DAG-aware Transformer for Causal Effect Estimation

Anonymous authors
Paper under double-blind review

Abstract

Causal inference is a critical task across fields such as healthcare, economics, and the social sciences. While recent advances in machine learning, especially those based on the deep-learning architectures, have shown potential in estimating causal effects, existing approaches often fall short in handling complex causal structures and lack adaptability across various causal scenarios. In this paper, we present a novel transformer-based method for causal inference that overcomes these challenges. The core innovation of our model lies in its integration of causal Directed Acyclic Graphs (DAGs) directly into the attention mechanism, enabling it to accurately model the underlying causal structure. This allows for flexible estimation of both average treatment effects (ATE) and conditional average treatment effects (CATE). Extensive experiments on both synthetic and real-world datasets demonstrate that our approach surpasses existing methods in estimating causal effects across a wide range of scenarios. The flexibility and robustness of our model make it a valuable tool for researchers and practitioners tackling complex causal inference problems.

1 Introduction

The estimation of Average Treatment Effect (ATE) and Conditional Average Treatment Effect (CATE) plays a pivotal role across various disciplines, significantly impacting decision-making processes and policy formulation. In medicine, these estimations guide treatment selections and personalized healthcare strategies Hernán & Robins (2024); Glass et al. (2013); Wager & Athey (2018). Within the realm of public policy, they inform the design and evaluation of interventions, from education reforms to social welfare programs Imbens & Rubin (2015); Hill (2011). In economics, ATE and CATE estimations are crucial for understanding the impacts of economic policies, labor market interventions, and consumer behavior Angrist & Pischke (2008); Heckman & Vytlacil (2007).

A fundamental challenge in this field lies in the correct specification of propensity score and outcome models, particularly when employing methods such as Inverse Probability of Treatment Weighting (IPTW) and Doubly-Robust Estimator (or Augmented IPW) to control for confounding factors Hernán & Robins (2024); Robins et al. (1994); Bang & Robins (2005). These methods, while powerful, are sensitive to model misspecification, which can lead to biased estimates and potentially misleading conclusions Kang & Schafer (2007); Funk et al. (2011). The complexity of real-world scenarios, characterized by high-dimensional data and complex causal relationships, further exacerbates this challenge, necessitating more sophisticated and robust approaches to causal inference Chernozhukov et al. (2018); Wager & Athey (2018).

The integration of machine learning (ML) methods into causal inference has opened new avenues for addressing complex causal relationships in high-dimensional settings. Athey and Imbens Athey & Imbens (2016) introduced causal trees and forests, adapting random forest algorithms to estimate heterogeneous treatment effects with valid statistical inference. Expanding on this, Wager and Athey Wager & Athey (2018) developed generalized random forests, extending forest-based methods to a broader class of causal parameters. These approaches have shown promise in settings with many covariates and potential treatment effect heterogeneity. Concurrently, Chernozhukov et al. Chernozhukov et al. (2018) proposed the double machine learning framework, combining flexible ML methods with orthogonalization techniques to achieve valid inference on treatment effects in high-dimensional settings.

Deep learning methods have also made significant inroads in causal inference, offering powerful tools for modeling complex relationships. Early work by Louizos et al. Louizos et al. (2017) introduced deep latent-variable models for causal effect inference, demonstrating how variational autoencoders could be adapted to learn representations that separate instrumental variables from confounders. The theoretical foundations connecting neural networks and causal inference were further established by Xia et al. Xia et al. (2021), who analyzed the expressiveness and learnability of neural networks for causal effect estimation, showing that neural architectures can universally approximate causal functionals under certain conditions. Jung et al. Jung et al. (2020) proposed learning causal effects via weighted empirical risk minimization, providing a principled framework for using neural networks in causal inference tasks.

Shalit et al. Shalit et al. (2017) introduced representation learning techniques for estimating individual treatment effects, using neural networks to learn balanced representations of covariates, addressing the fundamental problem of unobserved counterfactuals. The emergence of graph neural networks (GNNs) has further expanded the possibilities in causal inference, particularly for networked data. Building on graph attention networks Veličković et al. (2017), Ma et al. Ma & Tresp (2020) demonstrated how GNN-based approaches can estimate heterogeneous treatment effects in the presence of spillover effects, capturing complex dependencies in networked experiments.

Recent work has also explored transformer architectures for causal inference tasks. Guo et al. Guo et al. (2021) introduced CETransformer, which leverages self-attention mechanisms to capture correlations between covariates while using adversarial networks to balance treatment and control group representations. This approach addresses the limitation of traditional methods that rely on hand-crafted metric functions for balancing. Melnychuk et al. Melnychuk et al. (2022) introduced the Causal Transformer for estimating counterfactual outcomes over time, effectively capturing long-range dependencies in longitudinal data through a combination of transformer subnetworks with cross-attention mechanisms. Huang et al. Huang et al. (2023) proposed a double-head transformer architecture that simultaneously estimates short-term and long-term causal effects, addressing temporal heterogeneity in treatment effects. Zhang et al. Zhang et al. (2023b) proposed TransTEE, a transformer-based model for Heterogeneous Treatment Effect (HTE) estimation that handles various types of treatments. Zhang et al. Zhang et al. (2023a) developed Causal Inference with Attention (CInA), enabling zero-shot causal inference on unseen tasks with new data. The self-attention mechanism in transformers, enhanced with relative position representations Shaw et al. (2018), offers a natural way to model complex dependencies between covariates and treatments.

Despite significant advancements, current machine learning (ML) and deep learning (DL) approaches to causal inference face notable challenges. A primary limitation is their ability to simultaneously model complex relationships and incorporate structural causal knowledge. Many existing methods excel at flexible modeling of either the outcome regression or propensity score model, but rarely both concurrently. Moreover, they often lack natural mechanisms to explicitly integrate causal knowledge into the learning process. While transformer-based methods like CETransformer and Causal Transformer have shown promise in capturing complex covariate relationships and temporal dependencies, they do not explicitly leverage the causal structure encoded in directed acyclic graphs (DAGs).

A particularly persistent challenge in the field is the incorporation of unmeasured confounding into modern DL models. Traditional causal inference methods rely on the assumption of no unmeasured confounding (exchangeability), which is often untenable in observational studies (Tchetgen Tchetgen et al., 2020). Proximal causal inference has emerged as a powerful framework to address this limitation by leveraging proxy variables that capture information about unmeasured confounders (Miao et al., 2018; Cui et al., 2024). This approach requires pairs of treatment and outcome proxy variables that satisfy certain completeness conditions, enabling identification of causal effects even when exchangeability fails. The proximal framework has been extended to handle complex longitudinal settings (Ying et al., 2023), offering solutions through proximal g-formulas and doubly robust estimators (Liu et al., 2024). However, integrating proximal inference methods with modern deep learning architectures, particularly transformers, remains an open challenge (Melnychuk et al., 2022; Zhang et al., 2023b;a).

To address these limitations, we propose a novel approach that harnesses the power of transformer models while explicitly incorporating causal structure through a DAG-aware attention mechanism. Our method

enables the estimation of crucial causal quantities including the propensity score model $P(A|\mathbf{X})$, the outcome regression model $P(Y|A, \mathbf{X})$, and importantly, the bridge function $h(A, W, X)$ required for proximal inference. Here, A represents the treatment, $\mathbf{X}$ denotes observed confounders, Y is the outcome, and W serves as a proxy for the outcome in scenarios with unmeasured confounding. By incorporating causal structure directly into the attention mechanism, our approach differs fundamentally from existing transformer-based causal methods that treat all covariates symmetrically without considering their causal roles.

This approach allows for seamless integration of these estimated models into G-formula, IPTW, doubly robust estimators, and proximal inference methods. The ability to handle unmeasured confounding through proximal inference is particularly crucial for real-world applications where complete measurement of all confounders is infeasible. By doing so, our work bridges the gap between cutting-edge machine learning techniques and classical causal inference methods, offering a more comprehensive framework for causal analysis in complex, real-world scenarios. Our contributions include: (1) a DAG-aware transformer architecture that explicitly encodes causal relationships in the attention mechanism; (2) unified estimation of multiple causal quantities within a single framework; (3) native support for proximal inference to handle unmeasured confounding; and (4) empirical validation on both synthetic and real-world datasets demonstrating superior performance compared to existing methods.

2 Preliminaries

2.1 ATE and CATE

Consider treatment A and its effect on outcome Y. Let $\mathbf{X}$ denote a vector of *observed* confounders. We define Y^a as the counterfactual outcome for each individual had they received ($a = 1$) or not received ($a = 0$) the treatment. The Average Treatment Effect (ATE), denoted as τ, is then defined as $\tau = \mathbb{E}[Y^1 - Y^0]$.

While the ATE provides an overall measure of the treatment effect across the entire population, in many cases, it's important to understand how the treatment effect varies across different subgroups or individuals. The CATE, denoted as $\tau(x)$, measures the average treatment effect for a subpopulation with a specific set of covariates $X = x$: $\tau(x) = \mathbb{E}[Y^1 - Y^0|X = x]$.

2.2 Confounding Control Methods assuming Unconfoundedness

In causal inference, several methods have been developed to control for *observed* confounding and estimate treatment effects. Our paper focuses primarily on three methods: Standardization (G-formula), Inverse Probability of Treatment Weighting (IPTW) and Augmented Inverse Probability Weighting (AIPW), a form of Doubly Robust estimator (Hernán & Robins, 2024).

1. **Standardization (G-formula)**: Standardization, also known as the G-formula, estimates the ATE by modeling the outcome as a function of treatment and confounders. It then averages over the confounder distribution to estimate the population-level effect. The ATE is estimated as:

$$\tau_G = \mathbb{E}_X[\mathbb{E}[Y|A = 1, X] - \mathbb{E}[Y|A = 0, X]] \tag{1}$$

   where $\mu(a, X) = \mathbb{E}[Y|A = a, X]$ is the conditional expectation of the outcome given treatment a and confounders X. This method is effective when the outcome model is correctly specified.

2. **Inverse Probability of Treatment Weighting (IPTW)**: IPTW uses the propensity score to create a pseudo-population in which the treatment assignment is independent of the measured confounders. The ATE is estimated as:

$$\tau_{IPTW} = \mathbb{E}\Big[\frac{AY}{\pi(X)} - \frac{(1 - A)Y}{1 - \pi(X)}\Big] \tag{2}$$

   where $\pi(X) = P(A = 1|X)$ is the propensity score. This method is effective when the propensity score model is correctly specified.

3. **Augmented Inverse Probability Weighting (AIPW)**: AIPW combines IPTW with an outcome regression model, providing robustness against misspecification of either the propensity score model or the outcome model. The ATE is estimated as:

$$\begin{aligned}\tau_{AIPW} = \mathbb{E}\Big[\big(\mu(1,X) + \frac{A}{\pi(X)}(Y - \mu(1,X))\big) \\ - \big(\mu(0,X) + \frac{1-A}{1-\pi(X)}(Y - \mu(0,X))\big)\Big]\end{aligned} \tag{3}$$

where $\mu(a,X) = \mathbb{E}[Y|A=a,X]$ is the outcome regression function and $\pi(X) = P(A=1|X)$ is the propensity score.

2.3 Proximal Inference

In proximal inference (Tchetgen et al., 2024), we aim to estimate the expected potential outcome $\mathbb{E}[Y^a]$ for each treatment level a, in the presence of unobserved confounders U, given a set of proxies (W,Z) and observed confounders X. The key assumptions are:

Assumption 1 (Independence). Given (A,U,W,X,Y,Z), $Y \perp\!\!\!\perp Z|A,U,X$ and $W \perp\!\!\!\perp (A,Z)|U,X$.

Assumption 2 (Completeness with respect to U). For all $f \in L^2$ and all $a \in \mathcal{A}, x \in \mathcal{X}$, $\mathbb{E}[f(U)|A=a,X=x,Z=z] = 0$ for all $z \in \mathcal{Z}$ if and only if $f(U) = 0$ almost surely.

Assumption 3 (Completeness with respect to Z). For all $f \in L^2$ and all $a \in \mathcal{A}, x \in \mathcal{X}$, $\mathbb{E}[f(Z)|A=a,W=w,X=x] = 0$ for all $w \in \mathcal{W}$ if and only if $f(Z) = 0$ almost surely.

Under these assumptions, there exists a bridge function h satisfying:

$$\mathbb{E}[Y|A=a,X=x,Z=z] = \int_{\mathcal{W}} h(a,w,x)p(w|a,x,z)dw \tag{4}$$

The expected potential outcomes are given by:

$$\mathbb{E}[Y^a] = \mathbb{E}_{W,X}[h(a,W,X)] \tag{5}$$

The ATE then can be derived from the empirical mean of $\hat{h}$ with a fixed to the value of interest, $\hat{\mathbb{E}}[Y^a] = \frac{1}{M}\sum_{i=1}^{M}\hat{h}(a,w_i,x_i)$.

3 Methodology

We propose a novel DAG-aware Transformer model for causal effect estimation that explicitly incorporates causal structure into the attention mechanism. Our approach is flexible and can accommodate various causal scenarios, including those with or without unmeasured confounding.

Given a dataset of N observations, we define a set of possible input nodes. These include: A, the treatment variable; $\mathbf{X}$, the observed confounding variables; U, representing unmeasured confounding variables; Y, the outcome variable; Z, the proxy variable for treatment; and W, the proxy variable for outcome. The specific combination of input nodes used depends on the causal structure being modeled. The output nodes of our model vary based on the estimation method employed:

- For standardization or proximal inference: $\hat{Y}$ (estimated outcome), which corresponds to $\hat{\mu}(a,X)$ for standardization or $\hat{h}(a,W,X)$ for proximal inference.
- For Inverse Probability of Treatment Weighting (IPTW): $\hat{A}$ (estimated treatment probability), which corresponds to $\hat{\pi}(X)$ (estimated propensity score)

- For Augmented Inverse Probability Weighting (AIPW): Both $\hat{A}$ and $\hat{Y}$, where $\hat{A}$ corresponds to $\hat{\pi}(X)$ and $\hat{Y}$ corresponds to $\hat{\mu}(a, X)$

This flexible framework allows our DAG-aware Transformer to adapt to different causal inference scenarios and estimation techniques while maintaining its core structure.

After estimating these quantities, we can plug them into the corresponding formulas from Section 2 to estimate the Average Treatment Effect (ATE) or Conditional Average Treatment Effect (CATE). Specifically, we replace the true (but unknown) functions $\pi(X)$ and $\mu(a, X)$ with their estimates $\hat{\pi}(X)$ and $\hat{\mu}(a, X)$ obtained from our model:

- **For standardization**: We estimate $\hat{\mu}(a, X) = \mathbb{E}[Y|A = a, X]$ using our model, then compute:

$$\hat{\tau}_G = \frac{1}{N}\sum_{i=1}^{N}[\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)] \tag{6}$$

- **For IPTW**: We estimate $\hat{\pi}(X) = P(A = 1|X)$ using our model, then compute:

$$\hat{\tau}_{IPTW} = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - A_i)Y_i}{1 - \hat{\pi}(X_i)}\Big] \tag{7}$$

- **For AIPW**: We estimate both $\hat{\pi}(X)$ and $\hat{\mu}(a, X)$ using our model. These estimates replace the true functions in Equation 3:

$$\begin{aligned}\hat{\tau}_{AIPW} = \frac{1}{N}\sum_{i=1}^{N}\Big[&\big(\hat{\mu}(1, X_i) + \frac{A_i}{\hat{\pi}(X_i)}(Y_i - \hat{\mu}(1, X_i))\big) \\ &- \big(\hat{\mu}(0, X_i) + \frac{1 - A_i}{1 - \hat{\pi}(X_i)}(Y_i - \hat{\mu}(0, X_i))\big)\Big]\end{aligned} \tag{8}$$

  where $\hat{\mu}(a, X)$ is our model's estimate of $\mathbb{E}[Y|A = a, X]$ and $\hat{\pi}(X)$ is our model's estimate of $P(A = 1|X)$.

- **For proximal inference**: We estimate the bridge function $\hat{h}(a, W, X)$ using our model, then compute:

$$\hat{\tau}_{proximal} = \frac{1}{N}\sum_{i=1}^{N}[\hat{h}(1, W_i, X_i) - \hat{h}(0, W_i, X_i)] \tag{9}$$

For CATE estimation, we can condition on specific values of X in these equations. This approach allows us to estimate both population-level and subgroup-level causal effects using our DAG-aware Transformer model, with the key advantage that all required functions ($\pi(X)$, $\mu(a, X)$, or $h(a, W, X)$) are estimated within a unified framework that respects the causal structure of the problem.

3.1 DAG-aware Transformer Architecture

Our DAG-aware Transformer architecture explicitly encodes causal relationships into the attention mechanism, enabling the model to respect the underlying causal structure during learning. Figure 1 illustrates our model architecture with a simple example DAG containing treatment (A), confounders (X), and outcome (Y) nodes. While this figure shows a basic scenario for clarity, our architecture is flexible and can accommodate more complex causal structures, including proximal inference settings with additional proxy nodes (W and Z) (See Figure 3 for an example), time-varying treatments, or high-dimensional confounder spaces.

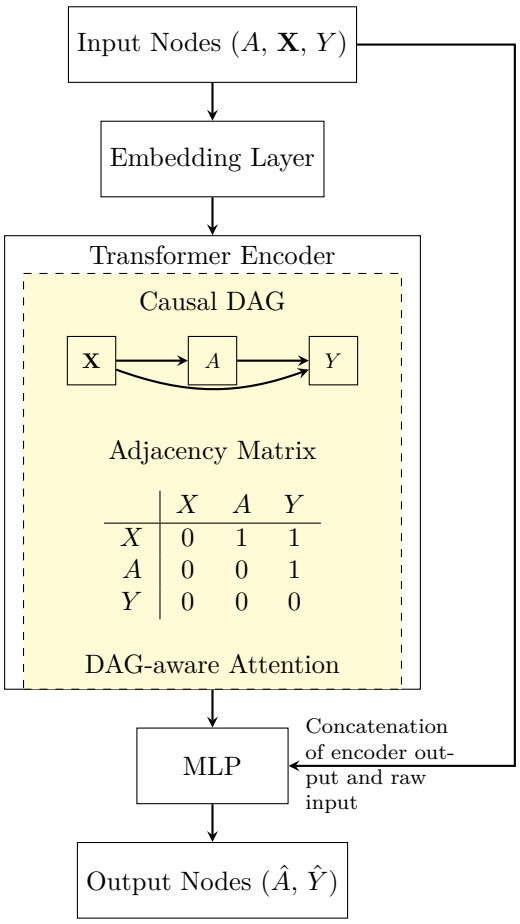

Figure 1: Architecture of the DAG-aware Transformer model. The Transformer Encoder integrates a DAG-aware attention mechanism (indicated by dashed lines), leveraging the causal structure outlined by the DAG. The adjacency matrix, derived from this causal DAG, guides the attention computations. The model merges the output from the transformer encoder with the raw input via a weighted average, which is subsequently processed by a multi-layer perceptron (MLP) to generate the final output. Note that we did not employ layer normalization, commonly used in transformers, as we found that it biased the causal effects empirically.

We encode the causal DAG into an adjacency matrix $\mathbf{M}^{adj} \in \{0,1\}^{D\times D}$, where D is the number of nodes in the graph. Each element $M_{ij}^{adj} = 1$ indicates a directed edge from node i to node j, representing a direct causal relationship. Importantly, we enforce $M_{ii}^{adj} = 0$ for all i, ensuring no self-loops exist in the causal graph, as nodes cannot cause themselves.

To incorporate this causal structure into the attention mechanism, we transform the adjacency matrix into an attention mask $\mathbf{M}$:

$$M_{ij} = \begin{cases} 0 & \text{if } M_{ij}^{adj} = 1 \\ 1 & \text{otherwise} \end{cases} \tag{10}$$

Our key innovation lies in incorporating the causal structure directly into the multi-head attention computation. For each attention head, we first compute the standard attention scores:

$$\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{E}} \tag{11}$$

where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N\times D\times E}$ are the query and key matrices respectively, N is the batch size, and E is the embedding dimension.

We then apply the DAG-based mask to these attention scores:

$$\mathbf{A}^{mask} = \mathbf{A} + \mathbf{M} \cdot (-\infty) \tag{12}$$

This masking operation effectively sets attention scores to negative infinity for node pairs that are not causally connected according to the DAG. After applying softmax, these masked positions will have zero attention weight, preventing information flow between causally unrelated nodes.

The final attention output is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{A}^{mask})\mathbf{V} \tag{13}$$

where $\mathbf{V} \in \mathbb{R}^{N\times D\times E}$ is the value matrix.

Figure 2 illustrates the visualization of the adjacency matrix for a simple scenario involving treatment A, two confounders X_1 and X_2, and the outcome Y. Additionally, it features the corresponding masking matrix and the attention weights, providing insights into the attention mechanism used in the model.

After processing inputs through multiple layers of DAG-aware attention, we combine the transformer encoder's output $\mathbf{H}$ with the original features $\mathbf{X}$ through a hybrid integration mechanism:

$$\mathbf{Z} = \alpha \cdot \mathbf{H} + \mathbf{X} \tag{14}$$

where α is a learnable weight parameter. This design, inspired by residual connections He et al. (2016), addresses an empirical phenomenon we observed: relying solely on the transformer's representations can lose

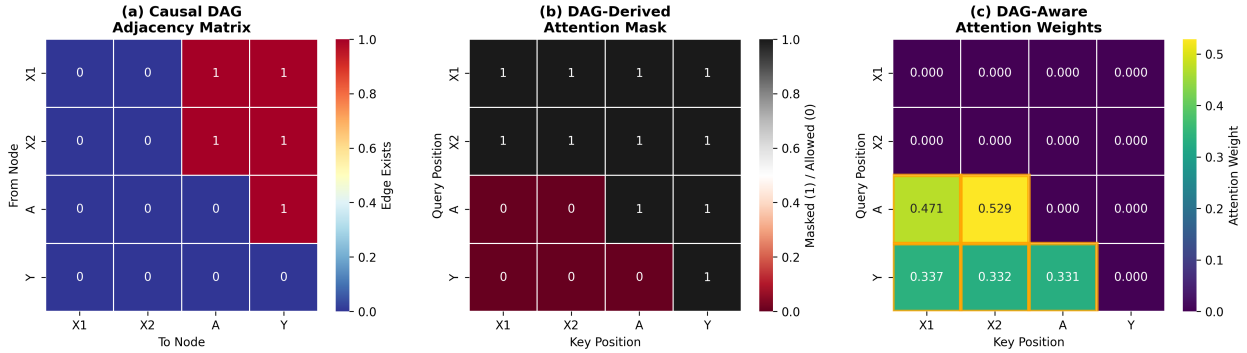

Figure 2: Visualization of key components in the DAG-aware Transformer model. (a) Causal DAG Adjacency Matrix: This matrix represents the directed edges between nodes, indicating the causal relationships among features. A value of 1 denotes the existence of an edge, while 0 signifies no direct relationship. (b) DAG-Derived Attention Mask: The attention mask, derived from the adjacency matrix, specifies which nodes can attend to one another during the attention computation. Here, a value of 0 indicates that attention is allowed, while 1 indicates masked positions, preventing information flow between non-causal nodes. (c) DAG-Aware Attention Weights: These weights illustrate the strength of attention assigned to each node pair in the attention mechanism, reflecting how the model weighs the importance of different nodes when updating information. The color scale indicates varying levels of attention, with brighter regions representing higher weights.

critical confounding information, causing the model to produce identical outcome predictions regardless of treatment assignment. By preserving both the learned representations and raw features, the hybrid approach ensures that essential confounding information remains accessible for downstream prediction.

The hybrid representation $\mathbf{Z}$ is then processed by a Multi-Layer Perceptron (MLP) to produce the final predictions:

$$\hat{A}/\hat{Y} = \text{MLP}(\mathbf{Z}) \tag{15}$$

The specific output depends on the causal inference task: $\hat{A}$ for propensity score estimation in IPTW, $\hat{Y}$ for outcome regression in standardization, or both for AIPW. In proximal inference settings with unmeasured confounding, the model estimates the bridge function $\hat{h}(a, W, X)$ instead.

3.2 Model Training and Objective Function

Our model employs different loss functions depending on the causal inference method used. We present the loss functions for standardization (G-formula), Inverse Probability of Treatment Weighting (IPTW), Augmented Inverse Probability Weighting (AIPW), and proximal inference. Here we assume Y (outcome) is continuous, and A (treatment) is a binary variable.

For standardization, we use Mean Squared Error (MSE) loss for maximum likelihood estimation:

$$\mathcal{L}_G = \text{MSE}(\hat{Y}, Y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{16}$$

where $\hat{Y}$ are the model outputs and Y are the true labels.

For IPTW, we use Binary Cross Entropy (BCE) loss for treatment/propensity score estimation:

$$\mathcal{L}_{\text{IPTW}} = \text{BCE}(\hat{A}, A) = \\ -\frac{1}{n}\sum_{i=1}^{n}[A_i \log(\hat{A}_i) + (1 - A_i)\log(1 - \hat{A}_i)] \tag{17}$$

where $\hat{A}$ are the model outputs (estimated propensity scores) and A are the true treatment assignments.

For AIPW, we combine MSE loss for outcome prediction and BCE loss for treatment assignment:

$$\mathcal{L}_{\text{AIPW}} = \frac{1}{2}(\text{MSE}(\hat{Y}, Y) + \text{BCE}(\hat{A}, A)) \tag{18}$$

For proximal inference, following the work of Kompa et al. (2022), we introduce two variants: NMMR-U and NMMR-V, based on U-statistics and V-statistics respectively. The empirical risk $\hat{R}_{k,n}$ given data $\mathcal{D} = \{(a_i, w_i, x_i, y_i, z_i)\}_{i=1}^{N}$ can be written as:

$$\hat{R}_{k,U,n}(h) = \frac{1}{n(n-1)} \sum_{i,j=1, i\neq j}^{n} (y_i - h_i)(y_j - h_j)k_{ij} \tag{19}$$

$$\hat{R}_{k,V,n}(h) = \frac{1}{n^2} \sum_{i,j=1}^{n} (y_i - h_i)(y_j - h_j)k_{ij} \tag{20}$$

where $h_i = h(a_i, w_i, x_i)$ and $k_{ij} = k((a_i, z_i, x_i), (a_j, z_j, x_j))$.

To prevent overfitting, we add an L2 penalty $\Lambda[h, \theta_h] = \sum_i \theta_{h,i}^2$, where θ_h are the model parameters. The penalized risk functions are:

$$\hat{R}_{k,U,\lambda,n}(h) = \hat{R}_{k,U,n}(h) + \lambda\Lambda[h, \theta_h] \tag{21}$$

$$\hat{R}_{k,V,\lambda,n}(h) = \hat{R}_{k,V,n}(h) + \lambda\Lambda[h, \theta_h] \tag{22}$$

The final loss function for training the neural networks is:

$$\mathcal{L}_{proximal} = (Y - h(A, W, X))^T K(Y - h(A, W, X)) + \lambda\Lambda[h, \theta_h] \tag{23}$$

where $(Y - h(A, W, X))$ is a vector of residuals from the neural network's predictions and K is a kernel matrix with entries k_{ij}. We use an RBF kernel for k. For the U-statistic variant (NMMR-U), we set the main diagonal of K to zero, while for the V-statistic variant (NMMR-V), we include the main diagonal elements.

4 Experiments

We evaluate our proposed DAG-aware Transformer method across four diverse datasets, demonstrating its flexibility in handling different causal structures and estimation tasks. Our experiments are designed to showcase the key advantage of our approach: the ability to flexibly encode causal DAGs and estimate the conditional probability distributions required for various causal inference methods within a unified framework.

Experimental Setup. We employ four datasets that span different causal inference scenarios. The LaLonde CPS and PSID datasets are used for Average Treatment Effect (ATE) estimation in standard unconfoundedness settings, providing a benchmark for evaluating our method's performance on classic causal inference

tasks. The ACIC dataset is employed for Conditional Average Treatment Effect (CATE) estimation, allowing us to assess our model's ability to capture heterogeneous treatment effects across different subpopulations. Additionally, we use the Demand dataset described in Kompa et al. (2022) for proximal inference scenarios with unmeasured confounding, demonstrating our model's ability to handle more complex causal structures involving proxy variables. The causal assumptions and DAG structures for each experiment are detailed in Appendix A.1. For AIPW, we investigate two training strategies to understand the benefits of joint learning. In the separate training approach, we train independent DAG-aware Transformers for the outcome and propensity score models. In the joint training approach, we use a single DAG-aware Transformer with multiple output heads to simultaneously learn both models. This comparison examines whether joint learning can leverage shared representations to better capture the relationships between confounders, treatment, and outcomes, potentially leading to more accurate estimates or improved computational efficiency.

Evaluation Metrics. We employ normalized root mean square error (NRMSE) as our primary evaluation metric across all experiments. For the LaLonde and ACIC datasets, we compute NRMSE between our estimated ATE or CATE and the true values. In the case of the Demand dataset with proximal inference, performance is evaluated using NRMSE computed across 10 equally-spaced price points between 10 and 30, comparing our estimated potential outcomes $\hat{\mathbb{E}}[Y^a]$ against Monte Carlo simulations of the true $\mathbb{E}[Y^a]$. This comprehensive evaluation allows us to assess our model's performance across different causal inference scenarios and estimation targets.

Baseline Models and Ablation stutdies. We compare our approach against established methods in causal inference to demonstrate its competitive performance. Generalized Random Forests (GRF) (Wager & Athey, 2018) serves as a state-of-the-art non-parametric baseline, known for its effectiveness in estimating heterogeneous treatment effects without strong parametric assumptions. We also include a standard Multilayer Perceptron (MLP) as a neural network baseline, which provides a direct comparison to our transformer architecture while using the same training framework. While transformer-based causal inference methods such as CETransformer (Guo et al., 2021) and Causal Transformer (Melnychuk et al., 2022) exist, they are designed for specific causal tasks (balanced representation learning and longitudinal counterfactual estimation, respectively) and cannot be directly applied to our diverse experimental settings that include both standard and proximal inference scenarios. Our focus is on demonstrating how explicit DAG encoding enables a single architecture to flexibly handle multiple causal inference frameworks, rather than comparing task-specific transformer implementations. To understand the contribution of our DAG-aware attention mechanism, we compare our full DAG-aware Transformer model, which encodes causal structure in attention masks, against an unconstrained transformer baseline that uses standard self-attention without causal constraints. Implementation details and hyperparameter settings are provided in Appendix A.2.

5 Results

We evaluate our DAG-aware Transformer across four datasets representing different causal inference scenarios: LaLonde-CPS and LaLonde-PSID for ATE estimation under unconfoundedness, ACIC for CATE estimation, and Demand for proximal inference with unmeasured confounding.

LaLonde Datasets. On the LaLonde-CPS dataset (Table 1), our DAG-aware Transformer achieves superior performance across all estimation methods. For G-formula estimation, our method achieves an NRMSE of 0.784 (SE: 0.074), outperforming GRF (0.925, SE: 0.003), MLP (0.935, SE: 0.069), and the unconstrained transformer (1.001, SE: 0.082). The performance gains are even more pronounced for IPW estimation, where our method achieves 4.279 (SE: 0.469) compared to 5.615 (SE: 0.509) for the unconstrained transformer, 6.419 (SE: 0.471) for MLP, and 6.342 (SE: 1.227) for GRF. Most notably, for AIPW estimation, our joint training approach achieves exceptional performance with an NRMSE of 0.169 (SE: 0.051), dramatically outperforming the separately trained transformer variant (0.961, SE: 0.204), MLP (1.362, SE: 0.216), and GRF (1.596, SE: 0.294).

Similar patterns emerge on the LaLonde-PSID dataset, where our method consistently achieves the lowest NRMSE across all estimators. The AIPW estimator again shows remarkable performance (0.118, SE: 0.068), while the unconstrained transformer struggles significantly (3.017, SE: 0.848), and both MLP (2.038, SE: 0.482) and GRF (2.517, SE: 0.242) show substantially higher error rates. For IPW estimation, our method

achieves 0.257 (SE: 0.093), dramatically outperforming GRF which exhibits poor performance at 9.408 (SE: 1.108).

ACIC Dataset. For conditional average treatment effect estimation on the ACIC dataset, our DAG-aware Transformer maintains competitive performance. The G-formula achieves an NRMSE of 0.315 (SE: 0.045), slightly outperforming GRF (0.346, SE: 0.044) and substantially better than MLP (0.558, SE: 0.062) and the unconstrained transformer (0.675, SE: 0.072). The IPW estimator shows significant improvement, with our method achieving 3.125 (SE: 0.704) compared to 4.281 (SE: 1.388) for GRF, 6.380 (SE: 1.700) for MLP, and 6.750 (SE: 1.640) for the unconstrained transformer. For AIPW, our joint training approach (0.781, SE: 0.031) outperforms the unconstrained transformer (1.162, SE: 0.135) and MLP (1.244, SE: 0.141), performing comparably to GRF (0.857, SE: 0.059), while the separately trained transformer variant achieves the lowest standard error (0.958, SE: 0.018).

Demand Dataset. Table 2 presents results for proximal inference on the Demand dataset across varying sample sizes. Our DAG-aware Transformer consistently outperforms baselines for both U-statistics and V-statistics estimators across all sample sizes. With smaller samples (N=1,000), our method achieves NRMSE of 0.109 (SE: 0.007) for U-statistics compared to 0.127 (SE: 0.008) for MLP and 0.164 (SE: 0.013) for the unconstrained transformer. Performance improvements become more pronounced with larger samples: at N=50,000, our U-statistics estimator achieves 0.082 (SE: 0.005) while MLP degrades to 0.181 (SE: 0.070) and the unconstrained transformer to 0.247 (SE: 0.089), suggesting that the DAG-aware attention mechanism provides crucial inductive bias that becomes increasingly valuable with more data.

Impact of DAG-aware Attention. The comparison between our full DAG-aware Transformer and the unconstrained transformer variant (no mask) provides clear evidence for the value of incorporating causal structure. Across all datasets and estimators, removing the DAG-aware attention mechanism leads to substantial performance degradation, with particularly dramatic effects observed for AIPW estimation on the LaLonde datasets and for larger sample sizes in proximal inference settings.

Joint vs Separate Training for AIPW. Our joint training approach for AIPW consistently outperforms separate training of propensity score and outcome models. On LaLonde-CPS, joint training achieves 0.169 (SE: 0.051) compared to 0.961 (SE: 0.204) for separate training. This suggests that jointly learning both components within a unified DAG-aware framework enables better representation sharing and more accurate estimation of treatment effects.

6 Conclusion

Our DAG-aware Transformer represents a fundamentally different approach to neural causal inference. Unlike task-specific methods such as CFR-Wass (Shalit et al., 2017) and CEVAE (Louizos et al., 2017), our framework encodes arbitrary causal structures as architectural constraints, enabling practitioners to seamlessly transition between estimation strategies while incorporating domain knowledge directly. This differs crucially from Graph Attention Networks (Veličković et al., 2017): while GATs learn attention weights from graph topology for node classification, we impose causal DAGs as hard constraints on attention patterns, ensuring information flows only along valid causal pathways.

Our approach provides value through three key aspects: encoding causal structure as inductive bias, offering a unified architecture for multiple estimation strategies (propensity scores, outcomes, bridge functions), and maintaining flexibility across different causal assumptions. The consistent improvements over both classical methods like GRF and unconstrained neural architectures demonstrate that explicitly incorporating causal structure into transformer-based models advances causal inference methodology. The contribution lies not in outperforming specialized methods on their target tasks, but in providing a principled framework that respects causal constraints while leveraging deep learning's representational capacity.

Future extensions could incorporate uncertainty quantification, handle time-varying treatments, or jointly learn causal structures alongside effect estimation. As causal inference grows increasingly central across domains, we believe architectures that bridge predictive modeling and causal reasoning will prove essential for principled decision-making from observational data.

Table 1: Performance comparison of causal inference estimators across datasets and model architectures.

Dataset	Estimator	Model	NRMSE (mean)	NRMSE (SE)
Lalonde-CPS	G-formula	GRF	0.925	**0.003**
		MLP	0.935	0.069
		Transformer (no mask)	1.001	0.082
		Transformer (Ours)	**0.784**	0.074
	IPW	GRF	6.342	1.227
		MLP	6.419	**0.471**
		Transformer (no mask)	5.615	0.509
		Transformer (Ours)	**4.279**	0.469
	AIPW	GRF	1.596	0.294
		MLP	1.362	0.216
		Transformer (no mask)	1.389	0.307
		Transformer (Sep)	0.961	0.204
		Transformer (Ours)	**0.169**	**0.051**
Lalonde-PSID	G-formula	GRF	1.009	0.021
		MLP	0.994	0.213
		Transformer (no mask)	1.205	0.199
		Transformer (Ours)	**0.938**	**0.008**
	IPW	GRF	9.408	1.108
		MLP	1.850	0.818
		Transformer (no mask)	1.937	0.142
		Transformer (Ours)	**0.257**	**0.093**
	AIPW	GRF	2.517	0.242
		MLP	2.038	0.482
		Transformer (no mask)	3.017	0.848
		Transformer (Sep)	3.479	0.831
		Transformer (Ours)	**0.118**	**0.068**
ACIC	G-formula	GRF	0.346	**0.044**
		MLP	0.558	0.062
		Transformer (no mask)	0.675	0.072
		Transformer (Ours)	**0.315**	0.045
	IPW	GRF	4.281	1.388
		MLP	6.380	1.700
		Transformer (no mask)	6.750	1.640
		Transformer (Ours)	**3.125**	**0.704**
	AIPW	GRF	0.857	0.059
		MLP	1.244	0.141
		Transformer (no mask)	1.162	0.135
		Transformer (Sep)	0.958	**0.018**
		Transformer (Ours)	**0.781**	0.031

Table 2: Performance comparison of U-statistics and V-statistics estimators on the Demand dataset across different sample sizes.

Sample Size	Estimator	Model	NRMSE (mean)	NRMSE (SE)
1,000	U-statistics	MLP	0.127	0.008
		Transformer (no mask)	0.164	0.013
		Transformer (Ours)	**0.109**	**0.007**
	V-statistics	MLP	0.138	0.008
		Transformer (no mask)	0.143	0.008
		Transformer (Ours)	**0.131**	**0.007**
5,000	U-statistics	MLP	0.114	0.006
		Transformer (no mask)	0.125	0.013
		Transformer (Ours)	**0.096**	**0.005**
	V-statistics	MLP	0.126	0.006
		Transformer (no mask)	0.136	0.016
		Transformer (Ours)	**0.108**	**0.006**
10,000	U-statistics	MLP	0.137	0.012
		Transformer (no mask)	0.292	0.105
		Transformer (Ours)	**0.102**	**0.004**
	V-statistics	MLP	0.132	0.007
		Transformer (no mask)	0.129	0.007
		Transformer (Ours)	**0.103**	**0.004**
50,000	U-statistics	MLP	0.181	0.070
		Transformer (no mask)	0.247	0.089
		Transformer (Ours)	**0.082**	**0.005**
	V-statistics	MLP	0.279	0.064
		Transformer (no mask)	0.249	0.056
		Transformer (Ours)	**0.146**	**0.022**

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

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

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

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.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric J Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1295–1308, 2024.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

Zhenyu Guo, Shuai Zheng, Zhizhe Liu, Kun Yan, and Zhenfeng Zhu. Cetransformer: Causal effect estimation via transformer based representation learning. *International Conference on Pattern Recognition*, pp. 524–530, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770–778, 2016.

James J Heckman and Edward J Vytlacil. Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics*, 6:4779–4874, 2007.

M. A. Hernán and J. M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2024.

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

Han Huang et al. Short-term and long-term causal effect estimation with double-head transformer. *arXiv preprint*, 2023.

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

Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33:12697–12709, 2020.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Benjamin Kompa, David Remy Bellamy, Tom Kolokotrones, James Robins, and Andrew Beam. Deep learning methods for proximal inference via maximum moment restriction. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=fRWwcgfXXZ`.

Jiewen Liu, Chan Park, Kendrick Li, and Eric J Tchetgen Tchetgen. Regression-based proximal causal inference. *American Journal of Epidemiology*, 2024. kwae370.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, 2020. URL https://api.semanticscholar.org/CorpusID:233236161.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15293–15329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

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

J. M. Robins, A. Rotnitzky, and L. Zhao. Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, September 1994. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1994.10476818. URL https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476818.

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

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An Introduction to Proximal Causal Inference. *Statistical Science*, 39(3):375 – 390, 2024. doi: 10.1214/23-STS911. URL https://doi.org/10.1214/23-STS911.

Eric J Tchetgen Tchetgen, Aolin Ying, Yifan Cui, Xu Shi, and Wang Miao. Introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

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.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Andrew Ying, Yifan Cui, and Eric J Tchetgen Tchetgen. Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):684–704, 2023.

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

YiFan Zhang, Hanlin Zhang, Zachary Chase Lipton, Li Erran Li, and Eric Xing. Exploring transformer backbones for heterogeneous treatment effect estimation. *Transactions on Machine Learning Research*, 2023b. ISSN 2835-8856. URL https://openreview.net/forum?id=1kl4YM2Q7P.

A Appendix

A.1 Causal Assumptions

To ensure valid causal inference, several key assumptions must hold. In this paper, we primarily focus on three fundamental assumptions:

1. **Positivity (or Overlap)**: For every $x \in \text{support}(X)$, and $\forall a \in \{0, 1\}$, $P(A = a|X = x) > 0$.

   This assumption ensures that there is a non-zero probability of receiving each treatment level for all possible values of the observed covariates. It is crucial for estimating treatment effects across the entire covariate space and prevents extrapolation to regions where we have no information about one of the treatment groups.

2. **Exchangeability (or Unconfoundedness)**: $Y^a \perp\!\!\!\perp A|X, \forall a \in \{0, 1\}$.

   This assumption implies that, conditional on the observed confounders X, the potential outcomes Y^a are independent of the treatment assignment A. In other words, after controlling for X, there are no unmeasured confounders that affect both the treatment assignment and the outcome. This is also known as the "no unmeasured confounding" assumption.

3. **Consistency**: If $A = a$, then $Y^a = Y$.

   This assumption states that the potential outcome under a particular treatment level is the same as the observed outcome if the individual actually receives that treatment level. It ensures that the observed outcomes can be used to estimate the potential outcomes.

For the Lalonde and ACIC experiments, we assume that all three of these assumptions hold. For the proximal inferernce experiment, we remove the strong assumption of no unmeasured confounding (Assumption 2).

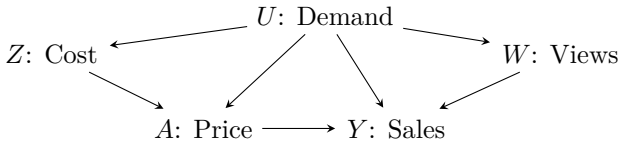

Figure 3: Causal DAG for the Demand experiment.

A.2 Hyperparameter Tuning

We performed extensive hyperparameter tuning for each dataset and estimation method. The hyperparameter spaces explored are detailed below. For each configuration, we report the range of values considered during the tuning process.

For each dataset and estimation method, we performed a grid search over these hyperparameter spaces. The final model for each configuration was selected based on the best performance on a held-out validation set.

Table 3: Hyperparameter tuning ranges for Lalonde-CPS dataset

Parameter	G-formula	IPW	AIPW
Number of epochs	80	20	20
Batch size	32	32	32
Learning rate	0.001	0.001	0.001
L2 penalty	3e-05, 3e-03	3e-05, 3e-03	3e-05, 3e-03
Network width (MLP)	80	80	80
Input layer depth (MLP)	2–4	1–2	2–6
Number of layers (encoder)	2–4	1–2	1–2
Dropout rate	0.0001	0.0001–0.001	0.0001
Embedding dimension (encoder)	40	40	40
Feedforward dimension (encoder)	80	80	80
Number of heads (encoder)	2	1–2	1–2
Encoder weight (alpha)	0.02	0.002–0.02	0.02

Table 4: Hyperparameter tuning ranges for Lalonde-PSID dataset

Parameter	G-formula	IPW	AIPW
Number of epochs	100	30	30
Batch size	32, 64	64	32, 64
Learning rate	0.001–0.01	0.001	0.001
L2 penalty	3e-05	3e-05	3e-05–3e-03
Network width (MLP)	80	10, 20	40
Input layer depth (MLP)	6–16	1	4–8
Number of layers (encoder)	1–2	1	1–2
Dropout rate	0.0001	0.0001	0.0001
Embedding dimension (encoder)	40	10–40	20–40
Feedforward dimension (encoder)	80	20–80	40–80
Number of heads (encoder)	1–2	1	2
Encoder weight (alpha)	0.002–0.02	0.02–0.2	0.02–0.2

Table 5: Hyperparameter tuning ranges for ACIC dataset

Parameter	G-formula	IPW	AIPW
Number of epochs	500	30	500
Batch size	64	64, 128, 256	64, 128, 256
Learning rate	1e-03	1e-03	1e-04
L2 penalty	3e-08	3e-05	3e-08, 3e-04
Network width (MLP)	40	40, 60, 80	40, 60, 80, 120
Input layer depth (MLP)	8–16	4–16	2–16
Number of layers (encoder)	16	2–4	2–8
Dropout rate	0.0001	0.0001	0.0001–0.003
Embedding dimension (encoder)	256	40, 60, 80	256
Feedforward dimension (encoder)	1024	80, 320	512, 1024
Number of heads (encoder)	4	2	2–4
Encoder weight (alpha)	0.02–0.2	0.002–0.2	0.1–2

Table 6: Hyperparameter tuning ranges for Demand dataset (Proximal Inference)

Parameter	Range
Number of epochs	1000
Batch size	32, 64
Learning rate	0.001
L2 penalty	3e-06
Network width (MLP)	160
Input layer depth (MLP)	8, 16
Number of layers (encoder)	1, 2
Dropout rate	0
Embedding dimension (encoder)	40
Feedforward dimension (encoder)	40, 80
Number of heads (encoder)	1, 2
Encoder weight (alpha)	0.001–0.025