Identifying and Estimating Causal Effects under Weak Overlap by Generative Prognostic Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

As an important problem of causal inference, we discuss the identification and estimation of treatment effects (TEs) under weak overlap, i.e., subjects with certain features all belong to a single treatment group. We use a latent variable to model a prognostic score (PGS), which is widely used in biostatistics and sufficient for TEs, i.e., we build a generative prognostic model. We prove that the latent variable recovers a PGS, and the model identifies individualized treatment effects. The model is then learned as the Intact-VAE, a new type of variational autoencoder (VAE). We derive counterfactual generalization bounds which motivate representation balanced for treatment groups conditioned on individualized features. The proposed method is compared with recent methods using (semi-)synthetic datasets.

1 Introduction

Causal inference [21, 34], i.e, inferring causal effects of interventions, is a fundamental problem. In this work, we focus on treatment effects (TEs), such as effects of public policies or a new drug, based on a set of observations consisting of binary labels for treatment / control (non-treated), outcome, and other covariates (e.g, patients' personal records). The fundamental difficulty of causal inference is that we never observe *counterfactual* outcomes, which would have been if we had made the other decision (treatment or control). While the ideal protocol for causal inference is randomized controlled trials (RCTs), they often have ethical and practical issues, or suffer from prohibitive costs. Thus, causal inference from observational data is indispensable. It introduces other challenges, however. The most crucial one is *confounding*: there may be variables (called *confounders*) that causally affect both the treatment and the outcome, and spurious correlation follows.

A large majority of works, including this work, rely on the *unconfoundedness*, which means that appropriate covariates are collected so that the confounding can be controlled by conditioning on covariates. That is, all the confounders are in essence observed. This is still challenging, due to systematic *imbalance* (difference) of the distributions of the covariates between the treatment and control groups, introducing bias in estimation. Among classical ways of dealing with imbalance are matching and re-weighting [44, 35]. Machine learning methods are also exploited; there are semi-parametric methods, e.g, [48, TMLE], which have better finite sample performance, and also non-parametric, tree-based methods, e.g., [49, Causal Forests (CF)]. Notably, starting from [23], there is a recent rise of interest in learning representation of covariates, which is independent of treatment groups, i.e., *balanced* representation learning (BRL) .

The most serious form of imbalance is that sample points with certain values of covariate are all belong to a single treatment group, which is called *weak overlap* of the covariate. Causal effects are not directly estimable at non-overlapped covariate values. There are lines of work that give robustness to weak overlap [3], trim non-overlapped sample points [52], or study convergence rate depending on overlap [19]. Weak overlap is particularly relevant to machine learning methods exploiting rich covariates, because, with higher-dimensional covariates, overlap is harder to satisfy and verify [10].

Our approach to the weak overlap issue is based on the prognostic score (PGS) [14], which is among the important concepts of sufficient scores. While the most well-known score is the propensity score (PPS) [36], PGSs have also long been known to improve methods using PPS [37, 5], and interests last in biostatistics [45, 2]. Prognostic modeling can benefit more from predictive systems and exploit richer literature than propensity modeling, particularly in Medicine and Health. A comparative study in [13] shows PGS-based methods perform better, or as well as, PPS methods. Thus, it is promising to combine the predictive powers of prognostic modeling and machine learning.

To solve the inverse problem of recovering PGSs, our method exploits also the recent advance of identifiable representation, particularly of VAE [26, iVAE]. *Identification* means parameters of interest (for us, representation function and causal effects) are uniquely determined and given by true *observational* distribution. Identification logically precedes estimation and inference. Without identification there is no hope of a consistent estimator, and a model would fail silently; it may fit perfectly but return an estimator that converges to the wrong one or does not converge [29, particularly Sec. 8]. Identification is even more important for causal inference, because, unlike usual (non-causal) model misspecification, causal assumptions are often unverifiable through observables [50]. Thus, it is critical to specify theoretical conditions for identification, and then the applicability of methods can be judged by knowledge of an application domain.

In this work, we study identification (Sec. 3) and estimation (Sec. 4) of TEs under weak overlap. We particularly discuss individualized treatment effects, conditioned on the covariates. Code and proofs are in Supplementary Materials. The main **contributions** of this paper are:
1) theory of TE identification under weak overlap of covariates, using PGS and identifiable model;
2) counterfactual generalization bounds on TE error, which motivates our *conditional* BRL;
3) a new regularized VAE to estimate TEs, with connections to identification and balancing;
4) experimental comparison to state-of-the-art methods on (semi-)synthetic datasets.

2 Setup and motivation

2.1 Counterfactuals, treatment effects, and identification

Following [21], we introduce *potential outcomes* (POs, or counterfactual outcomes) $\mathbf{y}(t) \in \mathbb{R}^d, t \in \{0, 1\}$. $\mathbf{y}(t)$ is the outcome that would have been observed, if treatment value $\mathrm{t} = t$ had been applied. Formally, this is the *consistency of counterfactuals*: $\mathbf{y} = \mathbf{y}(t)$ if $\mathrm{t} = t$, or simply $\mathbf{y} = \mathbf{y}(\mathrm{t})$. We see $\mathbf{y}(t)$ as the hidden variables that give *factual* $\mathbf{y}$ under *factual assignment* $\mathrm{t} = t$. The *fundamental problem of causal inference* is that, for a unit under research, we could observe only one of $\boldsymbol{y}(0)$ or $\boldsymbol{y}(1)$, corresponding the treatment value applied. That is, "factual" refers to $\mathbf{y}$ or t that is in principle *observable* in data, or statistical entities (e.g, estimators) built on them. We also observe relevant covariate(s) $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, which is associated with individuals, with distribution $\mathcal{D} \sim p(\mathbf{x}, \mathbf{y}, \mathrm{t})$. Note, we use Roman fonts for random variables (e.g., t) and italic for realization (e.g., t).

The expected PO is denoted by $\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathbf{y}(t)|\mathbf{x} = \boldsymbol{x})$, conditioned on $\mathbf{x} = \boldsymbol{x}$. The estimands in this work are the Conditional Average TE (CATE) and Average TE (ATE), defined respectively by

$$\tau(\boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}), \quad \nu = \mathbb{E}(\tau(\mathbf{x})). \tag{1}$$

CATE is an *individual-level*, personalized, treatment effect, given highly discriminative covariate.

Standard results [38][16, Ch. 3] give sufficient conditions for identification under general setting. They are *Exchangeability*: $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t}|\mathbf{x}$, and *Overlap*: $p(t|\boldsymbol{x}) > 0$ for any $\boldsymbol{x} \in \mathcal{X}$. Both are required for $t \in \{0, 1\}$. When t appears in a statement without quantification, we always mean "for both t". Often, *Consistency* is also listed, but, as above, it is better known as well-definedness of counterfactuals. Exchangeability means, just as in RCTs but additionally given $\mathbf{x}$, that there is no correlation between factual treatment t and counterfactual outcomes $\mathrm{y}(t)$. Overlap means that the supports of $p(\mathbf{x}|\mathrm{t} = 0)$ and $p(\mathbf{x}|\mathrm{t} = 1)$ should be the same, and this ensures it is valid to condition on any $(\boldsymbol{x}, t)$.

We relax overlapped covariate in Sec. 3.2, to allow some non-overlapped values $\boldsymbol{x}$, i.e., covariate $\mathbf{x}$ is *weakly overlapped*. In Sec. 2.2, we introduce a condition which gives exchangeability (and PGSs). In this paper, we also discuss overlap of variables other than $\mathbf{x}$ (e.g. PGSs), and we give a definition for any random variable $\mathbf{v}$ with support $\mathcal{V}$ as following.

Definition 1. *Overlap* of $\mathbf{v}$ means $p(t|\boldsymbol{v}) > 0$ for all $t \in \{0, 1\}, \boldsymbol{v} \in \mathcal{V}$. We also say $\mathbf{v}$ is *overlapped*. If the condition is violated at some value $\boldsymbol{v}$, then $\boldsymbol{v}$ is *non-overlapped* and $\mathbf{v}$ is *weakly overlapped*.

2.2 Prognostic scores

Our method is motivated by PGSs [14], adapted as Pt-score and P-score in Definition 2 in this paper, related to balancing scores $\boldsymbol{b}(\mathbf{x})$, which is defined by $\mathrm{t} \perp\!\!\!\perp \mathbf{x} | \boldsymbol{b}(\mathbf{x})$ [36]. The PPS $p(\mathrm{t}=1|\mathbf{x})$ is a special case of this. Both are sufficient scores for identification; PGSs are sufficient statistics of *outcome predictors* and $\boldsymbol{b}(\mathbf{x})$ is for the treatment (see Appendix for details).

Definition 2. A *Pt-score* (PtS) is two functions $\mathbb{P}_t(\mathbf{x})$ $(t=0,1)$ such that $\mathbf{y}(t) \perp\!\!\!\perp \mathbf{x} | \mathbb{P}_t(\mathbf{x})$. A PtS is called a *P-score* (PS) if $\mathbb{P}_0 = \mathbb{P}_1$.

Note that, a PtS is by definition two functions, thus overlapped $\mathbb{P}_t(\mathbf{x})$ means *both* $\mathbb{P}_0(\mathbf{x})$ and $\mathbb{P}_1(\mathbf{x})$ are overlapped. **Why PtS (PGS)?** PtS is more applicable than balancing score $\boldsymbol{b}(\mathbf{x})$ under weak overlap. Overlapped $\boldsymbol{b}(\mathbf{x})$ implies overlapped $\mathbf{x}$, which in turn implies overlapped PGS [10]. Lower-dimensional than $\mathbf{x}$, PtS is likely more overlapped than $\mathbf{x}$, and, moreover, there is evidence that PtS *maximizes overlap* among all sufficient scores for ATE [9].

Below is a direct corollary of Proposition 5 in [14]. Both of PtS and PS give CATE, but, as we will see, PS is better as a *conditionally balanced* representation, since $\mathbb{P}_{\mathrm{t}}(\mathbf{x}) \perp\!\!\!\perp \mathrm{t} | \mathbf{x}$ only when $\mathbb{P}_0 = \mathbb{P}_1$.

Proposition 1 (CATE by PtS)**.** *If $\mathbb{P}_t$ is a PtS, then CATE can be given by*

$$\mu_t(\boldsymbol{x}) = \mathbb{E}(\mathbb{E}(\mathbf{y}(t)|\mathbb{P}_t, \boldsymbol{x})) = \mathbb{E}(\mathbb{E}(\mathbf{y}|\mathbb{P}_t(\boldsymbol{x}), \mathrm{t}=t)) = \textstyle\int p(y|\mathbb{P}_t = \mathbb{P}_t(\boldsymbol{x}), t) y dy \tag{2}$$

With the knowledge of $\mathbb{P}_t$, we choose one of $\mathbb{P}_0, \mathbb{P}_1$ corresponding to the counterfactual outcome of interest. This ability of counterfactual assignment resolves the problem of non-overlap at $\boldsymbol{x}$.

PtSs exist under general settings when $\mathbf{y}(t)$ follows an additive noise model (ANM).

- **(G1)** (ANM) the data generating process (DGP) for $\mathbf{y}$ is $\mathbf{y} = \boldsymbol{f}^*(\mathbb{M}(\mathbf{x}), \mathrm{t}) + \mathbf{e}$ where $\boldsymbol{f}^*, \mathbb{M}$ are functions and $\mathbf{e}$ is a zero-mean exogenous (external) noise.

The DGP defines $\mathbf{y}(t)$ by setting $\mathrm{t} = t$ in the equation. And it also specifies how other variables causally affect $\mathbf{y}$. For example, $\mathbf{x}$ affects $\mathbf{y}$ through $\mathbb{M}$, so $\mathbb{M}(\mathbf{x})$ is the effect modifier [14], which is often components of $\mathbf{x}$ affecting $\mathbf{y}$ directly. Note **(G1)** also implies exchangeability given $\mathbf{x}$, through $\mathbf{y}(t) \perp\!\!\!\perp \mathrm{t} | \mathbb{M}(\mathbf{x})$. ANMs are also commonly used in nonparametric regression methods for TEs [6].

Under **(G1)**, 1) $\mathbb{P}_t := \boldsymbol{f}_t^*(\mathbb{M}(\mathbf{x})) = \mu_t(\mathbf{x})$[1] is a PtS[2] but not PS, 2) $\mathbb{M}$ is a PS ($\mathbf{x}$ is a trivial PS), and 3) $\mathbb{P} := (\mu_0(\mathbf{x}), \mu_1(\mathbf{x}))$ is a PS. We use the same symbol to denote a PtS and the random variable defined by it, when appropriate.

3 Identification under generative prognostic model

In Sec. 3.1, we introduce our generative prognostic model and VAE based on $p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \mathrm{t})$ and prove identifiability of our model. In Sec. 3.2, we prove identification of CATEs, one of our main contributions. The theoretical analysis involves only our generative model (i.e., prior and decoder), but not encoder, because model identifiability is a property of *model*, and causal identification is about *DGP* and model. The encoder is involved in *estimation* which is studied in Sec. 4.

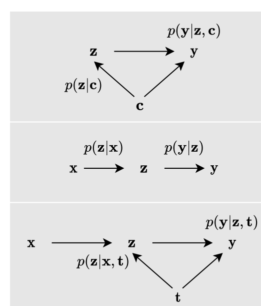

Figure 1: Graphical models of the decoders. From top: CVAE, iVAE, and Intact-VAE. The encoders are similar, taking all observables and build approximate posteriors, and thus are omitted.

3.1 Intact-VAE: model, architecture, and identifiability

Generative models are useful to solve the inverse problem of recovering Pt-score. Our goal is to build a model that can be learned by VAE from observational data to obtain a PtS, or more ideally PS, via the latent variable $\mathbf{z}$. That is, a generative prognostic model.

With the above goal, the generative model of our VAE is built as

$$p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \mathrm{t}) = p(\mathbf{y}|\mathbf{z}, \mathrm{t}) p(\mathbf{z}|\mathbf{x}, \mathrm{t}). \tag{3}$$

The first factor is our decoder which models $p(\mathbf{y}|\mathbb{P}_t, \mathrm{t})$ in (2), and the second factor is our *conditional* prior which models $\mathbb{P}_t(\mathbf{x})$. Conditioning on $\mathbf{x}$ in the joint model

[1] We often write t of function argument in subscript, which indicates possible counterfactual assignment.

[2] μ_t is the most common PGS, to the extent that some call it *the* PGS (e.g, [39, 9, 47]), even without ANMs.

$p(\mathbf{y}, \mathbf{z}|\mathbf{x}, \mathrm{t})$ reflects that our estimand is CATE given $\mathbf{x}$. Modeling the score by a conditional distribution rather than a deterministic function is more flexible. We parameterize our model by ANM outcome and factorized Gaussian prior as

$$p_{\boldsymbol{f}}(\mathbf{y}|\mathbf{z}, \mathrm{t}) = p_{\boldsymbol{\epsilon}}(\mathbf{y} - \boldsymbol{f}_{\mathrm{t}}(\mathbf{z})); p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}, \mathrm{t}) \sim \mathcal{N}(\mathbf{z}; \boldsymbol{h}_{\mathrm{t}}(\mathbf{x}), \mathrm{diag}(\boldsymbol{k}_{\mathrm{t}}(\mathbf{x}))) \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{k})$ are functional parameters and $\boldsymbol{\epsilon}$ is a noise. $\boldsymbol{\lambda}(\mathbf{x}) := \mathrm{diag}(\boldsymbol{k}_{\mathrm{t}}^{-1}(\mathbf{x}))(\boldsymbol{h}(\mathbf{x}), -\frac{1}{2})^T$ is the natural parameter of the Gaussian prior, and we also use it as a shorthand for both $\boldsymbol{h}, \boldsymbol{k}$.

The ELBO of our model can be derived from standard variational lower bound

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{x}, \mathrm{t}) &\geq \log p(\mathbf{y}|\mathbf{x}, \mathrm{t}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \| p(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t})) \\ &= \mathbb{E}_{\boldsymbol{z} \sim q} \log p(\mathbf{y}|\boldsymbol{z}, \mathrm{t}) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{y}, \mathrm{t}) \| p(\mathbf{z}|\mathbf{x}, \mathrm{t})). \end{aligned} \tag{5}$$

Our encoder q, which conditions on all the observables, is standard, and we will see its importance later. We name this architecture *Intact-VAE* (*Id*entifiable *t*re*a*tment-*c*ondi*t*ional VAE).

We naturally have an identifiable conditional VAE (CVAE), as the name suggests. Note that (3) has a similar factorization with the generative model of iVAE [26], that is $p(\mathbf{y}, \mathbf{z}|\mathbf{x}) = p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\mathbf{x})$; the first factor does not depend on $\mathbf{x}$. Further, since we have the conditioning on t in both the factors of (3), our VAE architecture is a combination of iVAE and CVAE [42, 28], with t as the conditioning variable. See Figure 1 for the comparison in terms of graphical models. The core idea of iVAE is reflected in our model identifiability (Lemma 1 below). See Appendix for the basics of VAEs.

The following conditions on the model are used in theoretical analysis.

(M1) i) $\boldsymbol{f}_t$ is injective, ii) $\boldsymbol{f}_t$ is differentiable, and iii) $n := \dim(\mathbf{z}) = d(= \dim(\mathbf{y}))$.

Lemma 1 (Model identifiability)**.** *Given model* (3) *and* (4) *under **(M1)** i) and ii), for* $\mathrm{t} = t$*, assume*

(D1) (Linear independence of $\boldsymbol{\lambda}$) there exist $2n+1$ points $\boldsymbol{x}_0, ..., \boldsymbol{x}_{2n} \in \mathcal{X}$ such that the $2n$-square matrix $\boldsymbol{L}_t := [\gamma_{t,1}, ..., \gamma_{t,2n}]$ is invertible, where $\gamma_{t,k} := \boldsymbol{\lambda}_t(\boldsymbol{x}_k) - \boldsymbol{\lambda}_t(\boldsymbol{x}_0)$.

Then, given $\mathrm{t} = t$*, the family is* identifiable *up to an equivalence class. That is, if* $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t} = t) = p_{\boldsymbol{\theta}'}(\mathbf{y}|\mathbf{x}, \mathrm{t} = t)$*, we have the relation between parameters: for any* $\boldsymbol{y}_t$ *in the image of* $\boldsymbol{f}_t$*,*

$$\boldsymbol{f}_t^{-1}(\boldsymbol{y}_t) = \mathrm{diag}(\boldsymbol{a})\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t) + \boldsymbol{b} =: \mathcal{A}(\boldsymbol{f}_t'^{-1}(\boldsymbol{y}_t)) \tag{6}$$

where $\mathrm{diag}(\boldsymbol{a})$ *is an invertible* n*-diagonal matrix and* $\mathbf{b}$ *is a* n*-vector, both depend on* $\boldsymbol{\lambda}_t$*.*

The conditions are inherited from iVAE. **(D1)** holds easily in practice, if the components of $\boldsymbol{\lambda}_t(\mathbf{x})$ are *linearly independent*; if **(D1)** fails, then the support of $\boldsymbol{\lambda}_t(\mathbf{x})$ is in a $(2n-1)$-dimensional space.

The essence of the result is $\boldsymbol{f}_t' = \boldsymbol{f}_t \circ \mathcal{A}_t$, that is, $\boldsymbol{f}_t$ can be identified (learned) up to an affine transformation defined by $\boldsymbol{\lambda}_t$. This is achieved by combining the techniques from [26] and [43], and essentially the same results can be proved for other exponential family priors [43]. In this paper, symbol $'$ (prime) always indicates another parameter (variable, etc.).

3.2 Nonparametric identifications under weakly-overlapped covariate

In this subsection, we give two identification results based on (partial) recovery of PS or PtS, respectively. Since PtSs are functions of $\mathbf{x}$, the recovery is achieved by a noiseless prior, that is, $\boldsymbol{k}(\mathbf{x}) = \mathbf{0}$; the prior $\mathbf{z}_{\boldsymbol{\lambda},t} \sim p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}, \mathrm{t} = t)$ degenerates to deterministic function $\boldsymbol{h}_t(\mathbf{x})$.

PtSs with dimensionality lower than or equal to $\mathbf{y}$ are essential to work for weak overlap of $\mathbf{x}$, to the extent that, from now on, we simply say PSs / PtSs when referring to this kind of *low-dimensional PSs / PtSs*, unless particularly indicated. **(M1)** iii), i.e. $n = d$, is not restrictive because μ_t is a PtS of the same dimension as $\mathbf{y}$ under **(G1)**. Also, in practice, $n = d$ means that we seek a low-dimensional representation of $\mathbf{x}$. In fact, to make the dimensionality explicit in **(G1)**, we introduce an alternative **(G1')** which includes **(G1)** with $\mathbb{P}_t = \mu_t$ and $\boldsymbol{j}_t$ is identity.

(G1') (Low-dimensional PtS) Under **(G1)**, $\mu_t(\mathbf{x}) = \boldsymbol{j}_t(\mathbb{P}_t(\mathbf{x}))$ for some $\mathbb{P}_t$ and injective $\boldsymbol{j}_t$.

We use **(G1')** afterwards. Clearly, $\mathbb{P}_t$ in **(G1')** is a PtS, and injectivity and $n = d$ ensure $n = \dim(\mathbf{y}) \geq \dim(\mathbb{P}_t)$. Similarly, the next **(G2)** reduces unverifiable $n \geq \dim(\mathbb{P})$ to $n = d$, for PS.

(G2) (Low-dimensional PS) Under **(G1)**, $\mu_t(\mathbf{x}) = \boldsymbol{j}_t(\mathbb{P}(\mathbf{x}))$ for some $\mathbb{P}$ and injective $\boldsymbol{j}_t$.

(G2) means that CATEs are given by μ_0 and an invertible function $\boldsymbol{i} := \boldsymbol{j}_1 \circ \boldsymbol{j}_0^{-1}$. See Appendix for more discussion and a (closely related) real world example. In Sec. 4.1, we argue that *there often exist equivalent PSs under **(G1')***, at least approximately.

With **(G1')** or **(G2)**, overlapped $\mathbf{x}$ can be relaxed to overlapped P(t)S plus the following.

(M2) (Score partition preserving) For any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\mathbb{P}_t(\boldsymbol{x}) = \mathbb{P}_t(\boldsymbol{x}') \implies \boldsymbol{h}_t(\boldsymbol{x}) = \boldsymbol{h}_t(\boldsymbol{x}')$.

Note that **(M2)** is in fact required for optimal $\boldsymbol{h}$, in the sense specified in Proposition 1 and Theorem 1 below. The intuition is that, $\mathbb{P}_t$ maps non-overlapped $\boldsymbol{x}$ to an overlapped value, and $\boldsymbol{h}_t$ preserves this property, through learning. In fact, **(M2)** is trivially satisfied if $\mathbb{P}_t$ and $\boldsymbol{h}_t$ are *linear*, and this is still challenging and considered by many works [32, 9], or some with linear outcome models [11, 39].

Our first identification, Proposition 2, relies on **(G2)** and our generative model, *without* model identifiability (so differentiable $\boldsymbol{f}_t$ is not needed). This is a nonparametric[3] identification under shape restriction [7], because $\boldsymbol{f}, \boldsymbol{h}$ are functional parameters, and injectivity is monotonicity if $\boldsymbol{j}_t$ is on $\mathbb{R}$.

Proposition 2 (Identification with PS)**.** *Given **(G2)** and model* (3) *and* (4) *under **(M1)** i) and iii), and* **(M3)** (PS matching) $\boldsymbol{h}_0(\mathbf{x}) = \boldsymbol{h}_1(\mathbf{x})$ and $\boldsymbol{k}(\mathbf{x}) = \mathbf{0}$. *Then, if* $\mathbb{E}_{p_{\boldsymbol{\theta}}}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = \mathbb{E}(\mathbf{y}|\mathbf{x}, \mathrm{t})$*, we have*[4]

1) *(Recovery of PS)* $\boldsymbol{z}_{\boldsymbol{\lambda},t} = \boldsymbol{h}_t(\boldsymbol{x}) = \boldsymbol{v}(\mathbb{P}(\boldsymbol{x}))$ *on overlapped* $\boldsymbol{x}$*,*
   where $\boldsymbol{v} : \mathcal{P} \to \mathbb{R}^n$ *is an injective function and* $\mathcal{P} := \{\mathbb{P}(\boldsymbol{x})|\textit{overlapped } \boldsymbol{x}\}$
2) *(Identification) if* $\mathbb{P}$ *in **(G2)** is overlapped, and **(M2)** is satisfied, then* $\mu_t(\boldsymbol{x}) = \hat{\mu}_t(\boldsymbol{x})$
   for any $t \in \{0, 1\}, \boldsymbol{x} \in \mathcal{X}$*, where* $\hat{\mu}_t(\boldsymbol{x}) := \mathbb{E}_{p_{\boldsymbol{\lambda}}(\mathbf{z}|\boldsymbol{x},t)}\mathbb{E}_{p_{\boldsymbol{f}}}(\mathbf{y}|\mathbf{z}, t) = \boldsymbol{f}_t(\boldsymbol{h}_t(\boldsymbol{x}))$.

The essence is, i) the true DGP is identified up to an invertible mapping $\boldsymbol{v}$, so that $\boldsymbol{f}_t = \boldsymbol{j}_t \circ \boldsymbol{v}^{-1}$ and $\boldsymbol{h}_t = \boldsymbol{v} \circ \mathbb{P}_t$, and ii) $\mathbb{P}_t$ is recovered up to $\boldsymbol{v}$ and $\mathbf{y}(t) \perp\!\!\!\perp \mathbf{x}|\mathbb{P}_t$ is preserved, with *same* $\boldsymbol{v}$ for both t.

PS is preferred since it satisfies overlap more easily and **(M2)** than PtS which refers to two functions. However, the existence of low-dimensional PS is uncertain in practice when our knowledge of the DGP is limited. Thus, we need Theorem 1 to work under PtS which generally exists.

Theorem 1 (Identification with PtS)**.** *Given the DGP **(G1')** and model* (3)&(4) *under **(M1)** and* **(M3')** (Noise matching) $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ and $\boldsymbol{k}(\mathbf{x}) = k\boldsymbol{k}'(\mathbf{x}), k \to 0$*, assume **(D1)** and*

(D2) (Spontaneous balance) There exist $2n+1$ points $\boldsymbol{x}_0, ..., \boldsymbol{x}_{2n} \in \mathcal{X}$, $2n$-square matrix $\boldsymbol{C}$, and $2n$-vector $\boldsymbol{d}$, such that $\boldsymbol{L}_0^{-1}\boldsymbol{L}_1 = \boldsymbol{C}$ and $\boldsymbol{\beta}_0 - \boldsymbol{C}^{-T}\boldsymbol{\beta}_1 = \boldsymbol{d}/k$ for optimal $\boldsymbol{\lambda}_t$ (see below), where $\boldsymbol{L}_t$ is defined in **(D1)**, $\boldsymbol{\beta}_t := (\alpha_t(\boldsymbol{x}_1) - \alpha_t(\boldsymbol{x}_0), ..., \alpha_t(\boldsymbol{x}_{2n}) - \alpha_t(\boldsymbol{x}_0))^T$, and $\alpha_t(\mathbf{x}; \boldsymbol{\lambda}_t)$ is the log-partition function of the prior in (4).

Then, if $p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}, \mathrm{t}) = p(\mathbf{y}|\mathbf{x}, \mathrm{t})$*, conclusions 1) and 2) in Proposition 2 hold with* $\mathbb{P}$ *replaced with* $\mathbb{P}_t$ *in **(G1')**, and the domain of* $\boldsymbol{v}$ *becomes* $\mathcal{P} := \mathcal{P}_0 \cup \mathcal{P}_1, \mathcal{P}_t := \{\mathbb{P}_t(\boldsymbol{x})|\textit{overlapped } \boldsymbol{x}\}$.

Theorem 1 also achieves the two essential points, but in different and complementary ways. Proposition 2 starts from the prior by $\mathbb{P}_0 = \mathbb{P}_1$ and setting $\boldsymbol{h}_0 = \boldsymbol{h}_1$. Conversely, Theorem 1 starts from the decoder with $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ and strengthens model identifiability (6) by **(D2)**. **(D2)** restricts the discrepancy between $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1$ on $2n+1$ points of $\mathbf{x}$, thus is relatively easy to satisfy with high-dimensional $\mathbf{x}$.

We see more reasons to prefer PS. In general, to identify the mean function $\mu_t(\mathbf{x})$, a regression is enough, and $p_{\boldsymbol{\epsilon}} = p_{\mathbf{e}}$ is unnecessary as in Proposition 2. Also, **(D2)** is trivial if we have PS and set $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1$. See Appendix for more on the complementarity between the two identifications.

4 Estimation by β-Intact-VAE

4.1 Prior as PS, posterior as PtS, and β as regularization strength

In this subsection, we discuss our focus on balanced PtSs and give an estimation method to realize it.

In learning the Intact VAE with data, we assume that there is a PtS and the decomposition of **(G1')** holds. Such a decomposition is not unique in general, however. Among possible PtSs, we wish to learn a balanced PtS, which is close to PS. This is based on the observations in Sec. 3.2: we saw that existence of PS is preferable in identifying the true DGP up to equivalent expression. Here, we introduce the notion of balanced PtS in a non-rigorous way: a PtS $\mathbb{P}_t$ is called *balanced* if the value of some measure for the conditional independence $\mathbb{P}_{\mathrm{t}}(\mathbf{x}) \perp\!\!\!\perp \mathrm{t}|\mathbf{x}$ is small. The idea is common in practice. For example, in a real-world nutrition study, [20] reduces 11 covariates to a 1-dimensional linear PS.

[3] Some references (e.g, [16]) only refer to identification without models as "nonparametric". However, with recent advances, we have identification under nonparametric models [29], also the case in this paper.

[4] Same results hold *without* "under **(G1)**" in **(G2)** (or **(G1')** for Theorem 1), except that $\mathbf{z}_{\boldsymbol{\lambda},t}$ is *not* necessarily a PS (or PtS for Theorem 1). This is because **(G1)** is required to ensure μ_t is a PtS.

Assuming that there is a balanced PtS, we consider two ways for estimating it with Intact VAE. One is to exploit a prior that does not depend on t. Namely, we set $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}_1 =: \boldsymbol{\lambda}$ in (4). The other is to introduce a hyperparameter β in the ELBO as in β-VAE [17]. More specifically, with the prior $p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})$, we use factorized Gaussian for the decoder and encoder:

$$p_{\boldsymbol{f},\boldsymbol{g}}(\mathbf{y}|\mathbf{z},\mathrm{t}) \sim \mathcal{N}(\mathbf{y}; \boldsymbol{f}_{\mathrm{t}}(\mathbf{z}), \mathrm{diag}(\boldsymbol{g}_{\mathrm{t}}(\mathbf{z}))); \quad q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y},\mathrm{t}) \sim \mathcal{N}(\mathbf{z}; \boldsymbol{r}_{\mathrm{t}}(\mathbf{x},\mathbf{y}), \mathrm{diag}(\boldsymbol{s}_{\mathrm{t}}(\mathbf{x},\mathbf{y}))). \tag{7}$$

The modified ELBO with β, up to additive constant, is derived as

$$\mathbb{E}_{\mathcal{D}}\{-\beta D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y},\mathrm{t}) \| p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})) - \mathbb{E}_{\boldsymbol{z}\sim q}[(\mathbf{y} - \boldsymbol{f}_{\mathrm{t}}(\boldsymbol{z}))^2/2\boldsymbol{g}_{\mathrm{t}}^2(\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim q} \log |\boldsymbol{g}_{\mathrm{t}}(\boldsymbol{z})|\}. \tag{8}$$

Here, for convenience, we omit the summation (also in $\mathcal{L}_{\boldsymbol{f}}$ in Sec. 4.2), as if $\mathbf{y}$ was univariate. The approximate posterior (or encoder) $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y},\mathrm{t})$ depends on t, which can realize a PtS. With β, we control the trade-off between the first and second term: the former is the divergence of the posterior from the balanced prior, and the latter is the reconstruction of the outcome. By choosing β in an appropriate way, such as by validation, the ELBO can find a solution that explains the outcome while keeping the balancedness of the posterior. Also note that, the parameters $\boldsymbol{g}$ and $\boldsymbol{k}$, which models the outcome noise and expresses uncertainty of the prior, respectively, are both learned by the ELBO. This deviates from the theoretical conditions in Sec. 3.2, but is more practical and gives better results in experiments. See Appendix for much more on ideas and connections behind the ELBO.

Once the encoder $q_{\boldsymbol{\phi}}$ is learned[5], the estimate of the expected POs is given by

$$\hat{\mu}_{\hat{t}}(\boldsymbol{x}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x}=\boldsymbol{x})}\boldsymbol{f}_{\hat{t}}(\mathbf{z}) = \mathbb{E}_{\mathcal{D}|\boldsymbol{x}\sim p(\mathbf{y},\mathrm{t}|\boldsymbol{x})}\mathbb{E}_{\mathbf{z}}\boldsymbol{f}_{\hat{t}}(\mathbf{z})q_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x},\mathbf{y},\mathrm{t}), \hat{t} \in \{0,1\}, \tag{9}$$

where $q(\mathbf{z}|\mathbf{x}) := \mathbb{E}_{p(\mathbf{y},\mathrm{t}|\boldsymbol{x})}q_{\boldsymbol{\phi}}(\mathbf{z}|\boldsymbol{x},\mathbf{y},\mathrm{t})$ is the aggregated posterior and $\mathcal{D} \sim p(\mathbf{x},\mathbf{y},\mathrm{t})$. In estimation, we consider the case where $\boldsymbol{x}$ is observed in the data, and the sample of $(\mathbf{y},\mathrm{t})$ are taken from the data given $\mathbf{x} = \boldsymbol{x}$ (when $\boldsymbol{x}$ is not in the data, we replace $q_{\boldsymbol{\phi}}$ with $p_{\boldsymbol{\lambda}}$ in (9), see Appendix for details and results). Note that $\hat{t}$ in (9) indicates counterfactual assignment which may not be the same as factual $\mathrm{t} = t$ in the data. That is, we set $\mathrm{t} = \hat{t}$ in the decoder. The assignment is not applied to the encoder, but it is learned from factual $\mathbf{x}, \mathbf{y}$ (see also Sec. 4.2, the explanation for $\epsilon_{CF,t}$). The overall **algorithm** steps are i) we train VAE by (8), ii) infer CATE $\hat{\tau}(\boldsymbol{x}) = \hat{\mu}_1(\boldsymbol{x}) - \hat{\mu}_0(\boldsymbol{x})$ by (9).

4.2 Conditional balanced representation learning

We formally justify our ELBO (8) from the viewpoint of BRL. Usually, particularly in ATE estimation, balance means covariate balance, i.e., $\mathbf{x} \perp\!\!\!\perp \mathrm{t}$ [44]. Influenced by this, most BRL methods learn balanced covariate representation $\mathbf{z}$ such that $\mathbf{z} \perp\!\!\!\perp \mathrm{t}$ [40, 31] and usually $\mathbf{z}$ is a function of $\mathbf{x}$. From Sec. 4.1, we understand that larger β in ELBO (8) encourages $\mathbf{z} \perp\!\!\!\perp \mathrm{t}|\mathbf{x}$ which is given by the prior, corresponding to $\mathbb{P}_{\mathrm{t}}(\mathbf{x}) \perp\!\!\!\perp \mathrm{t}|\mathbf{x}$ for a balanced PtS. Here, we show that, this *conditional balance* of representation $\mathbf{z}$ is natural for CATE estimation, and CATE error due to bad recovery of $\boldsymbol{j}_t$ in **(G1')** is controlled by ELBO (8). In Appendix, we detail novel implications of our bounds, compared to those in [40, 31].

Using (9) to estimate CATE, $\hat{\tau}_{\boldsymbol{f}}(\boldsymbol{z}) = \boldsymbol{f}_1(\boldsymbol{z}) - \boldsymbol{f}_0(\boldsymbol{z})$ is marginalized on $q(\mathbf{z}|\mathbf{x})$. The bounds below motivate both prior and posterior balancing. Let us first consider errors defined by the aggregated prior $p(\mathbf{z}|\mathbf{x}) := \sum_t p(t|\mathbf{x})p_t(\mathbf{z}|\mathbf{x})$ (denote $p_t(\mathbf{z}|\mathbf{x}) := p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}, \mathrm{t} = t)$), and $q(\mathbf{z}|\mathbf{x})$ afterwards. We introduce the objective we bound. The *true CATE*, given covariate $\mathbf{x} = \boldsymbol{x}$ or score $\boldsymbol{z}$, is

$$\tau(\boldsymbol{x}) = m_1(\mathbb{P}_1(\boldsymbol{x})) - m_0(\mathbb{P}_0(\boldsymbol{x})); \quad \tau_m(\boldsymbol{z}) = m_1(\boldsymbol{z}) - m_0(\boldsymbol{z}) \tag{10}$$

where $m_t(\boldsymbol{z}) := \mathbb{E}(\mathbf{y}(t)|\mathbb{P}_t = \boldsymbol{z})$ and $\mathbb{P}_t$ is a balanced PtS in **(G1')**. Accordingly, given $\boldsymbol{x}$, the *error of prior CATE*, with or without knowing $\mathbb{P}_t$, is naturally defined as

$$\epsilon_{\boldsymbol{f}}^{*,p}(\boldsymbol{x}) := \mathbb{E}_{p(\mathbf{z}|\boldsymbol{x})}(\hat{\tau}_{\boldsymbol{f}}(\mathbf{z}) - \tau(\boldsymbol{x}))^2; \quad \epsilon_{\boldsymbol{f}}^{p}(\boldsymbol{x}) := \mathbb{E}_{p(\mathbf{z}|\boldsymbol{x})}(\hat{\tau}_{\boldsymbol{f}}(\mathbf{z}) - \tau_m(\mathbf{z}))^2. \tag{11}$$

We bound $\epsilon_{\boldsymbol{f}}^{p}$ instead of $\epsilon_{\boldsymbol{f}}^{*,p}$ because the error between $\tau(\mathbf{x})$ and $\tau_m(\mathbf{z})$ is small if balanced $\mathbb{P}_t$ is recovered (then $\boldsymbol{z} \approx \mathbb{P}_0(\boldsymbol{x}) \approx \mathbb{P}_1(\boldsymbol{x})$ in (10), see Appendix for details). Instead, we consider the error between $\hat{\tau}_{\boldsymbol{f}}$ and τ_m below. We define the risks of outcome regression, into which ϵ^p is decomposed.

Definition 3 (PO Risks)**.** The *expected loss of PO* at $(\boldsymbol{z}, t)$, *factual risk*, and *counterfactual risk* are

$$\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t) := \boldsymbol{g}_t^{-2}\mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t=\boldsymbol{z})}(\mathbf{y}(t) - \boldsymbol{f}_t(\boldsymbol{z}))^2 = \boldsymbol{g}_t(\boldsymbol{z})^{-2}\int(\boldsymbol{y} - \boldsymbol{f}_t(\boldsymbol{z}))^2 p(\mathbf{y}(t) = \boldsymbol{y}|\mathbb{P}_t = \boldsymbol{z})d\boldsymbol{y};$$
$$\epsilon_{F,t}^{p}(\boldsymbol{x}) := \mathbb{E}_{p_t(\mathbf{z}|\boldsymbol{x})}\mathcal{L}_{\boldsymbol{f}}(\mathbf{z},t); \quad \epsilon_{CF,t}^{p}(\boldsymbol{x}) := \mathbb{E}_{p_{1-t}(\mathbf{z}|\boldsymbol{x})}\mathcal{L}_{\boldsymbol{f}}(\mathbf{z},t) = \int \mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z},t)p_{1-t}(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}.$$

[5] As usual, we expect variational inference and optimization procedure are (near) optimal, i.e., Consistency of VAE (see Appendix for formal statement). Consistent estimation using the prior is a direct corollary of consistent VAE. Under Gaussian models, it is possible to prove consistency of posterior estimation, as shown in [4].

With $\mathbf{y}(t)$ involved, $\mathcal{L}_{\boldsymbol{f}}$ is a PO error of $\boldsymbol{f}$ weighted by $\boldsymbol{g}$. Factual and counterfactual counterparts are defined accordingly, w.r.t factual p_t learned from data. Note, in $\epsilon_{F,t}$, unit $\boldsymbol{u} = (\boldsymbol{x}, \boldsymbol{y}, t)$ involves in the learning of $p_t(\mathbf{z}|\boldsymbol{x})$ (and $q_t(\mathbf{z}|\boldsymbol{x})$ afterwards), and also in $\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, t)$ since $\mathbf{y}(t) = \boldsymbol{y}$ for the unit. In $\epsilon_{CF,t}$, however, $\mathbf{y}(t) \neq \boldsymbol{y}' = \mathbf{y}(1-t)$ for $\boldsymbol{u}' = (\boldsymbol{x}, \boldsymbol{y}', 1-t)$ (particularly relevant for the posterior).

Thus, *the regression error (second) term in ELBO* (8) *controls* $\epsilon^p_{F,t}$ *via factual data*. On the other hand, $\epsilon^p_{CF,t}$ is not estimable due to unobservable $\mathbf{y}(1-\mathrm{t})$, but is bounded as below.

Lemma 2 (Counterfactual risk bound)**.** *Assume* $|\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, t)| \leq M$*, we have*

$$\epsilon^p_{CF}(\boldsymbol{x}) \leq \textstyle\sum_t p(1-t|\boldsymbol{x})\epsilon^p_{F,t}(\boldsymbol{x}) + M\mathbb{D}^p(\boldsymbol{x}) \tag{12}$$

where $\epsilon^p_{CF}(\boldsymbol{x}) := \sum_t p(1-t|\boldsymbol{x})\epsilon^p_{CF,t}(\boldsymbol{x})$*, and* $\mathbb{D}^p(\boldsymbol{x}) := \sum_t \sqrt{D_{\mathrm{KL}}(p_t \| p_{1-t})/2}$*.*

$\epsilon^p_{CF}(\boldsymbol{x})$ is bounded by $\epsilon^p_{F,t}$ plus $M\mathbb{D}^p(\boldsymbol{x})$, which measures the imbalance between $p_t(\mathbf{z}|\boldsymbol{x})$ and is *symmetric* for t. We can implicitly control ϵ^p_{CF} by making $\mathbb{D}^p$ small. Again, this means PS is preferred as a conditional balanced representation, and justifies our balanced prior $p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})$. Moreover, *the results (including Theorem 2 below) also hold for posterior estimation*, that is, replace $p_t(\mathbf{z}|\mathbf{x})$ with $q_t(\mathbf{z}|\mathbf{x}) := q(\mathbf{z}|\mathbf{x}, \mathrm{t} = t) = \mathbb{E}_{p(\mathbf{y}|\mathbf{x},t)} q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{y}, t)$, the results and proofs hold as it was. This implies that the imbalance between q_t should also be controlled. Correspondingly, *the symmetric KL term in ELBO* (8) *balances* $q_t(\mathbf{z}|\mathbf{x})$ by encouraging $\mathbf{z} \perp\!\!\!\perp \mathrm{t}|\mathbf{x}$ for the posterior.

Theorem 2 in turn bounds $\epsilon^p_{\boldsymbol{f}}$, by decomposing it to $\epsilon^p_{F,t}$, $\epsilon^p_{CF,t}$, and $\mathbb{V}^p_{\mathbf{y}}$.

Theorem 2 (Generalization bound)**.** *Assume* $|\mathcal{L}_{\boldsymbol{f}}(\boldsymbol{z}, t)| \leq M$ *and* $|\boldsymbol{g}_t(\boldsymbol{z})| \leq G$*, then,*

$$\epsilon_{\boldsymbol{f}}(\boldsymbol{x}) \leq 2(G^2(\epsilon_{F,0}(\boldsymbol{x}) + \epsilon_{F,1}(\boldsymbol{x})) + M\mathbb{D}(\boldsymbol{x}) - \mathbb{V}_{\mathbf{y}}(\boldsymbol{x})|p \tag{13}$$

where $\mathbb{V}^p_{\mathbf{y}}(\boldsymbol{x}) := \mathbb{E}_{p(\mathbf{z}|\boldsymbol{x})} \sum_t \mathbb{E}_{p(\mathbf{y}(t)|\mathbb{P}_t = \mathbf{z})}(\mathbf{y}(t) - m_t(\mathbf{z}))^2$*, and "*$|^p$*" collects all superscripts* q*.*

The new term $\mathbb{V}^p_{\mathbf{y}}(\boldsymbol{x})$ reflects the intrinsic variance in the DGP and can not be controlled, and it is negative because $\epsilon^p_{\boldsymbol{f}}$ is defined by mean functions $\boldsymbol{f}_t$ and m_t, not $\mathbf{y}(t)$. The other two terms, as we indicated, is estimated by our ELBO.

Estimating G, M is nontrivial. Instead, similarly to [40], we rely on β in the ELBO to weight the two terms in (13). We do not need two hyperparameters since G *is implicitly controlled by the third term in ELBO* (8), which is a norm constraint. As in matching methods, β is a trade-off between conditional balance of learned PtS (affected by $\boldsymbol{f}_t$) and precision / effective sample size of outcome regression, and can be seen as the probabilistic counterpart of [47, 25].

Finally, we note the bounds do not directly address non-overlap; in Lemma 2, when $p(1-t|\boldsymbol{x}) = 0$, $\epsilon^p_{F,1-t}$ in the r.h.s is unbounded since $p_{1-t}(\mathbf{z}|\boldsymbol{x})$ can not be learned from data. However, as we argued in Sec. 3.2, with more balanced $\mathbb{P}_t$ recovered as representation, overlap is more easily satisfied.

5 Related work

Weak overlap. Under (respective versions of) weak overlap, [32] estimates ATE by reducing covariates to a linear PGS, [11] estimates homogeneous (constant) TE under partial linear outcome model, and [9] studies identification of ATE by a general class of scores, given (linear) PPS and PGS. In machine learning, current focus is on finding overlap regions [33, 8], or indicating possible failure under weak overlap [22], but not remedies. An exception is [24] which provides bounds without overlap. [40, 31] are in line of [24] and have similar bounds to ours, without relating to overlap. Our method is the *first* in machine learning that gives identification without overlap.

Prognostic scores are recently combined with machine learning, mainly in biostatistics. For example, [39] trains a flexible PGS and fits a linear regression on the PGS among others, for constant TEs, and [12] models PGS in its Bayesian regression tree for CATE. More related, [20] estimates CATE by reducing covariates to a linear score that is a joint PPS and PGS, and [47] uses SVM to minimize the worst-case bias due to PGS imbalance. However, in machine learning, few methods consider PGSs; [55, 15] learn outcome predictors, without connection to PGS, while [24] conceptually, but not formally, connects BRL to PGS. Our work follows the recent boom in biostatistics and is the *first* to formally connect generative learning, PGS, and BRL (see below on BRL) for TE estimation.

Identifiable representation. Recently, independent component analysis (ICA) and representation learning, both ill-posed inverse problems, meet together to give nonlinear ICA and identifiable representation, e.g., using VAEs [26], and energy models [27]. The results are exploited in causal discovery [51] and out-of-distribution generalization [46]. This work is the *first* to explore identifiable representations in TE identification.

BRL and related methods amount to a major direction. Early BRL methods are BLR/BNN [23] and TARnet/CFR [40]. Adding to this, [53] also exploits the local similarity of between data points. [41] uses similar architecture to TARnet, considering the importance of treatment probability. There are also methods using GAN [54, GANITE] and Gaussian process [1]. Our method shares the idea of BRL, and further extends to conditional balancing more suitable for CATE.

Our work hopefully lays conceptual and theoretical foundations of **VAE methods for TEs** (e.g., [30, 31]), under unconfoundedness. Also, **monotonicity**, which is injectivity on $\mathbb{R}$, is important in causal inference, and some works consider it together with overlap [24, 56]. See Appendix for details.

6 Experiments

We compare the proposed method with existing methods on three types of datasets. Here we present two experiments, and the rest one, on the Pokec social network dataset, can be found in Appendix. As in previous works [40, 30], we report the absolute error of ATE $\epsilon_{ATE} \coloneqq |\mathbb{E}_{\mathcal{D}}(y(1)-y(0))-\mathbb{E}_{\mathcal{D}}\hat{\tau}(\boldsymbol{x})|$, and, as a surrogate of CATE, the empirical PEHE [18] $\epsilon_{PEHE} \coloneqq \mathbb{E}_{\mathcal{D}}((y(1)-y(0))-\hat{\tau}(\boldsymbol{x}))^2$.

Unless otherwise indicated, for each function $\boldsymbol{f}, \boldsymbol{h}, \boldsymbol{k}, \boldsymbol{r}, \boldsymbol{s}$ in (4)(7), we use a multilayer perceptron (MLP) that has 3*200 hidden units with ReLU activation, and $\boldsymbol{\lambda} = (\boldsymbol{h}, \boldsymbol{k})$ depends only on $\mathbf{x}$. We fix $\boldsymbol{g}(\mathbf{x}) = 1$ because the datasets have fixed noise scale, and results with learned $\boldsymbol{g}$ on synthetic dataset with dependent noise is in Appendix. The Adam optimizer with initial learning rate 10^{-4} and batch size 100 is employed. All experiments use early-stopping of training by evaluating the ELBO on a validation set, and results are reported on a testing set. Each set of running on synthetic dataset (a line in the figure) is within 1 hour on an 8-CPU machine, and it is within a day for IHDP. More details on hyper-parameters and settings are given in each experiment and Appendix.

6.1 Synthetic dataset

We generate synthetic datasets following (14). Both $\mathbf{x}, \mathbf{w}$ are factorized Gaussians. $\boldsymbol{\mu}, \boldsymbol{\sigma}$ are randomly sampled. The functions $\boldsymbol{h}, \boldsymbol{k}, l$ are linear. Outcome models f_0, f_1 are built by NNs with invertible activations. y is univariate, $\dim(\mathbf{x}) = 30$, and $\dim(\mathbf{w})$ ranges from 1 to 5. $\mathbf{w}$ is a PS, but is *not low-dimensional* when $\dim(\mathbf{w}) > 1$. We *control overlap* by ω which multiplies the logit value, and have 5 different overlap levels from strong overlap to very weak overlap. See Appendix for details.

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}); \mathbf{w}|\mathbf{x} \sim \mathcal{N}(\boldsymbol{h}(\mathbf{x}), \boldsymbol{k}(\mathbf{x})); \mathrm{t}|\mathbf{x} \sim \mathrm{Bern}(\mathrm{Logi}(\omega l(\mathbf{x}))); \mathrm{y}|\mathbf{w}, \mathrm{t} \sim \mathcal{N}(f_{\mathrm{t}}(\mathbf{w}), 1). \quad (14)$$

With the same $(\dim(\mathbf{w}), \omega)$, we evaluate our method and CFR on 10 random DGPs, with different sets of functions $f, \boldsymbol{h}, \boldsymbol{k}, l$ in (14). For each DGP, we sample 1500 data points, and split them into 3 equal sets for training, validation, and testing. We show our results for different hyperparameter β. For CFR, we try different balancing parameters and present the best results (see Appendix for details). We report ϵ_{PEHE}, see Appendix for ATE results.

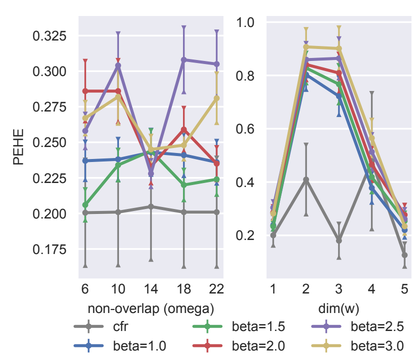

Figure 2: $\sqrt{\epsilon_{PEHE}}$ on synthetic dataset. Error bar on 10 random DGPs.

In each panel of Figure 2, we adjust one of $\omega, \dim(\mathbf{w})$ respectively, with the other fixed to the lowest. In the left panel, both our method and CFR are quite robust to overlap level, supporting respective theories ([24] gives bounds for CFR under weak overlap). Too large β seems to worsen the results, possibly because we already have an apparent PS ($\mathbf{w}$ with $\dim(\mathbf{w}) = 1$) and large β incurs sub-optimal ELBO with no gain.

In the right panel, when $\dim(\mathbf{w}) > 1$, f_t in (14) is non-injecitve and learning of PtS is necessary. Thus, larger β has a negative effect, and particularly, $\beta = 1$ is significantly better than $\beta = 3$. The drop of error for $\dim(\mathbf{w}) > 3$ is due to the randomness of f in (14). In Sec. 2.2, we saw that the 2-dimsensional PS $\mathbb{P} \coloneqq (\mu_0(\mathbf{x}), \mu_1(\mathbf{x}))$ always exists under ANMs. Thus, when $\dim(\mathbf{w}) > 2$, our method tries to recover that $\mathbb{P}$, and generally performs not worse than under $\dim(\mathbf{w}) = 2$, but still not better than under $\dim(\mathbf{w}) = 1$.

Our method is much robuster against different DGPs than CFR (see error bars), though it is worse than CFR when $\dim(\mathbf{w}) > 1$. This is unsurprising because our model has *1-dimensional* $\mathbf{z}$, while CFR uses 200-dimensional representation. Thus, the results already show the power of identification and recovery of scores (see Figure 3 also). In fact, we observed that our method outperforms or matches CFR with higher-dimensional $\mathbf{z}$ (see Appendix). Thus, we believe the performance gap with $\dim(\mathbf{z}) = 1$ is due to the capacity of NNs in Intact-VAE.

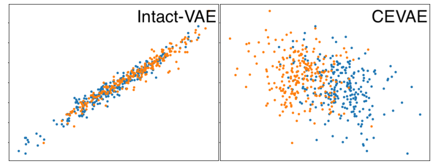

Figure 3: Plots of recovered - true latent. Blue: $t = 0$, Orange: $t = 1$.

When $\dim(\mathbf{w}) = 1$, there are no better PSs than $\mathbf{w}$, because f_t is invertible and no information can be dropped from $\mathbf{w}$. Thus, as shown in Figure 3, our method learns $\mathbf{z}$ as an approximate affine transformation of the true $\mathbf{w}$, showing identification. For comparison, we run [30, CEVAE] which is also based on VAE but without identification, and it shows much lower quality of recovery. As expected, both recovery and estimation are better with balanced prior $p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})$, and we can see an example of bad recovery using $p_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}, \mathrm{t})$ in Appendix. More latent plots can also be found there.

6.2 IHDP benchmark dataset

This experiment shows our conditional balancing matches state-of-the-art BRL methods, *and does not overfit to PEHE*. The IHDP dataset [18] is widely used to evaluate machine learning based methods, e.g. [40, 41]. It is also used in [24] which considers weak overlap, because *the covariates are weakly overlapped* due to their correlation to the artificial treatment assignment. Finally, there is a linear PS (linear combination of the covariates). See Appendix for details.

Note, most of covariates are binary, so the support of the PS is often on small and separated intervals and is possibly discrete. Thus, Gaussian latent $\mathbf{z}$ is misspecified. We use multivariate $\mathbf{z}$ in model to address this, similarly to [30]. We set $\beta = 1$ since it works well on synthetic dataset with weak overlap. To see our balancing property clearly, we modify our method and add two components for unconditional balancing from [40] (see Appendix), and compare this modified version to the original.

Table 1: Errors on IHDP over 1000 random DGPs. We report results with $\dim(\mathbf{z}) = 10$. **Bold** indicates method(s) that are *significantly* better. The results are taken from [40], except GANITE [54] and CEVAE [30].

Method	TMLE	BNN	CFR	CF	CEVAE	GANITE	Ours	Ours Mod.
ϵ_{ATE}	$.30_{\pm.01}$	$.37_{\pm.03}$	$.25_{\pm.01}$	$\mathbf{.18}_{\pm.01}$	$.34_{\pm.01}$	$.43_{\pm.05}$	$\mathbf{.178}_{\pm.006}$	$\mathbf{.167}_{\pm.005}$
$\sqrt{\epsilon_{PEHE}}$	$5.0_{\pm.2}$	$2.2_{\pm.1}$	$\mathbf{.71}_{\pm.02}$	$3.8_{\pm.2}$	$2.7_{\pm.1}$	$1.9_{\pm.4}$	$.859_{\pm.033}$	$.777_{\pm.026}$

As shown in Table 1, Intact-VAE outperforms or matches the state-of-the-art methods. Particularly, our method has the *best* ATE estimation, and is slightly worse than CFR for PEHE. This is possibly due to the fitting capacity (recall Sec. 6.1), and also we do *not* tune β. Notably, our method outperforms other generative models (CEVAE and GANITE) by large margins. The modified version is slightly improved, but we should note that the improvement for ϵ_{ATE} is barely significant. This indicates overfitting to PEHE. In fact, PEHE estimates the *marginalized* error $\mathbb{E}\epsilon(\mathbf{x})$ where $\epsilon(\mathbf{x}) = (\tau(\mathbf{x}) - \hat{\tau}(\mathbf{x}))^2$, and, compared with ϵ_{ATE}, it focuses on values $\boldsymbol{x}$ with high probability and / or large $\epsilon(\boldsymbol{x})$. The balancing in [40] is based on bounding $\mathbb{E}\epsilon(\mathbf{x})$, and thus tends to overly focus on the above values of $\mathbf{x}$, resulting in sub-optimal estimation of CATE and even of ATE. This tendency is more apparent with sub-optimal hyperparameter for the unconditional balancing (see Appendix).

7 Conclusion

In this work, we proposed a method for CATE estimation, under weak overlap. Our method exploits identifiable VAE, a recent advance in generative models, and is fully motivated and theoretically justified by causal considerations: identification, PGS, and balancing. We show that VAEs are suitable for *principled* causal inference thanks to its probabilistic nature, if not compromised by ad hoc heuristics. We believe it is possible to extend the bounds in Sec. 4.2 to weak overlap, just as [24] extends [40] to weak overlap, and leave this for future. Experiments show evidence that the injectivity of $\boldsymbol{f}$ in our model is possibly unnecessary because $\dim(\mathbf{z}) > \dim(\mathbf{y})$ often gives better results. Theoretical study of this is an interesting future direction. To avoid potential negative societal impact (e.g, bad prescriptions), practitioners should judge the conditions of the proposed method by their domain expertise, and careful trials are always recommended.

References

[1] Ahmed M Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pages 3424–3432, 2017.

[2] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[3] Timothy B. Armstrong and Michal Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *arXiv preprint arXiv:1712.04594v5*, 2021.

[4] Stéphane Bonhomme and Martin Weidner. Posterior average effects. *arXiv preprint arXiv:1906.06360v5*, 2021.

[5] M Alan Brookhart, Sebastian Schneeweiss, Kenneth J Rothman, Robert J Glynn, Jerry Avorn, and Til Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[6] Alberto Caron, Ioanna Manolopoulou, and Gianluca Baio. Estimating individual treatment effects using non-parametric regression models: a review. *arXiv preprint arXiv:2009.06472*, 2020.

[7] Denis Chetverikov, Andres Santos, and Azeem M Shaikh. The econometrics of shape restrictions. *Annual Review of Economics*, 10:31–63, 2018.

[8] Wangzhi Dai and Collin M Stultz. Quantifying common support between multiple treatment groups using a contrastive-vae. In *Machine Learning for Health*, pages 41–52. PMLR, 2020.

[9] Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

[10] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 2020.

[11] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

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

[13] David Hajage, Yann De Rycke, Guillaume Chauvet, and Florence Tubach. Estimation of conditional and marginal odds ratios using the prognostic score. *Statistics in medicine*, 36(4):687–716, 2017.

[14] Ben B Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

[15] Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

[16] Miguel A. Hernan and James M. Robins. *Causal Inference: What If*. CRC Press, 1st edition, 2020.

[17] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

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

[19] Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):32–47, 2020.

[20] Ming-Yueh Huang and Kwun Chuen Gary Chan. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 104(3):583–596, 2017.

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

[22] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

[23] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[24] Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

[25] Nathan Kallus, Brenton Pennicooke, and Michele Santacatterina. More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *arXiv preprint arXiv:1811.04274*, 2018.

[26] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

[27] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[29] Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

[30] 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*, pages 6446–6456, 2017.

[31] Danni Lu, Chenyang Tao, Junya Chen, Fan Li, Feng Guo, and Lawrence Carin. Reconsidering generative objectives for counterfactual reasoning. *Advances in Neural Information Processing Systems*, 33, 2020.

[32] Wei Luo, Yeying Zhu, and Debashis Ghosh. On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104(1):51–65, 2017.

[33] Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pages 788–798. PMLR, 2020.

[34] Judea Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

[35] Paul R Rosenbaum. Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7:143–176, 2020.

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

[37] Donald B Rubin. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127(8_Part_2):757–763, 1997.

[38] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[39] Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *arXiv preprint arXiv:2012.09935*, 2020.

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

[41] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pages 2507–2517, 2019.

[42] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[43] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2019.

[44] Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010.

[45] Elizabeth A Stuart, Brian K Lee, and Finbarr P Leacy. Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *Journal of clinical epidemiology*, 66(8):S84–S90, 2013.

[46] Xinwei Sun, Botong Wu, Chang Liu, Xiangyu Zheng, Wei Chen, Tao Qin, and Tie-yan Liu. Latent causal invariant model. *arXiv preprint arXiv:2011.02203*, 2020.

[47] Alexander Tarr and Kosuke Imai. Estimating average treatment effects with support vector machines. *arXiv preprint arXiv:2102.11926*, 2021.

[48] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[49] 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.

[50] Halbert White and Karim Chalak. Identification and identification failure for treatment effects using structural systems. *Econometric Reviews*, 32(3):273–317, 2013.

[51] Pengzhou Wu and Kenji Fukumizu. Causal mosaic: Cause-effect inference via nonlinear ica and ensemble method. In *International Conference on Artificial Intelligence and Statistics*, pages 1157–1167. PMLR, 2020.

[52] S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018.

[53] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[54] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[55] Weijia Zhang, Lin Liu, and Jiuyong Li. Treatment effect estimation with disentangled latent factors. *arXiv preprint arXiv:2001.10652*, 2020.

[56] Yao Zhang, Alexis Bellot, and Mihaela Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 1005–1014. PMLR, 2020.

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]
   (b) Did you describe the limitations of your work? [Yes] Particularly in Conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We mentioned possible misuse of our method and suggestions for practitioners in Introduction and Conclusion.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In Appendix.
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] In supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Possibly in Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (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]
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]
   (b) Did you mention the license of the assets? [Yes] In code project.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] In supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] In Appendix. And we refer to the original paper where this is discussed in details.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Refer to the original paper. And we believe there are no privacy issues.
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? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]