration. An auxiliary Gaussian sample of size $N_{\text{cal}} = 50{,}000$ is used to calibrate

$$\tau_w(x) = 0.5 + 0.5\frac{q_w(x) - \bar{q}_w}{s_{q,w}},$$

where $\bar{q}_w$ and $s_{q,w}$ are the empirical mean and standard deviation of q_w on the calibration sample. The same auxiliary sample is also used to calibrate the treatment intercept b_T by bisection so that the average treatment propensity is approximately 0.5.

C.5 Full SCM for Partial Identification

Compared with the point-estimation SCM, the partial-identification companion omits the mediator and proxies and replaces the continuous outcome mechanism with binary potential outcomes. It retains Gaussian covariates, a binary latent confounder and instrument, and shared-threshold logistic treatment assignment, but generates 40 new worlds with separately calibrated parameters.
Algorithm 2 summarizes the construction. The outcome amplitudes satisfy $A_b, A_\tau > 0$ and $A_b + A_\tau/2 \le 0.45$, ensuring $p_t(x, u) \in [0.05, 0.95]$ without clipping. Since $p_1(x, u) - p_0(x, u) = \tau_0(x)$, the target remains $\mathbb{E}[Y(1) - Y(0) \mid X = x] = \tau_0(x)$, but the realized unit-level effect need not equal this conditional mean. Hidden-confounding strength is calibrated on separate development seeds, with the MSM restriction at $\Gamma = 2$ required to hold after marginalizing over the instrument. These generator-specific restrictions are not added to the identified-set assumptions in Appendix B.5.
Manski and sensitivity analysis use identical observational data but different maintained assumptions. IV additionally observes I. All three settings preserve the realized (X, T, Y), split, query units, and target CATE. Sensitivity oracles use $\Gamma \in \{1, 1.25, 1.5, 2, 3, 5\}$. Oracle quantities are reserved for evaluation and are not supplied to estimators.

Table 4: Main SCM numerical configuration.

Quantity	Value
The number of world generation seeds	40
Baseline dimension	50
Context / query units per world	1,024 / 100
Calibration sample size	50,000
Calibration CATE mean / standard deviation	0.5 / 0.5
Treatment covariate / latent / instrument loading	0.55 / 0.75 / 1.50
Outcome-confounding base / modulation	0.80 / 0.75
Mediator shift / mediator noise / outcome noise	1.0 / 0.75 / 0.75
Proxy covariate / latent loading	0.20 / 2.00
Severity multiplier κ	1.0
Stored oracle samples per arm/query	8,192

Algorithm 2 Partial-identification SCM and matched-view construction

Input: World seed; calibrated parameters with $a_I > 0$; $N_c = 1024$, $N_q = 100$
Output: Matched contexts, shared query covariates and CATE targets, and population-oracle bounds

1: Select one of five effect families and sample its raw function q_f, baseline function f_b, and unit-normalized treatment direction v_T.
2: $e_j(x,u) \leftarrow \mathrm{expit}(b_T + a_X x^\top v_T + a_U u + a_I j)$, $j \in \{0,1\}$
3: $b(x,u) \leftarrow \frac{1}{2} + A_b \tanh\{f_b(x) + c_U u\}$
4: $\tau_0(x) \leftarrow A_\tau \tanh\{q_f(x)\}$
5: **for** $i = 1, \ldots, N_c + N_q$ **do**
6: Draw independent $X_i \sim \mathcal{N}(0, I_{50})$, $U_i \sim \mathrm{Unif}\{-1,+1\}$, and $I_i \sim \mathrm{Bernoulli}(1/2)$.
7: Draw independent $V_{T,i}, V_{Y,i} \sim \mathrm{Uniform}(0,1)$.
8: $T_i^I(j) \leftarrow \mathbf{1}\{V_{T,i} \le e_j(X_i,U_i)\}$, $j \in \{0,1\}$; $T_i \leftarrow T_i^I(I_i)$
9: $p_{it} \leftarrow b(X_i,U_i) + (t - \frac{1}{2})\tau_0(X_i)$, $t \in \{0,1\}$
10: $Y_i(t) \leftarrow \mathbf{1}\{V_{Y,i} \le p_{it}\}$, $t \in \{0,1\}$; $Y_i \leftarrow Y_i(T_i)$
11: **end for**
12: Assign one disjoint context–query split $(\mathcal{C}, \mathcal{Q})$.
13: $\mathcal{D}_{\mathrm{M}} = \mathcal{D}_{\mathrm{S}} \leftarrow \{(X_i, T_i, Y_i) : i \in \mathcal{C}\}$
14: $\mathcal{D}_{\mathrm{IV}} \leftarrow \{(X_i, I_i, T_i, Y_i) : i \in \mathcal{C}\}$
15: Reuse $\{X_i : i \in \mathcal{Q}\}$ and $\{\tau_0(X_i) : i \in \mathcal{Q}\}$ across all three settings.
16: Compute population probabilities analytically at query covariates and apply the Manski, monotone-IV, and binary-MSM bound functionals in Appendix B.5.

D Data Validation and Construction Audits

We audit the 40 saved synthetic worlds, each containing 1,024 context and 100 query units. All checks use existing benchmark artifacts and evaluator-only SCM quantities, without regenerating data or running estimators. Production-input hashes match before and after the audit. Copied quantities are compared by exact equality, whereas deterministic numerical identities are checked using an absolute tolerance of 10^{-10}.

D.1 Construction Integrity

Paired-view integrity. We compare the five observational views within each world, covering 200 view tables and 400 within-world view pairs. The comparisons include world and unit identifiers, X, T, Y, context–query membership, ordered query identifiers, and stored query targets. Each view is also checked against its saved master-world projection, including its retained auxiliary variables. Pairing is verified from stored values and row identities, rather than inferred from shared generation seeds.

Target consistency. Using the saved potential outcomes and targets, we evaluate $Y_i(1) - Y_i(0) - \tau_0(X_i)$ over all 44,960 units. The maximum absolute discrepancy is 1.55×10^{-15}, with no violations of the 10^{-10} tolerance. This verifies the benchmark-specific identity that each unit's treatment effect equals the target CATE evaluated at its covariates.

Stay/Move targets. We check the evaluation targets stored in the completed stay and move experiment. Repeated target entries across estimator–regime configurations agree exactly, and all conditions share the same ordered query identifiers. The Original targets exactly match the production-

Table 5: **Construction integrity.** Pairing checks cover all 40 worlds and their five observational views. The potential-outcome target check uses all 44,960 units. Saved stay and move target checks use 4,000 paired queries. Maximum errors are taken over all applicable worlds and units.

Check	Result
Shared-field mismatches	0
Master-projection mismatches	0
Query-order mismatches	0
Target identity: $\max_i \lvert Y_i(1) - Y_i(0) - \tau_0(X_i)\rvert$	1.55×10^{-15}
Saved Stay target: $\max_i \lvert \tau_0^{\mathrm{S}}(X_i) - \tau_0(X_i)\rvert$	0
Saved Move target: $\max_i \lvert \tau_0^{\mathrm{M}}(X_i) - 2\tau_0(X_i)\rvert$	0

world targets. Across the 4,000 paired queries, the saved targets satisfy $\tau_0^{\mathrm{S}}(X_i) = \tau_0(X_i)$ and $\tau_0^{\mathrm{M}}(X_i) = 2\tau_0(X_i)$ with zero discrepancy. Table 5 summarizes these results.

D.2 Regime-wise Functional Consistency and Non-degeneracy

Analytic functional consistency. At the same 4,000 query locations, we reconstruct the CATE through back-door adjustment, the front-door functional, the conditional Wald ratio, and the proximal bridge functional. These calculations use saved structural quantities and SCM-derived conditional means and probabilities. The stored targets are used only for comparison after reconstruction, not as inputs to the identifying functionals. All reconstructed values are finite, and no absolute discrepancy exceeds 10^{-10}. Table 6 reports the RMSE and maximum absolute error for each regime. These results establish numerical consistency between the analytic SCM-based functionals and the stored targets.

Numerical construction diagnostics. We reproduce the existing certification gates using their original evaluation populations and aggregation rules. Complete treatment propensities $P(T = 1 \mid X, U, I)$ are evaluated at every stored (X, U) pair under both instrument values, giving 2,248 probabilities per world. These are distinguished from the marginal back-door propensity $P(T = 1 \mid X, U)$. The front-door diagnostics are computed from the structural mediator shift δ and noise scale σ_M.

For IV, we compute $\Delta_T(x) = E[T \mid I = 1, X = x] - E[T \mid I = 0, X = x]$ at all generated covariate locations per world and apply the gate to the within-world fifth percentile. We evaluate the bridge operator

$$A_t(x)_{z,w} = P(W_p = w \mid Z_p = z, T = t, X = x), \qquad z, w \in \{0, 1\},$$

at the same covariate locations under both treatment arms in PX. We record each world's minimum singular value and maximum condition number. All 40 worlds satisfy the recorded gates on these populations.

Hidden-confounding strength. For the confounding-bias function $b_{\mathrm{HC}}(x)$ defined in Appendix B, we compute

$$R_{\mathrm{HC},w} = \frac{\mathrm{SD}_{Q_w}[b_{\mathrm{HC}}(X)]}{\mathrm{SD}_{Q_w}[\tau_0(X)]},$$

where both standard deviations use the 100 queries in world w and divisor N_q. The ratio ranges from 0.3699 to 0.5999, and every world passes the recorded threshold of 0.35. This measures covariate-varying distortion of the observational contrast relative to CATE variation. HC is not assigned an oracle point-identification reconstruction.

E Experimental Setup

E.1 Metrics

E.1.1 Point-Effect and Heterogeneity Metrics

We specify the metrics and aggregation rules used in Figures 2 and 3. Metrics are first computed within each held-out query set and then summarized across paired worlds or IHDP realizations. The synthetic comparisons use 40 worlds with 100 queries per world; the IHDP comparison uses 100 realizations with 75 test queries each.

Table 6: **Regime diagnostics and analytic functional consistency.** Diagnostics use the certification populations described in the text. Bracketed values give ranges of world-level statistics, except for complete propensity, which gives the global range over checked probabilities. FD diagnostics are identical across worlds. Oracle reconstruction uses 4,000 shared queries per regime.

Diagnostic	Criterion	Result
Complete propensity $P(T = 1 \mid X, U, I)$	$[0.01, 0.99]$	$[0.02998, 0.98115]$
FD standardized shift $\lvert\delta\rvert/\sigma_M$	≥ 0.75	1.3333
FD support score $\exp\{-\delta^2/(8\sigma_M^2)\}$	≥ 0.60	0.8007
IV first-stage fifth percentile	≥ 0.10	$[0.2601, 0.2737]$
PX minimum singular value	≥ 0.01	$[0.4139, 0.4655]$
PX maximum condition number	≤ 250	$[2.2484, 2.5821]$
HC bias-to-effect SD ratio R_{HC}	≥ 0.35	$[0.3699, 0.5999]$

Oracle functional	Queries	RMSE	Max. abs. error
Back-door	4,000	7.75×10^{-17}	4.44×10^{-16}
Front-door	4,000	1.05×10^{-16}	6.66×10^{-16}
Conditional Wald	4,000	2.85×10^{-16}	1.67×10^{-15}
Proximal bridge	4,000	1.91×10^{-16}	8.88×10^{-16}

Our point-error metrics from Section 5 are

$$\mathrm{sPEHE} = \sqrt{\frac{1}{N_q}\sum_{i=1}^{N_q}\big(\widehat{\tau}(x_i) - \tau(x_i)\big)^2},$$

$$\mathrm{ATEerror} = \left|\widehat{\tau}_{\mathrm{ATE},Q} - \tau_{\mathrm{ATE},Q}\right|,$$

$$\text{c-sPEHE} = \left[\frac{1}{N_q}\sum_{i=1}^{N_q}\Big\{\big(\widehat{\tau}(x_i) - \widehat{\tau}_{\mathrm{ATE},Q}\big) - \big(\tau(x_i) - \tau_{\mathrm{ATE},Q}\big)\Big\}^2\right]^{1/2}.$$

Exact decomposition. Write $e_i = \widehat{\tau}(x_i) - \tau(x_i)$ and define its query mean as

$$\bar{e} = \frac{1}{N_q}\sum_{i=1}^{N_q} e_i = \widehat{\tau}_{\mathrm{ATE},Q} - \tau_{\mathrm{ATE},Q}.$$

Since the centered residuals sum to zero,

$$\begin{aligned}\frac{1}{N_q}\sum_{i=1}^{N_q} e_i^2 &= \frac{1}{N_q}\sum_{i=1}^{N_q}\big\{(e_i - \bar{e}) + \bar{e}\big\}^2 \\ &= \frac{1}{N_q}\sum_{i=1}^{N_q}(e_i - \bar{e})^2 + 2\bar{e}\underbrace{\frac{1}{N_q}\sum_{i=1}^{N_q}(e_i - \bar{e})}_{0} + \bar{e}^2.\end{aligned}$$

Consequently, the decomposition follows:

$$\boxed{\mathrm{sPEHE}^2 = \big(\widehat{\tau}_{\mathrm{ATE},Q} - \tau_{\mathrm{ATE},Q}\big)^2 + \text{c-sPEHE}^2}.$$

All terms use the same queries and uniform weights. The first term measures the global mean-effect discrepancy, whereas the second measures error in the centered CATE function. A smaller total error alone therefore does not determine which component improved. Figure 2 stacks the world means of these two squared terms, rather than their square roots or centered quantities computed after pooling queries across worlds.

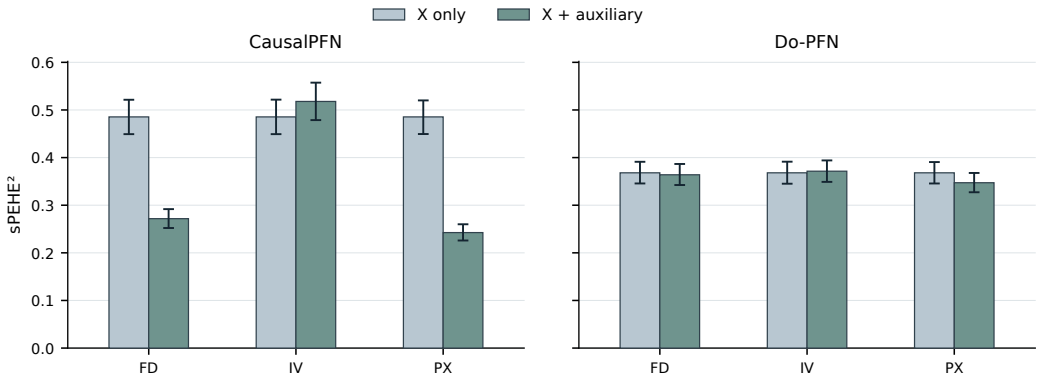


Figure 5: **Effect of auxiliary-variable inputs for CausalPFN and Do-PFN.** We compare X-only inputs with inputs augmented by the regime-specific auxiliary variables in the front-door (FD), instrumental-variable (IV), and proximal (PX) views. The auxiliary-input FD configurations condition on the factual mediator and are included as input-sensitivity diagnostics, not as front-door estimators of the total effect.

E.1.2 Sensitivity Metrics

Let $\widetilde{\mathcal{I}}^{\Gamma}_{wi} = [\widetilde{L}^{\Gamma}_{wi}, \widetilde{U}^{\Gamma}_{wi}]$ denote the stored oracle-reference sensitivity interval. For predicted endpoints $(\widehat{L}^{\Gamma}_{wi}, \widehat{U}^{\Gamma}_{wi})$, the combined endpoint error is

$$\mathrm{EndpointRMSE}_w(\Gamma) = \left[\frac{1}{2n_w}\sum_{i\in\mathcal{Q}_w}\left\{(\widehat{L}^{\Gamma}_{wi} - \widetilde{L}^{\Gamma}_{wi})^2 + (\widehat{U}^{\Gamma}_{wi} - \widetilde{U}^{\Gamma}_{wi})^2\right\}\right]^{1/2}.$$

We additionally report the fraction of query-unit oracle effects contained within the estimated bounds:

$$\mathrm{Containment}_w(\Gamma) = \frac{1}{n_w}\sum_{i\in\mathcal{Q}_w}\mathbf{1}\left\{\widehat{L}^{\Gamma}_{wi} \le \tau_{wi} \le \widehat{U}^{\Gamma}_{wi}\right\}.$$

E.2 Baselines

E.2.1 Causal foundation models

We evaluate CausalPFN, Do-PFN, and CausalFM (Balazadeh Meresht et al., 2026; Ma et al., 2026; Robertson et al., 2026) using their released checkpoints without fine-tuning. CausalPFN and CausalFM return CATE predictions directly, whereas Do-PFN contrasts predictions under the two treatment values. Their FD configurations omit M from both context and query to avoid conditioning total-effect predictions on a factual mediator. They are mediator-free transfer evaluations, not front-door estimators. Figure 5 compares these models under X-only inputs and inputs augmented with the regime-specific auxiliary variables, illustrating the sensitivity of their predictions to the input configuration. CausalFM uses its native FD and IV checkpoints in the corresponding regimes while CausalPFN and Do-PFN receive the auxiliary variables as ordinary covariates without role annotations or identification wrappers.

These comparisons concern the evaluated model–input configurations, not equally supported native estimators in every regime.

E.2.2 Modular approaches

Our modular estimators combine supervised predictive models with explicit causal estimation procedures matched to each identification regime. The causal procedure determines which nuisance quantities are required and how their predictions are combined into a CATE estimate, while the predictive backbone estimates the corresponding conditional means, probabilities, or pseudo-outcome regressions. When multiple predictive backbones are evaluated within the same procedure, we keep the causal estimation step fixed and change only the supervised models used for these predictive components.

Depending on the regime, we instantiate the predictive components with TabPFN-v3.5 (Jäger et al., 2026), XGBoost (Chen and Guestrin, 2016), feed-forward neural networks, or ExtraTrees (Geurts

Algorithm 3 Front-door plug-in estimator

Input: Context data $\{(X_i, T_i, M_i, Y_i)\}_{i=1}^n$, query covariate x
Output: Front-door estimate $\widehat{\tau}_{\mathrm{FD}}(x)$

Fit treatment model $\widehat{e}(x) \approx P(T = 1 \mid X = x)$
Fit outcome model $\widehat{Q}(x, m, s) \approx \mathbb{E}[Y \mid X = x, M = m, T = s]$
for $a \in \{0, 1\}$ **do**
    Split the context observations with $T_i = a$ into five folds
    for each fold k **do**
        Fit $\widehat{g}_{a,-k}(x) \approx \mathbb{E}[M \mid T = a, X = x]$ on the other four folds
        Predict the held-out fold and compute raw residuals

$$r_{ai}^{\mathrm{raw}} = M_i - \widehat{g}_{a,-k(i)}(X_i), \qquad T_i = a$$

    end for
    Center residuals within arm a:

$$r_{ai} = r_{ai}^{\mathrm{raw}} - \frac{1}{n_a} \sum_{j:T_j=a} r_{aj}^{\mathrm{raw}}$$

    Refit $\widehat{g}_a(x) \approx \mathbb{E}[M \mid T = a, X = x]$ on all observations with $T_i = a$
    Form the arm-specific mediator support $M_a(x) = \{\widehat{g}_a(x) + r_{ai} : T_i = a\}$
    Compute $\widehat{\theta}_a(x) = \frac{1}{|\mathcal{M}_a(x)|} \sum_{m \in \mathcal{M}_a(x)} \sum_{s=0}^{1} \widehat{p}_s(x) \widehat{Q}(x, m, s)$
end for
return $\widehat{\tau}_{\mathrm{FD}}(x) = \widehat{\theta}_1(x) - \widehat{\theta}_0(x)$

et al., 2006). For TabPFN-v3.5, continuous targets use `TabPFNRegressor` and binary probabilities use `TabPFNClassifier`. We implement XGBoost with 300 estimators, maximum tree depth 3, and learning rate 0.05, using the histogram tree method. The neural-network backbone is a two-layer MLP with 128 hidden units per layer and ReLU activations. It is trained with Adam using a learning rate of 10^{-3}, weight decay 10^{-4}, and batch size 128. ExtraTrees uses 300 trees with `max_features`=0. The minimum leaf size is 8 for classification, 6 for nuisance regression, and 10 for the final CATE regression. All remaining parameters use the corresponding package defaults and we perform no hyperparameter optimization.

Back-door (BD). We evaluate S-, X-, and DR-Learners using the observed adjustment vector $V = (X, U)$. The S-Learner (Künzel et al., 2019) fits a single outcome regression $\widehat{g}(V, T)$ on the full context and returns $\widehat{g}(v, 1) - \widehat{g}(v, 0)$.
The X-Learner (Künzel et al., 2019) uses five-fold treatment-stratified cross-fitting to obtain arm-specific out-of-fold outcome predictions. For controls, it constructs $D_i^0 = \widehat{m}_{1,-k(i)}(V_i) - Y_i$, whereas for treated observations it uses $D_i^1 = Y_i - \widehat{m}_{0,-k(i)}(V_i)$. Separate regressions of D^0 and D^1 produce $\widehat{f}_0(v)$ and $\widehat{f}_1(v)$. A propensity model fitted on the full context estimates $\widehat{e}(v) = \widehat{P}(T = 1 \mid V = v)$. The final X-Learner estimate is

$$\widehat{\tau}_{\mathrm{X}}(v) = \widehat{e}(v)\widehat{f}_0(v) + \{1 - \widehat{e}(v)\}\widehat{f}_1(v).$$

The DR-Learner (Kennedy, 2023) cross-fits both arm-specific outcome models and the propensity model. It constructs the doubly robust pseudo-outcome

$$\Gamma_i = \widehat{m}_1(V_i) - \widehat{m}_0(V_i) + \frac{T_i\{Y_i - \widehat{m}_1(V_i)\}}{\widehat{e}(V_i)} - \frac{(1 - T_i)\{Y_i - \widehat{m}_0(V_i)\}}{1 - \widehat{e}(V_i)},$$

and fits a final regression of Γ_i on V_i.
All three estimators receive the factual query U and do not marginalize over it. Their evaluation against $\tau_0(X)$ relies on the benchmark restriction $\mathbb{E}[Y(1) - Y(0) \mid X, U] = \tau_0(X)$.

Front-door (FD). We use a common front-door plug-in procedure (Guo et al., 2023; Ma et al., 2026; Pearl, 2022) with neural-network, XGBoost, or TabPFN-v3.5 predictive components. Across these variants, the front-door functional and numerical integration procedure are identical. Only the supervised models used for the treatment, mediator, and outcome regressions change. Algorithm 3 summarizes the estimator.

Algorithm 4 Conditional Wald estimator

Input: Context data $\{(X_i, I_i, T_i, Y_i)\}_{i=1}^n$, query covariate x
Output: Conditional Wald estimate $\widehat{\tau}_{\mathrm{IV}}(x)$
1: Fit outcome model $\widehat{\mu}_Y(j,x) \approx \mathbb{E}[Y \mid I = j, X = x]$
2: Fit treatment model $\widehat{\mu}_T(j,x) \approx P(T = 1 \mid I = j, X = x)$
3: **for** $j \in \{0,1\}$ **do**
4: Evaluate $\widehat{\mu}_Y(j,x)$ and $\widehat{\mu}_T(j,x)$
5: **end for**
6: Compute $\widehat{\Delta}_Y(x) = \widehat{\mu}_Y(1,x) - \widehat{\mu}_Y(0,x), \quad \widehat{\Delta}_T(x) = \widehat{\mu}_T(1,x) - \widehat{\mu}_T(0,x)$
7: **return** $\widehat{\tau}_{\mathrm{IV}}(x) = \frac{\widehat{\Delta}_Y(x)}{\widehat{\Delta}_T(x)}$

Our implementation approximates the conditional mediator distribution in each treatment arm by shifting the complete pool of centered out-of-fold mediator residuals around the predicted mediator mean. Thus, rather than fitting a separate conditional-density model, it uses an arm-specific location-shift approximation. The complete residual pool is averaged deterministically for every query, without Monte Carlo mediator sampling. Only mediator residual construction is cross-fitted; the treatment and outcome models use the full context.

Instrumental variable (IV). For IV, the modular conditional Wald estimator (Angrist and Imbens, 1995; Wald, 1940; Wang and Tchetgen Tchetgen, 2018) uses supervised models to estimate the conditional reduced-form outcome and treatment responses to the instrument. In the reported Wald configuration, these predictive components use TabPFN-v3.5. Algorithm 4 gives the complete procedure.
Both predictive models are fitted once on the full context and are not cross-fitted. We preserve the sign of the first-stage contrast and apply no denominator floor, additive epsilon, fallback prediction, or final-effect clipping. If $|\widehat{\Delta}_T(x)| \leq 10^{-12}$, the estimate remains undefined and the run fails the subsequent finiteness check. The benchmark's within-X gain restriction makes the conditional Wald estimand coincide with $\tau_0(x)$.

Proximal (PX). We implement a common P-Learner (Sverdrup and Cui, 2023) with ExtraTrees, neural networks, or TabPFN-v3.5 as the predictive backbone. The backbone is changed jointly for all nuisance models and the final pseudo-outcome regression, while the cross-fitting layout, bridge construction, and proximal score remain fixed. Algorithm 5 summarizes the procedure.
Because both proxies are binary, we solve the outcome and treatment bridges using explicit ridge-regularized two-point equations with ridge parameter 0.01 rather than fitting separate bridge models directly. All nuisance quantities entering Γ_i are out of fold.
Across the ExtraTrees, neural-network, and TabPFN-v3.5 variants, the same bridge equations and pseudo-outcome are retained. Only the predictive backbone used for the nuisance estimates and the final regression of Γ on X changes. The query input therefore contains X only. The proxies contribute through the context-level bridge and pseudo-outcome construction.

E.2.3 Additional point-estimation baselines

ForestDRIV combines the doubly robust instrumental-variable (DRIV) procedure (Syrgkanis et al., 2019) with a regression forest as the final effect model. The procedure constructs a residual-corrected regression target from estimated nuisance functions and a preliminary effect model, then regresses this target on the effect-modifying covariates. KIV (Singh et al., 2019) is a nonparametric IV estimator based on two-stage kernel ridge regression. It estimates a conditional mean embedding of the structural inputs given the instruments and then uses these embeddings to estimate the structural outcome response.
Causal Tree (Athey and Imbens, 2016) recursively partitions the covariate space to capture treatment-effect heterogeneity. It estimates a treatment effect within each terminal node and assigns that estimate to query observations falling in the node, yielding a piecewise-constant effect function. Generalized random forests (GRF) (Athey et al., 2019) estimate quantities defined by local moment equations using adaptive weights learned by a forest. The causal-forest specialization uses splits designed to capture treatment-effect heterogeneity and estimates conditional effects within the resulting forest-weighted neighborhoods.
The treatment-agnostic representation network (TARNet) (Shalit et al., 2017) learns a shared covariate representation with separate outcome heads for the treated and control groups. Training uses

Algorithm 5 Proximal P-Learner

Input: Context data $\{(X_i, Z_{p,i}, W_{p,i}, T_i, Y_i)\}_{i=1}^n$, query covariate x
Output: Proximal estimate $\widehat{\tau}_{\mathrm{PX}}(x)$

Split the context into five treatment-stratified folds
for fold $k = 1, \ldots, 5$ **do**
  Fit $e_w(x) \approx P(T = 1 \mid X = x, W_p = w)$ on the other four folds
  for $a \in \{0, 1\}$ **do**
    Fit $\mu_{az}(x) \approx \mathbb{E}[Y \mid X = x, Z_p = z, T = a]$
    Fit $p^W_{az}(x) \approx P(W_p = 1 \mid X = x, Z_p = z, T = a)$
    Fit $p^Z_{aw}(x) \approx P(Z_p = 1 \mid X = x, W_p = w, T = a)$
    Solve the regularized binary outcome bridge $h_a(w, x)$
    Solve the regularized binary treatment bridge $q_a(z, x)$
  end for
  for held-out observation i in fold k **do**
    Clip the factual treatment-bridge evaluation:

$$\widetilde{q}_{T_i}(Z_{p,i}, X_i) = \mathrm{clip}_{[-25,25]}\big(q_{T_i}(Z_{p,i}, X_i)\big)$$

    Compute the proximal pseudo-outcome

$$\Gamma_i = (2T_i - 1)\,\widetilde{q}_{T_i}(Z_{p,i}, X_i)\{Y_i - h_{T_i}(W_{p,i}, X_i)\} + h_1(W_{p,i}, X_i) - h_0(W_{p,i}, X_i)$$

  end for
end for
Fit the final regression $\widehat{\tau}_{\mathrm{PX}}(x) \approx \mathbb{E}[\Gamma \mid X = x]$ using $\{(X_i, \Gamma_i)\}_{i=1}^n$
return $\widehat{\tau}_{\mathrm{PX}}(x)$

observed outcomes through the corresponding treatment head without an explicit representation-balancing penalty. CFRNet (Shalit et al., 2017) augments the shared-representation, treatment-specific-head architecture with a penalty on the distributional discrepancy between treated and control representations. Its objective combines factual outcome prediction with this balancing term. As with TARNet, treatment effects are predicted by contrasting the two outcome heads at the same query covariates.

E.2.4 Models for Partial Identification

We evaluate bound estimators on the matched binary companion worlds in Appendix C.5. For Manski and IV bounds, multinomial logistic regression, XGBoost (Chen and Guestrin, 2016), and frozen TabPFN-v3.5 (Jäger et al., 2026) estimate $P(T, Y \mid X)$ and $P(T, Y \mid X, I)$, respectively. The Manski formula (Manski, 2003) and the IV response-type linear program (Balke and Pearl, 1997) are held fixed across backbones, with treatment monotonicity imposed in the latter. We apply a common feasibility projection to infeasible estimated IV cells. For sensitivity bounds, we evaluate CSA-PFN (MSM) (Javurek et al., 2026), B-Learner (Oprescu et al., 2023), and NeuralCSA (Frauen et al., 2024) at $\Gamma \in \{1, 1.25, 1.5, 2, 3, 5\}$. In our implementation, CSA-PFN uses its frozen checkpoint with a context-fitted 50-to-10 PCA adapter, while B-Learner uses five-fold cross-fitting with random forests. For the B-Learner variants, "all" uses the indicated predictive backbone for all learned components, whereas "final" replaces only the final bound regressions with TabPFN and retains RF-based nuisance estimators. NeuralCSA is trained separately in each world, applying its continuous-density flow to binary outcomes. All estimators use only context data for fitting and are evaluated against the corresponding SCM oracle bounds on shared query units.

E.3 Semi-Synthetic Evaluation Settings

For ACIC 2016 (Dorie et al., 2019), we select 40 parameter settings by rounding 40 equally spaced indices from 1 to 77, using simulation 1 for each setting. For the LaLonde-PSID and LaLonde-CPS datasets (Dehejia and Wahba, 1999; 2002; LaLonde, 1986) generated using RealCause (Neal et al., 2020), we use official samples 0–39. From each realization, we select 1,024 context units and 100 disjoint query units, with identical context and query units across estimators. Within each RealCause dataset, covariate rows and split indices are also preserved across realizations when the archived samples support exact pairing. ACIC retains its original 58-column schema, with the three string-valued columns encoded ordinally using sorted category levels, whereas RealCause uses its

eight native numeric covariates. The CATE targets are conditional-mean contrasts from the official ACIC generator or the pre-trained RealCause outcome generator.

E.4 Real-World Evaluation Settings

The control–treatment contrasts are defined as AZT monotherapy versus AZT+ddI combination therapy in ACTG 175 (Hammer et al., 1996), group 0 versus group 6 among records satisfying `dem_inel=1` and `revasamp=1` in Pennsylvania (Corson et al., 1992), and the original control group versus the job-search incentive group in Illinois (Woodbury and Spiegelman, 1987). The respective negative-control outcomes are `cd40`, `wages`, and `avprearn`. Context sampling is treatment-stratified to approximately preserve the eligible sample's treatment proportion, with 768, 1,024, and 2,048 context units for ACTG 175, Pennsylvania, and Illinois, respectively. Each split contains 256 query units sampled without replacement from the remaining records, and all estimators share the same context and query units. All pre-processing is fitted on the context only. Numeric features undergo median imputation, categorical features undergo mode imputation and one-hot encoding, and outcomes are standardized using the context mean and population standard deviation. We compute metrics relative to the null target $\tau(x) = 0$ on the standardized outcome scale and average them over the 40 splits. These repetitions assess split sensitivity within a fixed dataset, rather than variation across independent trials.

Although the original trials contain multiple treatment arms, we form each benchmark contrast by selecting two randomized arms and designating one as control and the other as treatment. Let $A \in \mathcal{A}$ denote the original treatment assignment. For selected arms $a_0, a_1 \in \mathcal{A}$, define $S = 1\{A \in \{a_0, a_1\}\}$ and $T = 1\{A = a_1\}$. Since the negative-control outcome is measured before treatment, it cannot be affected by treatment assignment (Arnold and Ercumen, 2016). In the potential-outcome formulation (Ashby et al., 2025),

$$Y^{\mathrm{pre}}(a_0) = Y^{\mathrm{pre}}(a_1) = Y^{\mathrm{pre}},$$

and therefore

$$\tau(x) = \mathbb{E}[Y^{\mathrm{pre}}(a_1) - Y^{\mathrm{pre}}(a_0) \mid X = x, S = 1] = 0.$$

Moreover, randomization of the original arm assignment implies

$$A \perp \{Y^{\mathrm{pre}}(a) : a \in \mathcal{A}\} \mid X,$$

and hence, after restricting to the selected pair,

$$T \perp \left\{Y^{\mathrm{pre}}(a_0), Y^{\mathrm{pre}}(a_1)\right\} \mid X, S = 1.$$

E.5 IHDP Full versus IHDP-HC

We adopt the Infant Health and Development Program (IHDP) benchmark to evaluate treatment-effect estimation under confounder omission. Following the hidden-confounding construction of Jesson et al. (2021) and its Quince implementation, we combine empirical covariates and treatment assignments with synthetically generated outcomes. The resulting dataset contains 747 units with 25 covariates. The Full view retains all covariates, including $U = \texttt{b.marr}$, the mother's marital status at childbirth. The IHDP-HC view removes only U from the model inputs, leaving 24 covariates. The omitted variable remains part of the outcome-generating process and can also modify treatment effects. Unlike the main synthetic benchmark, the IHDP-HC comparison also removes an effect modifier. Its error increase therefore combines confounding-related degradation with loss of information about effect heterogeneity.

We evaluate 100 paired trials using the Quince data preprocessing and splitting procedure. Each trial contains 470 training, 202 validation, and 75 test units. For the main comparison, the training and validation subsets are pooled into a context of 672 units, with the 75 test units used as queries. Within each trial, Full and Hidden share the same unit identities, treatment assignments, factual and potential outcomes, split, and oracle effect values. Both views are evaluated against the same $\tau_{\mathrm{true}}(X, U)$ values, without redefining the target after removing U. This paired comparison serves as a semi-synthetic stress test of point estimation when a confounder is withheld from the estimator.

F Ablation Studies

F.1 Hidden Confounder Track

Figure 6 examines point-estimator behavior under unobserved confounding and the accuracy of estimated partial-identification bounds. The synthetic HC view retains only (X, T, Y), without the confounder or auxiliary identifying variables. Point predictions in this HC view are evaluated without a point-identification claim.

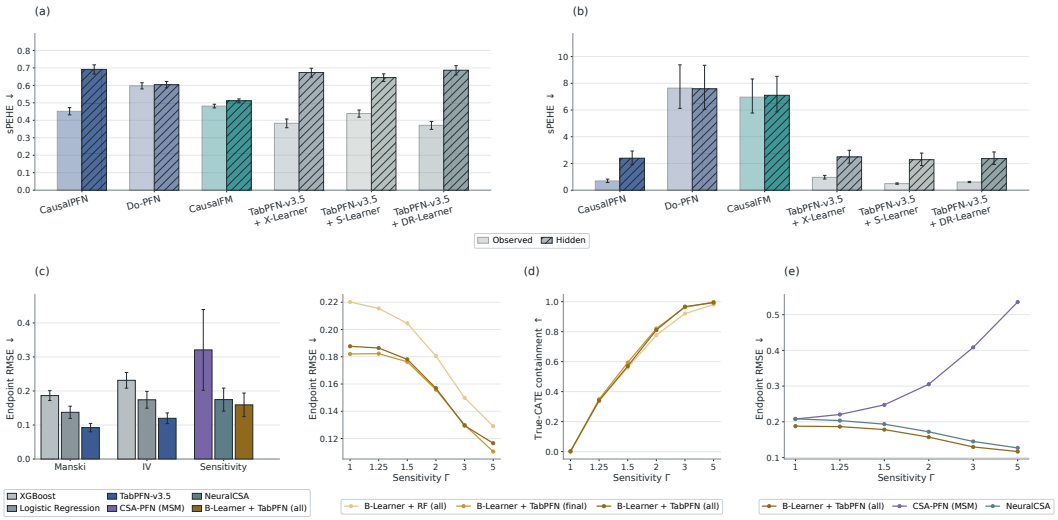


Figure 6: **Hidden-confounding and partial-identification evaluation.** (a) Synthetic paired comparison of point-estimation error when the confounder is observed (BD) or unobserved (HC). (b) Fully Observed versus Hidden comparisons on IHDP-HC. (c) Endpoint RMSE for representative partially identified settings, including bounded-outcome, IV-based, and sensitivity-model leaves. The sensitivity comparison in panel (c) is evaluated at $\Gamma = 2$. (d) Performance of B-Learner variants across sensitivity levels Γ. (e) Endpoint RMSE of sensitivity-aware estimators across Γ.

Confounder omission affects models differently. Panels (a) and (b) compare point-estimation performance with the designated confounder observed or unobserved on synthetic and IHDP-HC data, respectively. CausalPFN and all three TabPFN-based modular approaches exhibit pronounced increases in sPEHE after confounder omission in both settings whereas Do-PFN and CausalFM show smaller changes.

TabPFN improves bound-estimation accuracy. Panel (c) evaluates three representative partial-identification settings. TabPFN-v3.5 achieves the lowest endpoint RMSE among the compared estimators in the Manski and IV settings, while B-Learner + TabPFN (all) achieves the lowest error in the sensitivity setting. Panel (d) further compares RF- and TabPFN-based B-Learner variants. Both TabPFN configurations obtain lower endpoint RMSE than the RF-based configuration at every evaluated sensitivity level Γ. Their true-CATE containment curves increase with Γ. These results distinguish accurate recovery of the reference bounds from inclusion of the oracle CATE and support the use of TabPFN within bound-estimation procedures.

Sensitivity-level dependence varies across estimators. In panel (e), the endpoint RMSE of CSA-PFN (MSM) increases monotonically over the evaluated sensitivity grid. In contrast, B-Learner + TabPFN (all) and NeuralCSA exhibit decreasing endpoint errors, with the TabPFN-based B-Learner achieving the lowest error at every evaluated Γ. Its advantage over CSA-PFN becomes larger as the maintained sensitivity restriction is relaxed. This contrast highlights the importance of evaluating sensitivity-bound estimators across multiple sensitivity levels.

F.2 Stress Test for Synthetic Data

Regime-specific stress. Table 7 highlights both the strengths and limitations of modular estimation under regime-specific stress. In BD and FD, the TabPFN-based DR-Learner and front-door plug-in, respectively, retain the lowest mean sPEHE^2 despite weaker treatment overlap and reduced mediator support. All evaluated plug-in estimators in FD comparison outperform the CFM configurations throughout the stress sweep. However, this advantage does not extend to IV and PX. Under weaker instruments, the TabPFN-based Wald estimator and ForestDRIV exhibit sharply increasing errors and between-world variability, while CausalFM maintains the lowest mean error. In PX, worsening bridge conditioning erodes the initial advantage of the TabPFN-based P-Learner, with CausalFM becoming the lowest-error configuration at Severe. These contrasting results show that the empirical benefits of modular estimation depend not only on the identification regime but also on the conditions under which its estimation procedure is applied.

Table 7: **Regime-specific stress tests across four identification regimes.** We report sPEHE^2 (↓) as the mean ± standard deviation across 40 SCM worlds. Default, Mild, Moderate, and Severe denote increasing levels of regime-specific stress. These levels are not calibrated across regimes. **Bold** means indicate the lowest mean error within each regime and severity level.

	Stress severity			
Method	Default	Mild	Moderate	Severe
Back-door (BD): Overlap				
CausalPFN	0.208 ± 0.062	0.233 ± 0.068	0.253 ± 0.074	0.291 ± 0.084
Do-PFN	0.360 ± 0.071	0.338 ± 0.066	0.322 ± 0.064	0.310 ± 0.061
CausalFM	0.233 ± 0.033	0.232 ± 0.033	0.234 ± 0.034	0.235 ± 0.034
TabPFN-v3.5 + X-Learner	0.153 ± 0.060	0.166 ± 0.066	0.182 ± 0.068	0.212 ± 0.078
TabPFN-v3.5 + S-Learner	0.197 ± 0.057	0.199 ± 0.058	0.205 ± 0.054	0.209 ± 0.054
TabPFN-v3.5 + DR-Learner	$\mathbf{0.143} \pm 0.054$	$\mathbf{0.154} \pm 0.060$	$\mathbf{0.171} \pm 0.063$	$\mathbf{0.202} \pm 0.070$
Front-door (FD): Mediator support				
CausalPFN	0.485 ± 0.117	0.480 ± 0.115	0.476 ± 0.114	0.473 ± 0.113
Do-PFN	0.368 ± 0.074	0.365 ± 0.074	0.364 ± 0.075	0.364 ± 0.075
CausalFM	0.551 ± 0.073	0.550 ± 0.073	0.549 ± 0.073	0.549 ± 0.073
NN + FD Plug-in	0.227 ± 0.031	0.230 ± 0.033	0.232 ± 0.033	0.235 ± 0.032
XGBoost + FD Plug-in	0.233 ± 0.037	0.227 ± 0.041	0.234 ± 0.042	0.251 ± 0.054
TabPFN-v3.5 + FD Plug-in	$\mathbf{0.154} \pm 0.057$	$\mathbf{0.181} \pm 0.051$	$\mathbf{0.198} \pm 0.049$	$\mathbf{0.212} \pm 0.050$
Instrumental variable (IV): Instrument strength				
CausalPFN	0.518 ± 0.127	0.528 ± 0.124	0.542 ± 0.120	0.540 ± 0.118
Do-PFN	0.371 ± 0.073	0.369 ± 0.072	0.366 ± 0.071	0.365 ± 0.071
CausalFM	$\mathbf{0.258} \pm 0.041$	$\mathbf{0.256} \pm 0.039$	$\mathbf{0.253} \pm 0.039$	$\mathbf{0.253} \pm 0.039$
ForestDRIV	0.460 ± 0.264	1.770 ± 4.671	26.980 ± 43.796	111.043 ± 156.843
KIV	0.420 ± 0.187	0.457 ± 0.189	0.468 ± 0.187	0.491 ± 0.209
TabPFN-v3.5 + Wald	0.336 ± 0.105	0.389 ± 0.174	1.145 ± 2.578	$1.94 \times 10^7 \pm 1.22 \times 10^8$
Proximal (PX): Bridge conditioning				
CausalPFN	0.243 ± 0.056	0.287 ± 0.069	0.352 ± 0.080	0.445 ± 0.101
Do-PFN	0.347 ± 0.066	0.351 ± 0.067	0.356 ± 0.070	0.362 ± 0.071
CausalFM	0.240 ± 0.033	0.250 ± 0.034	0.269 ± 0.035	$\mathbf{0.301} \pm 0.043$
ExtraTrees + P-Learner	0.233 ± 0.049	0.284 ± 0.065	0.318 ± 0.062	0.441 ± 0.094
NN + P-Learner	0.662 ± 1.199	0.471 ± 0.247	0.365 ± 0.121	0.536 ± 0.213
TabPFN-v3.5 + P-Learner	$\mathbf{0.181} \pm 0.054$	$\mathbf{0.193} \pm 0.056$	$\mathbf{0.217} \pm 0.052$	0.442 ± 0.108

Context length. Table 8 shows that smaller contexts generally reduce estimation accuracy and can change which configuration performs best. While the TabPFN-based FD plug-in and CausalFM in IV retain the lowest mean errors throughout the sweep, the leading configurations in BD and PX change at the shortest context. The extent of degradation also differs substantially across estimators. CausalPFN's mean error more than doubles in BD and PX, and ForestDRIV exhibits pronounced increases in both error and between-world variability in IV. By contrast, Do-PFN's errors change relatively little across the four regimes. Such limited sensitivity should not, however, be interpreted as evidence of accurate estimation: CausalFM's FD error remains nearly unchanged yet consistently exceeds that of every FD plug-in estimator. Thus, context-length evaluation reveals both the dependence of relative performance on data availability and the distinction between maintaining a similar error and maintaining a low error.

F.3 Identification Beyond Additional Inputs

Figure 7 compares three TabPFN configurations to distinguish the effect of additional observed variables from that of explicit regime-specific causal estimation. The first configuration uses TabPFN to predict Y from (X, T) alone, excluding all auxiliary variables and providing a common refer-

Table 8: **Context-length stress test across four identification regimes.** We report sPEHE^2 (↓) as the mean ± standard deviation across 40 SCM worlds. Context lengths of 1,024, 512, 256, and 160 correspond to the Default, Mild, Moderate, and Severe settings, respectively. **Bold** means indicate the lowest mean error within each regime and context length.

	Context length n_{ctx}			
Method	Default	Mild	Moderate	Severe
Back-door (BD)				
CausalPFN	0.208 ± 0.062	0.249 ± 0.063	0.327 ± 0.085	0.451 ± 0.134
Do-PFN	0.360 ± 0.071	0.363 ± 0.071	0.370 ± 0.079	0.383 ± 0.082
CausalFM	0.233 ± 0.033	0.240 ± 0.035	0.253 ± 0.044	0.273 ± 0.055
TabPFN-v3.5 + X-Learner	0.153 ± 0.060	0.202 ± 0.056	0.261 ± 0.069	0.308 ± 0.089
TabPFN-v3.5 + S-Learner	0.197 ± 0.057	0.227 ± 0.039	0.246 ± 0.039	$\mathbf{0.268} \pm 0.053$
TabPFN-v3.5 + DR-Learner	$\mathbf{0.143} \pm 0.054$	$\mathbf{0.197} \pm 0.049$	$\mathbf{0.244} \pm 0.068$	0.289 ± 0.095
Front-door (FD)				
CausalPFN	0.485 ± 0.117	0.542 ± 0.116	0.580 ± 0.159	0.691 ± 0.225
Do-PFN	0.368 ± 0.074	0.372 ± 0.072	0.381 ± 0.082	0.391 ± 0.084
CausalFM	0.551 ± 0.073	0.551 ± 0.074	0.552 ± 0.074	0.552 ± 0.074
NN + FD Plug-in	0.227 ± 0.031	0.240 ± 0.039	0.282 ± 0.067	0.286 ± 0.066
XGBoost + FD Plug-in	0.233 ± 0.037	0.271 ± 0.043	0.299 ± 0.053	0.313 ± 0.057
TabPFN-v3.5 + FD Plug-in	$\mathbf{0.154} \pm 0.057$	$\mathbf{0.212} \pm 0.050$	$\mathbf{0.242} \pm 0.047$	$\mathbf{0.260} \pm 0.048$
Instrumental variable (IV)				
CausalPFN	0.518 ± 0.127	0.570 ± 0.127	0.593 ± 0.168	0.691 ± 0.224
Do-PFN	0.371 ± 0.073	0.374 ± 0.072	0.385 ± 0.081	0.394 ± 0.083
CausalFM	$\mathbf{0.258} \pm 0.041$	$\mathbf{0.262} \pm 0.039$	$\mathbf{0.273} \pm 0.043$	$\mathbf{0.288} \pm 0.055$
ForestDRIV	0.460 ± 0.264	1.561 ± 3.460	11.620 ± 38.073	29.283 ± 58.180
KIV	0.420 ± 0.187	0.432 ± 0.231	0.443 ± 0.192	0.461 ± 0.179
TabPFN-v3.5 + Wald	0.336 ± 0.105	0.372 ± 0.159	0.402 ± 0.095	0.477 ± 0.331
Proximal (PX)				
CausalPFN	0.243 ± 0.056	0.296 ± 0.059	0.354 ± 0.087	0.505 ± 0.149
Do-PFN	0.347 ± 0.066	0.351 ± 0.066	0.360 ± 0.074	0.373 ± 0.078
CausalFM	0.240 ± 0.033	0.251 ± 0.039	0.267 ± 0.051	$\mathbf{0.288} \pm 0.057$
ExtraTrees + P-Learner	0.233 ± 0.049	0.300 ± 0.069	0.383 ± 0.101	0.516 ± 0.275
NN + P-Learner	0.662 ± 1.199	0.699 ± 0.576	0.889 ± 1.914	0.851 ± 0.911
TabPFN-v3.5 + P-Learner	$\mathbf{0.181} \pm 0.054$	$\mathbf{0.222} \pm 0.048$	$\mathbf{0.258} \pm 0.080$	0.394 ± 0.382

ence across views. The second configuration additionally supplies all auxiliary variables available in the corresponding observational view as ordinary predictive features. Neither configuration provides annotations about the causal roles of these variables or applies a regime-specific identification procedure. The last integrates TabPFN into the corresponding modular estimator.

The modular configuration achieves the lowest mean sPEHE^2 across all four evaluated views. Adding auxiliary variables as ordinary features improves performance over the (X, T)-only reference in the back-door, front-door, and proximal views, but increases error in the IV view. Even though additional features improve performance, their errors remain higher than those of the corresponding modular estimators. These results support combining TabPFN's predictive capabilities with regime-specific identification.

F.4 Effect of the Predictive Backbone in Modular Approaches

We examine how the predictive backbone affects the modular estimators while keeping the regime-specific identification procedure fixed. For each regime, we replace the predictive model used within the corresponding X-Learner, front-door plug-in, Wald, and P-Learner pipeline with alternative tabular predictors.

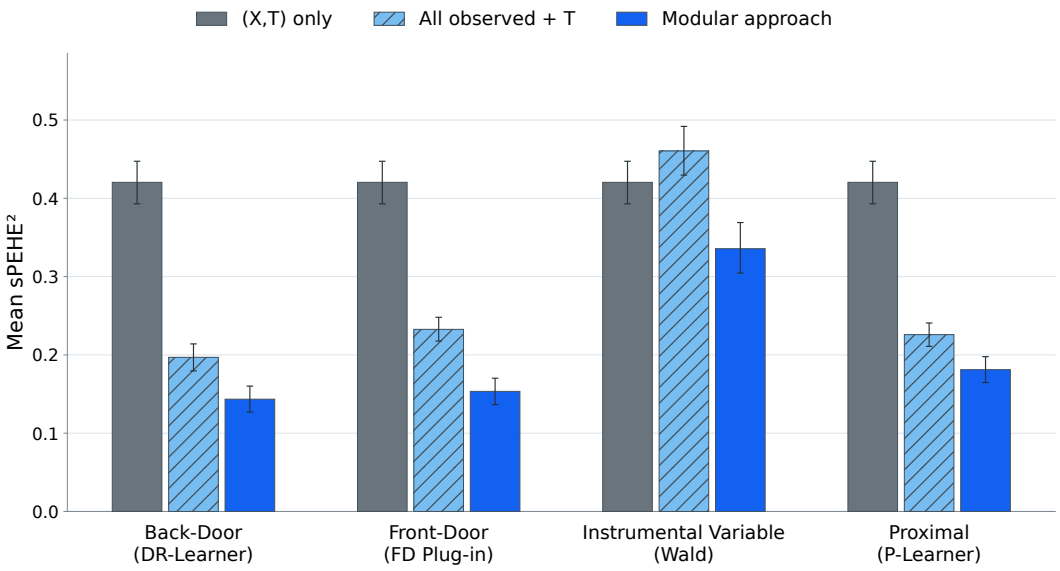


Figure 7: **Input and identification ablation for the modular approach.** We compare three TabPFN configurations: (X, T) only, all variables available in each observational view provided as ordinary predictive features, and the corresponding regime-specific modular estimator. The first two configurations provide no annotations or information about the causal roles of the input variables.

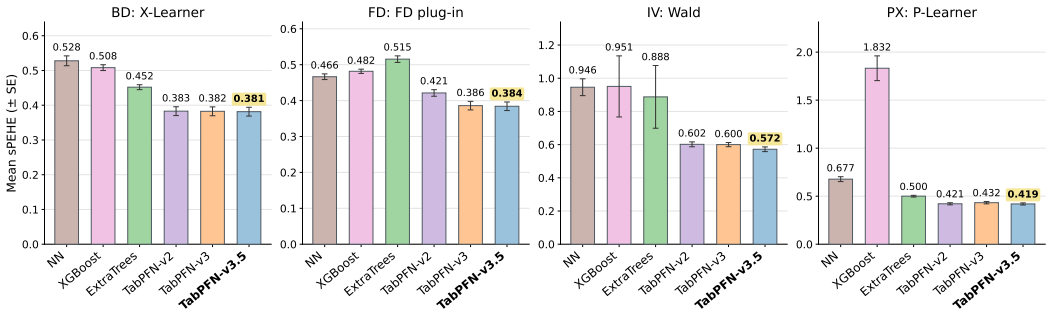


Figure 8: **Effect of the predictive backbone in modular approaches.** We replace the predictive model in modular approaches while keeping the regime-specific identification procedure fixed within each panel.

Figure 8 shows that the choice of predictive backbone can substantially affect causal estimation performance, even when the identification procedure is unchanged. The TabPFN variants generally achieve lower mean sPEHE than the conventional predictive models across the four regimes. Performance is relatively stable across versions of TabPFN. TabPFN-v3.5 (Jäger et al., 2026) attains the lowest mean error in BD, FD, IV, and PX. These results suggest that the quality of the predictive backbone remains consequential.