Exploring Transformer Backbones for Heterogeneous Treatment Effect Estimation

Yi-Fan Zhang*
University of Chinese Academy of Sciences

Hanlin Zhang*
Carnegie Mellon University

Zachary Lipton
Carnegie Mellon University

Li Erran Li
Amazon AWS

Eric Xing
Mohamed Bin Zayed University of Artificial Intelligence
Carnegie Mellon University, and Petuum Inc.

Abstract

Neural networks (NNs) are often leveraged to represent structural similarities of potential outcomes (POs) of different treatment groups to obtain better finite-sample estimates of treatment effects. However, despite their wide use, existing works handcraft treatment-specific (sub)network architectures for representing various POs, which limit their applicability and generalizability. To remedy these issues, we develop a framework called **Trans**formers as **T**reatment **E**ffect **E**stimators (TransTEE) where attention layers govern interactions among treatments and covariates to exploit structural similarities of POs for confounding control. Using this framework, through extensive experiments, we show that TransTEE can: (1) serve as a general purpose treatment effect estimator which significantly outperforms competitive baselines on a variety of challenging TEE problems (e.g., discrete, continuous, structured, or dosage-associated treatments.) and is applicable both when covariates are tabular and when they consist of structural data (e.g., texts, graphs); (2) yield multiple advantages: compatibility with propensity score modeling, parameter efficiency, robustness to continuous treatment value distribution shifts, interpretability in covariate adjustment, and real-world utility in debugging pre-trained language models.

1 Introduction

Recently, feed-forward neural networks have been adapted to model causal relationships and estimate treatment effects [34, 53, 40, 68, 8, 51, 43, 12], in part due to their flexibility to model nonlinear functions [28] and high-dimensional input [34]. Among them, the specialized NN's architecture plays a key role in learning representations for counterfactual inference [2, 12] such that treatment variables and covariates are well distinguished [53]. Despite these encouraging results, several key challenges make it difficult to adopt these methods as standard tools for treatment effect estimation. We argue that most current works based on subnetworks do not sufficiently exploit the structural similarities of potential outcomes for heterogeneous TEE and accounting for them needs complicated regularizations, reparametrization, or multitask architectures that are problem-specific [12]. Practically, their treatment-specific designs suffer several key weaknesses, including parameter inefficiency (Table 2), brittleness under different scenarios, such as when treatments or dosages shift slightly from the training distribution (Figure 4). We discuss these problems in detail in Sections 4.

To overcome the above challenges and be motivated by the observation that the model structure plays a crucial role in TEE [2, 12], we provide compelling evidence that transformers can outperform multilayer perceptrons and offer a promising alternative approach when lever-

*Equal Contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

aging deep learning to estimate treatment effects. Our work is based on the Transformer architecture [60] which has emerged as an architecture of choice for diverse domains, including natural language processing [60], image recognition [17], and multimodal processing [57].

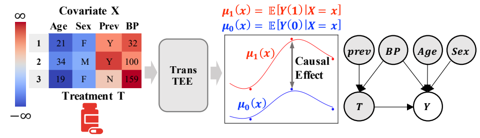

Figure 1: **A motivating example** with a corresponding causal graph. **Prev** denotes previous infection condition and **BP** denotes blood pressure. TransTEE adjusts an appropriate covariate set $\{\textbf{Prev}, \textbf{BP}\}$ with attention which is visualized via a heatmap.

In this paper, we investigate the following question: *can Transformers be similarly effective for treatment effect estimation in problems of practical interest?* Throughout, we adopt the notation of the Rubin-Neyman potential outcomes framework [47] and focus on conditional average treatment effect (CATE) estimation. In particular, we develop TransTEE, a method that builds upon the attention mechanisms and achieves state-of-the-art on a wide range of TEE tasks. Note that the Transformer is originally designed for sequence modeling, to utilize its power in TEE, three key design choices are proposed. First, *treatment and covariate embedding layer* is used to represent covariate and treatment variables separately through learnable embeddings. This design is parameter-efficient in comparison to related works and we show that it appears to perform better under some practically motivated treatment shifts. In summary, we make the following contributions.

1. We propose TransTEE to explore the design space of TEE, showing that Transformers, equipped with the proposed design choices, can be effective and versatile treatment effect estimators under the Rubin-Neyman potential outcome framework. TransTEE is empirically verified to be (i) a general framework applicable for a wide range of neural TEE settings; (ii) compatible with propensity score modeling; (ii) parameter-efficient; (ii) robust under treatment shifts; (iv) interpretable in covariate adjustment; (v) deliverable for real world utility beyond semi-synthetic settings.

2. Experiments are conducted on six benchmarks with four types of treatments in various scenarios to verify the effectiveness of TransTEE and propensity score regularized adversarial training in estimating treatment effects. We show that TransTEE produces covariate adjustment interpretation and significant performance gains given discrete, continuous or structured treatments on popular benchmarks including IHDP, News, TCGA. An empirical study on pre-trained language models is conducted to show the real-world utility of TransTEE that implies potential applications.

2 Problem Statement and Assumptions

Given N observed samples $(\mathbf{x}_i, t_i, s_i, y_i)_{i=1}^N$, each containing N pre-treatment covariates $\{\mathbf{x}_i \in \mathbb{R}^p\}_{i=1}^N$, the treatment variable t_i in this work has various support, e.g., $\{0, 1\}$ for binary settings, $\mathbb{R}$ for continuous settings, and graphs/words for structured settings. For each sample, the potential outcome (μ-model) $\mu(\mathbf{x}, t)$ or $\mu(\mathbf{x}, t, s)$ is the response of the i-th sample to a treatment t, where in some cases each treatment will be associated with a dosage $s_{t_i} \in \mathbb{R}$. The propensity score (π-model) is the conditional probability of treatment assignment given the observed covariates $\pi(T = t|X = \mathbf{x})$. The above two models can be parameterized as μ_θ and π_ϕ, respectively. The task is to estimate the Average Dose Response Function (ADRF): $\mu(\mathbf{x}, t) = \mathbb{E}[Y|X = \mathbf{x}, do(T = t)]$ [55], which includes special cases in discrete treatment scenarios that can also be estimated as the Average Treatment Effect (ATE): $ATE = \mathbb{E}[\mu(\mathbf{x}, 1) - \mu(\mathbf{x}, 0)]$ and its individual version ITE.

Assumption 2.1. We assume no hidden confounders such that $Y(T = t) \perp\!\!\!\perp T|X$. In the binary treatment case, $Y(0), Y(1) \perp\!\!\!\perp T|X$. Besides, the treatment assignment is non-deterministic such that, i.e. $0 < \pi(t|x) < 1, \forall x \in \mathcal{X}, t \in \mathcal{T}$

3 TransTEE: Transformers as Treatment Effect Estimators

Preliminary. The main module in TransTEE is the attention layer [60]: given d-dimensional query, key, and value matrices $Q \in \mathbb{R}^{d \times d_k}, K \in \mathbb{R}^{d \times d_k}, V \in \mathbb{R}^{d \times d_v}$, attention mechanism computes the outputs as $\mathcal{H}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$. In practice, multi-head attention is preferable to jointly attend to the information from different representation subspaces at different positions. $\mathcal{H}_M(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O$, where $head_i = \mathcal{H}(QW_i^Q, KW_i^K, VW_i^V)$, where $W_i^Q \in \mathbb{R}^{d \times d_k}, W_i^V \in \mathbb{R}^{d \times d_k}, W_i^V \in \mathbb{R}^{d \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d}$ are learnable matrices.

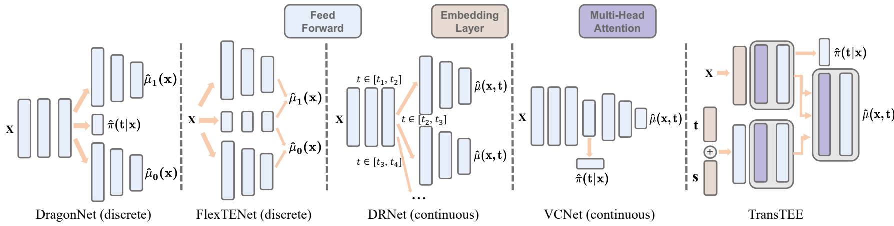

Figure 2: **A schematic comparison** of TransTEE and recent works including DragonNet[54], FlexTENet[12], DRNet[51] and VCNet[43]. TransTEE handles all the scenarios without handcrafting treatment-specific architectures and any additional parameter overhead.

Covariate and Treatment Embedding Layers. *(1) Treatment Embedding Layer.* We use two learnable linear layers to project scalar treatments and dosages to d-dimension vectors separately: $M_t = \text{Linear}(t), M_s = \text{Linear}(s)$, where $M_t \in \mathbb{R}^d$. $M_s \in \mathbb{R}^d$ exists just when each treatment has a dosage parameter, otherwise, only treatment embedding is needed. When multiple (n) treatments act simultaneously, the projected matrix will be $M_t \in \mathbb{R}^{d\times n}$, $M_s \in \mathbb{R}^{d\times n}$ and when facing structural treatments (languages, graphs), the embedding of the treatment will be projected by language models and graph neural networks respectively. By using the treatment embeddings, TransTEE is shown to be (i) *robust under treatment shifts* (Proposition 2 in Appendix D), and (ii) *parameter-efficient* (Figure 2 and Table. 2). *(2) Covariates Embedding Layer.* Different from previous works that embed all covariates by one fully connected layer, where the differences between covariate tend to be lost, and is hard to study the function of an individual covariate in a sample. TransTEE learns different embeddings for each covariate, namely $M_x = \text{Linear}(\mathbf{x})$, and $M_x \in \mathbb{R}^{d\times p}$, where p is the number of covariate. Covariates embedding enables us to study the effect of the individual covariates on the outcome.

Covariate and Treatment Self-Attention For covariates, prevalent methods represent covariates as a whole feature using MLPs, where pairwise covariate interactions are lost when adjusting covariates. Therefore, we cannot study the effect of each covariate on the estimated result. In contrast, TransTEE processes each covariate embedding independently and models their interactions by self-attention layers. Namely, $\hat{M}_x^l = \mathcal{H}_M(M_x^{l-1}, M_x^{l-1}, M_x^{l-1}) + M_x^{l-1}, M_x^l = \text{MLP}(\text{BN}(\hat{M}_x^l)) + \hat{M}_x^l$, where M_x^l is the output of l layer and BN is the BatchNorm layer. Simultaneously, the treatments and dosages embeddings are concatenated and projected to the latent dimension by a linear layer, which generates a new embedding $M_{st} \in \mathbb{R}^d$. Then self-attention is applied $M_{st}^l = \mathcal{H}_M(M_{st}^{l-1}, M_{st}^{l-1}, M_{st}^{l-1}) + M_{st}^{l-1}, M_{st}^l = \text{MLP}(\text{BN}(\hat{M}_{st}^l)) + \hat{M}_{st}^l$.

The self-attention layer for treatments enables treatment interactions, an important desideratum for S- and T-learners. Namely, TransTEE can *model the scenario where multiple treatments are applied and attains strong practical utility*, e.g., multiple prescriptions in healthcare or different financial measures in economics. This is an effective remedy for existing methods which are limited to settings where various treatments are not used simultaneously.

Treatment-Covariate Cross-Attention One of the fundamental challenges of causal metalearners is modeling treatment-covariate interactions. TransTEE realizes such a goal using a cross-attention module, treating M_{st} as a query and M_x as both the key and the value $\hat{M}^l = \mathcal{H}_M(M_{st}^{l-1}, M_x^{l-1}, M_x^{l-1}) + M^{l-1}, M^l = \text{MLP}(\hat{M}^l) + \hat{M}^l, \hat{y} = \text{MLP}(\text{Pooling}(M^L))$, where M^L is the output of the last cross-attention layer and $M^0 = M_{st}^L$. The above interactions are particularly important for adjusting proper covariate or confounder sets for estimating treatment effects [59], which empirically yields *suitable covariate adjustment principles (the Disjunctive Cause Criteria) [14, 59] about pre-treatment covariates and confounders* as intuitively illustrated in Figure 1 and corroborated in our experiments.

Denote $\hat{y} := \mu_\theta(\mathbf{x}, t)$ and the training objective is the mean square error (MSE) of the outcome regression is $\mathcal{L}_\theta(\mathbf{x}, y, t) = \sum_{i=1}^n (y_i - \mu_\theta(\mathbf{x}_i, t_i))^2$.

In summary, thanks to the designs described above for modeling treatments and covariates, when combined with strong modeling capacity of Transformers, *TransTEE can be extended to high-dimensional data easily and effectively* on the tabular, graph, and textual data. The generalizability of the TransTEE also allows new applications like auditing language models beyond semi-synthetic settings as shown in the next section. We include an illustration of the TransTEE workflow using a concrete example in Appendix B.

Table 1: **Experimental results comparing NN based methods on the IHDP datasets.** We report the results based on 100 repeats, and numbers after $\pm$ are the estimated standard deviation. For Extrapolation ($h = 2$), models are trained with $t \in [0.1, 2.0]$ and tested in $t \in [0, 2.0]$. For Extrapolation ($h = 5$), models are trained with $t \in [0.25, 5.0]$ and tested in $t \in [0, 5]$.

Methods	Vanilla (Binary)	Vanilla ($h = 1$)	Extrapolation ($h = 2$)	Vanilla ($h = 5$)	Extrapolation ($h = 5$)
TARNet	0.3670 ± 0.61112	2.0152 ± 1.07449	12.967 ± 1.78108	5.6752 ± 0.53161	31.523 ± 1.5013
DRNet	0.3543 ± 0.60622	2.1549 ± 1.04483	11.071 ± 0.99384	3.2779 ± 0.42797	31.524 ± 1.50264
FlexTENet	0.2700 ± 0.10000	——	——	——	——
VCNet	0.2098 ± 0.18236	0.7800 ± 0.61483	NAN	NAN	NAN
TransTEE	$\mathbf{0.0983} \pm \mathbf{0.15384}$	0.1151 ± 0.10289	0.2745 ± 0.14976	0.1621 ± 0.14443	0.2066 ± 0.23258
TransTEE+MLE	0.1721 ± 0.40061	0.0877 ± 0.03352	0.2685 ± 0.17552	0.2079 ± 0.17637	0.1476 ± 0.07123
TransTEE+TR	0.1913 ± 0.29953	0.0781 ± 0.03243	0.2393 ± 0.08154	$\mathbf{0.1143} \pm \mathbf{0.03224}$	$\mathbf{0.0947} \pm \mathbf{0.0824}$
TransTEE+PTR	0.2193 ± 0.34667	$\mathbf{0.0762} \pm \mathbf{0.07915}$	$\mathbf{0.2352} \pm \mathbf{0.17095}$	0.1363 ± 0.08036	0.1363 ± 0.08035

4 Experimental Results

We elaborate basic experimental settings, results, analysis and empirical studies in this section. See Appendix E for full details of all experimental settings and detailed definition of metrics. See Appendix F for many more results and remarks.

Case study on treatment distribution shifts We start by conducting a case study on treatment distribution shifts (Figure 4), and exploring an extrapolation setting in which treatment can subsequently be administered at values never seen before during training. Surprisingly, we find that while standard results rely on constraining the values of treatments [43] and dosages [51] to a specific range, our methods perform surprisingly well when extrapolating beyond these ranges as assessed on several empirical benchmarks. By comparison, many other methods appear to be comparatively brittle in the same settings. See Appendix D for a detailed discussion and analysis.

Case study of propensity modeling. TransTEE is conceptually simple and effective. However, when the sample size is small, it becomes important to account for selection bias [2]. However, most existing regularizations can only be used when treatments are discrete [7, 37, 18]. Thus we propose two regularization variants for continuous treatment/dosages, which are termed Treatment Regularization (TR, $\mathcal{L}_{\phi}^{TR}(\mathbf{x}, t) = \sum_{i=1}^{n} \left(t_i - \pi_{\phi}(\hat{t}_i|\mathbf{x}_i)\right)^2$) and its probabilistic version Probabilistic Treatment Regularization (PTR, $\mathcal{L}_{\phi}^{PTR} = \sum_{i=1}^{n} \left[\frac{(t_i - \pi_{\phi}(\mu|\mathbf{x}_i))^2}{2\pi_{\phi}(\sigma^2|\mathbf{x}_i)} + \frac{1}{2}\log \pi_{\phi}(\sigma^2|\mathbf{x}_i)\right]$) respectively. The overall model is trained in an adversarial pattern, namely $\min_{\theta} \max_{\phi} \mathcal{L}_{\theta}(\mathbf{x}, y, t) - \mathcal{L}_{\phi}(\mathbf{x}, t)$. Specifically, a propensity score model $\pi_{\phi}(t|\mathbf{x})$ parameterized by an MLP is learned by minimizing $\mathcal{L}_{\phi}(\mathbf{x}, t)$, and then the outcome estimators $\mu_{\theta}(\mathbf{x}, t)$ is trained by $\min_{\theta} \mathcal{L}_{\theta}(\mathbf{x}, y, t) - \mathcal{L}_{\phi}(\mathbf{x}, t)$. To overcome selection biases, the bilevel optimization enforces effective treatment effect estimation while modeling the discriminative propensity features to partial out parts of covariates that cause the treatment but not the outcome and dispose of nuisance variations of covariates [36].

Continuous dosage. In Table 3, we compare TransTEE against baselines on the TCGA (D) dataset with default treatment selection bias 2.0 and dosage selection bias 2.0. As the number of treatments increases, TransTEE and its variants (with regularization term) consistently outperform the baselines by a large margin on both training and test data. TransTEE's effectiveness is also shown in Appendix Figure 6, where the estimated ADRF curve of each treatment considering continuous dosages is plotted. Compared to baselines, TransTEE attains better results over all treatments. Stronger selection bias in the observed data makes estimation more difficult because it becomes less likely to see certain treatments or particular covariates. Considering different dosage and treatment selection biases, Appendix Figure 5 shows that as biases increase, TransTEE consistently performs the best.

Structured treatments. We compared the performance of TransTEE to baselines on the training and test set of the SW and TCGA datasets with varying degrees of treatment selection bias. The numerical results are shown in Appendix Table 9. The performance gain between GNN and Zero indicates that taking into account graph information significantly improves estimation. The results suggest that, overall, the performance of TransTEE is the best due to the strong modeling ability and advanced model structureto process high-dimensional treatments.

5 Concluding Remarks

In this work, we show that transformers can be effective and versatile treatment effect estimators. Extensive experiments well verify the effectiveness and utility of TransTEE, which also imply that a more challenging and unified evaluation alternatives of TEE with domain experts are needed.

References

[1] Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 2016.

[2] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *ICML*, 2018.

[3] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *NeurIPS*, 2017.

[4] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *arXiv*, 2017.

[5] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 2011.

[6] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 2010.

[7] Ioana Bica, Ahmed M Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *arXiv*, 2020.

[8] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. 2020.

[9] Kyle Chang, Chad J Creighton, Caleb Davis, Lawrence Donehower, Jennifer Drummond, David Wheeler, Adrian Ally, Miruna Balasundaram, Inanc Birol, Yaron SN Butterfield, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 2013.

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv*, 2021.

[11] Jonathan Crabbé, Alicia Curth, Ioana Bica, and Mihaela van der Schaar. Benchmarking heterogeneous treatment effect models through the lens of interpretability. *arXiv preprint arXiv:2206.08363*, 2022.

[12] Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In *NeurIPS*, 2021.

[13] Ralph B D'Agostino. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine*, 1998.

[14] Xavier De Luna, Ingeborg Waernbaum, and Thomas S Richardson. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, 2011.

[15] Peng Ding, TJ VanderWeele, and James M Robins. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 2017.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[18] Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *DMKD*, 2021.

[19] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 2021.

[20] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 2002.

[21] 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*, 2011.

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016.

[23] Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *EMNLP Findings*, 2020.

[24] Zhenyu Guo, Shuai Zheng, Zhizhe Liu, Kun Yan, and Zhenfeng Zhu. Cetransformer: Casual effect estimation via transformer based representation learning. In *PRCV*, 2021.

[25] Shonosuke Harada and Hisashi Kashima. Graphite: Estimating individual effects of graph-structured treatments. In *CIKM*, 2021.

[26] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2011.

[27] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

[28] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 1989.

[29] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *arXiv*, 2022.

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

[31] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

[32] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2022.

[33] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.

[34] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *ICML*, 2016.

[35] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv*, 2020.

[36] Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. 2021.

[37] Nathan Kallus. Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. In *ICML*, 2020.

[38] 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*, 2007.

[39] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[40] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard S Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, 2017.

[41] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv*, 2016.

[42] David Newman. Bag of words data set. *UCI Machine Learning Respository*, 2008.

[43] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. 2021.

[44] Sonali Parbhoo, Stefan Bauer, and Patrick Schwab. Ncore: Neural counterfactual representation learning for combinations of treatments. *arXiv*, 2021.

[45] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 2014.

[46] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *JASA*, 1984.

[47] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 2005.

[48] Donald B Rubin. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 2007.

[49] Donald B Rubin and Neal Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 1996.

[50] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[51] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *AAAI*, 2020.

[52] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv*, 2018.

[53] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, 2017.

[54] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. 2019.

[55] Brian K Shoichet. Interpreting steep dose-response curves in early inhibitor discovery. *Journal of medicinal chemistry*, 2006.

[56] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. 2019.

[57] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.

[58] Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2006.

[59] Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 2019.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. 2021.

[62] Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *ICML*, 2020.

[63] Haohan Wang, Zeyi Huang, Hanlin Zhang, Yong Jae Lee, and Eric Xing. Toward learning human-aligned cross-domain robust models by countering misaligned features. 2022.

[64] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 1998.

[65] Yifan Wu, Ezra Winston, Divyansh Kaushik, and Zachary Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, 2019.

[66] Guoqiang Xu, Cunxiang Yin, Yuchen Zhang, Yuncong Li, Yancheng He, Jing Cai, and Zhongyu Wei. Learning discriminative representation base on attention for uplift. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 200–211. Springer, 2022.

[67] Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? 2021.

[68] Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*, 2018.

[69] Shuxi Zeng, Serge Assaad, Chenyang Tao, Shounak Datta, Lawrence Carin, and Fan Li. Double robust representation learning for counterfactual prediction, 2020.

[70] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. *CVPR*, 2022.

[71] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *AISTATS*, 2020.

[72] Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. Adversarial multiple source domain adaptation. 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See the scope summarized in the abstract and introduction. The contributions are summarized point by point in Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read the ethics review guidelines and ensured that our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 2.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix C
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have included them in the Appendix E.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We have included them in the Appendix E.
(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We have reported our error bars in terms of standard deviation in the quantitative experiments.
(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We have included them in the Appendix E.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
(a) If your work uses existing assets, did you cite the creators? [Yes] We have cited the datasets (as well as the domain splits) we used in the **Datasets** and **Baselines** paragraphs in Section 4.
(b) Did you mention the license of the assets? [Yes] We have mentioned the lincense in Appendix E.
(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We have included the code, data, and instructions needed to reproduce the main experimental results in the supplemental material.
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We have mentioned that we used the open-sourced datasets (as well as the domain splits) and cited them we used in the **Datasets** paragraph in Section 4.
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] We didn't use any crowdsourcing or conduct research with human subjects.
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] We didn't include any human participant.
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We didn't include any human participant.