CRA-PFN: TREATMENT EFFECT PREDICTION WITH MIXED-ROLE COVARIATES VIA PRIOR-DATA FITTED NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional average treatment effect (CATE) estimation from observational data usually presumes that valid adjustment covariates are known. In practice, however, treatment and outcome are designated while the remaining covariates may mix pre-treatment causes with prognostic covariates, mediators, descendants, and colliders, potentially hindering CATE identification and estimation by existing methods. To address this, we introduce the Causal Role-Agnostic Prior-data Fitted Network (CRA-PFN), which is pretrained under a role-heterogeneous prior over structural causal models (SCMs) and predicts task-specific CATEs without causal-role labels or a designated adjustment set being supplied to the estimator. We formalize role-agnostic CATE identifiability and derive conditions for mean-square consistency of CATE prediction as the mixed-role observational data grow. We further characterize the population behavior of the training objective relative to the prior-induced Bayes estimator, and derive a CATE error decomposition that separates estimator error from residual uncertainty. Extensive experiments across diverse datasets and out-of-distribution SCMs demonstrate that CRA-PFN achieves strong performance and generalizes effectively across mechanism shifts.

1 INTRODUCTION

Conditional average treatment effect (CATE) estimation from observational data is an important task in empirical research across economics (Athey & Wager, 2021), medicine (Feuerriegel et al., 2024), and marketing (Ellickson et al., 2023). CATE characterizes how treatment effects vary across individuals and is therefore important for individualized decision-making (Kern et al., 2024). Existing CATE estimators span a range of modeling paradigms, including tree-based, meta-learning, orthogonal, Bayesian, and neural approaches (Athey et al., 2019; Künzel et al., 2019; Nie & Wager, 2021; Hill, 2011; Shalit et al., 2017). These methods flexibly model nonlinear relationships and treatment effect heterogeneity, but domain experts are typically required to manually select an appropriate estimator and fit a separate model for each new application or task.

Recently, inspired by in-context learning, CATE estimation has been extended to pretrained estimators that can be directly applied to new tasks without task-specific retraining. These methods (Balazadeh Meresht et al., 2025; Robertson et al., 2025; Ma et al., 2026), built on prior-data fitted networks (PFNs) (Müller et al., 2022), amortize CATE estimation across synthetic causal tasks during pretraining. Among them, Do-PFN (Robertson et al., 2025) considers effect estimation without requiring knowledge of the underlying causal graph, but does not provide identifiability of the causal query. In contrast, CausalFM (Ma et al., 2026) explicitly separates causal identification from effect estimation and restricts pretrained estimation to identifiable causal tasks.

However, ensuring identifiability does not necessarily require that the causal roles of observed covariates be supplied to the estimator. In practice, observational data may mix valid pre-treatment covariates with mediators, treatment descendants, outcome descendants, and colliders. These roles determine not only which variables should be used for causal adjustment, but also which variables should define the conditioning information for CATE estimation. Consequently, directly adjusting for all observed covariates would introduce bias or change the estimand (Peters et al., 2017; Spirtes

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.

et al., 2001). This raises a question: can a pretrained estimator recover CATE when the specific causal role of observed covariates is unknown?

To address this question, we introduce the Causal Role-Agnostic Prior-data Fitted Network (CRA-PFN), a pretrained estimator for CATE estimation from observational data with unknown covariate roles. CRA-PFN is pretrained over synthetic structural causal models with varying covariate roles and causal mechanisms, and can be directly applied to new observational datasets without role annotations or a prespecified adjustment set. Our main contributions can be summarized as follows:

- We study CATE estimation with unknown covariate roles, and characterize when CATEs remain identifiable from the observational data without explicitly recovering these roles or the underlying causal graph.
- We develop CRA-PFN, pretrained on tasks sampled from an SCM prior with varying covariate roles and causal mechanisms across tasks. We further provide theoretical guarantees for its learning objective and establish conditions for the consistent CATE estimation.
- We evaluate CRA-PFN on standard benchmarks, structural variants, synthetic SCMs, and four out-of-distribution SCM families. Experimental results show that CRA-PFN performs effectively across both in-distribution and out-of-distribution tasks.

2 Related Work

CATE estimation. A broad class of methods has been developed for estimating CATEs, including tree-based estimators, meta-learners, orthogonal objectives, Bayesian estimators, and neural approaches (Athey et al., 2019; Künzel et al., 2019; Nie & Wager, 2021; Chernozhukov et al., 2018; Hill, 2011; Hahn et al., 2020; Shalit et al., 2017; Shi et al., 2019). Despite their different estimation strategies, these methods typically assume that the covariates only contain pre-treatment covariates for confounding adjustment. Another line of work leverages causal structure or variable selection to identify appropriate covariates before effect estimation (Tran & Zheleva, 2022; Henckel et al., 2022; Li et al., 2024). In contrast, we consider settings where the observed covariates may have heterogeneous causal roles and the variables appropriate for treatment effect estimation are not specified in advance.

Prior-data fitted networks. Prior-data fitted networks (PFNs) learn amortized inference by pretraining on large-scale synthetic tasks sampled a specified prior, enabling inference on new tasks without task-specific model fitting (Müller et al., 2022). TabPFN (Hollmann et al., 2023; 2025) applies this paradigm to tabular prediction and has demonstrated strong generalization across diverse tabular tasks. Subsequent work extends prior-data fitting to simulation-based Bayesian inference (Vetter et al., 2025), time series data (Dooley et al., 2023; Fang et al., 2026), and so on, demonstrating its applicability beyond standard supervised prediction. This ability to amortize inference across heterogeneous data generating processes provides a natural foundation for extending prior-data fitting to causal tasks.

Causal prior-data fitting. Recent work extends prior-data fitting to causal effect estimation. CausalPFN learns to predict treatment effects under observational processes satisfying conditional exchangeability (Balazadeh Meresht et al., 2025). Do-PFN uses simulated interventions to learn conditional interventional prediction from observational context (Robertson et al., 2025). CausalFM (Ma et al., 2026) formalizes SCM priors for identified causal problems and develops back-door, front-door, and instrumental-variable instances. Most closely related to our setting, recent work TC-LD (Ham et al., 2026) also studies post-treatment covariates in causal foundation models and proposes a pre-inference filter for post-treatment variables. In contrast, we study when CATE remains identifiable from the observational distribution when covariate roles are unobserved, and develop CRA-PFN to predict it without requiring causal-role labels, a causal graph, or a designated adjustment set at inference time.

3 PROBLEM FORMULATION

Let $\mathbb{P}_{\text{obs}}$ be the observational distribution of covariates $\mathbf{X} \in \mathbb{R}^p$, binary treatment T, and scalar outcome Y. The latent vector $\mathbf{R} = (R_1, \ldots, R_p)$ records the covariate roles. Each task specifies a pre-treatment covariate set $\mathbf{C} \subseteq \mathbf{X}$ that defines the CATE and is sufficient for adjustment. The estimator is not told which observed covariates belong to $\mathbf{C}$. Concretely, $\mathbf{C}$ includes pre-treatment common causes of treatment and outcome, treatment predictors, and prognostic covariates, while excluding mediators and other post-treatment descendants of T, descendants of Y, and colliders. We denote potential outcome as $Y(t)$, and further define potential outcome mean and CATEs as

$$\mu_t(\mathbf{c}) := \mathbb{E}[Y(t) \mid \mathbf{C} = \mathbf{c}], \qquad \tau(\mathbf{c}) := \mathbb{E}[Y(1) - Y(0) \mid \mathbf{C} = \mathbf{c}] = \mu_1(\mathbf{c}) - \mu_0(\mathbf{c}). \tag{1}$$

We impose the following standard identification assumptions.

Assumption 1 (CATE identification)**.** Suppose (1) consistency holds, $Y = Y(T) = TY(1) + (1 - T)Y(0)$, and units do not interfere, (2) positivity holds, $0 < \mathbb{P}_{\text{obs}}(T = t \mid \mathbf{C}) < 1$ almost surely for each $t \in \{0, 1\}$, and (3) conditional exchangeability holds, $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid \mathbf{C}$.

Under Assumption 1, the task-specific CATE is identified from the observational distribution by conditioning on the valid pre-treatment covariates $\mathbf{C}$:

$$\tau(\mathbf{c}) = \mathbb{E}_{\mathbb{P}_{\text{obs}}}[Y \mid T = 1, \mathbf{C} = \mathbf{c}] - \mathbb{E}_{\mathbb{P}_{\text{obs}}}[Y \mid T = 0, \mathbf{C} = \mathbf{c}]. \tag{2}$$

However, the estimator receives the full covariates $\mathbf{X}$ instead of pre-treatment covariates $\mathbf{C}$ or the causal roles among $\mathbf{X}$. This creates the central challenge considered in our work, i.e., estimating CATEs defined by an unknown valid covariate subset from inputs containing covariates with mixed causal roles.

At prediction time, the input contains context $\mathcal{D}_{\text{obs}} = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n_{\text{obs}}}$ and query covariates indexed by $\mathcal{Q}$. For query j, let $\mathbf{c}_j$ denote the values of the valid pre-treatment covariates $\mathbf{C}$ in $\mathbf{x}_j$. Query treatments and outcomes are masked. We write $W_j = (\mathcal{D}_{\text{obs}}, \mathbf{x}_j)$, and let $W = (\mathcal{D}_{\text{obs}}, \mathbf{X})$ denote a generic input. In the full-input setting, all query covariates remain visible, including post-treatment or outcome-descendant measurements when present. Covariates outside $\mathbf{C}$ therefore remain available to the estimator, although they do not define the CATE.

Definition 1 (Role-agnostic CATE prediction)**.** A role-agnostic estimator uses W_j to predict the CATE $\tau(\mathbf{c}_j)$. It receives no causal-role labels, causal graph, or specification of $\mathbf{C}$.

Thus, the learning problem is to estimate $\tau(\mathbf{c}_j)$ from the observational context and query covariates W_j when the estimator receives full observed covariates without role information.

4 CRA-PFN

CRA-PFN addresses this problem by amortizing prediction under a prior Π over causal tasks with heterogeneous latent roles and mechanisms. During pretraining, each sampled task provides observational inputs together with arm-specific supervision, and the contrast between the two predicted arm means forms the CATE prediction.

Let $\mathcal{S} \sim \Pi$ denote a task. We regard all task-level quantities that remain fixed across factual and intervention draws as components of $\mathcal{S}$, including the sampled structural mechanisms and other task-specific parameters. The prior and sampling procedure induce a joint distribution $\mathbb{P}_\Pi$ over tasks, contexts, queries, and supervision draws, with expectations and variances denoted by $\mathbb{E}_\Pi$ and Var_Π. We next describe the task prior and supervision, finite-context prediction, and the arm-specific prediction heads. Figure 1 summarizes the procedure.

4.1 ROLE-HETEROGENEOUS SCM PRIOR

The task prior varies latent covariate roles while keeping the CATE defined by the task-specific valid pre-treatment covariates. Each task specifies latent roles, a task-specific directed acyclic graph (DAG), structural mechanisms, and noise distributions. Seven covariate-role labels define the mandatory relations listed in Table 5. Additional covariate edges preserve the causal ordering implied by these roles, acyclicity, and the specified treatment/outcome reachability relations. Covariate types and treatment-pathway modes also vary across tasks. Details are provided in Appendix C.

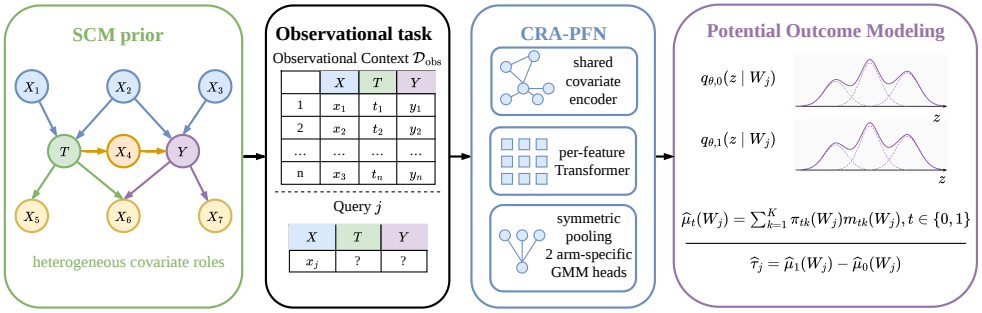

Figure 1: CRA-PFN pretraining and prediction. Each task supplies observational context and two arm labels. Query treatment and outcome entries are masked, and queries cannot attend to one another. Two arm-specific prediction heads produce the predicted arm means, and their contrast gives the CATE prediction.

For a sampled task $\mathcal{S}$, define the arm means and CATE as

$$\mu_t^{\mathcal{S}}(\mathbf{c}) := \mathbb{E}[Y(t) \mid \mathcal{S}, \mathbf{C} = \mathbf{c}], \qquad \tau^{\mathcal{S}}(\mathbf{c}) := \mu_1^{\mathcal{S}}(\mathbf{c}) - \mu_0^{\mathcal{S}}(\mathbf{c}). \tag{3}$$

At each covariate value $\mathbf{c}_j$, simulated interventions provide arm-specific supervision labels $Z_{t,j}$ while holding the task and pre-treatment covariates fixed. Under the arm-label centering condition in Appendix C.6, the conditional mean of each label equals the corresponding task-specific arm mean.

4.2 Bayes Estimator

Because neither the task $\mathcal{S}$ nor the role vector $\mathbf{R}$ is observed by the estimator, a finite observational input W may remain compatible with multiple task-specific CATE values. The CATE remains $\tau^{\mathcal{S}}(\mathbf{C})$. The task prior therefore defines

$$\psi(W) := \mathbb{E}_{\Pi}[\tau^{\mathcal{S}}(\mathbf{C}) \mid W]. \tag{4}$$

We use $\psi(W)$ as the Bayes estimator under the task prior. It depends only on the observable input W and does not require identifying the latent roles $\mathbf{R}$ or the covariates of $\mathbf{C}$.

CRA-PFN uses a Transformer over the observational context and queries. A shared covariate encoder and separate treatment and outcome encoders produce $p + 2$ tokens per row. The architecture alternates between feature attention and row attention, followed by symmetric mean pooling. The row mask lets each query access only the context and itself, ensuring that its prediction depends only on W_j. No covariate-position or role embeddings are used. Together with the shared covariate encoder and symmetric pooling, these design choices yield covariate-permutation invariance under consistent permutations of the observed covariates. Appendix D.1 provides the architecture details, and Appendix B.4 provides the formal invariance result and proof.

4.3 Arm-Specific Prediction Heads

Each arm head parameterizes a K-component Gaussian mixture model (GMM). For arm t and mixture component k, let $\pi_{tk}(W_j)$, $m_{tk}(W_j)$, and $\sigma_{tk}(W_j) > 0$ denote the mixture weight, component mean, and component standard deviation, respectively. For an arm-label value z, the predictive density and its mean are

$$\begin{aligned} q_{\theta,t}(z \mid W_j) &= \sum_{k=1}^{K} \pi_{tk}(W_j) \mathcal{N}\big(z \mid m_{tk}(W_j), \sigma_{tk}^2(W_j)\big)\,, \\ \widehat{\mu}_t(W_j) &:= \sum_{k=1}^{K} \pi_{tk}(W_j) m_{tk}(W_j). \end{aligned} \tag{5}$$

The mixture models the conditional distribution of each arm label, while its first moment supplies the arm-mean prediction. Taking the difference gives

$$\widehat{\tau}_j = \widehat{\tau}(W_j) := \widehat{\mu}_1(W_j) - \widehat{\mu}_0(W_j). \tag{6}$$

This follows the same two-arm structure as the task-specific CATE in Eq. 3: each head predicts an arm mean, and their contrast yields the CATE prediction.

The training objective combines negative log-likelihood (NLL) and mean squared error (MSE):

$$\widehat{\mathcal{L}}(\theta) = \frac{1}{2|\mathcal{Q}|} \sum_{j\in\mathcal{Q}} \sum_{t=0}^{1} -\log q_{\theta,t}(Z_{t,j} \mid W_j) + \frac{\lambda}{2|\mathcal{Q}|} \sum_{j\in\mathcal{Q}} \sum_{t=0}^{1} \{\widehat{\mu}_t(W_j) - Z_{t,j}\}^2, \tag{7}$$

where $\lambda = 0.5$ is the arm-mean loss weight. NLL fits the arm-label distributions, while MSE directly trains the first moments used to form the CATE prediction. Section 5.2 analyzes how distributional fitting and arm-mean accuracy contribute to the resulting contrast.

5 Theoretical Analysis

Our theoretical analysis distinguishes role-agnostic CATE identifiability from prediction under finite mixed-role observational context. We characterize the population target of the two-arm GMM-MSE objective and establish conditions for consistent CATE prediction as the observational data grow. All proofs are provided in Appendix B.

Throughout, $\tau^{\mathcal{S}}(\mathbf{C})$, $\psi(W_n)$, and $\widehat{\tau}(W_n)$ denote the task-specific CATE, Bayes estimator, and CATE prediction. Unless another distribution is explicitly specified, moment conditions and almost-sure statements are understood with respect to $\mathbb{P}_\Pi$.

5.1 Role-Agnostic CATE Identifiability

We first formalize role-agnostic CATE identifiability over the class of causal processes considered here. Consider two such processes with distributions $\mathbb{P}$ and $\mathbb{P}'$ over $(\mathbf{X}, T, Y)$. Their query CATE functions are $g(\mathbf{x}) := \tau(\mathbf{c}(\mathbf{x}))$, $g'(\mathbf{x}) := \tau'(\mathbf{c}'(\mathbf{x}))$.

Definition 2 (Role-agnostic CATE functional)**.** Let $\mathbb{P}_{\mathbf{X}}$ denote the $\mathbf{X}$-marginal of $\mathbb{P}$. A functional Γ from observational distributions to CATE functions is called role-agnostic if

$$g(\mathbf{x}) = \Gamma(\mathbb{P})(\mathbf{x}), \qquad \mathbb{P}_{\mathbf{X}}\text{-a.e.}, \tag{8}$$

for every admissible causal process in the task class.

Proposition 1 (Role-agnostic CATE identification)**.** *Suppose Assumption 1 holds. A role-agnostic CATE functional exists if and only if*

$$\mathbb{P} = \mathbb{P}' \quad \Longrightarrow \quad g(\mathbf{x}) = g'(\mathbf{x}) \quad \mathbb{P}_{\mathbf{X}}\text{-a.e.} \tag{9}$$

Proposition 1 characterizes that the task-specific CATE is identifiable from the observational data whereas observationally equivalent causal processes induce the same query-level CATE function. Importantly, such identification does not require recovering the latent role assignment $\mathbf{R}$, the coordinate identities of $\mathbf{C}$, or the underlying causal structure.

Theorem 2 (CATE non-identifiability lower bound)**.** *Suppose two admissible causal data-generating processes satisfy $\mathbb{P} = \mathbb{P}'$. Let ν denote the common distribution of the observable input and query covariates. For any predictor $\widehat{\tau}(W)$ based only on the observable input, the following bound holds whenever the displayed squared risks are finite*

$$\max\Big\{\mathbb{E}_\nu[\{\widehat{\tau}(W) - g(\mathbf{X})\}^2], \mathbb{E}_\nu[\{\widehat{\tau}(W) - g'(\mathbf{X})\}^2]\Big\} \geq \frac{1}{4}\,\mathbb{E}_\nu[\{g(\mathbf{X}) - g'(\mathbf{X})\}^2]. \tag{10}$$

If $g(\mathbf{X}) \neq g'(\mathbf{X})$ on a set of positive ν-probability, the right-hand side is strictly positive.

Thus, role-agnostic identifiability does not require the latent roles themselves to be identifiable. It only requires that observationally equivalent processes agree on the CATE function. When this condition fails, the lower bound shows that no predictor using only observable input can have arbitrarily small squared error for all observationally equivalent processes.

5.2 Population Characterization of the Training Objective

We next connect the two-arm training objective to the CATE prediction target under a finite observational context. For context size n, let $\mathcal{D}_{\mathrm{obs},n} = \{(\mathbf{X}_i, T_i, Y_i)\}_{i=1}^n$ and $W_n = (\mathcal{D}_{\mathrm{obs},n}, \mathbf{X})$, where the final $\mathbf{X}$ denotes the query covariates. Each $\mathbf{X}_i$ contains the full observed mixed-role covariates. The estimator receives no causal-role labels, causal graph, designated adjustment set, or indicator specifying which covariates belong to $\mathbf{C}$.

Define the Bayes arm estimators by

$$\psi_t(W_n) := \mathbb{E}_\Pi[\mu_t^{\mathcal{S}}(\mathbf{C}) \mid W_n], \qquad t \in \{0,1\}, \tag{11}$$

The corresponding Bayes estimator is $\psi(W_n) := \mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n]$. Linearity of conditional expectation gives $\psi(W_n) = \psi_1(W_n) - \psi_0(W_n)$. Under the arm-label centering condition in Appendix C.6, iterated conditional expectation further gives

$$\mathbb{E}_\Pi[Z_t \mid W_n] = \psi_t(W_n). \tag{12}$$

This identity links the simulated arm-specific supervision to the same arm means whose contrast defines the Bayes estimator. Its derivation is given in Appendix B.2.

Let $f_t(\cdot \mid W_n)$ denote the conditional density of $Z_t \mid W_n$. For a candidate predictive density $q_t(\cdot \mid W_n)$, let $\widehat{\mu}_t(W_n) = \int z q_t(z \mid W_n)\, dz$. The population risk is

$$\mathcal{R}_n(q_0, q_1) := \frac{1}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[-\log q_t(Z_t \mid W_n)] + \frac{\lambda}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - Z_t\}^2]. \tag{13}$$

Let $\mathcal{R}_n^\star$ denote the infimum of this risk over measurable conditional densities with finite first moments and population risk.

Theorem 3 (Population objective alignment)**.** *Under Assumption 4, suppose the true conditional densities exist with respect to a common dominating measure, the candidate densities have finite first moments, and all terms below are finite. Then*

$$\begin{aligned}\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star = \frac{1}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[D_{\mathrm{KL}}(f_t(\cdot \mid W_n) \| q_t(\cdot \mid W_n))] \\ + \frac{\lambda}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - \psi_t(W_n)\}^2].\end{aligned} \tag{14}$$

The decomposition shows that the population objective jointly controls conditional arm-label distribution fitting and Bayes arm-mean prediction. Therefore the two-arm objective targets the same arm means whose contrast is $\psi(W_n)$.

Corollary 4 (Contrast control)**.** *For* $\widehat{\tau}(W_n) = \widehat{\mu}_1(W_n) - \widehat{\mu}_0(W_n)$*,*

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] \le \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\}. \tag{15}$$

Thus, small population excess risk controls the squared discrepancy between the estimator contrast and the Bayes estimator. This result concerns finite-context prediction and does not yet imply equality with the task-specific CATE, because W_n can leave residual task uncertainty.

5.3 CATE Consistency under Increasing Mixed-Role Context

We now study whether the residual uncertainty under a finite mixed-role observational context vanishes as the context size increases. The causal task and query are kept fixed while additional observational rows are revealed.

For the asymptotic analysis, let $\{(\mathbf{X}_i, T_i, Y_i)\}_{i\ge1}$ denote an infinite context sequence for a sampled task $\mathcal{S}$. For each n, define $\mathcal{D}_{\mathrm{obs},n} = \{(\mathbf{X}_i, T_i, Y_i)\}_{i=1}^n$ and $W_n = (\mathcal{D}_{\mathrm{obs},n}, \mathbf{X})$. The query covariates $\mathbf{X}$ are drawn once from the corresponding $\mathbf{X}$-marginal and are kept fixed as n increases. Hence $\sigma(W_n) \subseteq \sigma(W_{n+1})$. Because the task and query remain unchanged, $\tau^{\mathcal{S}}(\mathbf{C})$ is the same causal target for all n.

Assumption 2 (Conditional i.i.d. context sampling)**.** Conditional on task $\mathcal{S}$, the context rows $(\mathbf{X}_i, T_i, Y_i)$ are independently and identically distributed according to the task-specific observational distribution $\mathbb{P}_{\text{obs}}$.

Assumption 3 (Finite second moment)**.** The task-specific CATE satisfies $\mathbb{E}_\Pi[\{\tau^{\mathcal{S}}(\mathbf{C})\}^2] < \infty$.

The second condition ensures that the conditional expectations and squared prediction errors below are well defined in L^2.

Theorem 5 (Consistency of the Bayes estimator)**.** *Suppose a role-agnostic CATE functional Γ exists as in Definition 2, and the evaluation map $(\mathbb{P}, \mathbf{x}) \mapsto \Gamma(\mathbb{P})(\mathbf{x})$ is measurable. Under Assumptions 2 and 3, $\psi(W_n) \longrightarrow \tau^{\mathcal{S}}(\mathbf{C})$ almost surely and in mean square as $n \to \infty$. In particular,*

$$\mathbb{E}_\Pi[\{\psi(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \longrightarrow 0. \tag{16}$$

Theorem 5 shows that observing the latent roles is unnecessary for asymptotic CATE prediction when the observational distribution determines the CATE function. As the mixed-role context tends to infinity, the residual uncertainty of the Bayes estimator vanishes and it therefore converges to the task-specific CATE.

Corollary 6 (Monotone decrease of residual uncertainty)**.** *Under the increasing-context construction and Assumption 3,*

$$\mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]$$

is nonincreasing in n. Under the conditions of Theorem 5, it converges to zero.

Appendix B.3 gives the corresponding one-step variance identity.

Theorem 7 (Finite-context CATE prediction error)**.** *Under Assumption 3, let $\widehat{\tau}(W_n)$ be any W_n-measurable predictor satisfying $\mathbb{E}_\Pi[\{\widehat{\tau}(W_n)\}^2] < \infty$. Then, for any finite n,*

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] = \mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] + \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \tag{17}$$

For the two-arm predictor $\widehat{\tau}(W_n) = \widehat{\mu}_1(W_n) - \widehat{\mu}_0(W_n)$ induced by q_0 and q_1, under the conditions of Theorem 3,

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \leq \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\} + \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \tag{18}$$

The decomposition separates finite-context CATE error into two components: prediction error relative to the Bayes estimator and residual uncertainty about the task-specific CATE given W_n. The former is controlled by the training excess risk through Corollary 4, and the latter decreases with additional observational context through Corollary 6.

Corollary 8 (Mean-square consistency of CATE prediction)**.** *For the two-arm predictor $\widehat{\tau}(W_n) = \widehat{\mu}_1(W_n) - \widehat{\mu}_0(W_n)$, suppose the conditions of Theorems 3 and 5 hold. If $\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star \longrightarrow 0$, then*

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \longrightarrow 0. \tag{19}$$

Hence, if the learned predictor approaches the population optimum and the residual uncertainty from the finite observational context vanishes, CRA-PFN is consistent.

6 Experiments

We organize the empirical evaluation around three research questions:

- **RQ1:** Can CRA-PFN accurately predict task-specific CATEs across benchmarks without causal-role labels, a causal graph, or a designated adjustment set?
- **RQ2:** How does CRA-PFN generalize under OOD-SCM mechanism shifts?
- **RQ3:** How do the role-heterogeneous pretraining prior and prediction-head design affect CATE prediction across evaluation settings?

We report additional results about pre-treatment-only evaluation, covariate-permutation diagnostic of CRA-PFN, and running time in Appendix E.6, E.8, and E.9, respectively.

Table 1: PEHE on benchmarks from prior work, structural variants, and Synthetic SCM (lower is better).

(a) Benchmarks from prior work (All covariates are pre-treatment covariates)

Method	CausalFM-BD	HMBART	IHDP	Twins	Jobs	Avg. Rank
CRA-PFN	**0.451 ± 0.05**	**0.940 ± 0.19**	2.404 ± 0.11	0.326	0.523	**2.40**
CausalFM	0.459 ± 0.05	1.394 ± 0.23	5.719 ± 0.21	**0.324**	**0.478**	4.20
CausalPFN	0.454 ± 0.05	1.026 ± 0.20	**0.567 ± 0.02**	0.330	0.557	3.60
Do-PFN	0.530 ± 0.06	1.648 ± 0.44	6.573 ± 0.30	0.329	0.496	6.60
CF-DML	0.462 ± 0.05	1.155 ± 0.17	3.728 ± 0.17	0.325	0.546	4.60
ForestDR	0.471 ± 0.05	1.136 ± 0.16	3.342 ± 0.17	0.325	0.527	3.80
S-Learner	0.491 ± 0.05	1.288 ± 0.20	2.926 ± 0.14	0.326	0.640	5.40
T-Learner	0.664 ± 0.05	1.289 ± 0.23	2.408 ± 0.11	0.343	0.865	7.40
X-Learner	0.558 ± 0.05	1.186 ± 0.21	3.418 ± 0.16	0.332	0.657	7.00

(b) Structural variants and Synthetic SCM

Method	IHDP-C	IHDP-M	IHDP-MC	Twins-C	Twins-M	Twins-MC	Synthetic SCM	Avg. Rank
CRA-PFN	2.376 ± 0.11	**0.242 ± 0.00**	**0.250 ± 0.00**	0.328	**0.095**	**0.080**	**2.609 ± 0.52**	**1.29**
CausalFM	5.622 ± 0.20	0.335 ± 0.01	0.346 ± 0.01	**0.324**	0.220	0.385	4.141 ± 0.65	5.43
CausalPFN	**0.692 ± 0.02**	0.343 ± 0.01	0.351 ± 0.01	0.338	0.210	0.344	3.646 ± 0.74	4.57
Do-PFN	6.518 ± 0.29	0.347 ± 0.01	0.366 ± 0.01	0.340	0.122	0.524	4.322 ± 0.76	6.86
CF-DML	3.763 ± 0.17	0.312 ± 0.01	0.316 ± 0.01	0.330	0.118	0.202	3.658 ± 0.74	4.00
ForestDR	3.512 ± 0.17	0.301 ± 0.01	0.305 ± 0.01	0.329	0.122	0.142	3.702 ± 0.75	3.57
S-Learner	3.161 ± 0.15	0.369 ± 0.01	0.372 ± 0.01	0.329	0.126	0.115	4.003 ± 0.75	4.71
T-Learner	2.593 ± 0.12	0.693 ± 0.01	0.679 ± 0.01	0.354	0.366	0.262	6.359 ± 1.09	7.71
X-Learner	3.474 ± 0.15	0.532 ± 0.01	0.523 ± 0.01	0.340	0.259	0.172	4.848 ± 0.91	6.86

6.1 Experimental Setup

Datasets and baselines. We evaluate CRA-PFN on benchmarks from prior work, including CausalFM-BD (Ma et al., 2026), HMBART (Liu et al., 2026), IHDP (Hill, 2011), Twins (Louizos et al., 2017), and Jobs (Ma et al., 2026), together with structural variants, a separately generated Synthetic SCM benchmark, four OOD-SCM families, and pre-treatment-only inputs. Comparisons include CausalFM (Ma et al., 2026), CausalPFN (Balazadeh Meresht et al., 2025), Do-PFN (Robertson et al., 2025), CausalForestDML (CF-DML) (Chernozhukov et al., 2018; Wager & Athey, 2018), ForestDR (Kennedy, 2023), and the S-, T-, and X-Learners (Künzel et al., 2019). PFN-based methods are evaluated without benchmark-specific parameter updates, whereas the task-specific estimators are fitted separately to each evaluation task. Appendix E provides details.

Evaluation. We measure prediction error using precision in estimation of heterogeneous effect (PEHE) over query set $\mathcal{Q}$, defined as $\text{PEHE} := \sqrt{\frac{1}{|\mathcal{Q}|}\sum_{j\in\mathcal{Q}}(\tau_j - \widehat{\tau}_j)^2}$, where τ_j is the ground-truth CATE and $\widehat{\tau}_j$ is the predicted CATE for query j. Lower is better. We report task-level means and standard errors, with point estimates for single-task benchmarks.

6.2 RQ1: CATE Prediction across Benchmarks

Table 1 evaluates CATE prediction on benchmarks from prior work, structural variants, and a separately generated Synthetic SCM benchmark. Panel (a) reports five benchmarks in the forms used in prior work. CRA-PFN achieves an average rank of 2.40 and the lowest PEHE on CausalFM back-door (CausalFM-BD) and Heterogeneous Mediation with Bayesian Additive Regression Trees (HMBART). Performance varies across individual benchmarks. CausalPFN achieves the lowest PEHE on IHDP, while CausalFM achieves the lowest PEHE on Twins and Jobs.

Panel (b) reports mediator and collider variants of IHDP and Twins, together with a separately generated Synthetic SCM benchmark. CRA-PFN achieves an average rank of 1.29 and the lowest PEHE on five of the seven benchmarks in Panel (b). These results show that CRA-PFN can predict task-specific CATEs from mixed-role observational inputs without causal-role labels, a causal graph, or a designated adjustment set. In the pre-treatment-only evaluation, inputs are restricted to the task-

Table 2: PEHE on four OOD-SCM families for all compared methods (lower is better). OOD-SCM Rank is the average rank across the four families computed from unrounded means.

Method	Fourier	GP-RFF	Heavy-tailed	Piecewise	OOD-SCM Rank
CRA-PFN	**1.262 ± 0.05**	**1.026 ± 0.04**	**0.872 ± 0.05**	**0.940 ± 0.05**	**1.00**
CausalFM	1.277 ± 0.05	1.233 ± 0.07	1.037 ± 0.06	1.105 ± 0.07	3.00
CausalPFN	1.526 ± 0.07	1.470 ± 0.13	1.216 ± 0.09	1.201 ± 0.12	7.25
Do-PFN	1.297 ± 0.05	1.108 ± 0.04	1.129 ± 0.07	1.164 ± 0.10	3.75
CF-DML	1.460 ± 0.07	1.453 ± 0.13	1.194 ± 0.09	1.081 ± 0.11	5.25
ForestDR	1.349 ± 0.05	1.319 ± 0.09	1.103 ± 0.06	1.005 ± 0.07	3.75
S-Learner	1.318 ± 0.05	1.316 ± 0.08	1.134 ± 0.07	1.087 ± 0.06	4.25
T-Learner	1.644 ± 0.06	1.656 ± 0.08	1.350 ± 0.07	1.247 ± 0.07	9.00
X-Learner	1.561 ± 0.07	1.608 ± 0.12	1.259 ± 0.08	1.180 ± 0.09	7.75

Table 3: Ablation ranks in the expanded pool of four variants and eight external baselines. Lower is better. Macro Avg. weights the evaluation settings equally.

Variant	Prior Work	Struct. + SCM	OOD-SCM	Pre-Tr. Only	Macro Avg.
CRA-PFN (GMM-MSE)	3.60	**2.29**	1.75	4.14	**2.94**
CRA-PFN (Arm-MSE)	6.00	3.71	4.50	4.00	4.55
CRA-PFN (Pre-treatment Prior)	**2.20**	3.29	8.50	**1.86**	3.96
CRA-PFN (Direct-CATE)	6.80	2.86	**1.25**	4.57	3.87

specific valid pre-treatment covariates while the benchmark CATEs and evaluation splits remain unchanged. CRA-PFN retains the best average rank of 2.29. Appendix E.6 provides detailed results.

6.3 RQ2: OOD-SCM Mechanism Shifts

Table 2 evaluates CRA-PFN under four OOD-SCM families different from the prior SCM. CRA-PFN achieves the lowest mean PEHE among all compared methods in all four families, with an average rank of 1.00. These results show that CRA-PFN generalizes effectively across the evaluated OOD-SCM mechanism shifts. Please refer to Appendix E.5 for detailed mechanism setting.

6.4 RQ3: Pretraining Prior and Prediction Heads

Table 3 compares four validation-selected variants. Pre-treatment Prior disables non-pre-treatment roles during pretraining, Arm-MSE uses deterministic arm-mean heads, and Direct-CATE uses a single effect-distribution head. GMM-MSE achieves the best macro-average rank of 2.94 and outperforms Arm-MSE in three of four settings. Pre-treatment Prior performs best on the benchmarks from prior work and with pre-treatment-only inputs, while Direct-CATE performs best under OOD-SCM. Overall, these results support the role-heterogeneous prior and GMM-MSE for role-agnostic CATE prediction from mixed-role observational inputs. Appendix D.4 provides details.

7 Conclusion

We introduce CRA-PFN for CATE prediction from observational data with mixed and unobserved covariate roles. By pretraining over a role-heterogeneous SCM prior, CRA-PFN directly predicts task-specific conditional average treatment effects without requiring causal-role labels or a prespecified adjustment set. Theoretically, we establish role-agnostic CATE identifiability and show that, when the observational distribution determines the task-specific CATE and the learned predictor approaches the population optimum, CATE prediction is consistent as the observational data grows. Across standard benchmarks, synthetic SCMs, and unseen mechanism families, CRA-PFN performs effectively and remains robust under substantial changes in covariate roles and causal mechanisms. Together, these results suggest that explicit recovery of covariate roles is not always necessary for amortized treatment-effect prediction.

REFERENCES

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: 10.3982/ECTA15732.

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

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, 2025.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A python package for ml-based heterogeneous treatment effects estimation. `https://github.com/py-why/EconML`, 2019. Version 0.15.0.

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*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

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

Samuel Dooley, Gurnoor Singh Khurana, Chirag Mohapatra, Siddartha V Naidu, and Colin White. Forecastpfn: Synthetically-trained zero-shot forecasting. *Advances in Neural Information Processing Systems*, 36:2403–2426, 2023.

Paul B. Ellickson, Wreetabrata Kar, and James C. Reeder. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 42(4):704–728, 2023. doi: 10.1287/mksc.2022.1401.

Juntao Fang, Shifeng Xie, Ruichu Cai, Shengji Zheng, Zijian Li, Keli Zhang, Lujia Pan, Themis Palpanas, and Zhifeng Hao. Chorustic: Training-free multivariate time series classification via chorus in-context learning. *arXiv preprint arXiv:2608.24033*, 2026.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4):958–968, 2024. doi: 10.1038/s41591-024-02902-1.

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

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. URL `https://openreview.net/forum?id=ULoLF1aOo1`.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):579–599, 2022.

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

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*, 2023.

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

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

Christoph Kern, Michael P Kim, and Angela Zhou. Multi-accurate cate is robust to unknown covariate shifts. *Transactions on Machine Learning Research*, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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.

Haoxuan Li, Yue Liu, Zhi Geng, and Kun Zhang. A local method for satisfying interventional fairness with partially known causal graphs. *Advances in Neural Information Processing Systems*, 37:135415–135436, 2024.

Chen Liu, Xu Qin, Victor B Talisa, and Jiebiao Wang. Heterogeneous causal mediation analysis using bayesian additive regression trees. *Biometrics*, 82(2):ujag079, 2026.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Craig T Ramey, Donna M Bryant, Barbara H Wasik, Joseph J Sparling, Kaye H Fendt, and Lisa M La Vange. Infant health and development program for low birth weight, premature infants: Program elements, family participation, and child intelligence. *Pediatrics*, 89(3):454–465, 1992.

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, 2025.

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

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT Press, 2001.

Christopher Tran and Elena Zheleva. Improving data-driven heterogeneous treatment effect estimation under structure uncertainty. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1787–1797, 2022.

Julius Vetter, Manuel Gloeckler, Daniel Gedon, and Jakob H Macke. Effortless, simulation-efficient bayesian inference using tabular foundation models. *Advances in Neural Information Processing Systems*, 38:82685–82724, 2025.

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

AI USE STATEMENT

Generative AI tools were used for language polishing, LaTeX editing, and consistency checks. The authors reviewed all AI-assisted edits, executed the reported experiments, checked cited sources, and take responsibility for the final paper.

A NOTATIONAL DESCRIPTION

Table 4 summarizes the main notation. Task causal quantities omit the superscript $\mathcal{S}$ in Section 3. Notation specific to the generator, architecture, and proofs is defined locally where it is used.

Symbol	Meaning and first definition
$\mathbf{X}, T, Y$	Observed covariates, treatment, and factual outcome.
$\mathbf{C}, \mathbf{R}$	Valid pre-treatment covariates and latent role assignment.
$\mathcal{D}_{\text{obs}}, \mathcal{Q}$	Observational context and query index set.
W_j, W, W_n	Query-specific, generic, and context-size-indexed prediction inputs.
$\mathcal{S}, \Pi$	Sampled task and pretraining task prior.
$\mathbb{P}_\Pi, \mathbb{E}_\Pi$	Induced joint distribution and its expectation.
$\mu_t^{\mathcal{S}}, \tau^{\mathcal{S}}$	True task-specific arm means and CATE.
$Z_{t,j}$	Simulated arm-specific supervision label.
$\psi(W), \psi(W_n)$	Bayes estimator for generic and context-size-indexed inputs.
$q_{\theta,t}$	Predictive density for arm labels.
$\widehat{\mu}_t, \widehat{\tau}$	Arm-mean and CATE predictions.
$\psi_t(W_n)$	Bayes arm estimator.
$\mathcal{R}_n, \mathcal{R}_n^\star$	Population risk and unrestricted optimum.
τ_j	Benchmark CATE.

Table 4: Main notation used in CRA-PFN and its analysis.

The remaining appendices proceed from theory to implementation and evaluation. Appendix B gives the theoretical proofs, Appendix C specifies the SCM prior and supervision construction, Appendix D gives the estimator and training implementation, and Appendix E documents benchmark construction and evaluation.

B THEORETICAL PROOFS

Unless another distribution is specified, expectations, variances, and moment conditions in these proofs are under $\mathbb{P}_\Pi$.

B.1 ROLE-AGNOSTIC CATE IDENTIFIABILITY

Proof of Proposition 1. **Necessity.** Suppose a role-agnostic CATE functional Γ exists. For two admissible processes satisfying $\mathbb{P} = \mathbb{P}'$, Definition 2 gives

$$g(\mathbf{x}) = \Gamma(\mathbb{P})(\mathbf{x}), \qquad g'(\mathbf{x}) = \Gamma(\mathbb{P}')(\mathbf{x}).$$

Because $\mathbb{P} = \mathbb{P}'$, $\Gamma(\mathbb{P}) = \Gamma(\mathbb{P}')$. Hence

$$g(\mathbf{x}) = g'(\mathbf{x}) \qquad \mathbb{P}_{\mathbf{X}}\text{-a.e.},$$

which proves necessity.

Sufficiency. Suppose Eq. 9 holds. For any observational distribution $\mathbb{P}$ induced by the task class, all admissible processes that induce $\mathbb{P}$ have the same query CATE function $\mathbb{P}_{\mathbf{X}}$-almost everywhere. We can therefore define

$$\Gamma(\mathbb{P})(\mathbf{x}) := g(\mathbf{x}), \qquad \mathbb{P}_{\mathbf{X}}\text{-a.e.}$$

where g is the query CATE function of any admissible process inducing $\mathbb{P}$. Equation 9 ensures that this definition does not depend on which admissible process induces $\mathbb{P}$. Thus Γ is a well-defined role-agnostic CATE functional. □

Proof of Theorem 2. To work with finite squared risks, it is sufficient to assume $\mathbb{E}_\nu[g(\mathbf{X})^2] < \infty$, $\mathbb{E}_\nu[g'(\mathbf{X})^2] < \infty$, and $\mathbb{E}_\nu[\widehat{\tau}(W)^2] < \infty$. These finite-second-moment conditions ensure integrability of the squared-error terms below.

Step 1: Pointwise squared-error inequality. For the two query CATE functions g and g',

$$\begin{aligned}
&\{\widehat{\tau}(W) - g(\mathbf{X})\}^2 + \{\widehat{\tau}(W) - g'(\mathbf{X})\}^2 \\
&\quad = 2\left\{\widehat{\tau}(W) - \frac{g(\mathbf{X}) + g'(\mathbf{X})}{2}\right\}^2 + \frac{1}{2}\{g(\mathbf{X}) - g'(\mathbf{X})\}^2 \\
&\quad \geq \frac{1}{2}\{g(\mathbf{X}) - g'(\mathbf{X})\}^2.
\end{aligned} \tag{20}$$

Step 2: Take expectation under the common distribution. Taking expectation of Eq. 20 under the common distribution ν of the observable input and query covariates gives

$$\mathbb{E}_\nu[\{\widehat{\tau}(W) - g(\mathbf{X})\}^2] + \mathbb{E}_\nu[\{\widehat{\tau}(W) - g'(\mathbf{X})\}^2] \geq \frac{1}{2}\mathbb{E}_\nu[\{g(\mathbf{X}) - g'(\mathbf{X})\}^2]. \tag{21}$$

Step 3: Pass from the sum to the larger risk. The maximum of the two nonnegative squared risks in Eq. 21 is at least one half of their sum. Therefore

$$\max\Big\{\mathbb{E}_\nu[\{\widehat{\tau}(W) - g(\mathbf{X})\}^2], \mathbb{E}_\nu[\{\widehat{\tau}(W) - g'(\mathbf{X})\}^2]\Big\} \geq \frac{1}{4}\mathbb{E}_\nu[\{g(\mathbf{X}) - g'(\mathbf{X})\}^2]. \tag{22}$$

Equation 22 is Eq. 10.

Step 4: Strict positivity under non-identifiability. If $g(\mathbf{X}) \neq g'(\mathbf{X})$ on a set with positive ν-probability, then $\{g(\mathbf{X}) - g'(\mathbf{X})\}^2$ is positive with positive probability. Therefore the expectation on the right-hand side of Eq. 22 is strictly positive. Hence at least one of the two squared prediction risks is bounded away from zero. □

B.2 Population Characterization of the Training Objective

Proof of Theorem 3. **Step 1: Identify the conditional arm-label mean.** By Assumption 4 and iterated conditional expectation,

$$\begin{aligned}
\mathbb{E}_\Pi[Z_t \mid W_n] &= \mathbb{E}_\Pi\left[\mathbb{E}_\Pi[Z_t \mid \mathcal{S}, \mathbf{C}, W_n] \mid W_n\right] \\
&= \mathbb{E}_\Pi[\mu_t^{\mathcal{S}}(\mathbf{C}) \mid W_n] \\
&= \psi_t(W_n).
\end{aligned} \tag{23}$$

Equation 23 establishes Eq. 12.

Step 2: Decompose the negative log-likelihood term. Conditionally on W_n, the cross-entropy identity gives

$$\begin{aligned}
&\mathbb{E}_\Pi[-\log q_t(Z_t \mid W_n) \mid W_n] \\
&\quad = \mathbb{E}_\Pi[-\log f_t(Z_t \mid W_n) \mid W_n] + D_{\mathrm{KL}}(f_t(\cdot \mid W_n) \| q_t(\cdot \mid W_n)).
\end{aligned} \tag{24}$$

Taking expectation of Eq. 24 over W_n gives

$$\begin{aligned}
&\mathbb{E}_\Pi[-\log q_t(Z_t \mid W_n)] \\
&\quad = \mathbb{E}_\Pi[-\log f_t(Z_t \mid W_n)] + \mathbb{E}_\Pi[D_{\mathrm{KL}}(f_t(\cdot \mid W_n) \| q_t(\cdot \mid W_n))].
\end{aligned} \tag{25}$$

The first term on the right-hand side of Eq. 25 does not depend on the candidate density.

Step 3: Decompose the arm-mean squared-error term. The arm-mean prediction error satisfies

$$\widehat{\mu}_t(W_n) - Z_t = \{\widehat{\mu}_t(W_n) - \psi_t(W_n)\} - \{Z_t - \psi_t(W_n)\}. \tag{26}$$

Equation 23 implies

$$\mathbb{E}_\Pi[Z_t - \psi_t(W_n) \mid W_n] = 0. \tag{27}$$

Squaring Eq. 26, taking expectation, and using Eq. 27 gives

$$\mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - Z_t\}^2] = \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - \psi_t(W_n)\}^2] + \mathbb{E}_\Pi[\{Z_t - \psi_t(W_n)\}^2]. \tag{28}$$

The second term on the right-hand side of Eq. 28 does not depend on the candidate density.

Step 4: Characterize the unrestricted population optimum. When $q_t = f_t$, the KL divergence in Eq. 25 is zero. The first moment of $f_t(\cdot \mid W_n)$ is the conditional arm-label mean in Eq. 23, so

$$\widehat{\mu}_t(W_n) = \mathbb{E}_\Pi[Z_t \mid W_n] = \psi_t(W_n). \tag{29}$$

Thus the arm-mean excess term in Eq. 28 is also zero when $q_t = f_t$. The true conditional densities are admissible under the theorem assumptions. Therefore the unrestricted infimum $\mathcal{R}_n^\star$ is attained by $q_t = f_t$ for both arms.

Step 5: Subtract the optimum. Substituting Eqs. 25 and 28 into Eq. 13, and subtracting the candidate-independent optimum characterized by Eq. 29, gives

$$\begin{aligned}\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star = \frac{1}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[D_{\mathrm{KL}}(f_t(\cdot \mid W_n) \| q_t(\cdot \mid W_n))] \\ + \frac{\lambda}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - \psi_t(W_n)\}^2].\end{aligned} \tag{30}$$

Equation 30 is Eq. 14. □

Proof of Corollary 4. **Step 1: Express the contrast error through the two arm errors.** Linearity of conditional expectation gives $\psi(W_n) = \psi_1(W_n) - \psi_0(W_n)$. Together with $\widehat{\tau}(W_n) = \widehat{\mu}_1(W_n) - \widehat{\mu}_0(W_n)$, this gives

$$\widehat{\tau}(W_n) - \psi(W_n) = \{\widehat{\mu}_1(W_n) - \psi_1(W_n)\} - \{\widehat{\mu}_0(W_n) - \psi_0(W_n)\}. \tag{31}$$

Step 2: Bound the squared contrast error. Applying $(x - y)^2 \le 2x^2 + 2y^2$ directly to Eq. 31 gives

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] \le 2\sum_{t=0}^{1} \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - \psi_t(W_n)\}^2]. \tag{32}$$

Step 3: Control the arm errors by population excess risk. Theorem 3 and nonnegativity of KL divergence imply

$$\frac{\lambda}{2}\sum_{t=0}^{1} \mathbb{E}_\Pi[\{\widehat{\mu}_t(W_n) - \psi_t(W_n)\}^2] \le \mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star. \tag{33}$$

Combining Eqs. 32 and 33 gives

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] \le \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\}, \tag{34}$$

which is Eq. 15. □

B.3 CATE Consistency under Increasing Mixed-Role Context

For the increasing-context analysis, define the observed information sets

$$\mathcal{F}_n := \sigma(W_n), \qquad \mathcal{F}_\infty := \sigma\left(\bigcup_{n \ge 1} \mathcal{F}_n\right). \tag{35}$$

The Bayes estimator in Section 5.2 can therefore be written as

$$\psi(W_n) = \mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) \mid \mathcal{F}_n]. \tag{36}$$

Under Assumption 3, $\tau^{\mathcal{S}}(\mathbf{C}) \in L^2$. The defining property of the conditional expectation in Eq. 36 gives the residual orthogonality identity

$$\mathbb{E}_\Pi\left[\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n) \mid \mathcal{F}_n\right] = 0. \tag{37}$$

The same conditional expectation also gives

$$\mathbb{E}_\Pi[\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n)\}^2] = \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \tag{38}$$

Equations 37 and 38 require only the finite second moment in Assumption 3. They do not require role-agnostic CATE identifiability or an infinite context.

Proof of Theorem 5. **Step 1: The infinite context determines the task-specific observational distribution.** Conditional on the sampled task $\mathcal{S}$, Assumption 2 states that $(\mathbf{X}_i, T_i, Y_i)_{i\geq 1}$ are i.i.d. from the task-specific observational distribution $\mathbb{P}_{\text{obs}}$. Because the covariate dimension p is fixed and observed within each task, the argument below is applied conditional on the realized task dimension p. For that realized p, define the empirical distribution

$$\widehat{\mathbb{P}}_{\text{obs},n} := \frac{1}{n}\sum_{i=1}^{n} \delta_{(\mathbf{X}_i, T_i, Y_i)}. \tag{39}$$

For the realized task dimension p, the observation space $\mathbb{R}^p \times \{0,1\} \times \mathbb{R}$ is Polish. Hence the empirical distributions in Eq. 39 satisfy

$$\widehat{\mathbb{P}}_{\text{obs},n} \Longrightarrow \mathbb{P}_{\text{obs}} \qquad \text{almost surely as } n \to \infty, \tag{40}$$

where $\Longrightarrow$ denotes weak convergence. Each $\widehat{\mathbb{P}}_{\text{obs},n}$ is measurable with respect to $\mathcal{F}_n$, and therefore with respect to $\mathcal{F}_\infty$. The space of Borel probability measures on the observation space is metrizable under the topology of weak convergence. Thus the almost-sure limit in Eq. 40, namely $\mathbb{P}_{\text{obs}}$, is $\mathcal{F}_\infty$-measurable.

Step 2: Role-agnostic CATE identifiability makes the task-specific CATE measurable from the limiting observable information. For the sampled task, Definition 2 gives

$$\tau^{\mathcal{S}}(\mathbf{C}) = \Gamma(\mathbb{P}_{\text{obs}})(\mathbf{X}) \qquad \text{almost surely for the sampled query.} \tag{41}$$

The almost-sure statement in Eq. 41 follows because the query covariates $\mathbf{X}$ are drawn from the corresponding $\mathbf{X}$-marginal. Equation 40 and the measurability argument in Step 1 show that $\mathbb{P}_{\text{obs}}$ is $\mathcal{F}_\infty$-measurable. The query covariates $\mathbf{X}$ are included in every W_n, so they are also $\mathcal{F}_\infty$-measurable. The measurability condition in Theorem 5, together with Eq. 41, therefore implies that $\tau^{\mathcal{S}}(\mathbf{C})$ is $\mathcal{F}_\infty$-measurable.

Step 3: Apply convergence of conditional expectations. Equation 35 gives $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. Under Assumption 3, $\tau^{\mathcal{S}}(\mathbf{C}) \in L^2$. Therefore, the convergence theorem for conditional expectations along increasing sigma-fields, applied to Eq. 36, gives

$$\psi(W_n) = \mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) \mid \mathcal{F}_n] \longrightarrow \mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) \mid \mathcal{F}_\infty] \tag{42}$$

almost surely and in L^2.

By Step 2, $\tau^{\mathcal{S}}(\mathbf{C})$ is $\mathcal{F}_\infty$-measurable. Hence

$$\mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) \mid \mathcal{F}_\infty] = \tau^{\mathcal{S}}(\mathbf{C}). \tag{43}$$

Substituting Eq. 43 into Eq. 42 yields

$$\psi(W_n) \longrightarrow \tau^{\mathcal{S}}(\mathbf{C}) \qquad \text{almost surely and in } L^2. \tag{44}$$

The L^2 part of Eq. 44 is exactly the mean-square convergence stated in Eq. 16. □

Proof of Corollary 6. The residual at context size n satisfies

$$\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n) = \{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_{n+1})\} + \{\psi(W_{n+1}) - \psi(W_n)\}. \tag{45}$$

Because $\psi(W_{n+1}) - \psi(W_n)$ is $\mathcal{F}_{n+1}$-measurable, Eq. 37 at context size $n+1$ gives

$$\begin{aligned} &\mathbb{E}_\Pi\Big[\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_{n+1})\}\{\psi(W_{n+1}) - \psi(W_n)\}\Big] \\ &\quad = \mathbb{E}_\Pi\Big[\{\psi(W_{n+1}) - \psi(W_n)\}\mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_{n+1}) \mid \mathcal{F}_{n+1}]\Big] = 0. \end{aligned} \tag{46}$$

Squaring Eq. 45, taking expectation, and using Eq. 46 gives

$$\begin{aligned} &\mathbb{E}_\Pi[\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n)\}^2] \\ &\quad = \mathbb{E}_\Pi[\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_{n+1})\}^2] + \mathbb{E}_\Pi[\{\psi(W_{n+1}) - \psi(W_n)\}^2]. \end{aligned} \tag{47}$$

Applying Eq. 38 at context sizes n and $n+1$ to Eq. 47 gives the one-step variance identity

$$\begin{aligned} &\mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)] \\ &\quad = \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_{n+1})] + \mathbb{E}_\Pi[\{\psi(W_{n+1}) - \psi(W_n)\}^2]. \end{aligned} \tag{48}$$

The final term in Eq. 48 is nonnegative, so the expected residual conditional variance is nonincreasing in n. Under the conditions of Theorem 5, Eq. 38 together with Eq. 16 gives

$$\mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)] \longrightarrow 0. \tag{49}$$

This proves Corollary 6. □

Proof of Theorem 7. Let $\widehat{\tau}(W_n)$ be any square-integrable W_n-measurable predictor. The prediction error satisfies

$$\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C}) = \{\widehat{\tau}(W_n) - \psi(W_n)\} - \{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n)\}. \tag{50}$$

Because $\widehat{\tau}(W_n) - \psi(W_n)$ is $\mathcal{F}_n$-measurable, Eq. 37 gives

$$\begin{aligned} &\mathbb{E}_\Pi\Big[\{\widehat{\tau}(W_n) - \psi(W_n)\}\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n)\}\Big] \\ &\quad = \mathbb{E}_\Pi\Big[\{\widehat{\tau}(W_n) - \psi(W_n)\}\mathbb{E}_\Pi[\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n) \mid \mathcal{F}_n]\Big] = 0. \end{aligned} \tag{51}$$

Squaring Eq. 50, taking expectation, and using Eq. 51 gives

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] = \mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] + \mathbb{E}_\Pi[\{\tau^{\mathcal{S}}(\mathbf{C}) - \psi(W_n)\}^2]. \tag{52}$$

Applying Eq. 38 to the final term in Eq. 52 gives

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] = \mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] + \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \tag{53}$$

Equation 53 is Eq. 17.

For the two-arm predictor induced by q_0 and q_1, Corollary 4 gives

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \psi(W_n)\}^2] \le \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\}. \tag{54}$$

Substituting Eq. 54 into Eq. 53 yields

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \le \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\} + \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \tag{55}$$

Equation 55 is Eq. 18. □

Proof of Corollary 8. For the two-arm predictor induced by q_0 and q_1, Eq. 18 gives

$$\begin{aligned} 0 &\le \mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \\ &\le \frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\} + \mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)]. \end{aligned} \tag{56}$$

By the population excess-risk condition in Corollary 8,

$$\frac{4}{\lambda}\{\mathcal{R}_n(q_0, q_1) - \mathcal{R}_n^\star\} \longrightarrow 0. \tag{57}$$

Under Theorem 5, Eq. 38 and Eq. 16 give

$$\mathbb{E}_\Pi[\mathrm{Var}_\Pi(\tau^{\mathcal{S}}(\mathbf{C}) \mid W_n)] \longrightarrow 0. \tag{58}$$

Applying Eqs. 57 and 58 to Eq. 56 yields

$$\mathbb{E}_\Pi[\{\widehat{\tau}(W_n) - \tau^{\mathcal{S}}(\mathbf{C})\}^2] \longrightarrow 0. \tag{59}$$

Equation 59 is Eq. 19. □

B.4 Covariate-Permutation Invariance

Proposition 9 (Covariate-permutation invariance). *Let $\mathbf{P}_{\mathrm{cov}}$ consistently permute the p covariates in every context and query row while leaving treatment and outcome fixed. Suppose context-fitted preprocessing is applied covariatewise and commutes with this permutation. The covariate encoder and all per-covariate parameters are shared. No covariate-position, role, or graph embedding is used, and pooling is symmetric. Then, for $t \in \{0,1\}$ and $u \in \mathbb{R}$,*

$$q_{\theta,t}(u \mid \mathbf{P}_{\mathrm{cov}} \cdot W) = q_{\theta,t}(u \mid W), \qquad \widehat{\mu}_t(\mathbf{P}_{\mathrm{cov}} \cdot W) = \widehat{\mu}_t(W),$$

and the derived CATE prediction is invariant.

Proof of Proposition 9. We trace the same covariate permutation through preprocessing, Transformer layers, pooling, and the arm heads. Let L_{Tr} be the number of Transformer layers. Let $\mathbf{H}^{(\ell)}(W)$ denote the token representation after layer ℓ, for $\ell = 0, \ldots, L_{\mathrm{Tr}}$. The matrix $\mathbf{P}_{\mathrm{cov}}$ denotes the induced permutation of the complete token axis. Treatment and outcome token positions remain fixed.

Step 1: Preprocess and encode the permuted covariates. By assumption, the context-fitted preprocessing commutes with a consistent permutation of the covariates. Because all covariates use the same scalar encoder and no covariate-position or role embedding is added,

$$\mathbf{H}^{(0)}(\mathbf{P}_{\mathrm{cov}} \cdot W) = \mathbf{P}_{\mathrm{cov}} \mathbf{H}^{(0)}(W). \tag{60}$$

Thus a covariate permutation changes only the order of the encoded covariate tokens.

Step 2: Apply feature attention. For a representation $\mathbf{H}$, let $\mathbf{A}_Q, \mathbf{A}_K, \mathbf{A}_V$ denote the query, key, and value projection matrices, respectively, and let d_h denote the attention-head dimension. Feature attention is

$$\mathrm{Attn}(\mathbf{H}) = \mathrm{softmax}\left(\frac{(\mathbf{H}\mathbf{A}_Q)(\mathbf{H}\mathbf{A}_K)^\top}{\sqrt{d_h}} \right) \mathbf{H}\mathbf{A}_V.$$

After applying $\mathbf{P}_{\mathrm{cov}}$,

$$\begin{aligned}
\mathrm{Attn}(\mathbf{P}_{\mathrm{cov}}\mathbf{H}) &= \mathrm{softmax}\left(\mathbf{P}_{\mathrm{cov}} \frac{(\mathbf{H}\mathbf{A}_Q)(\mathbf{H}\mathbf{A}_K)^\top}{\sqrt{d_h}} \mathbf{P}_{\mathrm{cov}}^\top \right) \mathbf{P}_{\mathrm{cov}}\mathbf{H}\mathbf{A}_V \\
&= \mathbf{P}_{\mathrm{cov}}\, \mathrm{softmax}\left(\frac{(\mathbf{H}\mathbf{A}_Q)(\mathbf{H}\mathbf{A}_K)^\top}{\sqrt{d_h}} \right) \mathbf{P}_{\mathrm{cov}}^\top \mathbf{P}_{\mathrm{cov}}\mathbf{H}\mathbf{A}_V \\
&= \mathbf{P}_{\mathrm{cov}}\, \mathrm{Attn}(\mathbf{H}),
\end{aligned}$$

where the second equality follows from the equivariance of row-wise softmax under simultaneous permutations of both axes. The same argument applies independently to each attention head. Concatenation and the shared output projection also preserve permutation equivariance.

Step 3: Propagate equivariance through the backbone. Residual connections, tokenwise layer normalization, and shared feed-forward maps commute with $\mathbf{P}_{\mathrm{cov}}$. Row attention acts along the row axis with shared parameters. In contrast, $\mathbf{P}_{\mathrm{cov}}$ acts only on the covariate-token axis. Therefore Eq. 60 propagates through every layer, giving

$$\mathbf{H}^{(L_{\mathrm{Tr}})}(\mathbf{P}_{\mathrm{cov}} \cdot W) = \mathbf{P}_{\mathrm{cov}} \mathbf{H}^{(L_{\mathrm{Tr}})}(W). \tag{61}$$

The same permutation relation is preserved throughout the Transformer.

Step 4: Apply symmetric pooling and the arm heads. Mean pooling is applied over the $p+2$ native token positions. Let $\mathbf{1} \in \mathbb{R}^{p+2}$ denote the all-ones vector. Then

$$\begin{aligned}
\frac{1}{p+2} \mathbf{1}^\top \mathbf{H}^{(L_{\mathrm{Tr}})}(\mathbf{P}_{\mathrm{cov}} \cdot W) &= \frac{1}{p+2} \mathbf{1}^\top \mathbf{P}_{\mathrm{cov}} \mathbf{H}^{(L_{\mathrm{Tr}})}(W) \\
&= \frac{1}{p+2} \mathbf{1}^\top \mathbf{H}^{(L_{\mathrm{Tr}})}(W),
\end{aligned}$$

because $\mathbf{1}^\top \mathbf{P}_{\mathrm{cov}} = \mathbf{1}^\top$. Thus both arm heads receive the same pooled representation before and after permutation. Treatment and outcome are fixed, so both context outcome statistics are unchanged. Consequently, $q_{\theta,t}(u \mid \mathbf{P}_{\mathrm{cov}} \cdot W) = q_{\theta,t}(u \mid W)$ and $\widehat{\mu}_t(\mathbf{P}_{\mathrm{cov}} \cdot W) = \widehat{\mu}_t(W)$. The CATE prediction is therefore invariant:

$$\widehat{\tau}(\mathbf{P}_{\mathrm{cov}} \cdot W) = \widehat{\mu}_1(\mathbf{P}_{\mathrm{cov}} \cdot W) - \widehat{\mu}_0(\mathbf{P}_{\mathrm{cov}} \cdot W) = \widehat{\tau}(W). \qquad \square$$

C SCM PRIOR AND SIMULATION

This section specifies the pretraining generator in construction order. We first sample task size and latent roles, construct the relation graph, instantiate the structural mechanisms, and construct the outcomes. We then verify identification for the generated task and construct arm-specific supervision.

C.1 TASK SIZE AND ROLE PRIOR

The pretraining dataset contains 20,000 independently generated tasks. For each task, let p and n denote the numbers of covariates and rows, respectively. They are sampled as

$$p \sim \mathrm{Unif}\{10, \ldots, 60\}, \qquad n \sim \mathrm{Unif}\{512, 768, 1024, 1536, 2048\}. \tag{62}$$

These draws vary the number of observations and visible covariates across pretraining tasks. For $d \in \{1, \ldots, p\}$, let X_d denote the d-th covariate.

Let the seven roles be ordered as in Table 5. The mediator gate is sampled first with probability 0.40 and determines the treatment-pathway mode of the structural generator. If it is off, the mode is direct-only. If it is on, indirect-only and direct-plus-indirect are sampled with equal probability. This yields marginal path probabilities $0.60, 0.20, 0.20$.

In the structural generator, direct-only mode disables mediator roles and includes T in the outcome parent vector. Indirect-only mode enforces at least one mediator and excludes T from that vector. Treatment then enters the outcome mechanism through $T \to X_d \to Y$. Direct-plus-indirect mode enforces a mediator and also includes T as a direct outcome input.

A separate shared gate controls treatment descendants, outcome descendants, and colliders. It is sampled independently with probability 0.40.

Let $a \in \{1, \ldots, 7\}$ index the role categories. Let $\mathcal{A} \subseteq \{1, \ldots, 7\}$ denote the active role-index set after gating. Define the base-mass vector $\boldsymbol{\beta} = (0.35, 0.20, 0.20, 0.10, 0.05, 0.05, 0.05)$. We renormalize the active base masses as $\widetilde{\boldsymbol{\beta}}_{\mathcal{A}} := \frac{(\beta_a)_{a \in \mathcal{A}}}{\sum_{a' \in \mathcal{A}} \beta_{a'}}$. This redistributes the base probability mass over the roles enabled for the current task.

Let $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_7)$ denote the full role-probability vector, with active subvector $\boldsymbol{\omega}_{\mathcal{A}}$. Let N_a denote the number of covariates assigned to role a. The weights and initial role counts are drawn as

$$\begin{aligned} \boldsymbol{\omega}_{\mathcal{A}} \sim \mathrm{Dirichlet}\left(30\widetilde{\boldsymbol{\beta}}_{\mathcal{A}}\right), \qquad \omega_a = 0 \text{ for } a \notin \mathcal{A}, \\ (N_1, \ldots, N_7) \sim \mathrm{Multinomial}(p, \boldsymbol{\omega}). \end{aligned} \tag{63}$$

The Dirichlet draw varies role proportions across tasks, and the multinomial draw converts those proportions into covariate counts. If an indirect treatment-pathway mode is active and this initial draw contains no mediator, one covariate is reassigned to the mediator role before graph construction.

Each covariate receives exactly one generator label. Each label determines its mandatory local motif during graph sampling. These labels do not imply mutually exclusive global DAG properties after optional $X \to X$ edges are added.

Table 5: Covariate roles, mandatory relations in the structural generator, and base masses.

Role	Mandatory relation	Base mass
Pre-treatment common cause	$X_d \to T$ and $X_d \to Y$	0.35
Treatment predictor	$X_d \to T$	0.20
Prognostic covariate	$X_d \to Y$	0.20
Mediator	$T \to X_d \to Y$	0.10
Treatment descendant	$T \to X_d$	0.05
Outcome descendant	$Y \to X_d$	0.05
Collider	$T \to X_d \leftarrow Y$	0.05

The covariate profile is sampled independently of the role assignment. Continuous-heavy, binary-heavy, and mixed profiles have probabilities $0.30, 0.30, 0.40$, respectively. Their continuous/binary probabilities are $0.80/0.20$, $0.25/0.75$, and $0.50/0.50$.

Continuous covariates retain the bounded $\tanh$ output of the covariate multilayer perceptron (MLP). Binary covariates are obtained by thresholding the latent MLP output. The threshold corresponds to a dataset-level target prevalence sampled uniformly from $[0.10, 0.90]$.

For binary covariates, the threshold is fixed within each generated dataset and reused during subsequent intervention draws. For mediator and treatment-descendant covariates, the same threshold is applied under both treatment arms.

C.2 Graph Construction

Graph construction starts from the mandatory relations listed in Table 5. Let $\mathcal{G}_{\text{rel}}$ denote the auxiliary relation graph used to screen additional covariate edges. It contains the mandatory relations and accepted $X \to X$ edges. The treatment-pathway mode is sampled in Appendix C.1, while the corresponding structural mechanisms are defined in Appendix C.3. Write $Z \rightsquigarrow Z'$ for a directed path between two nodes of $\mathcal{G}_{\text{rel}}$. Let $\mathbb{I}\{\cdot\}$ denote the indicator function. For covariate d, define the treatment/outcome ancestor and descendant summary

$$\mathbf{s}_d(\mathcal{G}_{\text{rel}}) := \big(\mathbb{I}\{X_d \rightsquigarrow T\}, \mathbb{I}\{X_d \rightsquigarrow Y\}, \mathbb{I}\{T \rightsquigarrow X_d\}, \mathbb{I}\{Y \rightsquigarrow X_d\}\big). \tag{64}$$

This summary records the reachability relations that each accepted covariate edge must preserve.

Candidate $X \to X$ edges respect this causal ordering. They are proposed with probability 0.60 between covariates with the same temporal position relative to treatment and outcome, and with probability 0.35 otherwise. A proposal is accepted only if the covariate subgraph remains acyclic and the summaries in Eq. 64 remain equal to those induced by the mandatory relations. The generator labels therefore specify the required relations with treatment and outcome, while additional $X \to X$ edges may introduce indirect paths beyond these local relations. The construction preserves this ordering and keeps every observed ancestor of T among the pre-treatment covariates. The relation graph is used only to screen candidate covariate edges and is separate from the outcome mechanism defined below. All task-level quantities fixed across factual and intervention draws are included in $\mathcal{S}$, as defined in Section 4.

C.3 Structural Mechanisms

C.3.1 Generic Structural Node

For observation i, consider a scalar structural node represented by an MLP with parent vector $\mathbf{v}_i$. Let L_{MLP} be its number of hidden layers. Let $(\mathbf{A}_\ell, \mathbf{b}_\ell)$ and $(\mathbf{w}, b)$ denote its hidden- and output-layer parameters. Let ϕ_ℓ and ϕ_{out} denote the corresponding maps. The additive noises are $\boldsymbol{\varepsilon}_{i\ell}$ and $\varepsilon_i^{\text{out}}$. Collect them in $\boldsymbol{\varepsilon}_i$. The structural map is

$$\begin{aligned} \mathbf{h}_i^{(0)} &= \mathbf{v}_i, \\ \mathbf{h}_i^{(\ell)} &= \phi_\ell\Big(\mathbf{A}_\ell \mathbf{h}_i^{(\ell-1)} + \mathbf{b}_\ell + \boldsymbol{\varepsilon}_{i\ell}\Big), \quad \ell = 1, \dots, L_{\text{MLP}}, \\ f(\mathbf{v}_i, \boldsymbol{\varepsilon}_i) &= \phi_{\text{out}}\Big(\mathbf{w}^\top \mathbf{h}_i^{(L_{\text{MLP}})} + b + \varepsilon_i^{\text{out}}\Big). \end{aligned} \tag{65}$$

In the reported generator, covariate MLP outputs are bounded by $\tanh$. Continuous covariates retain these outputs, while binary covariates are obtained by thresholding them. Other sampling distributions and parameter ranges are summarized in Table 7.

C.3.2 Treatment Mechanism

For treatment, f_T is a real-valued logit network with a linear final output. The sigmoid is applied only in Eq. 66.

Let $\mathbf{x}_{\text{pa}(T),i}$ denote the realized pre-treatment parent vector of treatment for observation i. Let b_T be the task-level intercept calibrated to a mean-propensity target drawn from $\text{Unif}(0.1, 0.9)$. The propensity e_i and treatment are generated by

$$e_i = \text{clip}\big(\text{sigmoid}\{f_T(\mathbf{x}_{\text{pa}(T),i}) + b_T\}, 0.02, 0.98\big), \qquad T_i \sim \text{Bernoulli}(e_i). \tag{66}$$

The clipping step keeps both treatment arms possible at every generated configuration of pre-treatment covariates.

C.3.3 Outcome Mechanism

The outcome network uses a linear final output, with hidden-unit activations sampled according to Table 6. The generic hidden-layer equation in Eq. 65 is instantiated componentwise. For hidden unit k in layer ℓ and observation i, define $z_{i\ell k} := [\mathbf{A}_\ell \mathbf{h}_i^{(\ell-1)} + \mathbf{b}_\ell + \boldsymbol{\varepsilon}_{i\ell}]_k$. For the sampled activation $\phi_{\ell k}$, let $I_{\phi_{\ell k}}$ denote its scale interval in Table 6. Here $\mathrm{LogUnif}$ denotes the log-uniform distribution on the displayed interval. Then

$$\begin{aligned} h_{i\ell k} &= \gamma_{\ell k}\, \phi_{\ell k}(a_{\ell k}(z_{i\ell k} + c_{\ell k}))\,, \\ a_{\ell k} &\sim \mathrm{LogUnif}(0.5, 1.5), \qquad \gamma_{\ell k} \sim \mathrm{LogUnif}(I_{\phi_{\ell k}}), \\ c_{\ell k} &\sim \mathcal{N}(0, 0.3^2). \end{aligned} \tag{67}$$

These draws vary the response shape across tasks while keeping each sampled mechanism fixed within its task.

The scalar output of the outcome network is further transformed by a task-level affine map. Let $\widetilde{Y}_i(t)$ denote the linear network output before calibration. We sample $s_Y^{\mathrm{out}} \sim \mathrm{LogUnif}(0.8, 1.5)$ and $b_Y^{\mathrm{out}} \sim \mathcal{N}(0, 0.3^2)$, and define $Y_i^{\mathrm{raw}}(t) = s_Y^{\mathrm{out}} \widetilde{Y}_i(t) + b_Y^{\mathrm{out}}$. Both quantities are sampled once per task and remain fixed across factual and intervention draws. The resulting $Y_i^{\mathrm{raw}}(t)$ is the potential outcome before the calibration below.

Table 6: Hidden-unit activation prior for the outcome network. Scale gives $I_{\phi_{\ell k}}$, the sampling interval for the multiplicative output scale $\gamma_{\ell k}$.

Activation	Probability	Scale	Activation	Probability	Scale
LeakyReLU	0.18	$[0.8, 1.8]$	ELU	0.14	$[0.8, 1.6]$
SiLU	0.14	$[0.8, 1.6]$	Softplus	0.12	$[0.6, 1.5]$
SELU	0.12	$[0.8, 1.8]$	ReLU	0.10	$[0.8, 1.8]$
Identity	0.08	$[0.7, 1.4]$	$\lvert\cdot\rvert$	0.06	$[0.6, 1.3]$
Square	0.04	$[0.4, 1.0]$	Exp	0.02	$[0.3, 0.8]$

For numerical stability, the implemented Square and Exp activations use $\phi_{\mathrm{sq}}(z) = \mathrm{clip}(z, -3, 3)^2$ and $\phi_{\exp}(z) = \exp\{\mathrm{clip}(z, -1.5, 2)\}$. These bounded-input implementations affect only the corresponding nonlinear transformations and leave the remainder of the sampled activation prior unchanged.

C.4 Outcomes

For $t \in \{0, 1\}$, let $Y_i^{\mathrm{raw}}(t)$ be the potential outcome defined by the sampled outcome mechanism above under $\mathrm{do}(T = t)$. The calibrated potential outcomes define the factual outcome, arm-specific supervision, and the task-specific CATE. Define

$$B_i^{\mathrm{raw}} := \tfrac{1}{2}\left\{Y_i^{\mathrm{raw}}(0) + Y_i^{\mathrm{raw}}(1)\right\}, \qquad \Delta_i^{\mathrm{raw}} := Y_i^{\mathrm{raw}}(1) - Y_i^{\mathrm{raw}}(0). \tag{68}$$

The midpoint captures the shared baseline, and the difference captures the treatment effect component.

We avoid near-degenerate outcome scales and treatment effect contrasts by amplifying nondegenerate baseline and effect components whose empirical amplitudes are small. Degenerate components are handled separately. Let Quant_α denote the empirical α-quantile. Let $\varepsilon_{\mathrm{tol}} = 10^{-8}$. Draw $A_B \sim \mathrm{LogUnif}(6, 12)$, $A_\Delta \sim \mathrm{LogUnif}(1, 3)$, and $\xi_\Delta \sim \mathrm{Unif}\{-1, +1\}$. Set

$$\begin{aligned} c_B &:= \mathrm{median}_i\, B_i^{\mathrm{raw}}, & r_B &:= \mathrm{Quant}_{0.95}(|B_i^{\mathrm{raw}} - c_B|)\,, \\ a_B &:= \begin{cases} \max(1, A_B / r_B), & r_B > \varepsilon_{\mathrm{tol}}, \\ 1, & r_B \le \varepsilon_{\mathrm{tol}}, \end{cases} & r_\Delta &:= \mathrm{Quant}_{0.75}(|\Delta_i^{\mathrm{raw}}|)\,, \\ a_\Delta &:= \begin{cases} \max(1, A_\Delta / r_\Delta), & r_\Delta > \varepsilon_{\mathrm{tol}}, \\ 1, & r_\Delta \le \varepsilon_{\mathrm{tol}}. \end{cases} \end{aligned} \tag{69}$$

The calibrated components are

$$B_i^{\mathrm{cal}} := \begin{cases} a_B\,(B_i^{\mathrm{raw}} - c_B)\,, & r_B > \varepsilon_{\mathrm{tol}}, \\ 0, & r_B \le \varepsilon_{\mathrm{tol}}, \end{cases} \qquad \Delta_i^{\mathrm{cal}} := \begin{cases} a_\Delta \Delta_i^{\mathrm{raw}}, & r_\Delta > \varepsilon_{\mathrm{tol}}, \\ \xi_\Delta A_\Delta, & r_\Delta \le \varepsilon_{\mathrm{tol}}, \end{cases} \tag{70}$$

For nondegenerate components, this calibration is intentionally one-sided. It amplifies a component only when its empirical amplitude is below the sampled level A_B or A_Δ. It does not shrink a component whose amplitude already exceeds the sampled level. Thus A_B and A_Δ act as lower amplitude bounds for nondegenerate components. Degenerate branches are handled explicitly in Eq. 70.

Finally,

$$Y_i^{\mathrm{cal}}(t) := B_i^{\mathrm{cal}} + \left(t - \tfrac{1}{2}\right) \Delta_i^{\mathrm{cal}}. \tag{71}$$

The calibration coefficients are computed once from the generated raw outcomes and are then held fixed for the factual and intervention outcomes of that task. This calibration is used to avoid near-degenerate outcome scales and treatment-effect contrasts.

The factual outcome is $Y_i = Y_i^{\mathrm{cal}}(T_i)$. Outcome-descendant variables are then generated from Y_i, while collider variables are generated from (T_i, Y_i). The same calibration is used for the intervention outcomes in Monte Carlo supervision construction.

Table 7: Principal generator settings used for pretraining.

Component	Quantity	Setting
Covariate MLP	hidden layers / width	$\mathrm{Unif}\{1,2,3,4\}$ / $\mathrm{Unif}\{15,\dots,39\}$
	weights and biases	$\mathcal{N}(0, 0.35^2)$
	edge dropout	0.30
	noise scale	Gaussian/Uniform $[0.03, 0.45]$, Laplace/Logistic $[0.03, 0.25]$
Treatment MLP	hidden layers / width	$\mathrm{Unif}\{1,2\}$ / $\mathrm{Unif}\{8,\dots,19\}$
	weights and biases	$\mathcal{N}(0, 0.6^2)$
	additive MLP noise	none
Outcome MLP	hidden layers / width	$\mathrm{Unif}\{1,2,3\}$ / $\mathrm{Unif}\{10,\dots,24\}$
	weights and biases	$\mathcal{N}(0, 0.8^2)$
	edge dropout	0.40
	noise scale	Gaussian/Uniform $[0.1, 1.0]$, Laplace/Logistic $[0.1, 0.5]$
Monte Carlo	draws per arm	$M = 256$ with common random numbers across arms

C.5 Identification within a Generated Task

For generated tasks, $\mathbf{C}$ is the complete vector of generated pre-treatment covariates. Conditional on the fixed task $\mathcal{S}$ and $\mathbf{C}$, the Bernoulli draw used to generate treatment in Eq. 66 is independent of the exogenous noises in the mediator and outcome mechanisms that determine the potential outcomes. Consequently, $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid \mathcal{S}, \mathbf{C}$. For every fixed task $\mathcal{S} = s$, this implies $\{Y(0), Y(1)\} \perp\!\!\!\perp T \mid \mathbf{C}$ under the task-specific distribution, which is the conditional exchangeability condition in Assumption 1. The clipping in Eq. 66 ensures $0.02 \le \mathbb{P}(T = 1 \mid \mathcal{S}, \mathbf{C}) \le 0.98$, and hence positivity. Consistency holds because the factual outcome is set to the calibrated potential outcome at the realized treatment, $Y = Y(T)$. The potential outcomes do not depend on the Bernoulli treatment draw conditional on $(\mathcal{S}, \mathbf{C})$, so forming the factual outcome $Y = Y(T)$ does not alter this conditional-independence property.

C.6 Factual Table and Arm-Specific Supervision

Construction. Variables are generated according to the causal ordering described above, and the factual table is fixed before supervision is constructed. During intervention draws, the generated task and the query pre-treatment covariates are held fixed.

Let $Y^{\mathcal{S}}(t, \mathbf{c}, \boldsymbol{\varepsilon})$ denote the calibrated intervention outcome given pre-treatment covariates $\mathbf{c}$, where $\boldsymbol{\varepsilon}$ collects the mediator and outcome noises used in the intervention draw.

Arm-label centering.

Assumption 4 (Arm-label centering for population analysis)**.** The simulated arm labels are conditionally centered at the task-specific arm means:

$$\mathbb{E}_\Pi[Z_t \mid \mathcal{S}, \mathbf{C}, W_n] = \mu_t^{\mathcal{S}}(\mathbf{C}), \qquad t \in \{0, 1\}. \tag{72}$$

Assumption 4 is the arm-label centering condition used in the population analysis in Section 5.2. In particular, it links the simulated arm labels to the task-specific arm means without requiring independence between the two arm labels.

Population consequence. By iterated expectation,

$$\mathbb{E}_{\Pi}[Z_t \mid W_n] = \psi_t(W_n), \qquad \mathbb{E}_{\Pi}[Z_1 - Z_0 \mid W_n] = \psi(W_n). \tag{73}$$

Under the arm-label centering condition, these identities link the simulated arm labels to $\psi_t(W_n)$ and $\psi(W_n)$ used in Section 5.2. They do not require independence between the two arm labels. Factual downstream variables remain visible to the predictor, but intervention-based supervision conditions only on the pre-treatment covariates.

D Architecture and Training Details

The implementation follows the supervision construction above. The Transformer encodes the observational context and queries and maps its outputs to two arm-specific predictive distributions. This section specifies the architecture, optimization, and checkpoint selection.

D.1 Input Representation and Masking

Context treatment and outcome values are observed. Their query entries are masked while query covariates remain visible. Row attention uses a row mask. Let rows i and i' index an attention receiver and source, respectively. Attention is permitted when both are context rows. It is also permitted when i is a query row and i' is either a context row or i itself. Thus each query can attend to the full observational context and its own covariates. It cannot access any other query row. Its prediction depends only on W_j, through the context-fitted preprocessing and the network. Covariates share one scalar encoder. Treatment and outcome use separate scalar encoders with missing-value indicators. No covariate-position, role, graph, propensity, or adjustment set information is provided. Tasks are grouped only when they have the same row count and native covariate dimension. The estimator therefore uses no covariate padding. Mean pooling covers exactly $p + 2$ native token positions.

D.2 Transformer Backbone and Arm Heads

The backbone alternates feature attention and row attention. It then mean-pools the $p + 2$ native token positions. The pooled representation feeds two separately parameterized arm heads. In the reported estimator, each arm head is a $K = 5$ Gaussian mixture. Their arm-mean predictions $\widehat{\mu}_t$ are given in Eq. 5. Their difference gives the CATE prediction in Eq. 6. The reported GMM-MSE configuration uses no independent effect decoder. Architectural settings are summarized in Table 8. The shared covariate encoder, absence of covariate-position embeddings, and symmetric pooling yield the covariate-permutation invariance proved in Appendix B.4.

D.3 Validation, Optimization, and pretraining details

Only the two arm labels supervise training. Validation uses the difference between the two arm labels as the target and computes RMSE against the predicted arm contrast over all queries in the held-out split. This score determines early stopping and checkpoint selection. The reported run uses two GPUs. Table 8 lists the optimization and stopping settings.

Training stops at epoch 93, and the checkpoint from epoch 43 with the best validation score is selected and applied unchanged to all downstream benchmarks. The reported training run takes approximately 15 hours on two NVIDIA RTX 5880 Ada Generation GPUs, corresponding to about 30 GPU-hours.

D.4 Ablation Configurations

Each variant selects its checkpoint on a held-out split from the corresponding pretraining prior and is then frozen for downstream evaluation. **GMM-MSE** uses two arm-specific Gaussian-mixture

Table 8: Architecture and optimization settings for the reported CRA-PFN configuration.

Quantity	Setting
Training / validation tasks	$80\%/20\%$, deterministic split with seed 42
Context / query rows	first 80% / remaining 20% within each task
Embedding / feed-forward dimension	$128/512$
Transformer layers / attention heads	$10/4$
Arm heads	two separate 5-component GMMs
Component-scale floor / mixture temperature	$10^{-3}/1.0$
Reported objective	Eq. 7, $\lambda = 0.5$
Learning rate	10^{-4}
Optimizer	Adam (Kingma & Ba, 2015)
Maximum training length / early stopping patience	150 epochs / 50 validation epochs
Periodic snapshot interval	10 epochs
Per-GPU task microbatch by row count	512:8, 768:8, 1024:8, 1536:4, 2048:4
Effective task batch	16 tasks across two GPUs after accumulation
Precision	bfloat16 automatic mixed precision in the reported training configuration

heads trained with marginal negative log-likelihood and first-moment squared error. **Arm-MSE** uses two deterministic arm-specific mean heads trained only with squared error under the same role-heterogeneous SCM prior. **Pre-treatment Prior** keeps the GMM-MSE two-arm formulation and the same SCM generator, but disables mediator, treatment-descendant, outcome-descendant, and collider roles during pretraining. **Direct-CATE** keeps the role-heterogeneous SCM prior but replaces the two-arm decomposition with a single Gaussian-mixture head trained on a Monte Carlo CATE label using the corresponding negative-log-likelihood and first-moment squared-error terms.

E Benchmark Construction, Evaluation Details and additional results

Evaluation covers four settings: benchmarks from prior work, structural variants together with Synthetic SCM, independently generated OOD-SCM families, and pre-treatment-only inputs. The benchmarks from prior work are CausalFM-BD, HMBART, IHDP, Twins, and Jobs (Ma et al., 2026; Liu et al., 2026; Hill, 2011; Ramey et al., 1992; Louizos et al., 2017). Potential outcomes and CATEs are used only for scoring. Role metadata supports benchmark construction and the pre-treatment-only evaluation. Neither the CATEs nor the role metadata are supplied to an estimator.

Metric and reporting. For query set $\mathcal{Q}$, let τ_j be the benchmark CATE and $\widehat{\tau}_j$ its prediction. We evaluate $\mathrm{PEHE} := \sqrt{|\mathcal{Q}|^{-1} \sum_{j \in \mathcal{Q}} (\tau_j - \widehat{\tau}_j)^2}$. PEHE is computed separately for each task. Multiple-task benchmarks report the task-level mean and standard error. Single-task benchmarks report only a point estimate. Blue boldface and purple underlining mark the best and second-best unrounded means. Rankings also use unrounded values. The main comparison includes CRA-PFN and eight external baselines. Appendix E.7 defines the expanded pool used for the ablations.

E.1 Baseline Implementations

CausalFM, CausalPFN, and Do-PFN use their released pretrained estimators without downstream gradient updates (Ma et al., 2026; Balazadeh Meresht et al., 2025; Robertson et al., 2025). We instantiate S-, T-, and X-Learners, ForestDR, and CausalForestDML (CF-DML) through EconML (Künzel et al., 2019; Battocchi et al., 2019; Kennedy, 2023; Chernozhukov et al., 2018; Wager & Athey, 2018). A single fixed estimator profile is used across benchmarks with no benchmark-specific tuning, and all methods receive the same permitted covariates in each evaluation setting. Because mixed-role inputs may violate the adjustment assumptions of standard estimators, full-input comparisons are interpreted as predictive performance under the same full-input setting.

Table 9: Pre-treatment-only PEHE for the baselines (lower is better).

Method	IHDP-C	IHDP-M	IHDP-MC	Twins-C	Twins-M	Twins-MC	Synthetic SCM	Avg. Rank
CRA-PFN	2.404 ± 0.11	0.224 ± 0.00	0.224 ± 0.00	0.326	**0.103**	**0.085**	$\mathbf{2.537 \pm 0.51}$	**2.29**
CausalFM	5.719 ± 0.21	0.303 ± 0.00	0.306 ± 0.01	**0.324**	0.223	0.380	3.700 ± 0.55	5.57
CausalPFN	$\mathbf{0.567 \pm 0.02}$	0.237 ± 0.00	0.239 ± 0.00	0.330	0.244	0.265	2.735 ± 0.55	4.71
Do-PFN	6.573 ± 0.30	0.341 ± 0.01	0.350 ± 0.01	0.329	0.113	0.407	3.919 ± 0.70	6.86
CF-DML	3.728 ± 0.17	$\mathbf{0.211 \pm 0.00}$	$\mathbf{0.212 \pm 0.00}$	0.325	0.119	0.132	2.921 ± 0.54	3.43
ForestDR	3.342 ± 0.17	0.212 ± 0.00	0.215 ± 0.00	0.325	0.118	0.134	2.865 ± 0.53	3.14
S-Learner	2.926 ± 0.14	0.319 ± 0.00	0.325 ± 0.00	0.326	0.111	0.111	3.212 ± 0.50	4.14
T-Learner	2.408 ± 0.11	0.696 ± 0.00	0.704 ± 0.01	0.343	0.344	0.237	6.085 ± 1.07	7.71
X-Learner	3.418 ± 0.16	0.489 ± 0.00	0.494 ± 0.00	0.332	0.226	0.170	4.112 ± 0.78	7.14

For the meta-learners, we use XGBoost as the base learner for the S-, T-, and X-Learners throughout (Chen & Guestrin, 2016). ForestDR and CF-DML remain forest-based and use five-fold cross-fitting, with 200-tree nuisance outcome and treatment/propensity estimators in each fold, followed by a 1000-tree final effect forest.

E.2 Splits, Covariates, and Evaluation Preprocessing

IHDP and derivatives retain the supplied train/test splits. Twins and derivatives use treatment-stratified 80/20 splits. The remaining tasks use deterministic 80/20 context/query splits. Evaluation preprocessing follows each method's fixed adapter and is not tuned per benchmark.

E.3 Benchmarks from Prior Work

CausalFM-BD follows the released CausalFM benchmark (Ma et al., 2026). IHDP uses the standard semi-synthetic realizations (Hill, 2011). Twins uses the processed potential-outcome benchmark common in treatment effect studies (Louizos et al., 2017). Jobs uses the CausalFM-distributed semi-synthetic task. HMBART uses the prepared simulation cases associated with the released benchmark (Liu et al., 2026).

E.4 Structural Variant Construction

Collider variants (IHDP-C, Twins-C) retain the source covariates, treatment, outcome, and benchmark CATE and append 1–4 generated colliders. Mediator variants (IHDP-M, Twins-M) retain the source pre-treatment covariates and generate treatment, 1–4 mediators, and outcome from a held-out nonlinear SCM. Mediator-plus-collider variants (IHDP-MC, Twins-MC) use the same construction but include both 1–4 mediators and 1–4 colliders. Their benchmark CATEs are computed from 64 Monte Carlo intervention draws per treatment arm.

E.5 OOD-SCM Construction

The OOD-SCM evaluation tasks are generated independently of the pretraining data and checkpoint selection. Each family contains 20 independently seeded tasks with $n = 1024$ rows. The visible dimension is sampled uniformly from $\{10, \ldots, 50\}$. Each family changes the structural mechanism class in a different way. **GP-RFF** uses smooth radial-basis-function mechanisms approximated with random Fourier features. **Piecewise** uses thresholds, hinge terms, discontinuities, and selection-like collider mechanisms. **Heavy-tailed** uses Student-t structural noise with covariate-dependent heteroskedasticity. **Fourier** uses high-frequency periodic terms and multiplicative interactions. Treatment prevalence is randomized across tasks, and propensities are clipped. Their CATEs are computed from 256 paired Monte Carlo intervention draws per treatment arm, with the pre-treatment covariates held fixed. Role metadata is retained only for generation and evaluation and is never supplied to any estimator.

Table 10: Full CRA-PFN ablation results across the four evaluation settings (lower is better). Bold/underlined entries compare the four displayed variants using unrounded means. Avg. Rank is computed in the expanded comparison together with the external baselines.

(a) Benchmarks from prior work

Variant	CausalFM-BD	HMBART	IHDP	Twins	Jobs	Avg. Rank
GMM-MSE	0.451 ± 0.05	**0.940 ± 0.19**	**2.404 ± 0.11**	0.326	0.523	3.60
Arm-MSE	0.461 ± 0.05	1.080 ± 0.19	3.589 ± 0.16	0.324	0.559	6.00
Pre-treatment Prior	**0.440 ± 0.05**	1.074 ± 0.20	2.445 ± 0.12	**0.323**	**0.480**	**2.20**
Direct-CATE	0.514 ± 0.05	1.328 ± 0.21	3.369 ± 0.14	0.324	0.521	6.80

(b) Structural variants and Synthetic SCM

Variant	IHDP-C	IHDP-M	IHDP-MC	Twins-C	Twins-M	Twins-MC	Synthetic SCM	Avg. Rank
GMM-MSE	**2.376 ± 0.11**	**0.242 ± 0.00**	0.250 ± 0.00	0.328	0.095	0.080	**2.609 ± 0.52**	**2.29**
Arm-MSE	3.554 ± 0.16	0.252 ± 0.00	0.258 ± 0.00	0.327	0.103	**0.072**	2.759 ± 0.49	3.71
Pre-treatment Prior	2.623 ± 0.12	0.295 ± 0.01	0.304 ± 0.01	0.325	**0.075**	0.100	3.153 ± 0.66	3.29
Direct-CATE	3.363 ± 0.14	0.242 ± 0.00	**0.245 ± 0.00**	**0.324**	0.095	0.101	2.868 ± 0.48	2.86

(c) OOD-SCM

Variant	Fourier	GP-RFF	Heavy-tailed	Piecewise	Avg. Rank
GMM-MSE	**1.262 ± 0.05**	1.026 ± 0.04	0.872 ± 0.05	0.940 ± 0.05	1.75
Arm-MSE	1.374 ± 0.06	1.133 ± 0.06	1.001 ± 0.06	1.010 ± 0.06	4.50
Pre-treatment Prior	1.436 ± 0.06	1.456 ± 0.13	1.136 ± 0.08	1.173 ± 0.09	8.50
Direct-CATE	1.269 ± 0.05	**1.021 ± 0.04**	**0.872 ± 0.04**	**0.911 ± 0.05**	**1.25**

(d) Pre-treatment-only

Variant	IHDP-C	IHDP-M	IHDP-MC	Twins-C	Twins-M	Twins-MC	Synthetic SCM	Avg. Rank
GMM-MSE	**2.404 ± 0.11**	0.224 ± 0.00	0.224 ± 0.00	0.326	0.103	0.085	2.537 ± 0.51	4.14
Arm-MSE	3.589 ± 0.16	0.223 ± 0.00	0.223 ± 0.00	0.324	0.103	**0.075**	2.715 ± 0.50	4.00
Pre-treatment Prior	2.445 ± 0.12	**0.196 ± 0.00**	**0.197 ± 0.00**	**0.323**	**0.062**	0.095	**2.421 ± 0.47**	**1.86**
Direct-CATE	3.405 ± 0.15	0.242 ± 0.00	0.240 ± 0.00	0.324	0.098	0.081	2.672 ± 0.47	4.57

E.6 Pre-Treatment-Only Evaluation

In the pre-treatment-only evaluation, the six structural variants and the Synthetic SCM benchmark are each restricted to $\mathbf{C}$, the task-specific valid pre-treatment covariates that define the CATE. All observed covariates outside $\mathbf{C}$ are removed before evaluation. All methods are evaluated using only these covariates, while the benchmark CATEs and evaluation splits remain unchanged. This setting is distinct from the *Pre-treatment Prior* ablation, which changes the pretraining distribution.

E.7 Ablation Reporting

The main ablation summary reports four validation-selected configurations: GMM-MSE, Arm-MSE, Pre-treatment Prior, and Direct-CATE. Average ranks in Table 3 are computed over the expanded comparison pool containing these four variants and the external baselines. These ranks are used to compare the ablation variants. Table 10 reports the corresponding per-benchmark PEHE values across the four evaluation settings.

E.8 Covariate-Permutation Diagnostic

To check the numerical implementation of covariate-permutation invariance, we rerun CRA-PFN on the six structural variants and the Synthetic SCM benchmark after applying the same fixed covariate permutation to every context and query row. Treatment and outcome remain fixed. Table 11 compares the original and permuted PEHE values.

The maximum absolute PEHE change is 1.19×10^{-7}. All seven values are unchanged at the precision used in Panel (b) of Table 1.

Table 11: Covariate-permutation diagnostic on the structural variants and Synthetic SCM. The final entry reports the absolute difference between the original and consistently permuted unrounded PEHE means.

Dataset	Original PEHE	Permuted PEHE	Abs. diff.
IHDP-C	2.37591204	2.37591204	3.74×10^{-9}
IHDP-M	0.24187048	0.24187048	2.27×10^{-10}
IHDP-MC	0.25002861	0.25002861	4.05×10^{-9}
Twins-C	0.32772985	0.32772985	0
Twins-M	0.09525312	0.09525310	1.49×10^{-8}
Twins-MC	0.07958676	0.07958678	2.24×10^{-8}
Synthetic SCM	2.60907847	2.60907859	1.19×10^{-7}

Table 12: Overall computation time comparison in seconds per task (lower is better).

Method	Fit / context	Prediction	Total
CRA-PFN	0.0000	0.0195	0.0195
CausalFM	0.0000	0.0189	0.0189
Do-PFN	0.0019	0.1472	0.1491
CausalPFN	0.4245	0.0938	0.5182
S-Learner	0.1922	0.0032	0.1954
T-Learner	0.3049	0.0042	0.3091
X-Learner	0.6632	0.0063	0.6695
ForestDR	5.4153	0.0256	5.4409
CF-DML	5.8383	0.0259	5.8642

E.9 Computational Time

We report the average computation time per task. For pretrained estimators, Fit / context denotes context processing without downstream parameter updates. For task-specific estimators, it includes model fitting. Prediction denotes query prediction time, and Total is their sum. Table 12 shows that CRA-PFN is computationally efficient relative to the compared methods.

E.10 Limitations

Our analysis relies on standard causal identification assumptions, including unconfoundedness and positivity. Extending CRA-PFN to settings with unmeasured confounding or limited treatment overlap remains an important direction for future work. Our current study also focuses on treatment effect estimation accuracy, while broader considerations such as fairness in effect estimation and treatment decision-making warrant further investigation.