A Additional technical details

A.1 Additional details on 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 $L_{1:K}$.

A.2 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.

A.3 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. Also to estimate outcomes within a subgroup of similar samples, we simply compute the frequency of each outcome as the estimation.

A.4 Estimating ITE in the settings with continuous dose variables

In the presence of dose variables along with treatment variables, 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.

A.5 Further details on Deep Q learning

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 pseudo-code for the training phase is sketched in Algorithm 2 in Appendix A.6. The training algorithm for rule learning is outlined in Algorithm 2.

A.6 The overview of the training algorithm

B Additional related work

Model interpretability Model interpretability remains a significant concern for many high-stake areas, such as health care and finance. 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

Algorithm 2 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 $\Theta_0^{policy}$ 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 $\Theta_1^{policy}$ 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 $\Theta_2^{policy}$ 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 $\Theta^{target}$ and averaging the maximum $Q$ values from $\Theta_1^{target}$ and $\Theta_2^{target}$;
    $\boldsymbol{w}_i^{pred} \leftarrow Q^i_{k-1}$; $\boldsymbol{w}_i^{target} \leftarrow \gamma Q^i_k * + r^i_k$;
end for
Backpropogate and update $\Theta^{policy}$ using loss MSE($\boldsymbol{w}^{pred}$,$\boldsymbol{w}^{target}$)
if $len(cache) \% t == 0$ then
    $\Theta^{target} \leftarrow \Theta^{policy}$
end if
```

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 always blamed for their potential lack of faithfulness (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. Other alternatives include learning feature importance and using that to determine the model output (e.g., Neural Additive Model (Agarwal et al., 2021)).

C Detailed descriptions of 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.

IHDP-C : is built on top of IHDP dataset, except that the treatment variable becomes a continuous variable and we follow (Nie et al., 2020) to generate synthetic treatment and outcome values.

News : is composed of 3000 randomly sampled news items from the NY Times corpus (Newman, 2008) and their 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.

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

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

	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 2: Ablation studies on the reward function in DISCRET

(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. But it turns out that such tabular features fail to cover important information such as the neighborhood-level features and geographical context (Jerzak et al., 2022), which, however, are very 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. We used the default hyper-parameters provided by (Zhang et al., 2022).

D Ablation studies

We further perform some ablation studies to explore how different components in 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 2. As this table 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.

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.

E Feature extractions on 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,

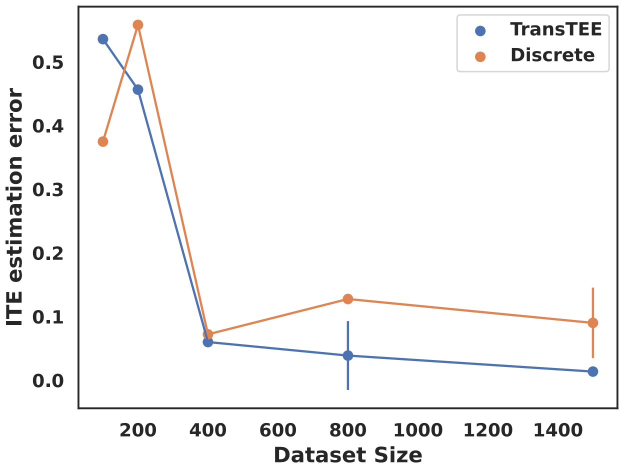

Figure 5: Ablation study on the effect of database size on the performance of DISCRET

we extract 20 concepts from the images of Uganda dataset, which are visually presented in Figure 6.

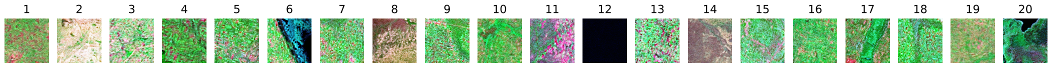

Figure 6: Extracted concepts from Uganda dataset

Various patterns of image patches are captured by Figure 6. 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 4.4, 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.

F ADDITIONAL DETAILS ON PERFORMANCE METRICS

F.1 ADDITIONAL NOTES FOR 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).

F.2 $AMSE$ 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.3 Faithfulness metrics for evaluating explanations

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:

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

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 7.

G Additional experimental results

	IHDP		TCGA		IHDP-C	News
	In-sample	Out-of-sample	In-sample	Out-of-sample	AMSE	AMSE
LR	3.366±2.189	2.497±1.814	31.737±0.001	57.541±0.001	36.640±16.455	NaN
DT	0.345±0.273	0.530±0.399	0.200±0.012	0.202±0.012	22.136±1.741	0.428±0.051
RF	0.739±0.284	0.737±0.383	0.263±0.057	0.264±0.058	21.348±1.222	0.452±0.048
NAM	0.225±0.221	0.519±0.512	4.201±0.232	4.211±0.152	24.706±0.756	0.653±0.026
ENRL	4.160±1.060	4.439±1.587	10.938±2.019	10.942±2.019	24.720±0.985	0.638±0.019
Dragonnet	0.177±0.139	0.219±0.143	-	-	-	-
TARNet	0.186±0.130	0.408±0.418	1.421 ± 0.078	1.421±0.078	12.967±1.781	0.079±0.021
Ganite	1.127±0.481	1.144±0.352	-	-	-	-
DRNet	0.188±0.132	0.407±0.422	1.374±0.086	1.374±0.085	11.071±0.994	0.080 ±0.025
VCNet	4.205±0.569	4.434±0.851	0.292±0.074	0.292 ± 0.074	-	-
TransTEE	**0.128±0.103**	**0.203±0.130**	**0.055±0.014**	**0.056 ± 0.013**	**0.488±0.288**	**0.074 ±0.037**
DISCRET	0.274±0.253	0.344±0.303	0.076±0.019	0.077±0.020	0.801±0.165	0.385±0.083

Table 3: ITE estimation errors in Tabular setting. We bold the smallest ITE estimation errors for each dataset and underline the smallest ITE estimation errors among the self-interpretable methods. DISCRET outperforms existing self-interpretable methods on 5 out of the 6 benchmarks.

H Additional experimental analysis

Likewise, 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

	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 4: Evaluations of the explanation consistency

	EEEC (Gender)	EEEC (Race)
ATE$_{GT}$	0.086±0.004	0.014±0.002
LR	0.000±0.000	0.000±0.000
DT	0.000±0.000	0.000±0.000
RF	0.544±0.559	0.539±0.557
NAM	0.139±0.027	0.166±0.025
Dragonnet	0.0263±0.0062	0.0253±0.0015
TARNet	0.0067±0.0009	0.005±0.002
Ganite	2.237±0.000	2.006±0.000
DRNet	0.0088±0.0002	0.006±0.002
VCNet	0.0085±0.0030	0.003±0.0005
TransTEE	0.013±0.005	0.0174±0.0006
DISCRET	0.011±0.002	0.0136±0.001

Table 5: Treatment effect estimation performance on NLP settings

	Uganda
LR	1.796±0.021
DT	1,796±0.021
RF	1.820±0.013
NAM	1.710±0.098
ProtoVAE	1.688±0.165
ENRL	1.800±0.143
Dragonnet	1.709±0.127
TARNet	1.743±0.135
Ganite	1.766±0.024
DRNet	1.748±0.127
VCNet	1.890±0.110
TransTEE	1.707±0.158
DISCRET	**1.662±0.136**

Table 6: Outcome errors on Image settings

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.

	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 7: The sufficiency score evaluation. Larger score indicates better sufficiency