DAG-aware Transformer for Causal Effect Estimation

Manqing Liu *
Department of Epidemiology, CAUSALab
Harvard School of Public Health
manqingliu@g.harvard.edu

David R. Bellamy
Flagship Labs 97 Inc.
dbellamy@fl97inc.com

Andrew L. Beam
Department of Epidemiology, CAUSALab
Harvard School of Public Health
andrew_beam@hms.harvard.edu

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 [9, 7, 26]. Within the realm of public policy, they inform the design and evaluation of interventions, from education reforms to social welfare programs [11, 10]. In economics, ATE and CATE estimations are crucial for understanding the impacts of economic policies, labor market interventions, and consumer behavior [1, 8].

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 [9, 22, 3]. These methods, while powerful, are sensitive to model misspecification, which can lead to biased estimates and potentially misleading conclusions [12, 6]. 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 [4, 26].

*Causal Representation Learning Workshop, 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

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 [2] introduced causal trees and forests, adapting random forest algorithms to estimate heterogeneous treatment effects with valid statistical inference. Expanding on this, Wager and Athey [26] 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. [4] 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. Shalit et al. [24] 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. Ma et al. [16] 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. Melnychuk et al. [18] introduced the Causal Transformer for estimating counterfactual outcomes over time, effectively capturing long-range dependencies in longitudinal data. Zhang et al. [28] proposed TransTEE, a transformer-based model for Heterogeneous Treatment Effect (HTE) estimation that handles various types of treatments. Zhang et al. [27] developed Causal Inference with Attention (CInA), enabling zero-shot causal inference on unseen tasks with new data. These deep learning methods offer new ways to handle the challenges of high-dimensional data and complex causal structures in modern causal inference problems.

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. In addition, a particularly persistent challenge in the field is the incorporation of unmeasured confounding into modern DL models, such as transformers [18, 28, 27].

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 the bridge function $h(A, W, X)$. 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. This approach allows for seamless integration of these estimated models into IPTW, doubly robust estimators, and proximal inference methods. 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.

The key contributions of this paper are:

- Development of a DAG-aware transformer model that explicitly incorporates causal structure into the attention mechanism, allowing for more accurate modeling of causal relationships.

- Concurrent estimation of propensity score and outcome models within a single, unified framework, improving efficiency and potentially reducing bias in causal effect estimation.

- Seamless integration of the proposed method with established causal inference techniques such as IPTW, doubly robust estimation and proximal inference, enhancing their performance in complex, high-dimensional settings.

- Empirical evaluation of the proposed method on both simulated and real-world datasets, demonstrating its effectiveness in estimating causal effects across various scenarios.

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 $X = x$: $\tau(x) = \mathbb{E}[Y^1 - Y^0 | X = x]$. The CATE allows us to capture heterogeneity in treatment effects across different subgroups defined by their covariate values. It's particularly useful in personalized medicine, targeted policy interventions, and other scenarios where the effect of a treatment may vary substantially across different segments of the population.

2.2 Confounding Control Methods assuming Unconfoundedness

In causal inference, several methods have been developed to control for *observed* confounding and estimate treatment effects. While we explored multiple approaches, our paper focuses primarily on two methods: Inverse Probability of Treatment Weighting (IPTW) and Augmented Inverse Probability Weighting (AIPW), a form of Doubly Robust estimator [9].

1. **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:

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

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

2. **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:

$$\tau_{AIPW} = \mathbb{E}\Big[\big(\mu(1,X) + \frac{A}{\pi(X)}(Y - \mu(1,X))\big) - \big(\mu(0,X) + \frac{1-A}{1-\pi(X)}(Y - \mu(0,X))\big)\Big] \tag{2}$$

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

2.3 Proximal Inference

In proximal inference [25], 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 *Given* (A, U, W, X, Y, Z), $Y \perp\!\!\!\perp Z | A, U, X$ *and* $W \perp\!\!\!\perp (A, Z) | U, X$.

Assumption 2 *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 *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:

$$\mathbb{E}[Y|A = a, X = x, Z = z] = \int_{\mathcal{W}} h(a, w, x) p(w|a, x, z) dw \tag{3}$$

The expected potential outcomes are given by:

$$\mathbb{E}[Y^a] = \mathbb{E}_{W,X}[h(a, W, X)] \tag{4}$$

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 assuming unconfoundedness

We propose a novel DAG-aware Transformer model for causal effect estimation that explicitly incorporates causal structure into the attention mechanism. Given a dataset of N observations, we define input nodes $\mathcal{X}$, which include the treatment A, *observed* confounding variables $\mathbf{X}$, and outcome Y. The output nodes are $\hat{A}$ and $\hat{Y}$.

3.1 DAG-aware Transformer Architecture

See Figure 1 for an illustration of our model architecture. 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. Each element $M_{ij}^{adj} = 1$ indicates that there is a directed edge from node i to node j. We then transform this into an attention mask $\mathbf{M}$:

$$M_{ij} = \begin{cases} 0 & \text{if } M_{ji}^{adj} = 1 \text{ or } i = j \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

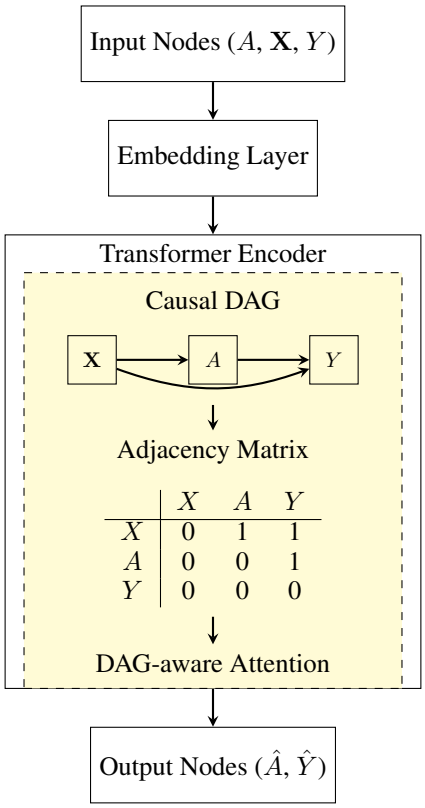

Figure 1: Architecture of the DAG-aware Transformer model. The Transformer Encoder incorporates the DAG-aware attention mechanism (highlighted with dashed lines), which utilizes the causal structure represented by the DAG. The adjacency matrix derived from the causal DAG informs the DAG-aware attention computation. For simplicity, layer normalization and feed-forward networks within the Transformer Encoder are not shown.

This mask ensures attention flows only along causal pathways and allows self-attention for each node. Our key innovation lies in incorporating the causal structure into the attention mechanism. In each multi-head attention layer, we compute attention scores $\mathbf{A} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{E}}$ and apply the DAG-based mask $\mathbf{A}^{mask} = \mathbf{A} + \mathbf{M}\cdot(-\infty)$, where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N\times D\times E}$ are the query and key matrices, and E is the embedding dimension. This operation effectively sets attention scores to zero (after softmax) for node pairs not causally linked in the DAG.

The masked attention scores are then normalized using softmax and used to compute the output:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\mathbf{A}^{mask})\mathbf{V} \tag{6}$$

where $\mathbf{V}$ is the value matrix.

A key feature of our approach is the simultaneous estimation of both the propensity score model and the outcome regression model within the same architecture. The propensity score model $\hat{\pi}(x) = P(A = 1|X = x)$ is estimated through the output head for node A. Given input features X, the model outputs a probability distribution over treatment assignment. Simultaneously, the outcome model $\hat{\mu}(a, x) = E[Y|A = a, X = x]$ is estimated through the output head for node Y, considering both X and A. These estimates are then plugged into the IPTW and AIPW estimators to compute the ATE and CATE.

3.2 Model Training and Objective Function

Building upon the work of [23], we train our DAG-aware Transformer model in an end-to-end manner. The model learns representations $\Phi(\mathbf{X})$ for the treatment and $h(\Phi(\mathbf{X}), A)$ for the outcome simultaneously. We employ the Adam optimizer with weight decay regularization [15] for training.

Our objective function combines a weighted empirical risk term with a regularization term based on the Integral Probability Metric (IPM). This formulation allows us to balance between fitting the observed data and ensuring that the learned representations ($\Phi(\mathbf{X})$) are similar for the treated and control groups. The objective function is defined as:

$$h^* = \min_h \left[\underbrace{\frac{1}{n}\sum_{i=1}^{n} w'_a(x_i) L(h(\Phi(x_i), a_i), y_i)}_{\text{empirical weighted risk}} + \alpha \underbrace{\text{IPM}_G\left(\{\Phi(x_i)\}_{i:a_i=0}, \{\Phi(x_i)\}_{i:a_i=1}\right)}_{\text{distributional distance}} \right] \quad (7)$$

In this equation, $w'_a(x_i) = \frac{w_a(x_i)}{2}\left(\frac{a_i}{\hat{\pi}_1} + \frac{1-a_i}{\hat{\pi}_0}\right)$ represents the sample weight, where $\hat{\pi}_a = \frac{1}{n}\sum_{i=1}^{n} I\{a_i = a\}$ is the empirical probability of treatment assignment. The stabilizing weight $w_a(x_i)$ is defined as $w_a(x_i) = \frac{a(1-2e(x_i))+e(x_i)^2}{e(x_i)(1-e(x_i))}$, with $e(x_i)$ being the propensity score. The trade-off hyperparameter α controls the balance between the empirical risk and the distributional distance. For the Integral Probability Metric IPM_G, we use the Wasserstein distance in our experiments. It's important to note that while the IPM term encourages balancing between the treated and control groups, excessive balancing could potentially increase the Representation-Induced Confounding Bias (RICB) [19]. This occurs because strong balancing might lead to a loss of information about confounders in the learned representations. Therefore, careful tuning of the hyperparameter α is crucial to achieve an optimal trade-off between balancing and preserving important confounding information.

3.3 Hyperparameter Tuning and Model Selection

Hyperparameter tuning and model selection present unique challenges in causal inference, as unlike traditional machine learning tasks with observed labels and cross-validation procedures, we cannot directly observe counterfactual potential outcomes [23, 17]. To address this issue, several methods have been developed [21, 23]. In our work, we adopt the approach proposed by [23], which aligns well with our training scheme and offers a straightforward implementation.

Given the unobservable nature of true causal effects, we estimate surrogate metrics that approximate these effects. Our procedure is as follows:

1. We train a plug-in estimator $\hat{\tau}$ on the validation set using generalized random forests [26].
2. We then select models and tune hyperparameters by finding estimators $\tilde{\tau}$ that minimize the difference between $\hat{\tau}$ and $\tilde{\tau}$.

Formally, we select the optimal estimator $\tilde{\tau}^*$ according to $\tilde{\tau}^* = \arg\min_{\tilde{\tau}\in\mathcal{T}} \text{NRMSE}(\hat{\tau}, \tilde{\tau})$, where $\mathcal{T}$ is the set of candidate estimators, and Normalized Root Mean Squared Error (NRMSE) is defined as $\text{NRMSE} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\tau}(X_i)-\tilde{\tau}(X_i))^2}{\hat{V}(\hat{\tau}(X))}}$. Here, $\{\tilde{\tau}(X_i)\}_{i=1}^{n}$ is a set of ATE or CATE predictions by $\tilde{\tau}(\cdot)$, and $\hat{V}(\hat{\tau}(X))$ is the empirical variance of the ground-truth ATE or CATE approximated by $\hat{\tau}(\cdot)$.

4 Methods for Proximal Inference

Building upon [13], we use our DAG-aware transformer to estimate the bridge function h. For proximal inference, in addition to A and $\mathbf{X}$, $\mathcal{X}$ also includes proxy of treatment Z and proxy of outcome W, and output node is $\hat{Y}$. Our empirical investigations found that the direct application of a transformer encoder to our low-dimensional data ($D = 3$) resulted in suboptimal performance compared to the benchmark Multilayer Perceptron (MLP) model. We hypothesize that this performance degradation is attributable to information loss during the encoding process, a phenomenon particularly pronounced

in low-dimensional settings. To address this, we propose a two-stage architecture: 1. A transformer encoder with DAG-masked self-attention layers, ensuring nodes attend only to their causal parents and themselves; 2. A Multilayer Perceptron (MLP) that combines the encoder output with raw inputs (treatment A and proxies W) to preserve information potentially lost during encoding. The weight of encoder output is a hyperparameter we can tune. The model output $\hat{Y}$ estimates the bridge function $\hat{h}$. We compute its empirical mean using both U-statistics and V-statistics kernel from [13].

5 Experiments

We evaluate our proposed method on four diverse datasets. We use the LaLonde datasets to assess Average Treatment Effect (ATE) estimation, and the ACIC dataset for Conditional Average Treatment Effect (CATE) estimation. We also extend our evaluation to proximal inference using the Demand dataset described in [13]. The specifications of baseline models are provided in Appendix B.1. Causal assumptions of these experiments were listed in Appendix A.

5.1 Lalonde dataset for ATE estimation

The LaLonde datasets, derived from the National Supported Work (NSW) Demonstration program, include the Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID) [14]. We define the treatment (A) as participation in the NSW program, the outcome (Y) as earnings in 1978, and the covariates (X) as age, education, race, marital status, earnings in 1974, and earnings in 1975. The LaLonde CPS dataset comprises 15,992 control units and 185 treated units, while the LaLonde PSID dataset consists of 2,490 control units and the same 185 treated units. The true ATE for both datasets is 1,794.34. To ensure robust evaluation, we create 10 distinct samples from each dataset using bootstrap with replacement with different random seeds. Figures 2 and 3 present the Normalized Root Mean Squared Error (NRMSE) results with standard errors for the LaLonde CPS and PSID datasets, respectively. Tables 1 and 2 in Appendix B.2 provide the corresponding numerical values.

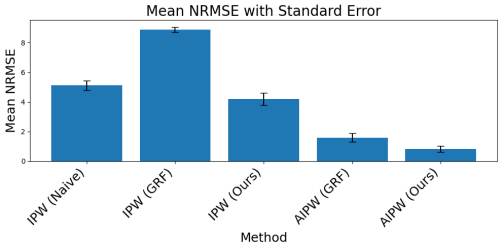

Figure 2: Mean NRMSE with Standard Error for LaLonde CPS Dataset

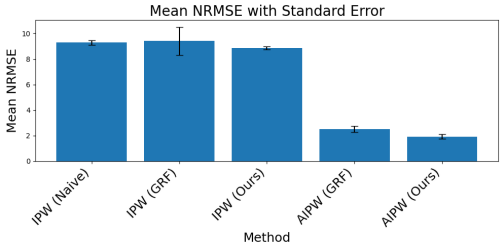

Figure 3: Mean NRMSE with Standard Error for LaLonde PSID Dataset

Our proposed AIPW method achieves the lowest NRMSE on both datasets, with 48.2% and 23.2% reductions compared to AIPW (GRF) for LaLonde CPS and PSID, respectively. The performance gap is more pronounced in the LaLonde CPS dataset, suggesting better generalization to diverse control populations. Notably, our method consistently achieves lower NRMSE with smaller standard errors, indicating more accurate and stable causal effect estimations in challenging real-world scenarios. We chose not to use the RealCause simulated LaLonde data [20] due to significant discrepancies in the true ATE, which could lead to misleading comparisons.

5.2 ACIC dataset for CATE estimation

The ACIC dataset from the 2016 Atlantic Causal Inference Conference data challenge [5] is based on real covariates with synthetically simulated treatment assignment and potential outcomes. We analyze 10 instances from different data-generating processes, each containing 58 pre-treatment variables, a binary treatment assignment, observed outcome, and ground truth potential outcomes. Figure 4 presents the NRMSE with standard errors with respect to the 10 instances of datasets for each method.

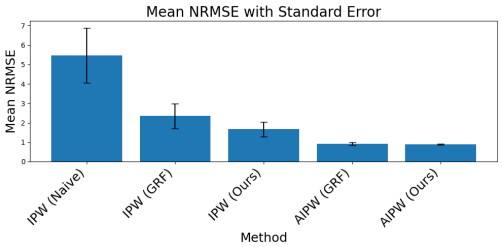

Figure 4: Mean NRMSE with Standard Error for ACIC Dataset

Our proposed AIPW method demonstrates superior performance, achieving the lowest NRMSE of 0.894, closely followed by AIPW (GRF) with 0.920. The IPW methods show higher error rates, with our IPW variant outperforming IPW (GRF) and IPW (Naive). Importantly, our proposed methods (both IPW and AIPW variants) consistently outperform their GRF counterparts, underscoring the effectiveness of our approach in CATE estimation tasks. Table 3 in Appendix B.3 provide the corresponding numerical values.

5.3 Demand for proximal inference

The goal of this experiment is to estimate the effect of airline ticket price A on sales Y, confounded by unobserved demand U. We use fuel cost Z as a treatment-inducing proxy and website views W as an outcome-inducing proxy (Figure 6). We train our model on simulated datasets with sample sizes ranging from 1,000 to 50,000. Performance is evaluated using causal mean squared error (c-MSE) across 10 equally-spaced price points between 10 and 30, comparing estimated potential outcomes $\hat{E}[Y^a]$ against Monte Carlo simulations of the true $E[Y^a]$. We use a heldout dataset of 1,000 draws from W to compute predictions. We perform 20 replicates for each sample size. Figure 5 summarizes the c-MSE distribution across sample sizes. Our model consistently outperforms the previous state-of-the-art model (MLP) reported in [13]. Table 4 in Appendix B.4 provides the corresponding numerical values.

Figure 5: Median (IQR) of c-MSE for Demand Dataset

6 Conclusion

In this paper, we introduced a novel transformer-based approach for causal inference that addresses key limitations in existing methods. Our model's primary innovation lies in its ability to encode any causal DAGs into the attention mechanism, allowing it to handle a wide range of causal scenarios. The causal-aware attention mechanism we developed explicitly models the encoded causal structure, leading to more accurate estimation of treatment effects. Our experimental results demonstrate the effectiveness of our approach across various synthetic and real-world datasets, showing improved performance compared to existing methods. While our work represents a significant step forward in causal inference using transformer architectures, there are several directions for future research. First, we recognize the importance of swift adaptability to a wider range of causal structures. Future work can explore developing a more generalized encoding mechanism that can quickly accommodate diverse causal graphs without requiring extensive retraining. This could involve creating a meta-learning framework that learns to adapt to new causal structures efficiently. Additionally, we acknowledge the need to investigate the robustness of our approach against potential noise or misspecifications in the input DAG. Future studies can systematically introduce perturbations to the causal graph to assess how our model's performance degrades under various levels of DAG uncertainty.

References

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

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

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

[4] 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.

[5] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

[6] 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.

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] 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, editors, *Advances in Neural Information Processing Systems*, 2022.

[14] Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[16] Yunpu Ma and Volker Tresp. Causal inference under networked interference and intervention policy enhancement. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[17] Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. Empirical analysis of model selection for heterogeneous causal effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024.

[18] 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, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15293–15329. PMLR, 17–23 Jul 2022.

[19] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *The Twelfth International Conference on Learning Representations*, 2024.

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

[21] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 2017.

[22] 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.

[23] Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pages 8398–8407. PMLR, 2020.

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

[25] 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.

[26] 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.

[27] 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*, 2023.

[28] 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*, 2023.

A 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).

B Experiments

B.1 Baseline models

For Lalonde and ACIC datasets, our baseline models include Naive Inverse Probability Weighting (IPW), which uses uniform weights, and IPW with Generalized Random Forests (GRF) [26] for propensity score estimation. Additionally, we consider Augmented IPW (AIPW) with GRF, which employs a doubly robust estimation approach. All baseline models (except the naive model) were fine-tuned using the `grf` package in R. For Demand dataset, we used the Multilayer Perceptron (MLP) as our baseline models (the hyper parameters of the MLP were fine-tuned in [13]).

B.2 LaLonde Dataset Results

Table 1: Results on LaLonde CPS Dataset

Method	NRMSE	SE of NRMSE
IPW (Naive)	5.106	0.340
IPW (GRF)	6.342	1.227
IPW (Ours)	4.209	0.414
AIPW (GRF)	1.596	0.294
AIPW (Ours)	**0.826**	**0.194**

Table 2: Results on LaLonde PSID Dataset

Method	NRMSE	SE of NRMSE
IPW (Naive)	9.280	0.168
IPW (GRF)	9.408	1.108
IPW (Ours)	8.857	0.118
AIPW (GRF)	2.517	0.242
AIPW (Ours)	**1.934**	**0.177**

B.3 ACIC Dataset Results

B.4 Proximal inference

Table 3: Results on ACIC Dataset

Method	NRMSE	SE of NRMSE
IPW (Naive)	5.473	1.421
IPW (GRF)	2.342	0.637
IPW (Ours)	1.669	0.378
AIPW (GRF)	0.920	0.082
AIPW (Ours)	**0.894**	**0.033**

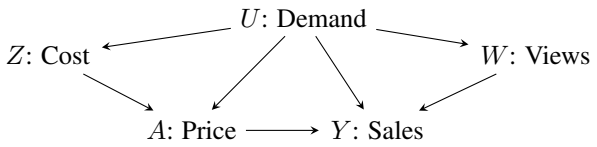

Figure 6: Causal DAG for the Demand experiment.

Table 4: Demand Median (IQR) values

Method	Training set size			
	1000	5000	10000	50000
NMMR-V (MLP)	23.41 (11.26)	30.74 (17.73)	42.88 (29.45)	62.18 (16.97)
NMMR-V (Ours)	21.54 (17.42)	24.46 (17.93)	21.37 (10.12)	27.50 (16.30)
NMMR-U (MLP)	23.68 (8.02)	16.21 (10.55)	14.25 (4.46)	14.27 (12.47)
NMMR-U (Ours)	**10.69 (14.72)**	**7.67 (6.70)**	**5.56 (6.72)**	**6.51 (5.90)**