DISCRET: Synthesizing Faithful Explanations For Treatment Effect Estimation

Yinjun Wu [1] **Mayank Keoliya** [2] **Kan Chen** [3] **Neelay Velingker** [2] **Ziyang Li** [2] **Emily J Getzen** [4] **Qi Long** [2 5] **Mayur Naik** [2] **Ravi B Parikh** [4] **Eric Wong** [2]

Abstract

Predicting individual treatment effect (ITE) is a vital problem across several domains. ITE prediction models deployed in critical settings such as healthcare should ideally be (i) accurate, and (ii) provide faithful explanations. However, current solutions are inadequate: state-of-the-art black-box models do not supply explanations, post-hoc explainers for black-box models lack faithfulness guarantees, and self-interpretable models greatly compromise accuracy. To address these issues, we propose DISCRET, a self-interpretable ITE framework that synthesizes faithful, rule-based explanations for each sample. A key insight behind DISCRET is that explanations can serve dually as *database queries* to identify similar subgroups of samples. We provide a novel RL algorithm to efficiently synthesize these explanations from a large search space. We evaluate DISCRET on diverse tasks involving tabular, image, and text data. DISCRET outperforms the best self-interpretable models and has accuracy comparable to the best black-box models while providing faithful explanations.

1. Introduction

Designing accurate and explainable AI models is a key challenge in solving a wide range of problems which require individualized explanations. In this paper, we tackle this challenge in the context of individual treatment effect (ITE) estimation. ITE quantifies the difference between one individual's outcomes with and without receiving treatment. Estimating ITE is a significant problem not only in healthcare (Basu et al., 2011) but also in other domains such as linguistics (Pryzant et al., 2021; Feder et al., 2021) and poverty alleviation (Jerzak et al., 2023a;b). A large body of literature has investigated accurately estimating ITE using various machine learning architectures, including GANs (Yoon et al., 2018) and transformers (Zhang et al., 2022), among others (Shalit et al., 2017; Liu et al., 2022).

ITE prediction models deployed in critical settings should ideally be (i) **accurate**, and (ii) provide **faithful explanations** in order to be trustable and usable. In this paper, we follow prior work on evaluating faithfulness of explanations in terms of *consistency*, which measures the degree to which samples with similar explanations have similar model predictions (Dasgupta et al., 2022; Nauta et al., 2023).

Current solutions for predicting ITE are either accurate or faithful, but not both, as illustrated in the first two rows of Figure 1. While self-interpretable models such as Causal Forest (Athey & Wager, 2019; Chen et al., 2022) produce consistent explanations, they struggle to provide sufficiently accurate ITE estimations. On the other hand, while black-box models like transformers are typically the most accurate, explanations generated by post-hoc explainers, such as Anchor (Ribeiro et al., 2018), are not provably consistent.

We therefore seek to answer the following central question: *Is it possible to design a faithfully explainable yet accurate learning algorithm for treatment effect estimation?* To this end, we propose DISCRET[1], the first provably-faithful, deep learning based ITE prediction framework. Given a sample x, DISCRET follows prior work and estimates ITE by computing the average treatment effect (ATE) of samples that are similar to x. However, in contrast to prior methods that discover similar samples through statistical matching (Anderson et al., 1980) or clustering (Xue et al., 2023), DISCRET finds similar samples by (i) synthesizing a logical rule that describes the key features of sample x (and hence *explains* the subgroup the sample belongs to) and then (ii) evaluating this rule-based explanation on a database of training samples (see Figure 2 for our pipeline). As shown in Figure 1, DISCRET produces consistent explanations for samples with similar predictions; in fact, it is guaranteed to

[1]School of Computer Science, Peking University, Beijing, China [2]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, United States [3]School of Public Health, Harvard University, Boston, MA, United States [4]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States [5]**AUTHORERR: Missing \icmlaffiliation.** . Correspondence to: Yinjun Wu <wuyinjun@pku.edu.cn>, Mayank Keoliya <mkeoliya@seas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]DIScovering Comparable items with Rules to Explain Treatment Effect

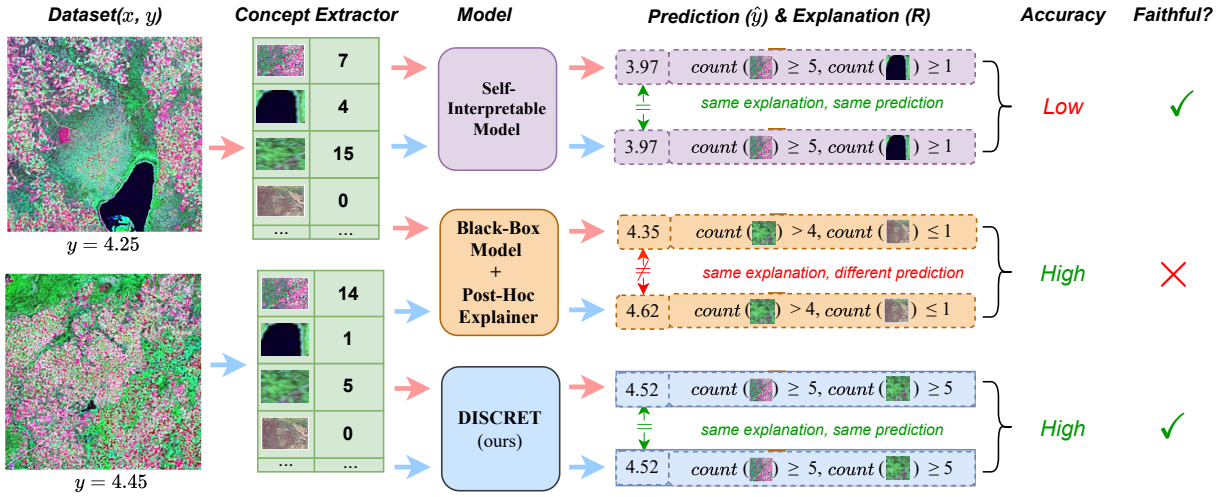

Figure 1. Motivating examples from the Uganda dataset. We predict how providing economic aid (the treatment) helps to develop remote regions of the country (the outcome) via satellite images. The task is to estimate the ITE for each sample x_1 and x_2. DISCRET predicts that, because both images have several indicators of rich soil and urbanization, they will have similar ITE if given aid. Self-interpretable models such as Causal Forest (Athey & Wager, 2019) produce *consistent* ITE estimates (i.e., samples with same explanations have same model predictions, *viz.* 3.97 and 3.97), but have poor accuracy ($\hat{ITE}_{x_1} \ll ITE_{x_1} = 4.25$). Black-box models such as TransTEE (Zhang et al., 2022), are accurate but do not produce similar predictions for samples x_1 and x_2 with similar explanations, when the explanations are sourced from post-hoc explainers such as Anchor (Ribeiro et al., 2018). DISCRET produces both consistent and accurate predictions.

be consistent by construction, as we show later.

How does DISCRET synthesize rules which correctly group similar samples, and thus lead to accurate predictions? Learning to synthesize rules is challenging, since the execution of database queries is non-differentiable and thus we cannot compute an end-to-end loss easily. To address this issue, we design a deep reinforcement learning algorithm with a novel and tailored reward function for dynamic rule learning. We also state the theoretical results of the convergence of DISCRET under some mild conditions suggesting if the ground-truth explanations are consistent, then our training algorithm can always discover them.

Due to the widely recognized trade-offs between interpretability and prediction performance (Dziugaite et al., 2020), DISCRET slightly underperforms the state-of-the-art black-box models (Zhang et al., 2022). In addressing this, we found that regularizing the training loss of black-box models such as TransTEE to penalize discrepancy with DISCRET predictions yields new state-of-the-art models.

We evaluate the capabilities of DISCRET through comprehensive experiments spanning four tabular, one image, and one text dataset, covering three different types of treatment variables. For tabular data, among others, we use the IHDP dataset (Hill, 2011) which tracks cognitive outcomes of premature infants. Other datasets used are TCGA (tabular) (Weinstein et al., 2013), IHDP-C (tabular), Uganda satellite-images for estimating poverty intervention (image), and the Enriched Equity Evaluation Corpus (text). Notably, our approach outperforms all self-interpretable methods, including by 34% on IHDP, is comparable to the accuracy of black-box models, and produces more faithful explanations than post-hoc explainers. In addition, regularizing the state-of-the-art black-box models with DISCRET reduces their ITE prediction error across tasks, including by 18% on TCGA.

Our contributions can be summarized as follows:

1. We introduce DISCRET, a self-interpretable framework that synthesizes faithful rule-based explanations, and apply it to the treatment effect estimation problem.
2. We present a novel deep Q-learning algorithm to automatically learn these rule-based explanations, and supplement it with theoretical results.
3. We conduct an extensive empirical evaluation which demonstrates that DISCRET outperforms existing self-interpretable models and is comparable to black-box models across tabular, image, and text datasets spanning a diverse range of treatment variable types. Moreover, regularizing the state-of-the-art black-box models with DISCRET further reduces their prediction error.

2. Preliminaries

2.1. Individual Treatment Effect (ITE) Estimation

Suppose each sample consists of (i) the pre-treatment covariate variable X, (ii) the treatment variable T, (iii) a dose variable S associated with T, and (iv) observed outcome Y under treatment T and dose S. We embrace a versatile framework throughout this study, where T can take on either

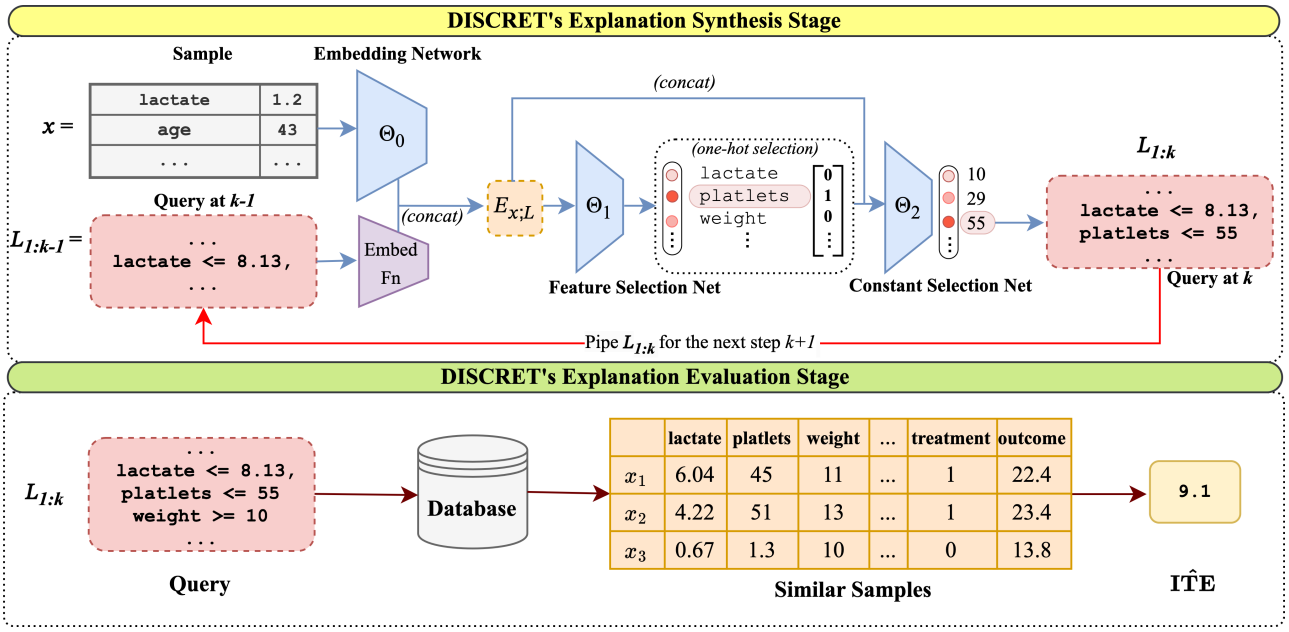

Figure 2. Illustration of DISCRET on the IHDP dataset, which tracks premature infants. DISCRET begins by generating an explanation; during explanation synthesis, DISCRET iteratively constructs each literal in the explanation by embedding the sample and any previously generated literals, passing that embedding to the feature selection network to pick the feature, and then passing the embedding and selected feature to the constant selection network to get the constant. The operator is auto-assigned based on the feature and sample. Second, DISCRET executes this explanation on the database to find relevant samples, from which ITE is calculated.

discrete or continuous values, S is inherently continuous but can be either present or absent, Y can be discrete or continuous, and X may incorporate structured features as well as unstructured features, such as text or image data. In the rest of the paper, we primarily explore a broadly studied setting where Y is a continuous variable, T is a binary variable ($T = 1$ and $T = 0$ represent treated and untreated respectively) and there is no dose variable. The goal is to estimate individual treatment effects (ITE), i.e., the difference of outcomes with $T = 1$ and $T = 0$. Typically, the average treatment effect (ATE), the average of ITE across all samples (i.e., ATE $= \mathbb{E}[\text{ITE}]$) is reported. Generalizations to other settings are provided in Appendix D.5.

Beyond the treatment effect definitions, the propensity score, represented as the probability of treatment assignment T conditioned on the observed covariates X, often plays a pivotal role in regularizing the treatment effect estimation. This propensity score is denoted as $\pi(T|X)$.

Unlike conventional prediction tasks, we are unable to directly observe the counterfactual outcomes during training, rendering the ground-truth treatment effect typically unavailable. To address this challenge and ensure the causal interpretability of our estimated treatment effect, we adhere to the standard assumptions proposed by Rubin (1974), which are formulated in Appendix D.1.

2.2. Syntax of Logic Rules

We assume that the covariate variable X is composed of m features, $X_1, X_2, \ldots, X_m$, which can be categorical or numeric attributes from tabular data or pre-processed features extracted from text data or image data. We then build logic rule-based explanations upon those features to construct our treatment effect estimator. Those logic rules are assumed to be in the form of K disjunctions of multiple conjunctions, i.e., $R_1 \vee R_2 \vee \cdots \vee R_H$ where each R_i is a conjunction of K literals: $l_{i1} \wedge l_{i2} \wedge l_{i3} \wedge \cdots \wedge l_{iK}$. Each $l_{ij} (j = 1, 2, \ldots)$ represents a literal of the form $l_{ij} = (A \; op \; c)$, where $A \in \{X_1, X_2, ..., X_m\}$; op is equality or inequality for categorical attributes, and $op \in \{<, >, =\}$ for numeric attributes; and c is a constant.

3. The DISCRET Framework

The DISCRET framework consists of a two-step process: *explanation synthesis* and *explanation evaluation*, depicted in Figure 2. Initially, a rule-based explanation is synthesized for a given sample, capturing pertinent characteristics. Subsequently, a subgroup of similar samples satisfying the explanation is selected from the training database, and the predicted ITE is computed over this subgroup.

This section outlines the two steps of DISCRET assuming a trained explanation synthesizer (§3.1 and §3.2). We then introduce our novel Q-learning algorithm to train the ex-

planation synthesizer (§3.3). Additionally, we explain how DISCRET is employed to regularize state-of-the-art deep learning models for maximal performance (§3.4). Note that we sometimes use "rule" and "explanation" interchangeably.

3.1. Explanation Synthesis

3.1.1. Desired Properties of Explanations

We first state four desired properties of a rule-based explanation. We refer to these properties in §3.1.2 and §3.3.

1. **Local interpretability**: We aim to synthesize a rule-based explanation R_x for *each* individual sample x rather than for a population of samples. Thus, explanations may differ for different samples.
2. **Satisfiability**: For any rule R_x for a given sample x, x's features must satisfy R_x. This guarantees that the sample x and any samples retrieved by R_x share the same characteristics.
3. **Low-bias**: We expect that R_x can retrieve a set of similar samples so that the bias between the estimated ATE over them and the ground-truth ITE is as small as possible.
4. **Non-emptiness**: There should be at least one sample from the database whose covariates satisfy R_x. In addition, for those samples satisfying R_x, their treatment variables should cover all essential treatment values for treatment effect estimations, e.g., containing both treated and untreated units in binary treatment settings.

3.1.2. Synthesis at Inference Phase

The explanation synthesizer consists of three models: a backbone model to encode features Θ_0, a feature-selector Θ_1, and a constant-value selector Θ_2. During inference, we assume that these are already trained. Now, given a sample x, Θ_0, Θ_1 and Θ_2, we want to synthesize a conjunctive rule R_x which takes the form of R_x :- $l_1 \wedge l_2 \wedge l_3 \wedge \cdots \wedge l_K$.

DISCRET synthesizes R_x recursively by generating all of $l_k, k = 1, 2, \cdots, K$ one after the other in multiple rounds. Since l_k takes the form of $(A\ op\ c)$, then at each round k, we determine a feature A to select using Θ_1, a constant c to be compared against using Θ_2, and an operator op.

Figure 2 depicts how a literal l_k is generated at the k_{th} round with a running example from the IHDP dataset. First, we encode the covariate x as E_x with the feature encoder of one backbone model (e.g., TransTEE model (Zhang et al., 2022)) parameterized by Θ_0, and also encode all the literals generated in the first $k-1$ rounds (see Appendix D.2), denoted by "$L_{1:k-1} = \{l_1, l_2, \cdots, l_{k-1}\}$", as E_L. E_x and E_L are then fed into a feed-forward neural network (parameterized by Θ_1) and *the most probable* feature from $\{X_1, X_2, \ldots, X_m\}$ is selected as A for l_k.

Selection of the constant c and the operator op for literal l_k depends on the type of the selected attribute A. If A is a categorical attribute, then we assign $c = x[A]$, i.e., the value of attribute A in R_x is the value of A in sample x; we assign op as $=$, which guarantees the **satisfiability** of R_x on x.

Instead, if A is a numeric attribute, we first discretize the range of A with a list of μ evenly distributed float numbers, $\{C_1, C_2, \ldots, C_\mu\}$, to facilitate the computation of the Q-value for the subsequent deep Q-learning process. We then design a feed-forward neural network (parameterized by Θ_2) to produce the probability distribution of these μ numbers, which takes the encoding of the covariates and $L_{1:k-1}$, $E_{x:L}$, and the one-hot encoding of feature A as the model input. *The most probable* C_j is then selected as c. After the feature A and the constant c are identified, the operator op is then deterministically chosen by comparing the value $x[A]$ and c. If $x[A]$ is greater than c, then op is assigned as $\geq$, and as $\leq$ otherwise, thus again guaranteeing the **satisfiability** of the rule R_x on x. To produce multiple disjunctions with DISCRET, multiple literals are generated simultaneously at each round, each of which is assigned to one disjunction respectively (see Appendix D.3 for details).

3.2. Explanation Evaluation

Querying of Similar Samples. Given a sample x (e.g., a patient) with (X, T, S, Y), and a rule R_x, the next step is to evaluate the rule R_x on a database $\mathcal{D}$ to retrieve a subgroup of similar samples, which is denoted by $R_x(\mathcal{D}) = \{(x_i^*, t_i^*, s_i^*, y_i^*)\}_{i=1}^n$.

ITE Estimation. The ITE of the sample x is then estimated by computing the average treatment effect (ATE) estimated within this subgroup. In this paper, we take the empirical mean by default for estimating ATE of $R_x(\mathcal{D})$, i.e., $\hat{y}(1) - \hat{y}(0)$, in which $\widehat{y}(t), (t = 0, 1)$ denotes the estimated outcome calculated with the following formula:

$$\widehat{y}(t) = \frac{1}{\sum \mathbb{I}(t_i^* = t)} \sum \mathbb{I}(t_i^* = t) \cdot y_i^* \tag{1}$$

We also estimate the propensity score for discrete treatment variables by simply calculating the frequency of every treatment within $R_x(\mathcal{D})$: $\widehat{\pi}(T = t|X = x) = \sum \mathbb{I}(t_i^* = t)/|R_x(\mathcal{D})|$.

3.3. RL problem formulation

To preserve the **low-bias** property, we need to guide the generation of rules such that the estimated ITE is as accurate as possible. We therefore need to optimize this objective by training the three models (Θ_0, Θ_1 and Θ_2) used in the rule generation process. One difficulty of learning $\Theta = [\Theta_0, \Theta_1, \Theta_2]$ is the non-differentiability issue caused by the step of evaluating R_x over the database. We overcome this issue by formulating this model training problem as a deep reinforcement learning (RL) problem, and propose to adapt the deep Q-learning (DQL) algorithm to solve this problem.

We first map the notations from Section 3.1.2 to classical RL terminology. An RL agent takes one *action* at one *state*, collects a *reward* from the environment, which is then transitioned to a new state. In our rule learning setting, a *state* is composed of the covariates x and the generated literals in the first $k-1$ rounds, $L_{1:k-1}$. With x and $L_{1:k-1}$, the model Θ_1 and Θ_2 collectively determine the k_{th} literal, l_k, which is regarded as one *action*. Our goal is then to learn a policy parameterized by Θ, which models the probability distribution of all possible l_k conditioned on the state $(x, L_{1:k-1})$, such that the value function calculated over all K rounds is maximized:

$$V_{1:K} = \sum_{k=1}^{K} r_k \gamma^{k-1}, \tag{2}$$

in which γ is a discounting factor. Note that there are only K horizons/rounds in our settings since the number of conjunctions in the generated rules is limited. To bias rule generation towards accurate estimation of ITE, we expect that the value function $V_{1:K}$ reflects how small the ITE estimation error is. However, since the counterfactual outcomes are not observed in the training phase, we therefore use the errors of the observed outcomes as a surrogate of the ITE estimation error. Also, we give a zero reward to the case where the retrieved subgroup, $L_{1:K}(\mathcal{D})$, violates the **non-emptiness** property. As a result, $V_{1:K}$ is formulated as

$$V_{1:K} = e^{-\alpha(y-\widehat{y}_{1:K})^2} \cdot \mathbb{I}(L_{1:K}(\mathcal{D}) \text{ is non-empty}), \tag{3}$$

in which $\widehat{y}_{1:K}$ represents the estimated outcome by using the generated rule composed of literals $L_{1:K}$ and α is a hyperparameter. As a consequence, the reward collected at the k_{th} round of generating l_k becomes $r_k = (V_{1:k} - V_{1:k-1})/\gamma^{k-1}$. We further discuss how to automatically fine-tune the hyperparameter α and incorporate the propensity score defined in Section 3.2 for regularization in Appendix D.8.

Next, to maximize the value function $V_{1:K}$, we employ Deep Q-learning (DQL) (Mnih et al., 2013) to learn the parameter Θ. To facilitate Q learning, we estimate the Q value with the output logits of the models given a state $(x, L_{1:k-1})$ and an action l_k. Recall that since DISCRET can generate consistent explanations by design, we can show that if Θ_0 is an identity mapping and Θ_1 is a one-layer neural network, the following theorem holds:

Theorem 3.1. *Suppose we have input data* $\{(x_i, t_i, s_i, y_i)\}_{i=1}^{N}$ *where* $x_i \in \mathbb{R}^m$ *and discrete,* $t_i \in \mathbb{R}, s_i \in \mathbb{R}$, *and* $y_i \in \mathbb{R}$, *then the* $\hat{ITE}_x$ *obtained from DISCRET converges to zero generalization error with probability 1 for ITE estimation (i.e.* $(ITE_x - \hat{ITE}_x)^2 \to 0$ *w.p. 1) for any fixed* $K \leq m$ *over the dataset with all discrete features under the data generating process* $y = f(\mathcal{X}_K) + c \cdot t + \epsilon$, *where* $\mathcal{X}_K \subseteq \{X_1, X_2, \cdots, X_m\}, c \in \mathbb{R}, t$ *is the treatment assignment, and* $\epsilon \sim \mathcal{N}(0, \sigma^2)$ *for some* $\sigma > 0$.

Intuitively, Theorem 3.1 suggests if the ground-truth explanations are consistent, then our training algorithm can perfectly discover them. We prove the theorem and explain our algorithm in detail in Appendix D.

3.4. Regularizing black-box models with DISCRET

Due to the widely recognized trade-offs between model interpretability and model performance (Dziugaite et al., 2020), self-interpretable models typically suffer from poorer performance than their neural network counterparts. To achieve a better balance between the performance and interpretability, we further propose to regularize the prediction of black-box models with that of DISCRET. Since DISCRET also leverages part of the black-box model such as TransTEE as the backbone Θ_0, we thus obtain the predictions of black-box models by reusing Θ_0. Specifically, starting from the encoded covariates E_x generated by Θ_0, we predict another outcome $\hat{y'}$ directly with E_x adhering to the mechanism employed by state-of-the-art neural models. This prediction is then regularized by the predicted outcome $\widehat{y}_{1:K}$ by DISCRET as follows:

$$\widehat{y'}_{1:K} = (\hat{y'} + \lambda \widehat{y}_{1:K})/(1+\lambda),$$

in which λ is a hyperparameter for controlling the impact of $\hat{y'}$. Afterward, $\widehat{y}_{1:K}$ is replaced with $\widehat{y'}_{1:K}$ in Equation 3 or Equation 9 for model training. In addition, to facilitate accurate $\hat{y'}$, we further minimize the loss involving $\hat{y'}$ and y along with the Deep Q-learning loss.

4. Experiments

In this section, we aim to answer the following research questions about DISCRET:

RQ1: Does DISCRET produce faithful explanations?
RQ2: How does the accuracy of DISCRET perform compare to existing self-interpretable models and black-box models?

4.1. Setup

Datasets. We evaluate across tabular, text, and image datasets, covering diverse categories of treatment variables. Specifically, we select IHDP (Hill, 2011), TCGA (Weinstein et al., 2013) IHDP-C (a variant of IHDP), and News for tabular setting, the Enriched Equity Evaluation Corpus (EEEC) dataset (Kiritchenko & Mohammad, 2018) for text setting and Uganda (Jerzak et al., 2023b;a) dataset for the image setting. We summarize the modality, categories of treatment and dose variables, and number of features for each dataset in Table 1, with more details in Appendix A.

Baselines. We use extensive baselines for neural network models, self-interpretable models, and post-hoc explainers.

Neural network models. For neural networks, we select the state-of-the-art models: TransTEE (Zhang et al., 2022), TVAE (Xue et al., 2023), Dragonnet (Shi et al., 2019), TARNet (Shalit et al., 2017), Ganite (Yoon et al., 2018), DRNet

Dataset	Type	Treatment	Dose	# Features
IHDP	Tabular	2	✗	25
TCGA	Tabular	3	✓	4000
IHDP-C	Tabular	cont.	✗	25
News	Tabular	cont.	✗	2000
EEEC	Text	2	✗	500
Uganda	Image	2	✗	20

Table 1. Datasets used for evaluation (cont. means continous)

(Schwab et al., 2020), and VCNet (Nie et al., 2020). Not all of these models support all categories of treatment variables, as discussed in Appendix C. Also, since our regularization strategy can be regarded as the integration of two models through weighted summation, we compare our regularized backbone (TransTEE) against the integration of TransTEE and another top-performing neural network model (Dragonnet for IHDP, EEEC, and Uganda dataset, VCNet for TCGA, DRNet for IHDP-C) in the same manner.

Self-interpretable models. We compare against classical self-interpretable models, e.g., Causal Forest (Athey & Wager, 2019), Bayesian Additive Regression Trees (BART), decision tree (DT), and random forests (RF), in which the latter two are integrated into R-learner (Nie & Wager, 2021) for treatment effect estimation. We also adapt three general-purpose self-interpretable models to treatment effect estimation—ENRL (Shi et al., 2022), ProtoVAE (Gautam et al., 2022)[2], and Neural Additive Model (NAM) (Agarwal et al., 2021), which generate rules, prototypes, and feature attributes as explanations respectively. For tree-based models among these methods, we maintain the same explanation complexity as DISCRET. For the sake of completeness we also conduct additional experiments to vary the complexity (e.g., the number of trees and tree depth) of all self-interpretable models, provided in Table 3 in Appendix F; DISCRET outperforms self-interpretable models even when they are configured to high complexity.

Post-hoc explainers. We apply several post-hoc explainers to the TransTEE model to evaluate the consistency of explanations. Thy include Lore (Guidotti et al., 2018), Anchor (Ribeiro et al., 2018), Lime (Ribeiro et al., 2016), Shapley values (Shrikumar et al., 2017), and decision tree-based model distillation methods (Frosst & Hinton, 2017) (hereinafter referred to as Model Distillation). We enforce the complexity of these explanations to be the same as DISCRET for fair comparison.

Evaluation metrics. We primarily evaluate faithfulness by measuring *consistency*, proposed by (Dasgupta et al., 2022); we also measure *sufficiency*, which is a generalization of consistency. Briefly, *consistency* quantifies how similar the model predictions are between samples with the same explanations, while *sufficiency* generalizes this notion to arbitrary samples *satisfying* the same explanations (but not necessarily *producing* the same explanations). Appendix E provides formal definitions of these two metrics.

We evaluate ITE estimation accuracy using different metrics for datasets to account for different settings. For the datasets with binary treatment variables, by following prior studies (Shi et al., 2019; Shalit et al., 2017), we employ the absolute error in average treatment effect, i.e., $\epsilon_{ATE} = |\frac{1}{n}\sum_{i=1}^{n} ITE(x_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{ITE}(x_i)|$. Both in-sample and out-of-sample ϵ_{ATE} are reported, i.e., ϵ_{ATE} evaluated on the training set and test set respectively. For the datasets with either continuous dose variables or continuous treatment variables, we follow (Zhang et al., 2022) to report the average mean square errors $AMSE$ between the ground-truth outcome and predicted outcome on the test set. For the image dataset, Uganda, since there is no ground-truth ITE, we therefore only report the average outcome errors between the ground-truth outcomes and the predicted outcomes conditioned on observed treatments, i.e., $\epsilon_{\text{outcome}} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$.

Configurations for DISCRET. We consider two variants of DISCRET: vanilla DISCRET and backbone models regularized with DISCRET (denoted as DISCRET + TransTEE). For both variants, we perform grid search on the number of conjunctions, K, and the number of disjunctions, H, and the regularization coefficient λ. We found that with $H = 1$ and $K = 6$, we can balance both explanation complexity and performance.

Extracting features from text and image data. For text data, we employ the word frequency features such as "Term Frequency-Inverse Document Frequency" (Baeza-Yates et al., 1999). For image data, we follow (Fel et al., 2023) to extract interpretable concepts as the features, which we further discuss in Appendix H. Note that we only extract features for DISCRET and self-interpretable baselines such as Causal Forest while all neural network model-based baselines still take raw images or text data as input.

4.2. RQ1: Faithfulness evaluation on explanations

We evaluate the *consistency* and *sufficiency* of explanations produced by DISCRET, the state-of-the-art self-interpretable models, and the post-hoc explainers. For those explainers producing feature-based explanations, we also follow (Dasgupta et al., 2022) to discretize the feature importance scores, say, by selecting the Top-K most important features, for identifying samples with exactly the same explanations. For fair comparison, we evaluate the explanations generated w.r.t. the same set of features extracted from NLP and image data.

[2]ProtoVAE is designed for image data. We therefore only compare DISCRET against this method on the Uganda dataset.

Modality →		**Tabular**					**Text**	**Image**
Dataset →		IHDP		TCGA		IHDP-C	EEEC	Uganda
Method ↓	Self-interp.?	ϵ_{ATE} (In-sample)	ϵ_{ATE} (Out-of-sample)	ϵ_{ATE} (In-sample)	ϵ_{ATE} (Out-of-sample)	AMSE	ϵ_{ATE}	$\epsilon_{\text{outcome}}$
Decision Tree	✓	1.507±0.061	1.420±0.066	0.200±0.012	0.202±0.012	22.136±1.741	0.014±0.016	1,796±0.021
Random Forest	✓	1.268±0.087	1.204±0.088	0.263±0.057	0.264±0.058	21.348±1.222	0.525±0.573	1.820±0.013
NAM	✓	0.260±0.031	0.250±0.032	-	-	24.706±0.756	0.152±0.041	1.710±0.098
ENRL	✓	4.104±1.060	3.759±0.087	10.938±2.019	10.942±2.019	24.720±0.985	-	1.800±0.143
Causal Forest	✓	0.177±0.027	0.240±0.024	-	-	-	0.011±0.001	-
BART	✓	1.335±0.159	1.132±0.125	230.74±0.312	236.81±0.531	12.063±0.410	0.014±0.016	1.676±0.042
DISCRET (ours)	✓	0.089±0.040	0.150±0.034	0.076±0.019	*0.098±0.007*	0.801±0.165	**0.001±0.017**	**1.662±0.136**
Dragonnet	✗	0.197±0.023	0.229±0.025	-	-	-	0.011±0.018	1.709±0.127
TVAE	✗	3.914±0.065	3.573±0.087	-	-	-	0.521±0.080	49.55±2.38
TARNet	✗	0.178±0.028	0.441±0.088	1.421±0.078	1.421±0.078	12.967±1.781	0.009±0.018	1.743±0.135
Ganite	✗	0.430±0.043	0.508±0.068	-	-	-	1.998±0.016	1.766±0.024
DRNet	✗	0.193±0.034	0.433±0.080	1.374±0.086	1.374±0.085	11.071±0.994	0.008±0.018	1.748±0.127
VCNet	✗	3.996±0.106	3.695±0.077	0.292±0.074	0.292±0.074	-	0.011±0.017	1.890±0.110
TransTEE	✗	**0.081±0.009**	0.138±0.014	0.070±0.010	0.067±0.008	0.112±0.008	0.003±0.017	1.707±0.158
TransTEE + NN	✗	0.224±0.022	0.300±0.035	0.093±0.013	0.094±0.013	0.363±0.033	0.006±0.008	2.001±0.425
TransTEE + DISCRET (ours)	✗	0.082±0.009	**0.120±0.014**	**0.058±0.010**	**0.055±0.009**	**0.102±0.007**	**0.001±0.017**	**1.662±0.136**

Table 2. ITE estimation errors (lower is better). We **bold** the smallest estimation error for each dataset, and underline the second smallest one. We show that DISCRET outperforms self-interpretable models across all datasets, particularly on text ($\epsilon_{ATE} = 0.001$ for DISCRET v/s 0.011 for causal forest). DISCRET is comparable to the performance of black-box models, with the exception of the IHDP-C dataset. Regularizing black-box models with DISCRET (shown here as TransTEE + DISCRET) outperforms *all* models.

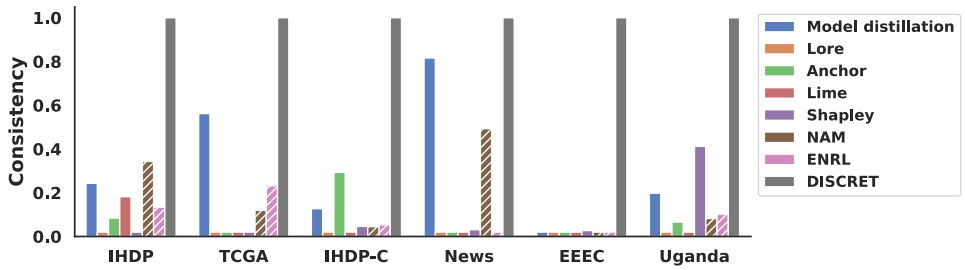

Figure 3. Consistency scores (higher is better) for DISCRET and a black-box model (TransTEE) combined with a post-hoc explainer. Our results confirm that DISCRET produces faithful explanations, and importantly, show that post-hoc explanations are rarely faithful, as evidenced by low consistency scores across datasets.

We graph the consistency scores in Figure 3; full consistency scores are provided in Table 5 in Appendix F. As Figure 3 indicates, DISCRET always achieves 100% consistency since the same explanations in DISCRET deterministically retrieve the same subgroup from the database, thus generating the same model predictions. In contrast, the baseline explanation methods generally have extremely low consistency scores in most cases. We also include the sufficiency score results in Table 6, which shows that DISCRET can still obtain higher sufficiency scores in most cases than other explanation methods.

4.3. RQ2: Accuracy Evaluation on ITE Predictions

We include the ITE estimation results for tabular setting, NLP setting, and image setting in Table 2. For brevity, the results on News dataset are not reported in Table 2, but are included in Table 7 in Appendix F.

As Table 2 shows, DISCRET outperforms all the self-interpretable methods, particularly on text ($\epsilon_{ATE} = 0.011$ for DISCRET v/s 0.0011 for causal forest). Compared to black-box models, DISCRET only performs slightly worse in most cases, and even outperforms them on the Uganda dataset. The outperformance is possibly caused by equivalent outcome values among most samples in this dataset as suggested by Figure 7 in Appendix F. Hence, consistent predictions (e.g., by DISCRET) between samples lead to a lower error rate. DISCRET underperforms TransTEE on IHDP-C, likely due to the complexity of the dataset; DISCRET still beats all other black-box models on this dataset.

Further, backbone models (TransTEE) regularized with DISCRET outperform the state-of-the-art neural network models, reducing their estimation errors by as much as 18% (TCGA dataset.) Interestingly, for the IHDPdataset, TransTEE outperforms its regularized version only on in-sample (i.e. training) error, but underperforms the regularized version when we consider out-of-sample (i.e. test) error. Intuitively, DISCRET's regularization incentivizes the underlying backbone's training (TransTEE) to focus on only a subset of the most important features, thereby reducing its variance and allowing it to perform better.

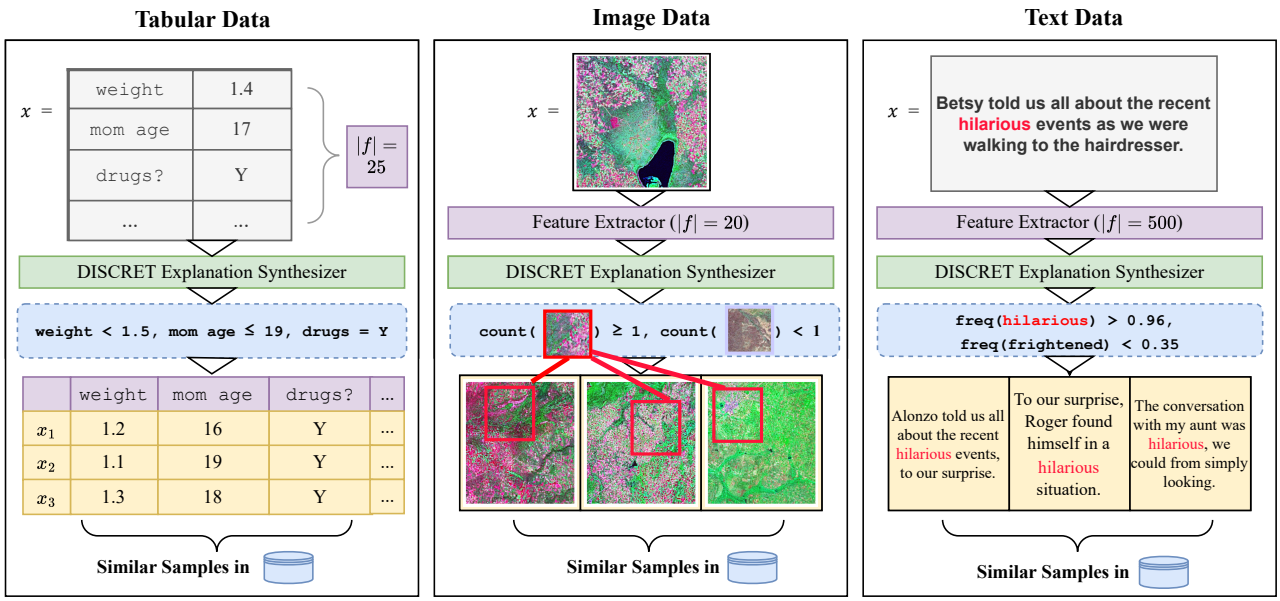

Figure 4. DISCRET identifies similar samples across diverse datasets —tabular (IHDP), image (Uganda), and text (EEEC). 1) In the first setting, given a tabular sample x describing a premature infant, DISCRET establishes a rule associating extremely underweight ($\texttt{weight} \leq 1.5$) infants born to teenage mothers ($\texttt{mom age} \leq 19$) with a history of drug use; such groups likely benefit from childcare visits (treatment), and will have highly improved cognitive outcomes. 2) In the second scenario on satellite images, for a sample x, DISCRET discerns a rule based on the presence of concepts like "high soil moisture" (reddish-pink pixels) and absence of minimal soil (brown pixels); thus characterizing areas with high soil moisture. DISCRET 's synthesized rule aligns with findings that government grants (treatment) are more effective in areas with higher soil moisture content (outcome) (Jerzak et al., 2023b). 3) Likewise, the text setting aims to measure the impact of gender (treatment) on the mood (outcome). Given a sentence x where the gendered noun ("Betsy") does not affect the semantic meaning, DISCRETś rule focuses on mood-linked words in the sentence, i.e., "hilarious".

5. Related Work

Treatment effect estimation. A substantial body of research has been dedicated to the estimation of treatment effects through machine learning. For instance, Shalit et al. (2017) introduced a novel theoretical analysis and a comprehensive family of algorithms to predict ITE from observational data. Additionally, Wager & Athey (2018) introduced a non-parametric causal forest approach tailored for the estimation of heterogeneous treatment effects. Bridging the gap between the predictive power of machine learning models and the need for interpretable decisions remains a pivotal challenge. To address the issue, (Kim & Bastani, 2019) proposed a framework for learning interpretable models from observational data for predicting ITE.

Model interpretability. There are two lines of work to address the model interpretability issues, one is for interpreting black-box models in a post-hoc manner while the other one is for building a self-interpretable model. Post-hoc explainers could explain models with feature importance (e.g., Lime (Ribeiro et al., 2016) and Shapley values (Shrikumar et al., 2017)) or logic rules (e.g., Lore (Guidotti et al., 2018), Anchor (Ribeiro et al., 2018) and model distillation methods (Frosst & Hinton, 2017)). However, post-hoc explanations are usually not faithful. (Rudin, 2019; Bhalla et al., 2023). To mitigate this issue, there are recent and ongoing efforts in the literature to develop self-interpretable models, meaning that such models perform predictions in a human-understandable manner. For example, ENRL (Shi et al., 2022) to learn tree-like decision rules and leverage those rules for predictions, ProtoVAE (Gautam et al., 2022) learns prototypes and predicts the label of one test sample by employing its similarity to prototypes.

We relegate the discussion on integrating rules into neural models and program synthesis to Appendix B.

6. Conclusion

In this work, we tackled the challenge of designing a faithful yet accurate AI model, DISCRET, in the context of ITE estimation. To achieve this, we developed a novel deep reinforcement learning algorithm that is tailored to the task of synthesized rule-based explanations which are database-executable. Extensive experiments across tabular, image, and text data demonstrate that DISCRET produces the most consistent (i.e. faithful) explanations, outperforms the accuracy self-interpretable models, is comparable in accuracy

to black-box models, and can be combined with existing black-box models to achieve state-of-the-art accuracy.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This work is supported by "The Fundamental Research Funds for the Central Universities, Peking University".

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34: 4699–4711, 2021.

Anderson, D. W., Kish, L., and Cornell, R. G. On stratification, grouping and matching. *Scandinavian Journal of Statistics*, pp. 61–66, 1980.

Athey, S. and Wager, S. Estimating treatment effects with causal forests: An application. *Observational studies*, 5 (2):37–51, 2019.

Baeza-Yates, R., Ribeiro-Neto, B., et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Basu, A., Polsky, D., and Manning, W. G. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology*, 11:1–26, 2011.

Bhalla, U., Srinivas, S., and Lakkaraju, H. Verifiable feature attributions: A bridge between post hoc explainability and inherent interpretability. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

Chen, K., Yin, Q., and Long, Q. Covariate-balancing-aware interpretable deep learning models for treatment effect estimation. *arXiv preprint arXiv:2203.03185*, 2022.

Dasgupta, S., Frost, N., and Moshkovitz, M. Framework for evaluating faithfulness of local explanations. In *International Conference on Machine Learning*, pp. 4794–4815. PMLR, 2022.

Dixit, A. K. *Optimization in economic theory*. Oxford University Press, USA, 1990.

Dziugaite, G. K., Ben-David, S., and Roy, D. M. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

Feder, A., Oved, N., Shalit, U., and Reichart, R. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

Fel, T., Boutin, V., Moayeri, M., Cadène, R., Bethune, L., Chalvidal, M., Serre, T., et al. A holistic approach to unifying automatic concept extraction and concept importance estimation. *arXiv preprint arXiv:2306.07304*, 2023.

Frosst, N. and Hinton, G. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.

Gautam, S., Boubekki, A., Hansen, S., Salahuddin, S., Jenssen, R., Höhne, M., and Kampffmeyer, M. Protovae: A trustworthy self-explainable prototypical variational model. *Advances in Neural Information Processing Systems*, 35:17940–17952, 2022.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

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

Jaakkola, T., Jordan, M., and Singh, S. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Jerzak, C. T., Johansson, F., and Daoud, A. Image-based treatment effect heterogeneity. *arXiv preprint arXiv:2206.06417*, 2022.

Jerzak, C. T., Johansson, F., and Daoud, A. Integrating earth observation data into causal inference: Challenges and opportunities. *ArXiv Preprint*, 2023a.

Jerzak, C. T., Johansson, F., and Daoud, A. Image-based treatment effect heterogeneity. *Proceedings of the Second Conference on Causal Learning and Reasoning (CLeaR), Proceedings of Machine Learning Research (PMLR)*, 213: 531–552, 2023b.

Kim, C. and Bastani, O. Learning interpretable models with causal guarantees. *arXiv preprint arXiv:1901.08576*, 2019.

Kiritchenko, S. and Mohammad, S. M. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

Kline, A. and Luo, Y. Psmpy: a package for retrospective cohort matching in python. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1354–1357. IEEE, 2022.

Law, M., Russo, A., and Broda, K. The ilasp system for inductive learning of answer set programs. *arXiv preprint arXiv:2005.00904*, 2020.

Liu, Q., Chen, Z., and Wong, W. H. Causalegm: a general causal inference framework by encoding generative modeling. *arXiv preprint arXiv:2212.05925*, 2022.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. 55 (13s), jul 2023. ISSN 0360-0300. doi: 10.1145/3583558. URL https://doi.org/10.1145/3583558.

Newman, D. Bag of Words. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C5ZG6P.

Nie, L., Ye, M., Nicolae, D., et al. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2020.

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

Pryzant, R., Card, D., Jurafsky, D., Veitch, V., and Sridhar, D. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4095–4109, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raghothaman, M., Mendelson, J., Zhao, D., Naik, M., and Scholz, B. Provenance-guided synthesis of datalog programs. In *Proceedings of the ACM Symposium on Principles of Programming Languages (POPL)*, 2020.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J. M., and Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5612–5619, 2020.

Seo, S., Arik, S., Yoon, J., Zhang, X., Sohn, K., and Pfister, T. Controlling neural networks with rule representations. *Advances in Neural Information Processing Systems*, 34: 11196–11207, 2021a.

Seo, S., Arik, S. Ö., Yoon, J., Zhang, X., Sohn, K., and Pfister, T. Controlling neural networks with rule representations. In *Neural Information Processing Systems*, 2021b. URL https://api.semanticscholar.org/CorpusID:235435676.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Shi, S., Xie, Y., Wang, Z., Ding, B., Li, Y., and Zhang, M. Explainable neural rule learning. In *Proceedings of the ACM Web Conference 2022*, pp. 3031–3041, 2022.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

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

Wang, J., Ren, Z., Han, B., Ye, J., and Zhang, C. Towards understanding cooperative multi-agent q-learning with value factorization. *Advances in Neural Information Processing Systems*, 34:29142–29155, 2021.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Xue, B., Said, A. S., Xu, Z., Liu, H., Shah, N., Yang, H., Payne, P., and Lu, C. Assisting clinical decisions for scarcely available treatment via disentangled latent representation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5360–5371, 2023.

Yang, F., Yang, Z., and Cohen, W. W. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30, 2017.

Yang, Y. and Song, L. Learn to explain efficiently via neural logic inductive learning. *arXiv preprint arXiv:1910.02481*, 2019.

Yoon, J., Jordon, J., and Van Der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International conference on learning representations*, 2018.

Zhang, Y.-F., Zhang, H., Lipton, Z. C., Li, L. E., and Xing, E. P. Exploring transformer backbones for heterogeneous treatment effect estimation. *arXiv preprint arXiv:2202.01336*, 2022.

A. Datasets

IHDP is a semi-synthetic dataset composed of the observations from 747 infants from the Infant Health and Development Program, which is used for the effect of home visits (treatment variable) by specialists on infants' cognitive scores (outcome) in the future.

TCGA. We obtain the covariates of TCGA from a real data set, the Cancer Genomic Atlas (Bica et al., 2020). We then follow the data generation process of (Zhang et al., 2022) to generate synthetic treatments, dosage values and outcomes.

IHDP-C is a variant of the IHDP dataset, where we modify the treatment variable to become continuous, and follow (Nie et al., 2020) to generate the synthetic treatment and outcome values.

News is composed of 3000 randomly sampled news items from the NY Times corpus (Newman, 2008). Bag-of-Word features are used for treatment effect estimation and we follow prior studies (Bica et al., 2020) to generate synthetic treatment and outcome values.

EEEC consists of 33738 English sentences. Each sentence in this dataset is produced by following a template such as "<Person> made me feel <emotional state word>" where <Person> and <emotional state word> are placeholders to be filled. To study the effect of race or gender on the mood state, placeholders such as <Person> are replaced with race-related or gender-related nouns (say an African-American name for <Person>) while the placeholder <emotional state word> is filled with one of the four mood states: *Anger*, *Sadness*, *Fear* and *Joy*. The replacement of those placeholders with specific nouns is guided by a pre-specified causal graph (Feder et al., 2021).

Uganda is composed of around 1.3K satellite images collected from around 300 different sites from Uganda. In addition to the image data, some tabular features are also collected such as age and ethnicity. However, as reported by (Jerzak et al., 2022), such tabular features often fail to cover important information such as the neighborhood-level features and geographical contexts, which, are critical factors for determining whether anti-poverty intervention for a specific area is needed.

Note that the generation of synthetic treatments and outcomes on IHDP-C, News and TCGA dataset relies on some hyper-parameters to specify the number of treatments or the range of dosage. For our experiments, we used the default hyper-parameters provided by (Zhang et al., 2022).

B. Additional related work

Integrate rules into neural models How to integrate logic rules into neural models has been extensively studied (Seo et al., 2021b;a). For instance, DeepCTRL (Seo et al., 2021b) has explored the use of *existing* rules to improve the training of deep neural networks; in contrast, DISCRET does not require existing rules; it effectively learns (i.e. synthesizes) rules from training data and can be incorporated into neural models as regularization.

Program synthesis Program synthesis concerns synthesizing human-readable programs out of data, which has been extensively studied in the past few decades. Initial solutions to this problem, e.g., ILASP (Law et al., 2020) and Prosynth (Raghothaman et al., 2020) leverage pure symbolic reasoning to search logic rules. Recently, people have started exploring neural-based solutions, such as NeuralLP (Yang et al., 2017) and NLIL (Yang & Song, 2019) for guiding the rule generation process. But note that these solutions are

C. Additional notes on baseline methods

TVAE and Ganite can only handle binary treatments without dose variables, which are thus not applicable to TCGA, IHDP-C, and News datasets. VCNet is not suitable for continuous treatment variables, and hence is not evaluated on IHDP-C and News datasets.

D. Additional Technical Details

D.1. Conventional Assumptions for Treatment Effect Estimation

Assumption 1. *(Strong Ignorability)* $Y(T=t) \perp T|X$. *In the binary treatment case,* $Y(0), Y(1) \perp T|X$.
Assumption 2. *(Positivity)* $0 < \pi(T|X) < 1, \forall X, \forall T$.

Assumption 3. *(Consistency) For the binary treatment setting,* $Y = TY(1) + (1-T)Y(0)$.

D.2. Encoding Rules

To encode a literal, $l_k = A\ op\ c$, we perform one-hot encoding on feature A and operator *op*, which are concatenated with the normalized version of c (i.e., all the values of A should be rescaled to $[0, 1]$) as the encoding for l_k. We then concatenate the encoding of all l_k to compose the encoding of $L_{1:K}$.

D.3. Generalizing to Disjunctive Rules

The above process of building a conjunctive rule can be viewed as generating *the most probable* conjunctive rules among all the possible combinations of A, op and c. This can be generalized to building a rule with multiple disjunctions, by generating the H most probable conjunctive rules instead, where H represents the number of disjunctions specified by users. Specifically, for the model Θ_1, we simply select the H most probable features from its model output while for the model Θ_2, we leverage beam search to choose the H most probable (A, c) pairs.

D.4. Generalizing to Categorical Outcome Variables

To generalize DISCRET to handle categorical outcome variables, by following (Feder et al., 2021), the treatment effect is defined by the difference between the probability distributions of all categorical variables. Additionally, to estimate outcomes within a subgroup of similar samples, we simply compute the frequency of each outcome as the estimation.

D.5. Generalizing to Other Categories of Treatment Variables

We first discuss general settings for various treatment variables and then discuss how to estimate the treatment effect for each of them.

The settings for all treatment variables that our methods can deal with:

1. Tabular data with a binary treatment variable T and no dose variables. In this setting, $T = 1$ represents treated unit while $T = 0$ represents untreated unit, and the ITE is defined as the difference of outcomes under the treatment and under the control, respectively (i.e., $\text{ITE}(x) = y_1(x) - y_0(x)$, where $y_1(x)$ and $y_0(x)$ represents the potential outcome with and without receiving treatment for a sample x). The average treatment effect, ATE, is the sample average of ITE across all samples (i.e., $\text{ATE} = \mathbb{E}[\text{ITE}]$).
2. Tabular data with a continuous treatment variable T. Following (Zhang et al., 2022), the average dose-response function is defined as the treatment effect, i.e., $\mathbb{E}[Y|X, do(T = t)]$.
3. Tabular data with a discrete treatment variable T with one additional continuous dose variable S. Following (Zhang et al., 2022), the average treatment effect is defined as the average dose-response function: $\mathbb{E}[Y|X, do(T = t, S = s)]$.

The treatment effect for each of the above settings is then estimated as follows:

1. With a binary treatment variable and no dose variable, we can estimate the ATE of $R_x(\mathcal{D})$ via arbitrary treatment effect estimation methods, such as the classical statistical matching algorithm (Kline & Luo, 2022), or state-of-the-art neural network models. In this paper, we adopt the K-Nearest Neighbor Matching by default for estimating the ATE of $R_x(\mathcal{D})$: $\text{ITE} = y_1(x) - y_0(x)$. We can also obtain the estimated outcome by averaging the outcome of samples from $R_x(\mathcal{D})$ with the same treatment as the sample x, i.e.:

$$\widehat{y}(t) = \frac{1}{\sum \mathbb{I}(t_i^* = t)} \sum \mathbb{I}(t_i^* = t) \cdot y_i^* \tag{4}$$

2. With a continuous treatment variable T but without dose variables, then as per 2, the ITE is represented by the outcome conditioned on the observed treatment. One straightforward way to estimate it is to employ the average outcome of samples within $R_x(\mathcal{D})$ that receive similar treatments to x, which is also the estimated outcome for this sample:

$$\hat{y} = \frac{\sum \mathbb{I}[(x_i^*, t_i^*, y_i^*) \in \text{top}_k(R_x(\mathcal{D}))] \cdot y_i^*}{\sum \mathbb{I}[(x_i^*, t_i^*, y_i^*) \in \text{top}_k(R_x(\mathcal{D}))]}, \tag{5}$$

in which $\text{top}_k(R_x(\mathcal{D}))$ is constructed by finding the top-k samples from $R_x(\mathcal{D})$ with the most similar treatments to x. But again, any existing treatment effect estimation methods for continuous treatment variables from the literature are applicable to estimate $\widehat{\text{ITE}}_x$.

3. With a discrete treatment variable T and one associated continuous dose variable S, ITE is estimated in a similar way to equation 5. Specifically, we estimate ATE over the subgroup of similar samples with the following formula:

$$\hat{y} = \frac{\sum \mathbb{I}[(x_i^*, t_i^*, s_i^*, y_i^*) \in \text{top}_k(R_x(\mathcal{D}))] \cdot y_i^*}{\sum \mathbb{I}[(x_i^*, t_i^*, s_i^*, y_i^*) \in \text{top}_k(R_x(\mathcal{D}))]}. \tag{6}$$

In the above formula, $\text{top}_k(R_x(\mathcal{D}))$ is constructed by first selecting the samples with the same treatment as the sample x and then only retaining the k samples with the most similar dose values to x.

D.6. Deep Q-learning and Training Algorithm

To facilitate Q-learning, we estimate the Q value with the output logits of the models given a state $(x, L_{1:k-1})$ and an action l_k, which is denoted by $Q(l_k, (x, L_{1:k-1}))$. Note that l_k is generated collaboratively by using two models, Θ_1 and Θ_2, we therefore need to collect two sub-Q values from these two models, and then aggregate (say average) them as the overall Q value, which follows prior multi-agent Q-learning literature (Wang et al., 2021). In the end, by following the classical DQL framework, we optimize the following objective function adapted from the Bellman equation (Dixit, 1990):

$$L_\Theta = \mathbb{E}[Q(l_k, (x, L_{1:k-1})) - (\gamma \cdot \max_{l_{k+1}} Q(l_{k+1}, (x, L_{1:k})) + r_k)]^2, \tag{7}$$

which is estimated over a sampled mini-batch of cached experience taking the form of $< (x, L_{1:k-1}), l_k, r_k, (x, L_{1:k}) >$ during the experience replay process. The training algorithm for rule learning is outlined in Algorithm 1 below.

Algorithm 1 The overview of Deep Q-Learning (DQL) algorithm for rule learning in DISCRET

Input: target model update: t, gamma: γ, batch size: b, target model parameters: Θ^{target}, policy model parameters: Θ^{policy}, experience replay cache: $cache =< (x, L_{1:k-1}), l_k, r_k, (x, L_{1:k}) >$ where x is a covariate, $L_{1:k-1}$ is the set of literals at step $k-1$, l_k is the literal synthesized at step k, r_k is the reward at step k, and $L_{1:k}$ is $L_{1:k-1} \cup l_k$
Output: $None$

Initialize $\boldsymbol{w}^{pred}$ and $\boldsymbol{w}^{target}$ of length b
Construct $batch$ by sampling b entries from $cache$
for $i, < (x^i, L^i_{1:k-1}), l^i_k, r^i_k, (x^i, L^i_{1:k}) >$ in *Enumerate*$(batch)$ **do**
    Use Θ_0^{policy} and a deterministic function to encode both x^i and $L^i_{1:k-1}$, respectively, to get E^i_{k-1};
    Forward pass E^i_{k-1} through Θ_1^{policy} and select the index of the feature from l^i_k to obtain Q^i_f;
    Append a one-hot encoding of the feature from l^i_k to E^i_{k-1} to get $E^i_{partial}$;
    forward pass $E^i_{partial}$ through Θ_2^{policy} and select the index of the constant from l^i_k to get Q^i_c;
    Obtain Q^i_{k-1} by averaging Q^i_f and Q^i_c;
    Obtain Q^i_k by forward passing x^i and $L^i_{1:k}$ through Θ^{target} and averaging the maximum Q values from Θ_1^{target} and Θ_2^{target};
    $\boldsymbol{w}^{pred}_i \leftarrow Q^i_{k-1}$; $\boldsymbol{w}^{target}_i \leftarrow \gamma Q^i_k * + r^i_k$;
end for
Backpropogate and update Θ^{policy} using loss *MSE*($\boldsymbol{w}^{pred}$,$\boldsymbol{w}^{target}$)
if *len*$(cache)\%t == 0$ **then**
    $\Theta^{target} \leftarrow \Theta^{policy}$
end if

D.7. Proof of Theorem 3.1

We first state some additional preliminary notations and settings for Q-learning. We denote the Markov decision process (MDP) as a tuple $(\mathcal{S}_k, \mathcal{L}_k, P_k, r_k)$ where

- $\mathcal{S}_k$ is the state space with a state $(x, L_{1:k})$;
- $\mathcal{L}_k$ is the action space with an action l_k;
- P_k represents the transition probability;
- r_k represents the reward function.

Theorem 3.1 is a direct implication of Lemma D.1 below.

Lemma D.1. *Suppose we have input data $\{(x_i, t_i, s_i, y_i)\}_{i=1}^N$ where $x_i \in \mathbb{R}^m$ and discrete, $t_i \in \mathbb{R}, s_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$, then the $\hat{y}$ obtained from DISCRET converges to zero generalization error with probability 1 (i.e. $(y - \hat{y})^2 \to 0$ w.p. 1) for any fixed $K \leq m$ over the dataset with all discrete features under the data generating process $y = f(\mathcal{X}_K) + c \cdot t + \epsilon$, where $\mathcal{X}_K \subseteq \{X_1, X_2, \cdots, X_m\}, c \in \mathbb{R}, t$ is the treatment assignment, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.*

To prove Lemma D.1, we need to use results from D.2.

Theorem D.2. *Given a finte Markov decision process $(\mathcal{S}_k, \mathcal{L}_k, P_k, r_k)$, given by the update rule*

$$Q(l_k, (x, L_{1:k})) = Q(l_{k-1}, (x, L_{1:k-1})) + \alpha_{k-1}(l_{k-1}, (x, L_{1:k-1}))\times \left(r_{k-1} + \gamma \max_{(x^*, L^*_{1:k-1}) \in \mathcal{S}_k \times \mathcal{L}_k} Q(l_{k-1}, (x^*, L^*_{1:k-1})) - Q(l_{k-1}, (x, L_{1:k-1})) \right) \tag{8}$$

converges with probability 1 to the optimal Q-function as long as

$$\sum_k \alpha_k(l_k, (x, L_{1:k-1})) = \infty, \quad \sum_k \alpha_k^2(l_k, (x, L_{1:k-1})) < \infty$$

for all $(l_k, (x, L_{1:k-1})) \in \mathcal{S}_k \times \mathcal{L}_k$.

Proof. We start rewriting equation (8) as

$$Q(l_k, (x, L_{1:k})) = (1 - \alpha_{k-1}(l_{k-1}, (x, L_{1:k-1}))) Q(l_{k-1}, (x, L_{1:k-1})) + \alpha_{k-1}(l_{k-1}, (x, L_{1:k-1}))\times \left(r_{k-1} + \gamma \max_{(x^*, L^*_{1:k-1}) \in \mathcal{S}_k \times \mathcal{L}_k} Q(l_{k-1}, (x^*, L^*_{1:k-1})) \right)$$

Denote the optimal Q function be $Q^*(l_k, (x, L_{1:k}))$, subtracting equation above from both sides the quantity $Q^*(l_k, (x, L_{1:k}))$ and letting

$$\Delta_k(l_k, (x, L_{1:k})) = Q(l_k, (x, L_{1:k})) - Q^*(l_k, (x, L_{1:k}))$$

yields

$$\Delta_k(l_k, (x, L_{1:k})) = (1 - \alpha_{k-1}(l_{k-1}, (x, L_{1:k-1}))) \Delta_k(l_k, (x, L_{1:k})) + \alpha_{k-1}(l_{k-1}, (x, L_{1:k-1})) \left(r_k + \gamma \max_{(x^*, L^*_{1:k-1}) \in \mathcal{S}_k \times \mathcal{L}_k} Q(l_{k-1}, (x^*, L^*_{1:k-1})) - Q^*(l_k, (x, L_{1:k})) \right).$$

If we write

$$F_k(l_k, (x, L_{1:k})) = r_k((x, L_{1:k}), l_k, \mathcal{S}(x, L_{1:k})) + \gamma \max_{(x^*, L^*_{1:k-1}) \in \mathcal{S}_k \times \mathcal{L}_k} Q(l_{k-1}, (x^*, L^*_{1:k-1})) - Q^*(l_k, (x, L_{1:k}))$$

where $\mathcal{S}(x, L_{1:k})$ is a random sample state obtained from the Markov chain $(\mathcal{S}_k, P_k)$, we have

$$\begin{aligned} &\mathbb{E}[F_k(l_k, (x, L_{1:k}))|\mathcal{F}_k] \\ &= \sum_{b \in \mathcal{S}_k} P_k((l_k, (x, L_{1:k}), b)[r_k((l_k, (x, L_{1:k}), l_k) + \gamma \max_{(x^*, L^*_{1:k-1}) \in \mathcal{S}_k \times \mathcal{L}_k} Q(l_{k-1}, (x^*, L^*_{1:k-1})) - Q^*(l_k, (x, L_{1:k}))] \\ &= (\mathbf{H}Q)(x, L_{1:k}) - Q^*(l_k, (x, L_{1:k})). \end{aligned}$$

Using the fact that $Q^* = (\mathbf{H}Q)(x, L_{1:k})$,

$$\mathbb{E}[F_k(l_k, (x, L_{1:k}))|\mathcal{F}_k] = (\mathbf{H}Q)(x, L_{1:k} - (\mathbf{H}Q^*)(x, L_{1:k} \leq \gamma \|Q - Q^*\| = \gamma \|\Delta_k\|_\infty.$$

We could also verify that

$$Var[F_k(l_k, (x, L_{1:k}))|\mathcal{F}_k] \leq C(1 + \|\Delta_k\|_W^2)$$

for some constant C. Then by the theorem below, Δ_k converges to zero with probability 1. Hence, Q converges to Q^* with probability 1. □

Theorem D.3 (Jaakkola et al. (1993)). *The random process* $\{\Delta_t\}$ *taking values in* $\mathbb{R}^n$ *and defined as*

$$\Delta_{t+1}(x) = (1 - \alpha_t(x))\Delta_t(x) + \alpha_t(x)F_t(x)$$

converges to zero with probability 1 under the following assumptions:

- $0 \leq \alpha_t \leq 1, \sum_t \alpha_t(x) = \infty$ *and* $\sum_t \alpha_t^2(x) < \infty$*;*
- $\|\mathbb{E}[F_t(x)|\mathcal{F}_t]\|_W \leq \gamma\|\Delta_t\|_W$*, with* $\gamma < 1$*;*
- $Var(\mathcal{F}_t(x)|\mathcal{F}_t) \leq C(1 + \|\Delta_t\|_W^2)$*, for* $C > 0$.

Proof. See Jaakkola et al. (1993) for the proof. □

Proof of Lemma D.1. Using Theorem D.2, we can see that Q obtained from DISCRET converges to optimal Q^*. As a result, $haty$ obtained from DISCRET converges to optimal y^*. We left to prove that y^* leads to a zero mean square error (i.e., $\|y - y^*\|_2^2$). We can prove this using the fact that all features are discrete. Since all features are discrete and the optimal feature being selected in each step leads to a zero mean square error and other features lead to non-zero mean square error, it turns out that y^* obtained from DISCRET leads to a zero mean square error. □

D.8. Additional Reward Function Optimizations

We further present some strategies to optimize the design of the cumulative reward function defined in equation 3, which includes incorporating estimated propensity scores into this formula and automatically fine-tuning its hyper-parameters.

Regularization by estimating propensity scores Similar to prior studies on ITE estimation (Shi et al., 2019; Zhang et al., 2022), we regularize the reward function r_k by integrating the estimated propensity score, $\widehat{\pi}(T = t|X = x)$. Specifically, for discrete treatment variables, we re-weight equation 9 with the propensity score as a regularized reward function, i.e.:

$$\begin{aligned} V_{1:K}^{reg} = [e^{-\alpha(y-\widehat{y}_{1:K})^2} + \beta \cdot \widehat{\pi_{1:k}}(T = t|X = x)] \\ \cdot \mathbb{I}(L_{1:K}(\mathcal{D}) \text{ is non-empty}), \end{aligned} \tag{9}$$

Automatic hyper-parameter fine-tuning We further studied how to automatically tune the hyper-parameter α and β in equation 9. For α, at each training epoch, we identify the training sample producing the median of $(y - \widehat{y}_{1:K})^2$ among the whole training set and then ensure that for this sample, equation 3 is 0.5 through adjusting α. This can guarantee that for those training samples with the smallest or largest outcome errors, equation 3 approaches 1 or 0 respectively.

We also designed an annealing strategy to dynamically adjust β by setting it as 1 during the initial training phase to focus more on treatment predictions, and switching it to 0 so that reducing outcome error is prioritized in the subsequent training phase.

E. Addendum on Performance Metrics

E.1. Faithfulness Metrics

We evaluate the faithfulness of explanations with two metrics, i.e., consistency and sufficiency from (Dasgupta et al., 2022). For a single sample x with local explanation e_x, the consistency is defined as the probability of getting the same model predictions for the set of samples producing the same explanations (denoted by C_x) as x while the sufficiency is defined in the same way, except that it depends on the set of samples satisfying e_x (denoted by S_x) rather than generating explanation e_x. These two metrics could be formalized with the following formulas:

$$\begin{aligned} \text{Consistency}(x) &= Pr_{x' \in_\mu C_x}(\hat{y}(x) == \hat{y}(x')) \\ \text{Sufficiency}(x) &= Pr_{x' \in_\mu S_x}(\hat{y}(x) == \hat{y}(x')) \end{aligned}$$

in which μ represents the probability distribution of C_x and S_x. To evaluate explanations with these two metrics, (Dasgupta et al., 2022) proposed an unbiased estimator for Consistency(x) and Sufficiency(x), i.e.,:

$$\widehat{\text{Consistency}}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(C_x > 1) \cdot \frac{C_{x,\hat{y}(x)} - 1}{C_x - 1}$$

$$\widehat{\text{Sufficiency}}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(S_x > 1) \cdot \frac{S_{x,\hat{y}(x)} - 1}{S_x - 1}$$

in which $C_{x,\hat{y}(x)}$ represents the set of samples sharing the same explanation and the same model predictions as the sample x while $S_{x,\hat{y}(x)}$ represents the set of samples that satisfy the explanation produced by x and share the same explanation as x. As the above formula suggests, both the consistency and sufficiency scores vary between 0 and 1.

But note that for typical ITE settings, the model output is continuous rather than discrete numbers. Therefore, we discretize the range of model output into evenly distributed buckets, and the model outputs that fall into the same buckets are regarded as having the same model predictions. As (Dasgupta et al., 2022) mentions, the sufficiency metric is a reasonable metric for evaluating rule-based explanations since it requires retrieving other samples with explanations. So we only report sufficiency metrics for methods that can produce rule-based explanations in Table 6.

E.2. Additional Notes for the EEEC Dataset

Note that for EEEC dataset, ϵ_{ATE} is used for performance evaluation but the ground-truth ITE is not observed, which is approximated by the difference of the predicted outcomes between factual samples and its ground-truth counterfactual alternative (Feder et al., 2021).

E.3. $AMSE$ for Continuous Treatment Variable or Dose Variable

To evaluate the performance of settings with continuous treatment variables or continuous dose variables, we follow (Zhang et al., 2022) to leverage $AMSE$ as the evaluation metrics, which is formalized as follows:

$$AMSE = \begin{cases} \frac{1}{N}\sum_{i=1}^{N}\int_t [\hat{y}(x_i,t) - y(x_i,t)]\pi(t)dt & \text{continuous treatment variable} \\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\int_s [\hat{y}(x_i,t) - y(x_i,t)]\pi(t)dt & \text{continuous dose variable,} \end{cases}$$

in which we compute the difference between the estimated outcome $\hat{y}$ and the observed outcome y conditioned on every treatment t, and average this over the entire treatment space and all samples for evaluations. Due to the large space of exploring all possible continuous treatments t or continuous dose values s, we collect sampled treatment or sampled dose rather than enumerate all s and t for the evaluations of $AMSE$.

F. Additional Experimental Results

F.1. Performance of Self-interpretable Models with Varying Complexity

On evaluating the performance of self-interpretable models when trained with a high depth, i.e number of conjunctive clauses ($K = 100$, as opposed to low-depth $K = 6$), we see that DISCRET ($K = 6$) outperforms these models despite having lower depth, and thus better interpretability.

F.2. Ablation Studies

We further perform ablation studies to explore how different components of DISCRET such as the database and featurziation process (for NLP and image data), affect the ITE estimation performance. In what follows, we analyze the effect of the size of the database, different featurization steps, and different components of the reward function.

Ablating the reward functions for DISCRET. Recall that in Section 3.3, the reward function used for the training phase could be enhanced by adding propensity scores as one regularization and automatically tuning the hyper-parameters, α and β. We removed these two components from the reward function one after the other to investigate their effect on the ITE estimation performance. We perform this experiment on Uganda dataset and report the results in Table 4. As this table

Modality →			**Tabular**				
Dataset →			IHDP		TCGA		IHDP-C
Method ↓	Trees	Depth	ϵ_{ATE} (In-sample)	ϵ_{ATE} (Out-of-sample)	ϵ_{ATE} (In-sample)	ϵ_{ATE} (Out-of-sample)	AMSE
Decision Tree	-	6	0.693±0.028	0.613±0.045	0.200±0.012	0.202±0.012	21.773±0.190
	-	100	0.638±0.031	0.549±0.052	0.441±0.004	0.445±0.004	23.382±0.342
Random Forest	1	6	0.801±0.039	0.666±0.055	19.214±0.163	19.195±0.163	21.576±0.185
	1	100	0.734±0.041	0.653±0.056	0.536±0.011	0.538±0.012	33.285±0.940
	10	100	0.684±0.033	0.676±0.034	0.536±0.011	0.538±0.012	38.299±0.841
NAM	-	-	0.260±0.031	0.250±0.032	-	-	24.706±0.756
ENRL	1	6	4.104±1.060	3.759±0.087	10.938±2.019	10.942±2.019	24.720±0.985
	1	100	4.094±0.032	4.099±0.107	10.938±2.019	10.942±2.019	24.900 ± 0.470
Causal Forest	1	6	0.144±0.019	0.275±0.035	-	-	-
	1	100	0.151±0.019	0.278±0.033	-	-	-
	100	max	0.124±0.015	0.230±0.031	-	-	-
BART	1	-	1.335±0.159	1.132±0.125	230.74±0.312	236.81±0.531	12.063±0.410
	N	-	0.232±0.039	0.284±0.036	-	-	4.323±0.342
DISCRET (ours)	-	6	<u>0.089±0.040</u>	<u>0.150±0.034</u>	<u>0.076±0.019</u>	<u>0.098±0.007</u>	<u>0.801±0.165</u>
TransTEE + DISCRET (ours)*	-	-	**0.082±0.009**	**0.120±0.014**	**0.058±0.010**	**0.055±0.009**	**0.102±0.007**

Table 3. ITE estimation errors (lower is better) at varying complexities for self-interpretable models. We **bold** the smallest estimation error for each dataset, and <u>underline</u> the second smallest one. Results in the first row for each method are duplicated from Table 2. For BART, we set $N = 200$ for IHDP, and $N = 10$ for TCGA and IHDP-C due to large feature number of features in the latter. We show that DISCRET outperforms self-interpretable models and has simpler rules regardless of the model complexity used. Asterisk (*) indicates model is not self-interpretable.

suggests, throwing away those two components from the reward function incurs higher outcome errors, thus justifying the necessity of including them for more accurate ITE estimation.

	Outcome error
DISCRET	**1.662±0.136**
DISCRET without propensity score	1.701±0.161
DISCRET without propensity score or auto-finetuning	1.742±0.151

Table 4. Ablation studies on the reward function in DISCRET

Ablating the database size. Since DISCRET estimates ITE through rule evaluations over a database. The size of this database can thus influence the estimation accuracy. We therefore vary the size of the database, i.e., the number of training samples, for IHDP dataset, and compare DISCRET against TransTEE with varied database size. The results are included in Figure 5, which shows that DISCRET is almost always on par with TransTEE on the ITE estimation errors. This thus indicates that with the influence of the database size (i.e, the training set size for TransTEE) on the relative performance between DISCRET and Neural Network models is insignificant. It is also worth noting that when the database size is reduced below certain level, e.g., smaller than 200 in Figure 5, DISCRET can even outperform TransTEE. This thus implies that DISCRET could be more data-efficient than the state-of-the-art Neural Network models for ITE estimations, which is left for future work.

F.3. Training Cost of DISCRET

We further plot Figure 6 to visually keep track of how the ATE errors on test set are evolved throughout the training process. As this figure suggests although the best test performance occurs after 200 epochs (ATE error is around 0.12). However, the performance in the first few epochs is already near-optimal (ATE error is around 0.14). Therefore, despite the slow convergence in typical reinforcement learning training processes, our methods obtain reasonable treatment effect estimation performance without taking too many epochs.

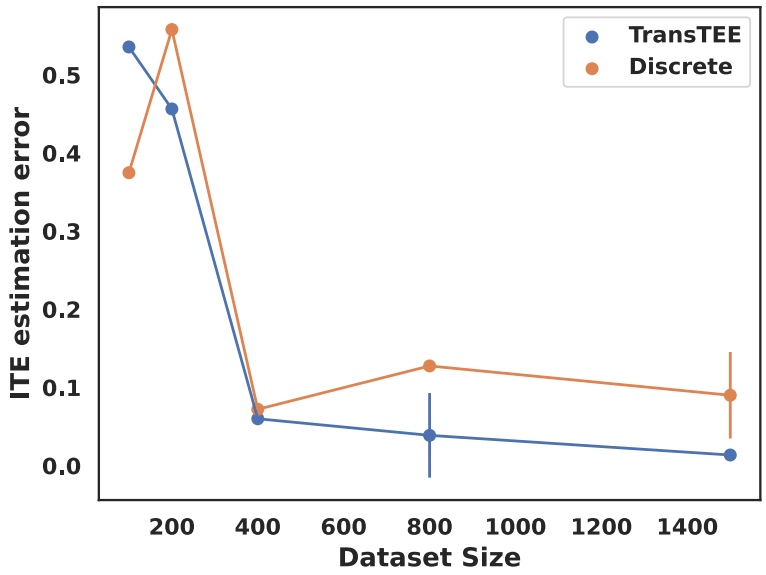

Figure 5. Ablation study on the effect of database size on the performance of DISCRET

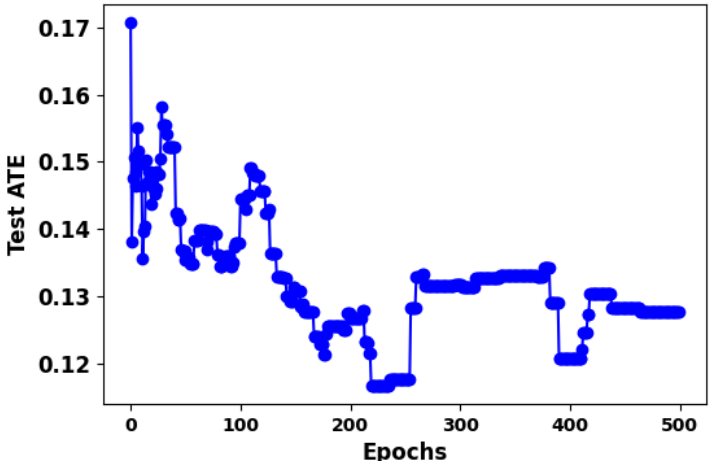

Figure 6. The curve of ATE errors on test split of IHDP by DISCRET

F.4. Consistency and Sufficiency Scores

We provide the full results of the consistency and sufficiency scores below.

F.5. Results for News dataset

Table 7 shows the results for the News dataset.

F.6. Consistent Ground-truth Outcomes in the Uganda Dataset

We observe that in Uganda dataset, the ground-truth outcome values are not evenly distributed, which is visually presented in Figure 7. As this figure suggests, the outcome value of most samples is -0.8816 while other outcome values rarely occur. This thus suggests that our method is preferable in such datasets due to its consistent predictions across samples, which can

	IHDP	TCGA	IHDP-C	News	EEEC (Gender)	EEEC (Race)	Uganda
Model distillation	0.243±0.126	0.562±0.026	0.127±0.008	0.816±0.032	0.004±0.001	0.013±0.002	0.198±0.008
Lore	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.001
Anchor	0.084±0.083	0.001±0.000	0.293±0.022	0.000±0.000	0.000±0.000	0.000±0.000	0.066±0.015
Lime	0.182±0.129	0.000±0.000	0.001±0.001	0.000±0.000	0.000±0.000	0.000±0.001	0.000±0.000
Shapley	0.009±0.017	0.005±0.002	0.046±0.027	0.031±0.035	0.034±0.003	0.027± 0.000	0.412±0.195
NAM	0.343±0.065	0.120±0.002	0.045±0.006	0.493±0.110	-	-	0.082±0.018
ENRL	0.134±0.002	0.231±0.043	0.053±0.002	0.002±0.000	-	-	0.102±0.032
DISCRET	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00

Table 5. Explanation consistency scores across datasets

	IHDP	TCGA	IHDP-C	News	EEEC (Gender)	EEEC (Race)	Uganda
Model distillation	0.243±0.126	0.529±0.001	0.029±0.003	**0.712±0.032**	0.004±0.001	0.013±0.002	0.198±0.008
Lore	0.320±0.084	0.034±0.013	0.030±0.009	0.142±0.012	0.002±0.001	0.002±0.001	**0.265±0.008**
Anchor	0.084±0.083	0.125±0.002	0.332±0.016	0.391±0.040	0.002±0.001	0.011±0.006	0.221±0.007
ENRL	0.452±0.012	0.512±0.005	0.032±0.018	0.053±0.020	-	-	0.004±0.002
DISCRET	**0.562±0.056**	**0.9999±0.000**	**0.588±0.019**	0.697±0.017	**0.926±0.067**	**0.996±0.001**	0.104±0.011

Table 6. Explanation sufficiency scores across datasets (larger score indicates better sufficiency)

	News
	AMSE
Decision Tree	0.428±0.051
Random Forest	0.452±0.048
NAM	0.653±0.026
ENRL	0.638±0.019
Causal Forest	0.829±0.042
BART	0.619±0.040
DISCRET(ours)	0.385±0.083
Dragonnet	-
TVAE	-
TARNet	0.073±0.020
Ganite	-
DRNet	0.065±0.021
VCNet	-
TransTEE	0.063±0.005
TransTEE + NN	0.383±0.041
DISCRET+ TransTEE (ours)	**0.043±0.005**

Table 7. ITE estimation errors for the News dataset

explain the performance gains of DISCRET over baseline methods.

G. Additional Qualitative Analysis

As shown in Figure 4, DISCRET generates one rule for one example image from Uganda dataset, which is defined on two concepts, i.e., one type of patches mainly containing reddish pink pixels that represent "soil moisture content" and the other type of patches mainly comprised of brown pixels indicating little soil. This rule thus represents the images from one type of location where there is plenty of soil moisture content that is suitable for agricultural development. Therefore, after the government grants are distributed in such areas, a more significant treatment effect is observed, i.e., 0.65. This is an indicator of significantly increasing working hours on the skilled jobs by the laborers in those areas. This is consistent to the conclusions from (Jerzak et al., 2023b;a) which states that government grant support is more useful for areas with more soil moisture content.

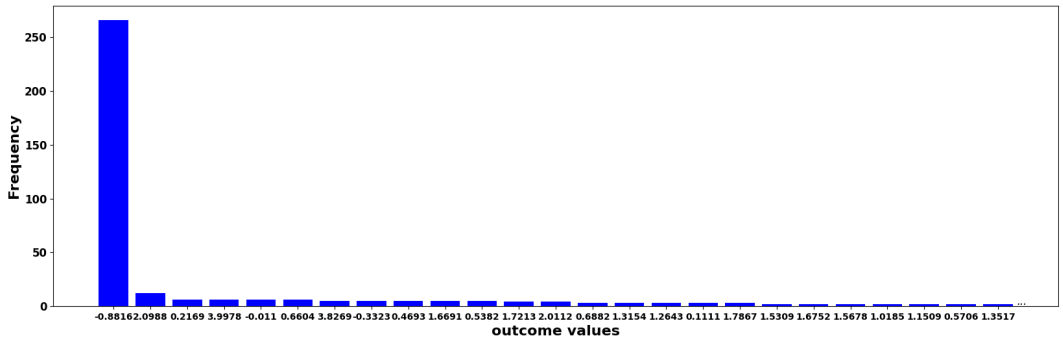

Figure 7. Frequency of the outcome values on Uganda dataset

H. Feature Extraction from Image Data

To extract concepts from images of Uganda dataset, we segment each image as multiple superpixels (Achanta et al., 2012), embed those superpixels with pretrained clip models (Radford et al., 2021), and then perform K-means on these embeddings. Each of the resulting cluster centroids is regarded as one concept and we count the occurrence of each concept as one feature for an image. Specifically, we extract 20 concepts from the images of Uganda dataset, which are visually presented in Figure 8.

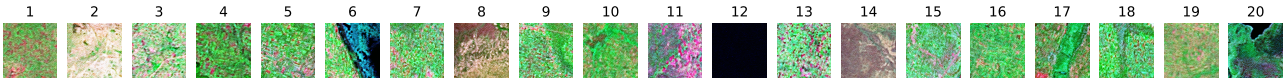

Figure 8. Extracted concepts from Uganda dataset

Various patterns of image patches are captured by Figure 8. For example, patch 12 is almost all black, which represents the areas with water, say, river areas or lake areas. Also, as mentioned in Section **??**, patch 11 with reddish pink pixels represents "soil moisture content", which is an important factor for determining whether to take interventions in the anti-poverty program conducted in Uganda.