CARD: Classification and Regression Diffusion Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Learning the distribution of a continuous or categorical response variable $\mathbf{y}$ given its covariates $\mathbf{x}$ is a fundamental problem in statistics and machine learning. Deep neural network-based supervised learning algorithms have made great progress in predicting the mean of $\mathbf{y}$ given $\mathbf{x}$, but they are often criticized for their ability to accurately capture the uncertainty of their predictions. In this paper, we introduce classification and regression diffusion (CARD) models, which combine a denoising diffusion-based conditional generative model and a pre-trained conditional mean estimator, to accurately predict the distribution of $\mathbf{y}$ given $\mathbf{x}$. We demonstrate the outstanding ability of CARD in conditional distribution prediction with both toy examples and real world datasets, the experimental results on which show that CARD in general outperforms state-of-the-art methods, including Bayesian neural network based one, designed for uncertainty estimation, especially when the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is multi-modal.

1 Introduction

A fundamental problem in statistics and machine learning is to predict the response variable $\mathbf{y}$ given a set of covariates $\mathbf{x}$. Generally speaking, $\mathbf{y}$ is a continuous variable for regression analysis and a categorical variable for classification. Denote $f(\mathbf{x}) \in \mathbb{R}^C$ as a deterministic function that transforms $\mathbf{x}$ into a C dimensional output. Denote $f_c(\mathbf{x})$ as the c-th dimension of $f(\mathbf{x})$. Existing methods typically assume an additive noise model: for regression analysis with $\mathbf{y} \in \mathbb{R}^C$, one often assumes $\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, while for classification with $y \in \{1, \ldots, C\}$, one often assumes $y = \arg\max\big(f_1(\mathbf{x}) + \epsilon_1, \ldots, f_C(\mathbf{x}) + \epsilon_C\big)$, where $\epsilon_c \overset{iid}{\sim} \mathrm{EV}_1(0, 1)$, a standard type-1 extreme value distribution. Thus we have the expected value of $\mathbf{y}$ given $\mathbf{x}$ as $\mathbb{E}[\mathbf{y} \,|\, \mathbf{x}] = f(\mathbf{x})$ in regression and $P(y = c \,|\, \mathbf{x}) = \mathbb{E}[y = c \,|\, \mathbf{x}] = \mathrm{softmax}_c(f(\mathbf{x})) = \frac{\exp(f_c(\mathbf{x}))}{\sum_{c'=1}^{C} \exp(f_{c'}(\mathbf{x}))}$ in classification.

These additive-noise models are primarily focusing on accurately estimating the conditional mean $\mathbb{E}[\mathbf{y} \,|\, \mathbf{x}]$, while paying less attention to whether the noise distribution can accurately capture the uncertainty of $\mathbf{y}$ given $\mathbf{x}$. For this reason, they may not work well if the distribution of $\mathbf{y}$ given $\mathbf{x}$ clearly deviates from the additive-noise assumption. For example, if $p(\mathbf{y} \,|\, \mathbf{x})$ is multi-modal, which commonly happens when there are missing categorical covariates in $\mathbf{x}$, then $\mathbb{E}[\mathbf{y} \,|\, \mathbf{x}]$ may not be close to any possible true values of $\mathbf{y}$ given that specific $\mathbf{x}$. More specifically, consider a person whose weight, height, blood pressure, and age are known but gender is unknown, then the testosterone level of this person is likely to follow a bimodal distribution and the chance of developing breast cancer is also likely to follow a bimodal distribution. Therefore, these widely used additive-noise models, which use a deterministic function $f(\mathbf{x})$ to characterize the conditional mean of $\mathbf{y}$, are inherently restrictive in their ability for uncertainty estimation.

In this paper, our goal is to accurately recover the full distribution of $\mathbf{y}$ conditioning on $\mathbf{x}$ given a set of N training data points, denoted as $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{1,N}$. To realize this goal, we consider the diffusion-based (a.k.a. score-based) generative models [23, 11, 24, 25] and inject covariate-dependence into

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

both the forward (inference) and reverse (generative) diffusion chains. Our method can model the conditional distribution of both continuous and categorical $\mathbf{y}$ variable, and the algorithms developed under this method will be collectively referred to as Classification And Regression Diffusion (CARD) models.

Diffusion-based generative models have received significant recent attention due to not only their ability to generate high-dimensional data, such as high-resolution photo-realistic images, but also their training stability. They can be understood from the perspective of score matching [12, 27] and Langevin dynamics [16, 28], as pioneered by Song and Ermon [23]. They can also be understood from the perspective of diffusion probabilistic models [22, 11], which first define a forward diffusion to transform the data into noise and then an inverse diffusion to regenerate the data from noise.

These previous methods mainly focus on unconditional generative modeling. While there exist guided-diffusion models [23, 25, 4, 17, 21] that target on generating high-resolution photo-realistic images that match the semantic meanings or content of the label, text, or corrupted-images, we focus on studying diffusion-based conditional generative modeling at a more fundamental level. In particular, our goal is to thoroughly investigate whether CARD can help accurately recovery $p(\mathbf{y} \mid \mathbf{x}, \mathcal{D})$, the predictive distribution of $\mathbf{y}$ given $\mathbf{x}$ after observing data $\mathcal{D}$. In other words, our focus is on regression analysis of continuous or categorical response variables given their corresponding covariates.

We summarize our main contributions as follows: 1) We show CARD, which injects covariate-dependence and a pre-trained conditional mean estimator into both the forward and reverse diffusion chains to construct a denoising diffusion probabilistic model, provides an accurate estimation of $p(\mathbf{y} \mid \mathbf{x}, \mathcal{D})$. 2) We provide a new metric to better evaluate how well a regression model captures the full distribution $p(\mathbf{y} \mid \mathbf{x}, \mathcal{D})$. 3) Experiments on standard benchmarks for regression analysis show that CARD achieves state-of-the-art results, using both existing metrics and the new one.

2 Methods and Algorithms for CARD

2.1 Problem Statement

Given the ground-truth response variable $\mathbf{y}_0$ and its covariates $\mathbf{x}$, and assuming a sequence of intermediate uncertain prediction $\mathbf{y}_{1:T}$ made by the diffusion model, the goal of supervised learning is to learn a good model such that the log-likelihood is maximized by optimizing the following ELBO:

$$\log p_\theta(\mathbf{x}, \mathbf{y}_0) = \log \int p_\theta(\mathbf{x}, \mathbf{y}_{0:T}) d\mathbf{y}_{1;T} \geq \mathbb{E}_{q(\mathbf{y}_{1:T}|\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{y}_{0:T}|\mathbf{x})}{q(\mathbf{y}_{1:T}|\mathbf{y}_0, \mathbf{x})} \right], \tag{1}$$

where $q(\mathbf{y}_{1:T}|\mathbf{y}_0, \mathbf{x})$ is called the forward process or diffusion process in the concept of diffusion models [11]. Denoting $D_{\mathrm{KL}}(q||p)$ as the Kullback–LeiblerKL (KL) divergence from distributions p to q. The above objective can be rewritten as:

$$\mathcal{L}_{\mathrm{ELBO}}(\mathbf{x}) \coloneqq \mathcal{L}_0(\mathbf{x}) + \sum_{t=2}^{T} \mathcal{L}_{t-1}(\mathbf{x}) + \mathcal{L}_T(\mathbf{x}), \tag{2}$$

$$\mathcal{L}_0(\mathbf{x}) \coloneqq \mathbb{E}_q \left[-\log p_\theta(\mathbf{y}_0 \mid \mathbf{y}_1, \mathbf{x}) \right], \tag{3}$$

$$\mathcal{L}_{t-1}(\mathbf{x}) \coloneqq \mathbb{E}_q \left[D_{\mathrm{KL}} \left(q(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{y}_0, \mathbf{x}) || p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x}) \right) \right], \tag{4}$$

$$\mathcal{L}_T(\mathbf{x}) \coloneqq \mathbb{E}_q \left[D_{\mathrm{KL}} \left(q(\mathbf{y}_T \mid \mathbf{y}_0, \mathbf{x}) || p(\mathbf{y}_T|\mathbf{x}) \right) \right]. \tag{5}$$

Here we follow the convention to assume $\mathcal{L}_T$ does not depend on any parameter and it will be close to zero by carefully diffusing the observed response variable $\mathbf{y}_0$ towards a pre-assumed distribution $p(\mathbf{y}_T|\mathbf{x})$. The rest of terms will make sure the model $p_\theta(\mathbf{y}_{t-1} \mid \mathbf{y}_t, \mathbf{x})$ predicts at different diffusion steps and we hope this gives us different scales in uncertainty measure. Different than vanilla diffusion model, we assume the endpoint of the diffusion as:

$$p(\mathbf{y}_T) = \mathcal{N}(f_\phi(\mathbf{x}), \boldsymbol{I}). \tag{6}$$

where $f_\phi(\mathbf{x})$ is pre-knowledge of the relation between $\mathbf{x}$ and $\mathbf{y}_0$, *e.g.,* pre-trained with $\mathcal{D}$ to approximate $\mathbb{E}[\mathbf{y} \mid \mathbf{x}]$, or $\mathbf{0}$ if we assume the relation is unknown. With a diffusion schedule $\{\beta_t\}_{t=1:T} \in (0, 1)^T$, we specify the forward diffusion process conditional distributions in a similar fashion as [18], but for all timesteps including $t = 1$:

$$q\big(\mathbf{y}_t \mid \mathbf{y}_{t-1}, f_\phi(\mathbf{x})\big) = \mathcal{N}\big(\mathbf{y}_t; \sqrt{1-\beta_t}\mathbf{y}_{t-1} + (1 - \sqrt{1-\beta_t}) f_\phi(\mathbf{x}), \beta_t \boldsymbol{I}\big), \tag{7}$$

which admits a closed form sampling distribution with arbitary timestep t:

$$q\big(\mathbf{y}_t \,|\, \mathbf{y}_0, f_\phi(\mathbf{x})\big) = \mathcal{N}\big(\mathbf{y}_t; \sqrt{\bar{\alpha}_t}\mathbf{y}_0 + (1-\sqrt{\bar{\alpha}_t})f_\phi(\mathbf{x}), (1-\bar{\alpha}_t)\boldsymbol{I}\big), \tag{8}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_t \alpha_t$. Note that Eq. (7) can be viewed as an interpolation between true data $\mathbf{y}_0$ and predicted conditional expectation $f_\phi(\mathbf{x})$, that gradually changes from the former to the latter throughout the forward process.

Such formulation corresponds to a tractable forward process posterior:

$$q(\mathbf{y}_{t-1} \,|\, \mathbf{y}_t, f_\phi(\mathbf{x})) = \mathcal{N}\big(\mathbf{y}_{t-1}; \tilde{\boldsymbol{\mu}}\big(\mathbf{y}_t, f_\phi(\mathbf{x})\big), \tilde{\beta}_t \boldsymbol{I}\big), \tag{9}$$

where

$$\tilde{\boldsymbol{\mu}}\big(\mathbf{y}_t, f_\phi(\mathbf{x})\big) \coloneqq \underbrace{\frac{\beta_t\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}}_{\gamma_0}\mathbf{y}_0 + \underbrace{\frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}}_{\gamma_1}\mathbf{y}_t + \underbrace{\left(1 + \frac{(\sqrt{\bar{\alpha}_t}-1)(\sqrt{\alpha_t}+\sqrt{\bar{\alpha}_{t-1}})}{1-\bar{\alpha}_t}\right)}_{\gamma_2} f_\phi(\mathbf{x}),$$

$$\tilde{\beta}_t \coloneqq \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.$$

We provide the derivation in Appendix A.1.

2.2 CARD for Regression

For regression problems, the goal of the reverse diffusion process is to gradually recover the distribution of the noise term, the aleatoric uncertainty inherent in the observations according to Kendall and Gal [13], enabling us to generate samples that match the true conditional $p(\mathbf{y} \,|\, \mathbf{x})$.

Following the reparameterization introduced by Ho et al. [11], we construct $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}\big(\boldsymbol{x}, \boldsymbol{y}_t, f_\phi(\mathbf{x}), t\big)$, which is a function approximator parameterized by a deep neural network that predicts the forward diffusion noise $\boldsymbol{\epsilon}$ sampled for $\boldsymbol{y}_t$. The training and inference procedure can be carried out in a standard DDPM manner.

Algorithm 1 Training (Regression)

1: pre-train $f_\phi(\mathbf{x})$ that predicts $\mathbb{E}(\mathbf{y} \,|\, \mathbf{x})$ with MSE
2: **repeat**
3: $\quad y_0 \sim q(y_0)$
4: $\quad t \sim \text{Uniform}(\{1 \dots T\})$
5: $\quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
6: $\quad$ compute noise estimation loss

$$\mathcal{L}_\epsilon = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(x, \sqrt{\bar{\alpha}_t}\boldsymbol{y}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} + \sqrt{1-\bar{\alpha}_t}f_\phi(\mathbf{x}), t)||^2$$

7: $\quad$ take numerical optimization step on:

$$\nabla_\theta \mathcal{L}_\epsilon$$

8: **until** convergence

Algorithm 2 Inference (Regression)

1: $\boldsymbol{y}_T \sim \mathcal{N}(f_\phi(\mathbf{x}), \boldsymbol{I})$
2: **for** $t = T$ to 1 **do**
3: $\quad \boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\boldsymbol{z} = \mathbf{0}$
4: $\quad$ reparameterized $\hat{\boldsymbol{y}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\boldsymbol{y}_t - (1-\sqrt{\bar{\alpha}_t})f_\phi(\mathbf{x}) - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}_t, f_\phi(\mathbf{x}), t)\sqrt{1-\bar{\alpha}_t}\big)$
5: $\quad \boldsymbol{y}_{t-1} = \gamma_0\hat{\boldsymbol{y}}_0 + \gamma_1\boldsymbol{y}_t + \gamma_2 f_\phi(\mathbf{x}) + \boldsymbol{z}\tilde{\beta}_t$
6: **end for**
7: **return** $\boldsymbol{y}_0$

2.3 CARD for Classification

We formulate the classification tasks in a similar fashion as in Section 2.2, where we:

1. replace the continuous response variable with one-hot encoded labels for $\mathbf{y}_0$;
2. replace the mean estimator with a pre-trained classifier that outputs softmax probabilities of the class labels for $f_\phi(\mathbf{x})$.

This construction no longer assumes $\boldsymbol{y}_0$ to be drawn from a categorical distribution, but instead treating each one-hot label as a class prototype. The sampling procedure would output reconstructed $\boldsymbol{y}_0$ in the range of real numbers for each dimension, instead of a vector in the probability simplex. We convert such output to a probability vector inspired by the Brier score [2], which computes the squared error between the prediction and a vector of 1s with length C, the number of classes:

$$\hat{\mathbf{y}} = \text{Softmax}\big(-(\mathbf{y}_0 - \mathbf{1}_C)^2\big). \tag{10}$$

Intuitively, this construction would assign the class whose raw output in the sampled $\mathbf{y}_0$ is closest to the true class, encoded by the value of 1 in the one-hot label, with the highest probability.

The stochasticity of generative model would give us a different class prototype reconstruction, which leads to the variation in the predicted probability for each class label, enabling us to prediction intervals. Such stochastic reconstruction is in a similar fashion as DALL-E 2 [20] that applies a diffusion prior to reconstruct the image embedding by conditioning on the text embedding during reverse diffusion process, which is a key step in generated image diversity.

3 Experiments

For the hyperparameters of CARD in both regression and classification tasks, we set the number of timesteps as $T = 1000$, a linear noise schedule with $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$, same as Ho et al. [11]. For network architecture, we adopt the settings described in [30, 32]. We provide a more detailed walkthrough of experimental setup in Appendix A.2.

3.1 Regression

Putting aside its statistical interpretation, the word *regress* indicates a direction opposite to *progress*, suggesting a less developed state. Such semantics in fact translates well into the statistical domain, in the sense that traditional regression analysis methods often only focus on estimating $\mathbb{E}(\mathbf{y} \mid \mathbf{x})$, while leaving out all other details about $p(\mathbf{y} \mid \mathbf{x})$. In recent years, Bayesian neural networks (BNNs) have emerged as a class of models that aims at estimating the uncertainty [10, 7, 14, 26], providing a more complete picture of $p(\mathbf{y} \mid \mathbf{x})$. The metric that they use to quantify uncertainty estimation, negative log-likelihood (NLL), is computed with a Gaussian density, implying their assumption such that the conditional distributions $p(\mathbf{y} \mid \mathbf{x} = x)$ for all x are Gaussian. However, this assumption is very difficult to verify for real world datasets: the covariates can be arbitrarily high-dimensional, making the feature space increasingly sparse with respect to the number of collected observations.

To accommodate the need for uncertainty estimation without imposing such restriction for the parametric form of $p(\mathbf{y} \mid \mathbf{x})$, we apply the following two metrics, both of which are designed to empirically evaluate the level of similarity between the learned and the true conditional distributions:

1. Prediction Interval Coverage Probability (PICP);
2. Quantile Interval Coverage Error (QICE).

PICP has been described in Yao et al. [31], whereas QICE is a new metric proposed by us. We describe both of them in the following section.

3.1.1 PICP and QICE

The PICP is computed as:

$$\text{PICP} \coloneqq \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}_{y_n \geq \hat{y}_n^{\text{low}}} \cdot \mathbb{1}_{y_n \leq \hat{y}_n^{\text{high}}}, \tag{11}$$

where $\hat{y}_n^{\text{low}}$ and $\hat{y}_n^{\text{high}}$ represents the low and high percentile of our choice for the predicted $\mathbf{y}$ outputs given the same $\mathbf{x}$ input. This metric measures the proportion of true observations that fall in the percentile range of the generated $\mathbf{y}$ samples given each $\mathbf{x}$ input. Intuitively, when the learned distribution represents the true distribution well, this measurement should be close to the difference between the selected low and high percentile. In this paper, we choose the 2.5% and 97.5% percentile, thus an ideal PICP value for the learned model should be 95%.

Meanwhile, there is a caveat for this metric: for example, imagine a situation where the 2.5% and 97.5% percentile of the learned distribution happens to cover the data between the 1% and 96% percentile from the true distribution. Given enough samples, we shall still obtain a PICP value close to 95%, but clearly there is a mismatch between the learned distribution and the true one.

Based on such reasoning, we propose a new empirical metric QICE, which by design can be viewed as PICP with finer granularity. To compute QICE, we first generate enough $\mathbf{y}$ samples given each $\mathbf{x}$, and divide them into M bins with roughly equal sizes. We would obtain the corresponding quantile values at each boundary. In this paper, we set $M = 10$, and obtain the following 10 quantile intervals (QIs) of the generated $\mathbf{y}$ samples: below 10% percentile, between 10% and 20% percentile, ..., between 80% and 90% percentile, and above 90% percentile. Optimally, when the learned conditional distribution is identical to the true one, given enough samples from both learned and true distribution we shall observe about 10% of true data falling into each of these 10 QIs.

We define *QICE* to be the mean absolute error between the proportion of true data contained by each QI and the optimal proportion, which is $1/M$ for all intervals:

$$\text{QICE} \coloneqq \frac{1}{M}\sum_{m=1}^{M}\left|r_m - \frac{1}{M}\right|, \text{ where } r_m = \frac{1}{N}\sum_{n=1}^{N}\mathbb{1}_{y_n \geq \hat{y}_n^{\text{low}_m}} \cdot \mathbb{1}_{y_n \leq \hat{y}_n^{\text{high}_m}}. \tag{12}$$

Intuitively, under optimal scenario with enough samples, we shall obtain a QICE value of 0. Note that each r_m is indeed the PICP for the corresponding QI with boundaries at $\hat{y}_n^{\text{low}_m}$ and $\hat{y}_n^{\text{high}_m}$. Since the true $\mathbf{y}$ for each $\mathbf{x}$ is guaranteed to fall into one of these QIs, we are thus able to overcome the mismatch issue described in the above example for PICP: fewer true instances falling into one QI would result in more instances captured by another QI, thus increasing the absolute error for both QIs.

QICE is similar to NLL in the sense that it also utilizes the summary statistics of the samples from the learned distribution conditional on each new $\mathbf{x}$ to empirically evaluate how well the model fits the true data. Meanwhile, it does not assume any parametric form on the conditional distribution, making it a much more generalizable metric to measure the level of distributional match between the learned and the underlying true conditional distributions, especially when the true conditional distribution is known to be multi-modal. We will demonstrate this point through the regression toy examples.

3.1.2 Toy Examples

To demonstrate the effectiveness of CARD in regression tasks for not only learning the conditional mean $\mathbb{E}(\mathbf{y} \mid \mathbf{x})$, but also recreating the ground truth data generating mechanism, we first apply CARD on 8 toy examples, whose data generating functions are designed to possess different statistical characteristics: some have a uni-modal symmetric distribution for their error term (linear regression, quadratic regression, sinusoidal regression[1]), others have heteroscedasticity (log-log linear regression, log-log cubic regression) or multi-modality (inverse sinusoidal regression[2], 8 Gaussians[3], full circle). We show that CARD can generate samples that are visually indistinguishable from the true response variables of the new covariates, as well as quantitatively match the true distribution in terms of some summary statistics.

The 8 toy examples are summarized by Table 7 in Appendix A.3. For each task, we create the dataset by sampling $10\,240$ data points from the data generating function, and randomly split them into training and test set with a $80\%/20\%$ ratio.

We first examine the performance of CARD visually, by making the scatter plots of both true test data and generated data conditional on test $\mathbf{x}$ for all 8 tasks. For tasks with uni-modal conditional

[1]The data generating function was originally proposed in [1].

[2]Swap x and the generated y from the sinusoidal regression task.

[3]We set the coordinates of 8 modes at $(\sqrt{2}, 0)$, $(-\sqrt{2}, 0)$, $(0, \sqrt{2})$, $(0, -\sqrt{2})$, $(1, 1)$, $(1, -1)$, $(-1, 1)$, $(-1, -1)$, and add a different noise sample to x and y coordinate of a mode to generate one instance.

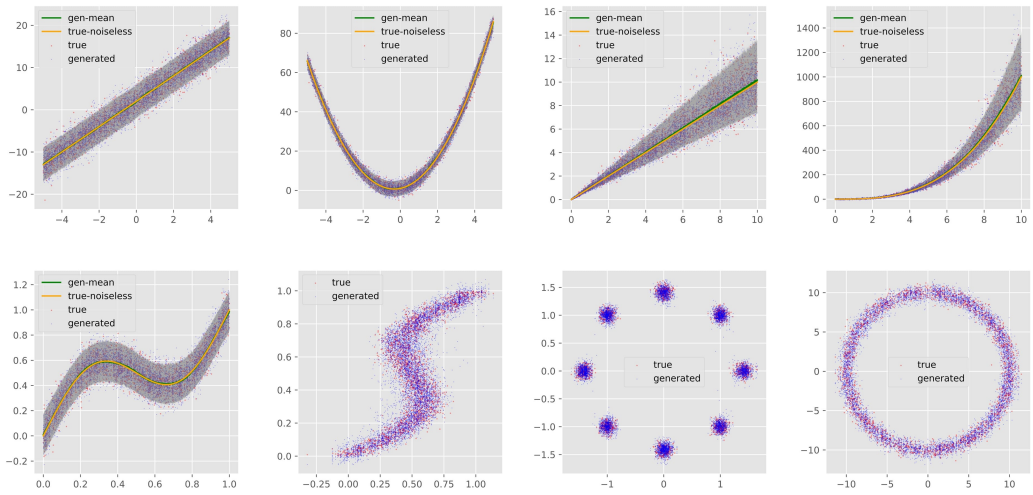

Figure 1: Regression toy example scatter plots. (**Top**) left to right: linear regression, quadratic regression, log-log linear regression, log-log cubic regression; (**Bottom**) left to right: sinusoidal regression, inverse sinusoidal regression, 8 Gaussians, full circle.

distribution, we fill the region between the 2.5-th and 97.5-th percentile of the generated $\mathbf{y}$'s, and plot the true $\mathbb{E}(\mathbf{y} \mid \mathbf{x})$ along with the sample means. We observe that for all 8 tasks, the generated samples (blue dots) have blended remarkably well with the true test instances (red dots), suggesting the CARD model is capable of reconstructing the underlying data generation mechanism. For all uni-modal tasks, we note that the empirical means of generated samples are close to the true conditional means for all test $\mathbf{x}$, as the green and orange line interlace with each other without a clear separation at any point. Note that we include all $\mathbf{y}$ samples for the computation of each conditional mean, instead of discarding any samples that might be deemed as extreme.

To quantitatively evaluate the performance of CARD, we generate $1\,000$ $\mathbf{y}$ samples for each $\mathbf{x}$ in the test set, and compute the corresponding metrics. We conduct such procedure over 10 runs, each applying a different random seed to generate the dataset, and report the mean and standard deviation over all runs for each metric. For all tasks regardless of the form of $p(\mathbf{y} \mid \mathbf{x})$, we compute PICP and QICE. For tasks with uni-modal $p(\mathbf{y} \mid \mathbf{x})$ distributions, we summarize the $1\,000$ samples for each test $\mathbf{x}$ by computing their mean, as an unbiased estimator to $\mathbb{E}(\mathbf{y} \mid \mathbf{x})$, and compute the root mean squared error (RMSE) between the estimated and true conditional mean. For all tasks, we obtain a mean PICP very close to the optimal 95%, and most of the tasks have a mean QICE value far less than 0.01 except log-log cubic regression, which also has a mean RMSE noticeably larger by an order of magnitude among cases with uni-modal conditional distributions. Note that the $\mathbf{y}$ samples here have a much wider range: as $\mathbf{x}$ increases from 0 to 10, $\mathbf{y}$ increases from 0 to over $1\,200$, resulting in a much more difficult task. Therefore, the metrics reported here can be viewed with relativity, and combined with the qualitative conclusions from 1. The metrics of all tasks are recorded to Table 8 in Appendix A.3.

3.1.3 UCI Regression Tasks

We continue to investigate our model through experiments on real world datasets. We adopt the same set of 10 UCI regression benchmark datasets [5] as well as the experimental protocol proposed in [10] and followed by [7] and [14]. The dataset information in terms of their size and number of features are summarized in 1.

Table 1: Dataset size (N observations, P features) of UCI regression tasks.

Dataset	Boston	Concrete	Energy	Kin8nm	Naval	Power	Protein	Wine	Yacht	Year
(N, P)	$(506, 13)$	$(1\,030, 8)$	$(768, 8)$	$(8\,192, 8)$	$(11\,934, 16)$	$(9\,568, 4)$	$(45\,730, 9)$	$(1\,599, 11)$	$(308, 6)$	$(515\,345, 90)$

We apply multiple train-test splits with $90\%/10\%$ ratio in the same way (20 folds for all datasets except 5 for Protein and 1 for Year), and report the metrics by their mean and standard deviation across all splits. We compare our method to all aforementioned BNN frameworks: PBP, MC Dropout, and Deep Ensembles, as well as another deep generative model that estimates a conditional distribution sampler, GCDS[33]. Similar to BNNs, we evaluate the accuracy and predictive uncertainty estimation of CARD by reporting RMSE and NLL. Furthermore, we also report QICE for all methods to evaluate distributional matching. Since this new metric was not applied in previous methods, we re-ran the experiments for all BNNs and obtained comparable or slightly better results reported in their literature. Further details about the experiment setup for these models can be found in the Appendix. The experiment results with corresponding metrics are shown in Table 2, 3, 4, with the number of times each model achieves the best corresponding metric reported in the last row.

Table 2: RMSE of UCI regression tasks. For both [1]Kin8nm and [2]Naval dataset, we multiply the response variable by 100 to match the scale of others.

Dataset	RMSE ↓				
	PBP	MC Dropout	Deep Ensembles	GCDS	CARD (ours)
Boston	2.89 ± 0.74	3.06 ± 0.96	3.17 ± 1.05	2.75 ± 0.58	$\mathbf{2.62 \pm 0.63}$
Concrete	5.55 ± 0.46	5.09 ± 0.60	$\mathbf{4.91 \pm 0.47}$	5.39 ± 0.55	4.92 ± 0.48
Energy	1.58 ± 0.21	1.70 ± 0.22	2.02 ± 0.32	$\mathbf{0.64 \pm 0.09}$	0.69 ± 0.20
Kin8nm[1]	9.42 ± 0.29	7.10 ± 0.26	8.65 ± 0.47	8.88 ± 0.42	$\mathbf{6.40 \pm 0.19}$
Naval[2]	0.41 ± 0.08	0.08 ± 0.03	0.09 ± 0.01	0.14 ± 0.05	$\mathbf{0.02 \pm 0.00}$
Power	4.10 ± 0.15	4.04 ± 0.14	4.02 ± 0.15	4.11 ± 0.16	$\mathbf{3.91 \pm 0.16}$
Protein	4.65 ± 0.02	4.16 ± 0.12	4.45 ± 0.02	4.50 ± 0.02	$\mathbf{3.71 \pm 0.01}$
Wine	0.64 ± 0.04	$\mathbf{0.62 \pm 0.04}$	0.63 ± 0.04	0.66 ± 0.04	$\mathbf{0.62 \pm 0.04}$
Yacht	0.88 ± 0.22	0.84 ± 0.27	1.19 ± 0.49	0.79 ± 0.26	$\mathbf{0.56 \pm 0.20}$
Year	$8.86\pm$ NA	$8.77\pm$ NA	$8.79\pm$ NA	$9.20\pm$ NA	$\mathbf{8.69}\pm$ NA
# best	0	1	1	1	**8**

Table 3: NLL of UCI regression tasks.

Dataset	NLL ↓				
	PBP	MC Dropout	Deep Ensembles	GCDS	CARD (ours)
Boston	2.53 ± 0.27	2.46 ± 0.12	2.35 ± 0.16	18.66 ± 8.92	$\mathbf{2.32 \pm 0.17}$
Concrete	3.19 ± 0.05	3.21 ± 0.18	$\mathbf{2.93 \pm 0.12}$	13.64 ± 6.88	2.99 ± 0.12
Energy	2.05 ± 0.05	1.50 ± 0.11	1.40 ± 0.27	1.46 ± 0.72	$\mathbf{1.12 \pm 0.15}$
Kin8nm	-0.83 ± 0.02	-1.14 ± 0.05	-1.06 ± 0.02	-0.38 ± 0.36	$\mathbf{-1.31 \pm 0.02}$
Naval	-3.97 ± 0.10	-4.45 ± 0.38	-5.94 ± 0.10	-5.06 ± 0.48	$\mathbf{-6.49 \pm 0.01}$
Power	2.92 ± 0.02	2.90 ± 0.03	2.89 ± 0.02	2.83 ± 0.06	$\mathbf{2.82 \pm 0.02}$
Protein	3.05 ± 0.00	2.80 ± 0.08	2.89 ± 0.02	2.81 ± 0.09	$\mathbf{2.47 \pm 0.01}$
Wine	1.03 ± 0.03	0.93 ± 0.06	0.96 ± 0.06	6.52 ± 21.86	$\mathbf{0.87 \pm 0.08}$
Yacht	1.58 ± 0.08	1.73 ± 0.22	1.11 ± 0.18	$\mathbf{0.61 \pm 0.34}$	0.89 ± 0.07
Year	$3.69\pm$ NA	$3.42\pm$ NA	$3.44\pm$ NA	$3.43\pm$ NA	$\mathbf{3.34}\pm$ NA
# best	0	0	1	1	**8**

Table 4: QICE (in $\%$) of UCI regression tasks.

Dataset	QICE ↓				
	PBP	MC Dropout	Deep Ensembles	GCDS	CARD (ours)
Boston	3.50 ± 0.88	3.82 ± 0.82	$\mathbf{3.37 \pm 0.00}$	11.73 ± 1.05	3.50 ± 0.89
Concrete	2.52 ± 0.60	4.17 ± 1.06	2.68 ± 0.64	10.49 ± 1.01	$\mathbf{2.38 \pm 0.61}$
Energy	6.54 ± 0.90	5.22 ± 1.02	$\mathbf{3.62 \pm 0.58}$	7.41 ± 2.19	4.82 ± 1.06
Kin8nm	1.31 ± 0.25	1.50 ± 0.32	1.17 ± 0.22	7.73 ± 0.80	$\mathbf{0.90 \pm 0.28}$
Naval	4.06 ± 1.25	12.50 ± 1.95	6.64 ± 0.60	5.76 ± 2.25	$\mathbf{2.57 \pm 0.71}$
Power	$\mathbf{0.82 \pm 0.19}$	1.32 ± 0.37	1.09 ± 0.26	1.77 ± 0.33	0.85 ± 0.18
Protein	1.69 ± 0.09	2.82 ± 0.41	2.17 ± 0.16	2.33 ± 0.18	$\mathbf{0.75 \pm 0.04}$
Wine	$\mathbf{2.22 \pm 0.64}$	2.79 ± 0.56	2.37 ± 0.63	3.13 ± 0.79	3.68 ± 0.82
Yacht	6.93 ± 1.74	10.33 ± 1.34	7.22 ± 1.41	$\mathbf{5.01 \pm 1.02}$	8.17 ± 0.85
Year	$2.96\pm$NA	$2.43\pm$NA	$2.56\pm$NA	$1.61\pm$NA	$\mathbf{0.56}\pm$NA
# best	2	0	2	1	**5**

We observe that CARD outperforms existing methods, often by a considerable margin (especially on larger datasets), in all metrics for most of the datasets, and is competitive with the best method for

the remaining ones: we obtain state-of-the-art results in 8 out of 10 datasets in terms of both RMSE and NLL, and 5 out of 10 for QICE. Note that although we do not explicitly optimize our model by MSE or by NLL, we still obtain better results than models trained with these objectives.

3.2 Classification

Similar to Lakshminarayanan et al. [14], our motivation for classification is not to achieve the state-of-the-art performance in terms of mean accuracy on the benchmark datasets, which is strongly related to network architecture design. Our goal is two-fold:

1. We aim to solve classification problems via a generative model, emphasizing its capability to improve the performance of a base classifier with deterministic outputs;
2. We intent to introduce the idea of uncertainty to classification predictions at instance level.

As another type of supervised learning problems, classification is different than regression mainly for the response variable being discrete class labels instead of continuous values. Conventionally, the output from a classification algorithm is a point estimate, a value being casted as between 0 and 1, with the intention to align with the human cognitive intuition of probability for each class label [3]. In other words, that point estimate should express a level of confidence, or the likelihood of correctness, in that prediction by the model. Metrics like Expected Calibration Error (ECE) and Maximum Calibration Error (MCE) have been proposed to evaluate the alignment between classifier predictions and the true correctness likelihood, and calibration methods like Platt scaling and isotonic regression have been developed to improve such alginment [9], but all are based on point estimate predictions. In the next section, we introduce an alternative in access model confidence, at the level of individual instances, specifically tailored for models with stochastic output.

3.2.1 Estimating Instance Level Prediction Confidence via Generative Models

We utilize the stochasticity of generative models to construct our framework of evaluating prediction uncertainty at instance level. As introduced in Section 2.3, we view each one-hot label as a class prototype, and we use the generative model to reconstruct this prototype in a stochastic fashion. The intuition is that if the learned model knows well about which class a particular instance belongs to, it would precisely reconstruct the prototype vector; otherwise if the model is unsure about a new instance based on the learned patterns, it would not be robust against the input noise sample: under the context of denoising diffusion models, the class prototype reconstruction would appear rather different from each other, given a different sample from the prior distribution at timestep T.

We propose the following framework to access model confidence: during test time, we sample N class prototype reconstructions for each instance, converting them into probability scale with Eq. (10), and make the following two computations:

1. We calculate the prediction interval width (PIW) between 2.5% and 97.5% percentile of the N predicted probabilities for the true class associated with each instance;
2. We apply paired two-sample t-test as an uncertainty estimation method proposed in [6]: we obtain the most and second most predicted class for each instance, and test whether the difference in their mean predicted probability is statistically significant.

3.2.2 Classification on FashionMNIST with Prediction Uncertainty Evaluation

We demonstrate our experiment results on the FashionMNIST dataset. Following the recipe in 2.3, we first pre-train a deterministic classifier with ResNet-18 architecture for 10 epochs with a batch size of 256, and achieve a test accuracy of 91.19%. After training CARD, we obtain our instance prediction through majority vote, i.e. the most predicted class label among its N samples, and achieve an improved test accuracy of 91.71%, showing its ability to improve the prediction accuracy from the base classifier. We contextualize such performance by reporting the mean accuracy of other BNNs with LeNet CNN architecture [26] in Table 5. CARD uses a much simpler network as its image feature encoder, yet still outperforms all other models.

In Table 6, we report the mean PIW among both correct and incorrect predictions, as well as the mean accuracy among both groups rejected and not-rejected by the paired two-sample t-test ($\alpha = 0.01$). We report these metrics for all test instances and for each class label along with their group accuracy.

Table 5: Comparison of mean accuracy for FashionMNIST classification task with other BNNs and ResNet-18 (our pre-trained classifier f_ϕ).

Model	CMV-MF-VI	CM-MF-VI	CV-MF-VI	CM-MF-VI OPT	MF-VI	MAP	MC Dropout	MF-VI EB	ResNet-18	CARD (ours)
Mean Accuracy	91.10%	90.95%	88.53%	90.67%	87.04%	88.06%	87.99%	87.04%	91.19%	**91.71**%

Table 6: PIW (in %) and t-test results for FashionMNIST classification task.

Class	Accuracy	PIW		Accuracy by t-test Status	
		Correct	Incorrect	Rejected	Not-Rejected (Count)
All	91.71%	2.39	22.05	92.05%	40.00% (65)
1	86.70%	4.35	21.75	87.13%	53.85% (13)
2	98.30%	0.77	13.79	98.30%	100% (1)
3	86.90%	3.65	20.41	87.35%	50.00% (12)
4	92.00%	2.80	27.58	92.25%	50.00% (6)
5	86.40%	3.99	33.36	87.21%	33.33% (15)
6	99.00%	0.83	18.36	99.00%	NA (0)
7	76.70%	5.57	16.00	77.56%	20.00% (15)
8	96.20%	1.32	29.15	96.30%	0.00% (1)
9	98.50%	0.57	10.79	98.50%	NA (0)
10	96.40%	1.29	15.30	96.49%	50.00% (2)

We observe from Table 6 that under the scope of the entire test set, mean PIW of the true class label among the correct predictions is narrower than that of the incorrect predictions by an order of magnitude, indicating that CARD is much more confident when making correct predictions; on the other hand, CARD is also aware of what it does not know. More revealing observations can be made when comparing mean PIWs across different classes: a class with more accurate predictions tends to have a sharper contrast between correct and incorrect predictions; additionally, the metric values of both types of predictions tend to be larger in a less accurate class. Meanwhile, we observe that the accuracy of test instances rejected by the t-test is much higher than that of the not-rejected ones, both among the entire set and within each class (except Class 2 with only 1 not-rejected instance).

We point out that these metrics can thus reflect how sure CARD is about its predictions, and can be used as an important indicator of whether the model prediction can be trusted or not. Therefore, it has the potential to be further applied in the human-machine collaboration domain [15, 19, 29, 8], such that one can apply such uncertainty measurement to decide if we can directly accept the model prediction, or we need to allocate the instance to humans for further evaluation.

4 Conclusion

In this paper, we propose Classification And Regression Diffusion (CARD) models, a class of generative models that approaches supervised learning problems from a generative modeling perspective. Without training with objectives directly related to the evaluation metrics, we achieve state-of-the-art results on benchmark regression tasks. Furthermore, CARD exhibits a strong ability to represent conditional distribution with multiple density modes. We also propose a new metric Quantile Interval Coverage Error (QICE), which can be viewed as a generalized version of negative log-likelihood in evaluating how well the model fits the data. Lastly, we introduce a framework to evaluate prediction uncertainty at instance level for classification tasks.

References

[1] Christopher Bishop. Mixture density networks. In *Aston University Neural Computing Research Group Report*, 1994.

[2] G. W. Brier. Verification of forecasts expressed in terms of probability. In *Monthly weather review*, 1950.

[3] Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. In *cognition*, volume 58(1), pages 1–73, 1996.

[4] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=AAWuCvzaVt`.

[5] D. Dua and C. Graff. UCI machine learning repository. 2017. URL `http://archive.ics.uci.edu/ml`.

[6] Xinjie Fan, Shujian Zhang, Korawat Tanwisuth, Xiaoning Qian, and Mingyuan Zhou. Contextual dropout: An efficient sample-dependent dropout module. In *International Conference on Learning Representations*, 2021.

[7] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

[8] Ruijiang Gao, Maytal Saar-Tsechansky, Maria De-Arteaga, Ligong Han, Min Kyung Lee, and Matthew Lease. Human-ai collaboration with bandit feedback. In *International Joint Conferences on Artificial Intelligence*, 2021.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[10] José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 2015.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[12] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[13] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *31st Conference on Neural Information Processing System*, 2017.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, 2017.

[15] David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.

[16] Radford M Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, page 113, 2011.

[17] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[18] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

[19] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. 2019.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

[22] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[23] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11918–11930, 2019.

[24] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

[25] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

[26] Marcin B. Tomczak, Siddharth Swaroop, Andrew Y. K. Foong, and Richard E. Turner. Collapsed variational bounds for bayesian neural networks. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, 2021.

[27] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[28] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[29] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *International Joint Conferences on Artificial Intelligence*, 2020.

[30] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

[31] Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. In *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, 2019.

[32] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Truncated diffusion probabilistic models. *arXiv preprint arXiv:2202.09671*, 2022.

[33] Xingyu Zhou, Yuling Jiao, Jin Liu, and Jian Huang. A deep generative approach to conditional sampling. *Journal of the American Statistical Association*, pages 1–28, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix B
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix B

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Supplement
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Supplement
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix C
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]